Seeking Certainty in An Uncertain World

Learning decisions that reduce uncertainty

John D. Martin
jmarti3@stevens.edu

May 1st, 2020

Relevant Scenerios



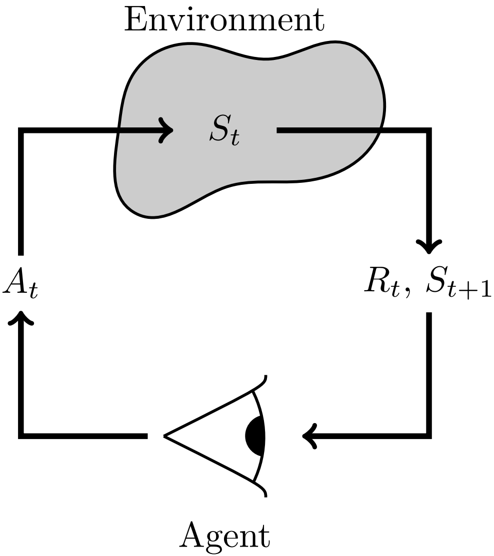
Today's Talk

Stochastically Dominant Distributional Reinforcement Learning

- Full paper: arXiv 1905.07318
- Measuring uncertainty involves computing statistics with hyperparameters.
- We can eliminate hyperparameters without sacrificing certainty.

Foundations

Reinforcement Learning



The RL Problem

- Markov Decision Process: $\langle \mathcal{S}, \mathcal{A}, p, \gamma \rangle$ (Puterman, 1994).
- States $\mathcal{S}$, Actions $\mathcal{A}$, Transition kernel $p\colon \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S} \times \mathbb{R})$, discount $\gamma \in [0, 1)$.
- Agent's goal: maximize expected sum of future rewards.

$$Z_\pi^{(s,a)} = R^{(s,a)} + \gamma R^{(S_1,A_1)} + \gamma^2 R^{(S_2,A_2)} + \cdots = \sum_{t=0}^{\infty} \gamma^t R^{(S_t,A_t)} \Big| S_0 = s, A_0 = a.$$

- Polices: $\Pi = \{\pi | \pi\colon \mathcal{S} \to \mathcal{P}(\mathcal{A})\}$
- Bellman's equations

$$v_\pi(s) = \sum_{a,r,s'} \pi(a|s) p(s', r|s, a) \left[r + \gamma v_\pi(s')\right].$$

$$q_\pi(s, a) = \sum_{r,s'} p(s', r|s, a) \left[r + \gamma v_\pi(s')\right] = \sum_{r,s',a'} p(s', r|s, a) \left[r + \gamma q_\pi(s', a') \pi(a'|s')\right].$$

- *Greedy policy* $\pi(s) = \arg\max_{a \in \mathcal{A}} q(s, a)$

Model-free Methods for Estimation

$$q_\pi(s,a) = \sum_{r,s',a'} p(s',r|s,a)\left[r + \gamma q_\pi(s',a')\pi(a'|s')\right].$$

- We assume the agent does not know the transition kernel $p(s',r|s,a)$.
- Directly estimate q_π or v_π from samples (s,a,r,s').
- Sarsa (Rummery and Niranjan, 1994)

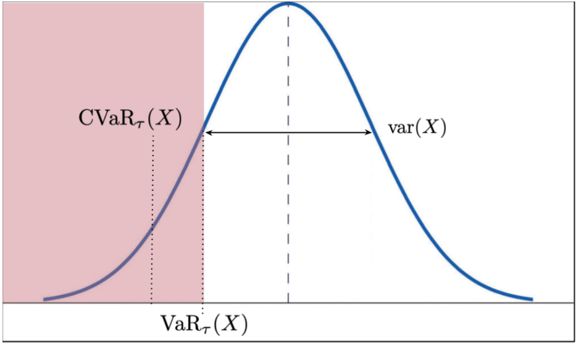
$$Q_\pi^{(s,a)} \leftarrow Q_\pi^{(s,a)} + \alpha\left(r + \gamma Q_\pi^{(s',a')} - Q_\pi^{(s,a)}\right).$$

- Q-learning (Watkins and Dayan, 1992)

$$Q^{(s,a)} \leftarrow Q^{(s,a)} + \alpha\left(r + \gamma \max_{a'\in\mathcal{A}} Q^{(s',a')} - Q^{(s,a)}\right).$$

- Minimize Temporal-Difference (TD) Error (Sutton, 1988): $\delta = R + \gamma Q^{(S',A')} - Q^{(S,A)}$

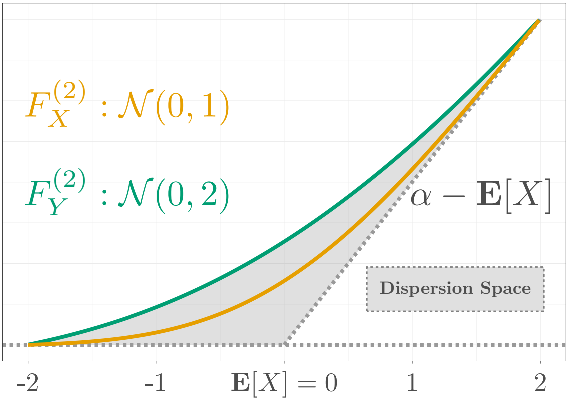
Measuring Uncertainty



Representing Uncertainty

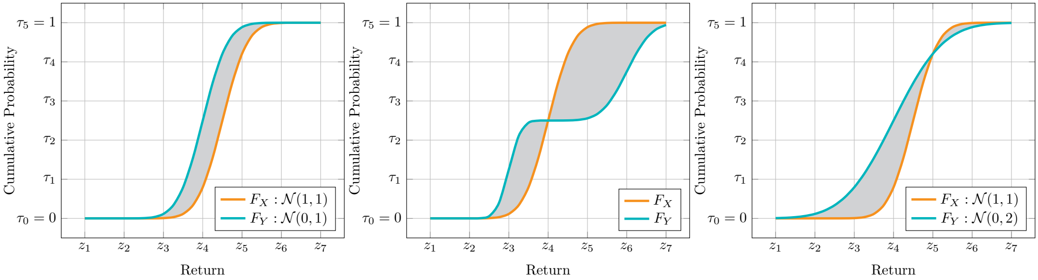
- Variance: $\mathrm{var}(X) = \mathbf{E}[(X - \mathbf{E}[X])^2]$
- Value at Risk: $\mathrm{VaR}_\tau(X) = F_X^{-1}(\tau)$
- Conditional Value at Risk: $\mathrm{CVaR}_\tau = \mathbf{E}[X|X \leq \xi_\tau]$, $\xi_\tau = \mathrm{VaR}_\tau(X)$
- Other measures of dispersion...

Dispersion Space



- For some CDF $F_X^{(1)}(\alpha) = P(X \leq \alpha)$, we define $F_X^{(2)}(\alpha) = \int_{-\infty}^{\alpha} F_X^{(1)}(z)dz$
- Volume of this space reflects the degree of uncertainty

A New Way to Compare Actions



- Equivalent to second-order stochastic dominance

$$X \succeq_{(2)} Y \leftrightarrow F_X^{(2)}(\alpha) \leq F_Y^{(2)}(\alpha) \; \forall \; \alpha \in \mathbb{R}$$

- Choose the action that induces a return with the smallest dispersion

$$\{a_* \in \mathcal{A}_s : Z^{(s,a_*)} \succeq_{(2)} Z^{(s,a')}, \forall \; a' \in \mathcal{A}_s \setminus \{a_*\}\}.$$

Learning the Distribution of Returns

Lemma (Fishburn (1980))

$X \succeq_{(2)} Y$ if, and only if $\mu_X^{(1)} \geq \mu_Y^{(1)}$ or $\mu_X^{(1)} = \mu_Y^{(1)}$ and $\mu_X^{(2)} \leq \mu_Y^{(2)}$, where $(\cdot)$ denotes a particular moment.

☞ SSD comparisons are valid when this ordering can be guaranteed

Distributional RL

- Learn the distribution of returns $\mu^{(s,a)} \in \mathcal{P}_2(\mathbb{R})$ s.t. $Q_\pi^{(s,a)} = \mathbf{E}_\mu[Z_\pi^{(s,a)}]$
- Satisfies a distributional Bellman equation Bellemare et al. (2017):

$$Z_\pi^{(s,a)} \overset{D}{=} R + \gamma Z_\pi^{(S,A)} \Big| \; R, S \sim p(\cdot|s,a), \;\; A \sim \pi(S)$$

- Distributional condition: $\mathcal{T} z^{(s,a)} = r + \gamma \max_{a' \in \mathcal{A}_{s'}} z^{(s',a')} \; \forall \; z \sim \mu^{(s,a)}$

Energy-based RL

Free-energy Minimization

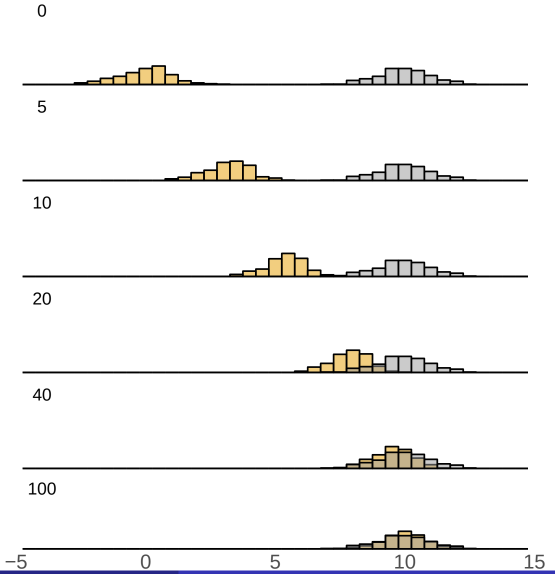
$$E(\mu) = \underbrace{\frac{1}{2}\int \left(\mathcal{T}z^{(s,a)} - z^{(s,a)}\right)^2 d\mu}_{F(\mu)} - \beta^{-1} H(\mu)$$

- Optimal μ is the solution of the Fokker-Planck equation

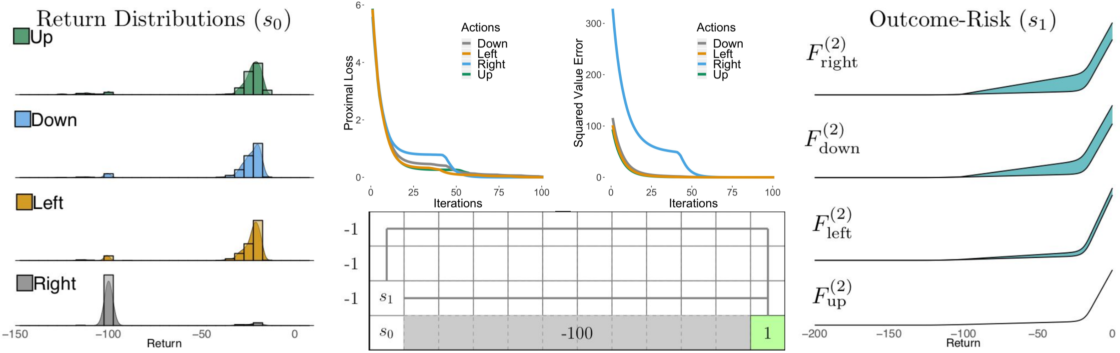
$$\partial_t \mu_t = \nabla \cdot \left(\mu_t \nabla\left(\frac{\delta F}{\delta \mu}\right)\right)$$

- Discrete-time updates are given by

$$\mu_{k+1} = \arg\min_{\mu} \mathcal{W}_2^2(\mu, \mu_k) + 2hE(\mu)$$



Cliffworld

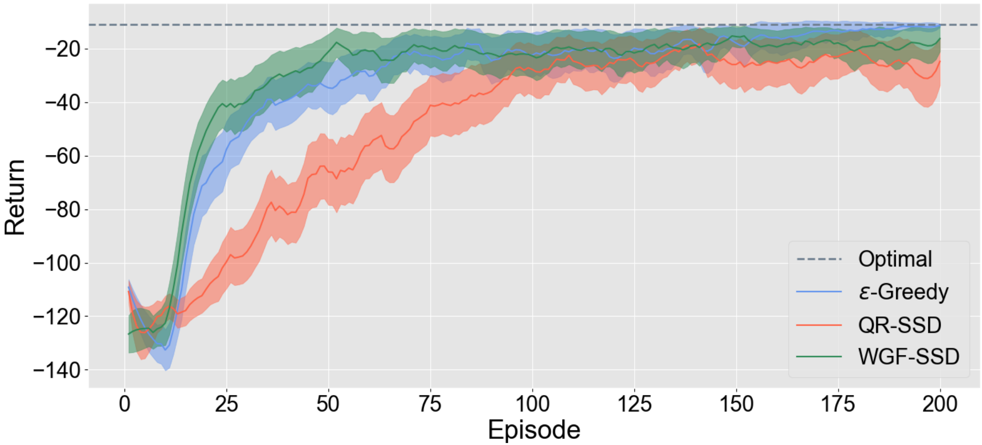


Return Distributions (s_0)
Up
Down
Left
Right
–150
–100
–50
0
Return
Proximal Loss
Iterations
Actions
Down
Left
Right
Up
Squared Value Error
Iterations
-1
-1
-1
s_1
s_0
-100
1
Outcome-Risk (s_1)
$F^{(2)}_{\text{right}}$
$F^{(2)}_{\text{down}}$
$F^{(2)}_{\text{left}}$
$F^{(2)}_{\text{up}}$
–200
–150
–100
–50
0
Return

An Experiment

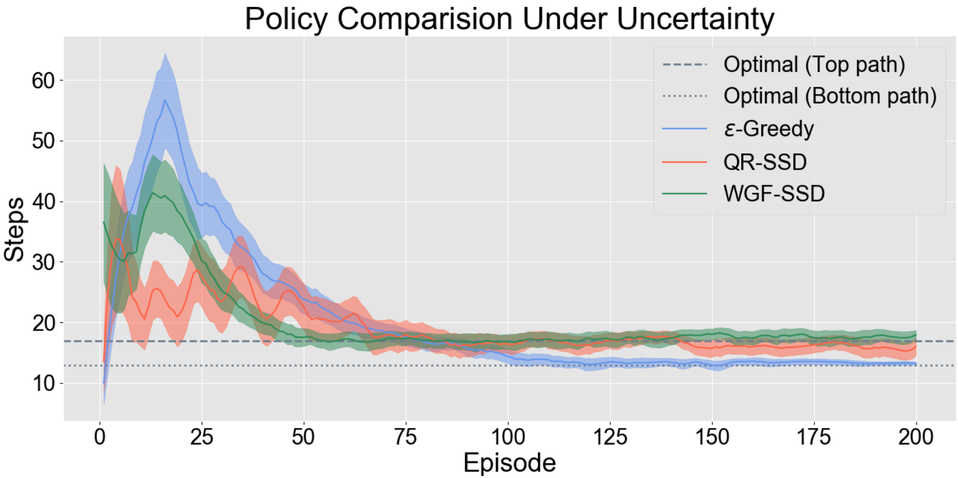
<table>
<tr><td>-1</td><td>-7/11</td><td>-7/11</td><td>-7/11</td><td>-7/11</td><td>-7/11</td><td>-7/11</td><td>-7/11</td><td>-7/11</td><td>-7/11</td><td>-7/11</td><td>-1</td></tr>
<tr><td>-1</td><td>-1</td><td>-1</td><td>-1</td><td>-1</td><td>-1</td><td>-1</td><td>-1</td><td>-1</td><td>-1</td><td>-1</td><td>-1</td></tr>
<tr><td>-1</td><td>$\mathcal{N}^{(-1)}_{10-3}$</td><td>$\mathcal{N}^{(-1)}_{10-3}$</td><td>$\mathcal{N}^{(-1)}_{10-3}$</td><td>$\mathcal{N}^{(-1)}_{10-3}$</td><td>$\mathcal{N}^{(-1)}_{10-3}$</td><td>$\mathcal{N}^{(-1)}_{10-3}$</td><td>$\mathcal{N}^{(-1)}_{10-3}$</td><td>$\mathcal{N}^{(-1)}_{10-3}$</td><td>$\mathcal{N}^{(-1)}_{10-3}$</td><td>$\mathcal{N}^{(-1)}_{10-3}$</td><td>-1</td></tr>
<tr><td>-1</td><td colspan="10">-100</td><td>1</td></tr>
</table>

Results

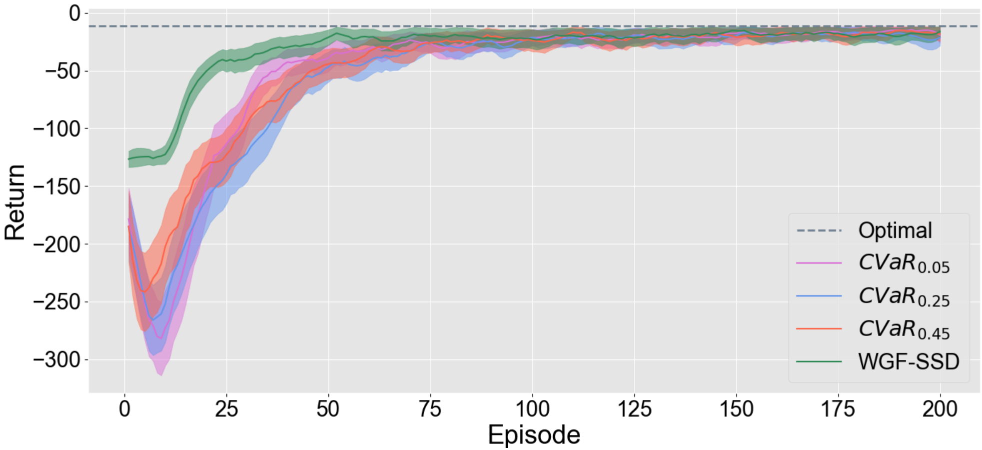


Return
−20
−40
−60
−80
−100
−120
−140
0
25
50
75
100
125
150
175
200
Episode
Optimal
ε-Greedy
QR-SSD
WGF-SSD

Results

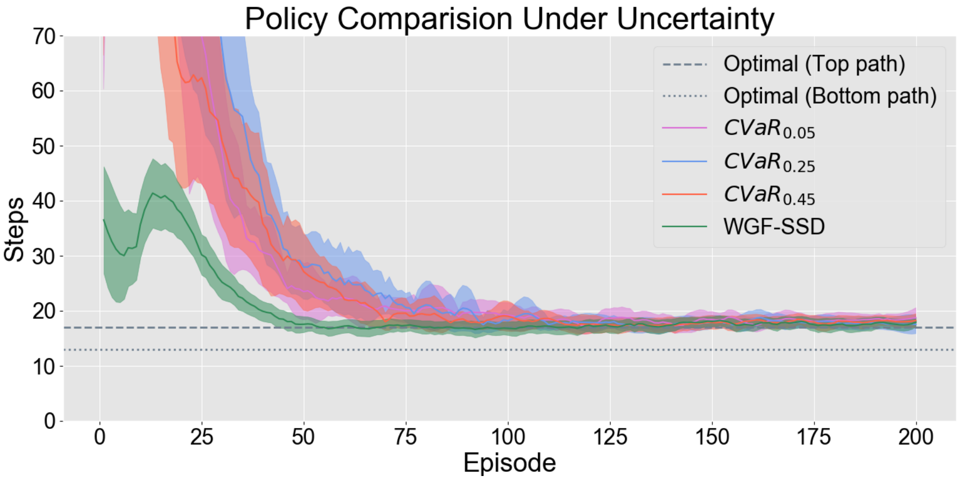

Policy Comparision Under Uncertainty
Optimal (Top path)
Optimal (Bottom path)
ε-Greedy
QR-SSD
WGF-SSD
Steps
60
50
40
30
20
10
0
25
50
75
100
125
150
175
200
Episode

Results



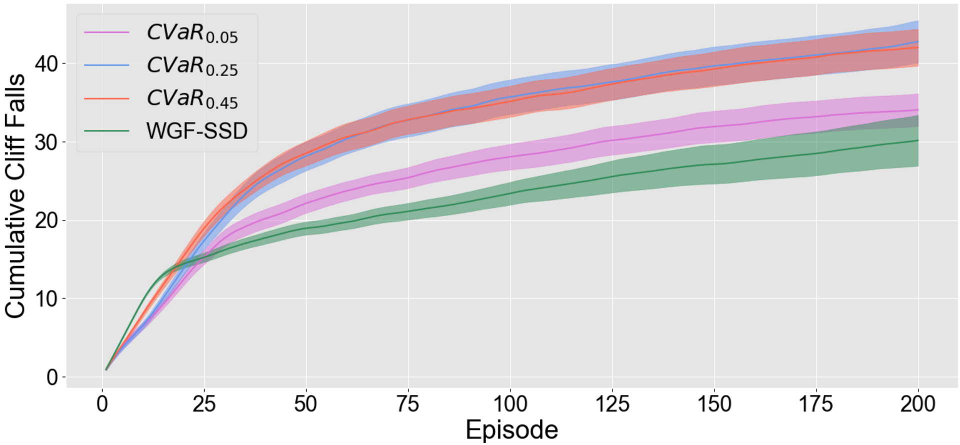
0
−50
−100
−150
−200
−250
−300
Return
0
25
50
75
100
125
150
175
200
Episode
Optimal
$CVaR_{0.05}$
$CVaR_{0.25}$
$CVaR_{0.45}$
WGF-SSD

Results



Policy Comparision Under Uncertainty
Steps
Episode
70
60
50
40
30
20
10
0
0
25
50
75
100
125
150
175
200
Optimal (Top path)
Optimal (Bottom path)
$CVaR_{0.05}$
$CVaR_{0.25}$
$CVaR_{0.45}$
WGF-SSD

Results



$CVaR_{0.05}$
$CVaR_{0.25}$
$CVaR_{0.45}$
WGF-SSD
Cumulative Cliff Falls
0
10
20
30
40
0
25
50
75
100
125
150
175
200
Episode

Discussion

Problem: Machines need to reason about the uncertainty in their environment.

- Aquire and exploit knowledge of environment uncertainty.
- Improve chance of aggregating rewards.

Investigate: Reducing hyperparameters in uncertainty statistics

- How to control aleatoric uncertainty during exploration.
- How to learn representations that capture aleatoric uncertainty.

Conclusion: Aleatoric uncertainty can be represented and exploited for decision making.

- Possible to learn distributional representations with WGF.
- Control uncertainty with SSD action selection.

Questions

Bibliography I

Bellemare, M., Dabney, W., and Munos, R. (2017). A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

Fishburn, P. (1980). Stochastic dominance and moments of distributions. *Math. Operations Research*.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition.

Rummery, G. A. and Niranjan, M. (1994). On-line q-learning using connectionist systems. Technical report, Cambridge University.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Mach. Learn.*, 3(1):9–44.

Watkins, C. J. C. H. and Dayan, P. (1992). Technical note: q-learning. *Mach. Learn.*, 8(3-4):279–292.