Reinforcement Learning Prediction Intervals with Guaranteed Fidelity

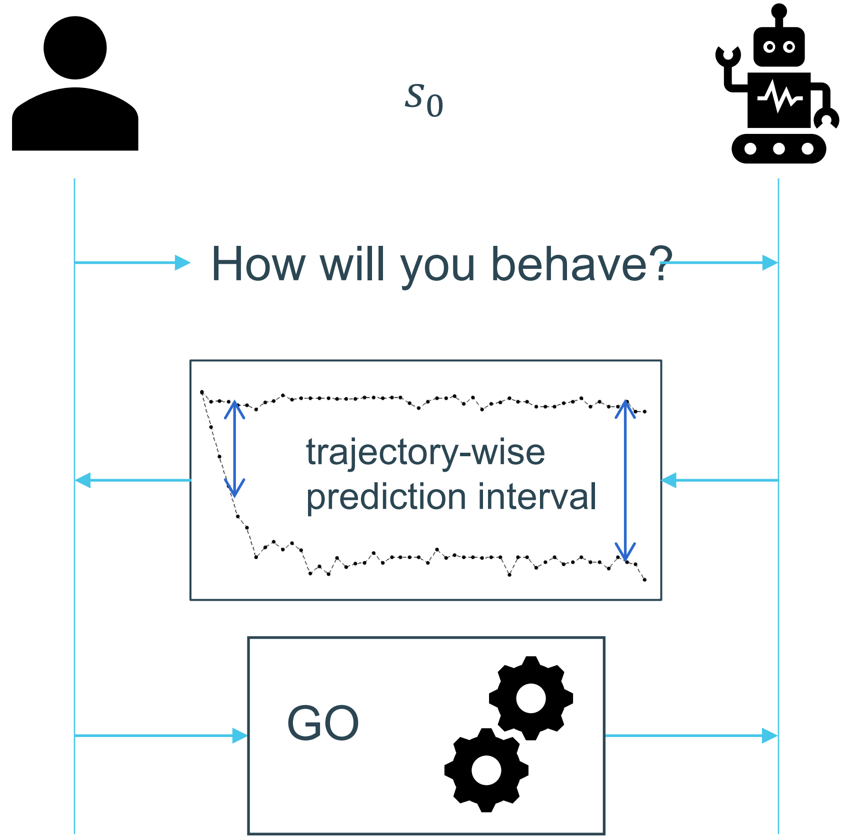
Tom Dietterich, Oregon State University

Jesse Hostetler, SRI International

Prospective MDP Performance Guarantee

Human decision maker must decide whether to command an AI assistant to execute policy π starting in state s_0 for H steps

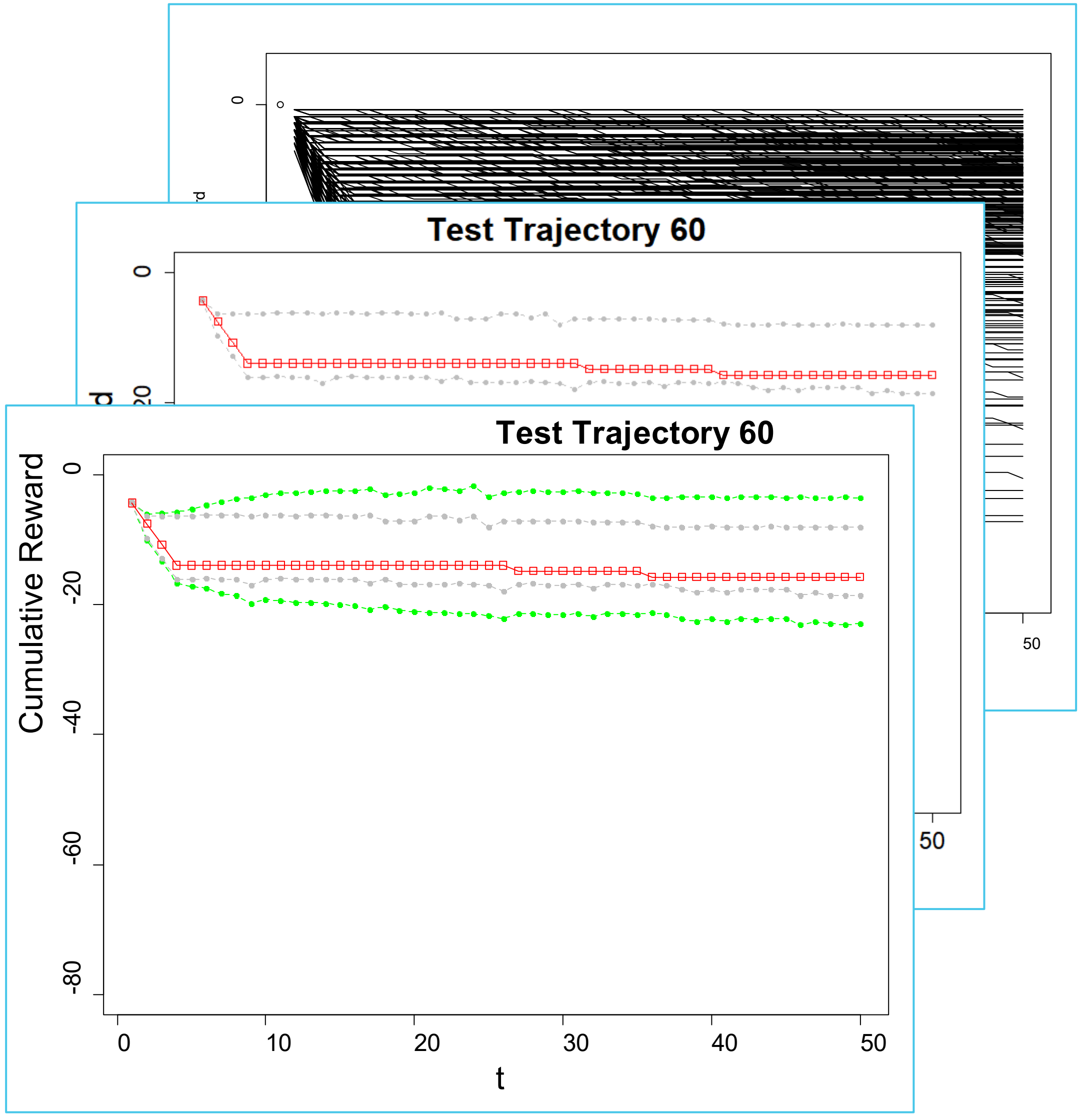
AI assistant provides a trajectory-wise prediction interval that guarantees with probability $1 - \delta$ that its behavior will be inside the interval



SRI International
Oregon State University

Summary of the Approach

- Generate a set of trajectories
  - Repeat N times
    - Sample a starting state $s_0 \sim P_0(\cdot)$
    - Execute π for h steps to obtain a trajectory
- Apply our new technique
  - Perform quantile regression to learn two functions
    - $F_t^{-1}\left(s_0, \frac{\delta}{2}\right)$ an estimate of the $\frac{\delta}{2}$ quantile of the return at time t
    - $F_t^{-1}\left(s_0, 1-\frac{\delta}{2}\right)$ an estimate of the $1-\frac{\delta}{2}$ quantile of the return at time t
  - Adjust these to obtain valid prediction intervals using a new method, SDSCALEDBOX



SRI International

Oregon State University

Outline

- Background: Conformal Prediction Intervals
- Core Problem: Multivariate Prediction Interval
- Prediction Intervals for MDP Trajectories
- Experimental Results
  - Tamarisk invasions
  - StarCraft battles
- Assessment

Background: Conformal Prediction Intervals
(Vovk, Gammerman, Shafer, 2005)

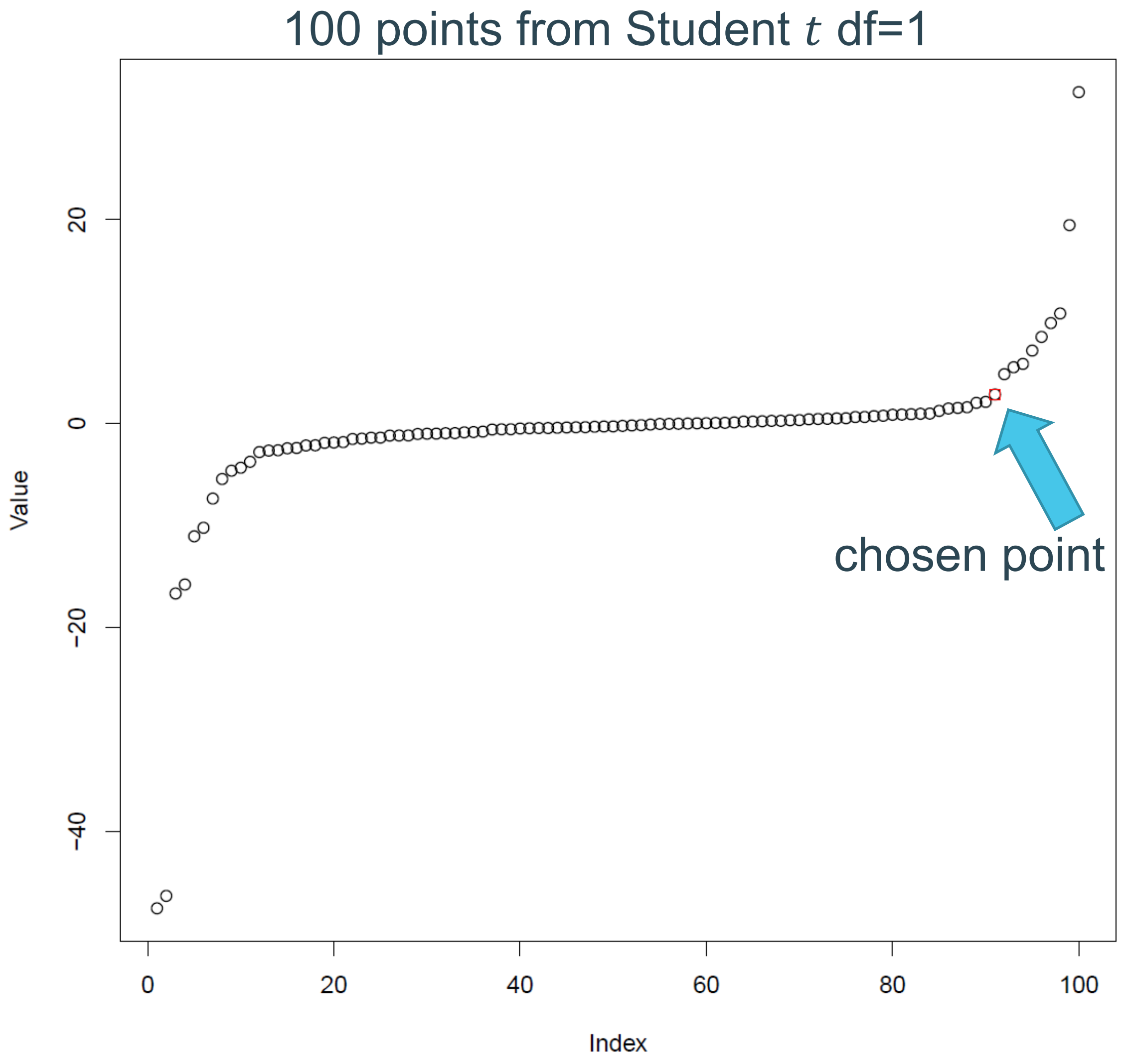
- Let $x_1, \ldots, x_n, x_{n+1} \sim P(\cdot)$ $x_i \in \mathbb{R}$ "exchangeable draws"
- Define $S = \{x_1, \ldots, x_n\}$ "training data"
- Goal:
  - Determine $hi(S)$ such that
    - $\Pr_{x_{n+1} \sim P}[x_{n+1} \leq hi(S)] \geq 1 - \delta$
- Method
  - Let $x_{(1)}, \ldots, x_{(n)}$ be the order statistics (sorted order) of $x_1, \ldots, x_n$
  - $hi(S) \coloneqq x_{(\lceil (1-\delta)(n+1) \rceil)}$

SRI International

Oregon State University

Proof (informal)

- Suppose we computed
  - $x_{(1)}, \ldots, x_{(n)}, x_{(n+1)}$
  - The rank of x_{n+1} will be uniformly distributed within these ranks (exchangeability)
  - The $1-\delta$ quantile will be $x_{(\lceil(1-\delta)(n+1)\rceil)}$
  - $\Pr\left[x_{n+1} \leq x_{(\lceil(1-\delta)(n+1)\rceil)}\right] \geq 1-\delta$
  - Where would the corresponding quantile be in $x_{(1)}, \ldots, x_{(n)}$?
  - At quantile $(1-\delta)\frac{n+1}{n}$, because we now have only n points
  - This will be position $\lceil(1-\delta)(n+1)\rceil$
- This works as long as $\delta \geq \frac{1}{n+1}$



Outline

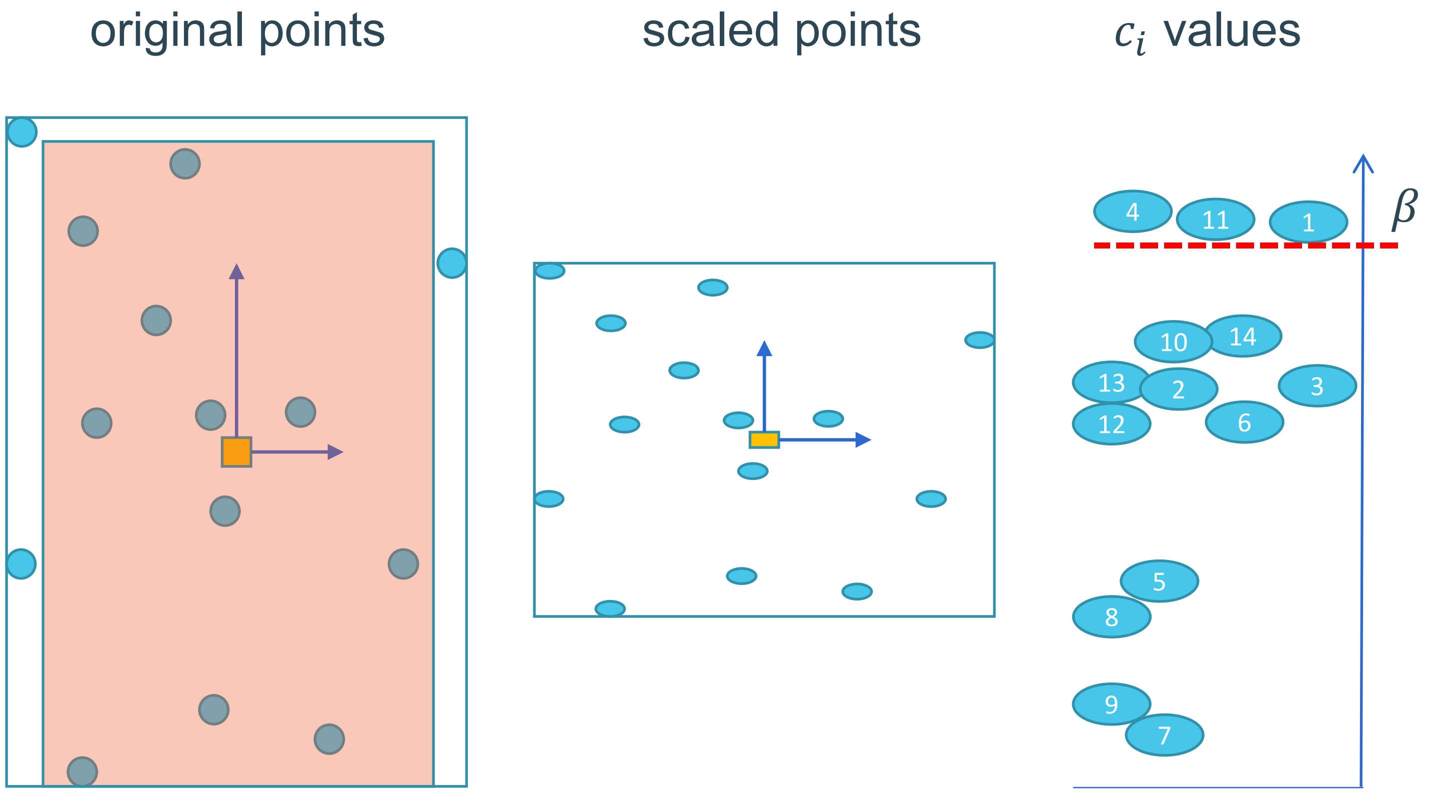
- Background: Conformal Prediction Intervals
- Core Problem: Multivariate Prediction Interval
- Prediction Intervals for MDP Trajectories
- Experimental Results
  - Tamarisk invasions
  - StarCraft battles
- Assessment

Multivariate Prediction Interval

- Given:
  - $D_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \sim P$ iid
  - $D_2 = \boldsymbol{x}_{m+1}, \ldots, \boldsymbol{x}_n \sim P$ iid
  - $\boldsymbol{x}_i \in \mathbb{R}^d \quad \forall i$
  - δ
- Find
  - $\boldsymbol{lo}, \boldsymbol{hi} \in \mathbb{R}^d$
  - $\Pr_{\boldsymbol{x}_{n+1} \sim P}[\boldsymbol{lo} \leq \boldsymbol{x}_{n+1} \leq \boldsymbol{hi}] \geq 1 - \delta$
- Challenge: Avoid thinking of this as d separate prediction intervals, as that requires a Bonferonni correction (or related method)
- Trick: Convert to a one-dimensional problem and apply conformal methods

SDScaledBox

- Compute mean and sample standard deviation from D_1
  - $\hat{\mu}_j \;\; \forall j$
  - $\hat{\sigma}_j \;\; \forall j$
- Proposed prediction interval:
  - $\hat{\mu}_j \pm \beta \hat{\sigma}_j$
  - β is our one-dimensional parameter
- Rescale the D_2 data along each dimension
  - $x'_{ij} := 0$ if $\hat{\sigma}_j = 0$
  - $x'_{ij} := \frac{|x_{ij} - \hat{\mu}_j|}{\hat{\sigma}_j}$ else
- Let $c_i := \max_j x'_{ij}$
  - "widest dimension of standardized $\boldsymbol{x}_i$"
- Sort to obtain $c_{(1)}, \ldots, c_{(14)}$
- $\beta = c_{(\lceil (1-\delta) \cdot 15 \rceil)}$



Theorem 1. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}_{n+1} \in \mathbb{R}^d$ be independent random variables with distribution P. Let $[\boldsymbol{lo}, \boldsymbol{hi}]$ be the multidimensional interval computed by SDScaledBox when applied to $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ with $2 \leq m < n$ and confidence parameter $\delta \in \left[\frac{1}{n-m}, 1\right)$. Then with probability $1 - \delta$

$$\boldsymbol{lo} \leq \boldsymbol{x}_{n+1} \leq \boldsymbol{hi}.$$

Outline

- Background: Conformal Prediction Intervals
- Core Problem: Multivariate Prediction Interval
- Prediction Intervals for MDP Trajectories
- Experimental Results
  - Tamarisk invasions
  - StarCraft battles
- Assessment

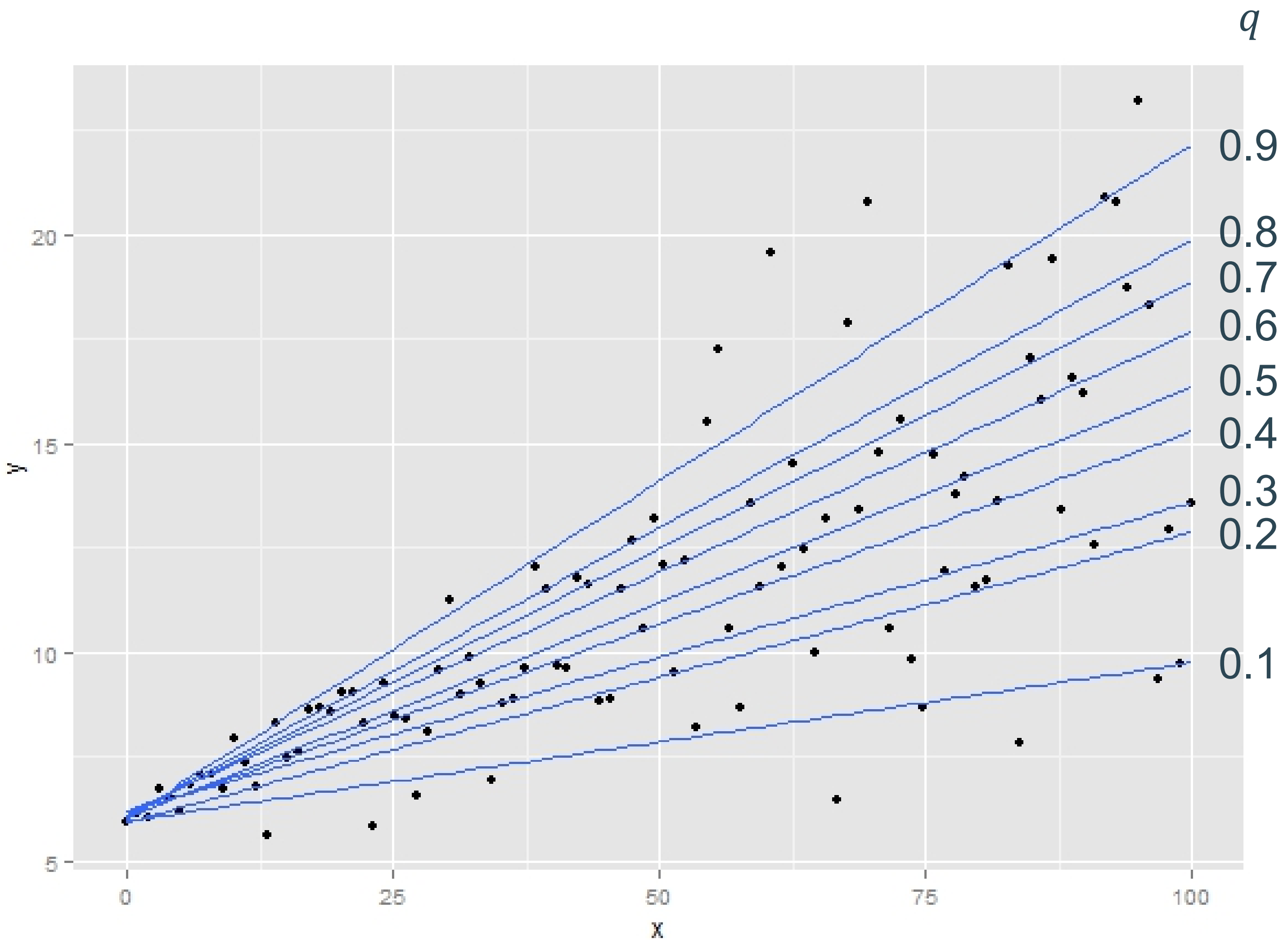
SRI International
Oregon State University

Prospective Prediction Intervals for MDP Policies

- Discrete time MDP with state space $\mathcal{S}$, starting state distribution P_0, and fixed policy π
- h-step trajectory τ
  - sample $s_0 \sim P_0$
  - execute π for h steps
  - collect states, actions, and rewards into τ
- Define a *behavior function* $B(\tau, t)$ to summarize the behavior of the policy at time t
  - some aspect of s_t
  - immediate reward
  - cumulative reward $r_1 + \cdots + r_{t-1}$
  - future reward $r_t + r_{t+1} + \cdots + r_{h-1}$
  - $\boldsymbol{b}(\tau) = \left(b_{\tau,1}, \dots, b_{\tau,h}\right)$ is the "behavior vector" of trajectory τ
- Prospective prediction interval
  - $\boldsymbol{lo}(s_0) \leq \boldsymbol{b}(\tau) \leq \boldsymbol{hi}(s_0)$ with probability $1 - \delta$

Quantile Regression

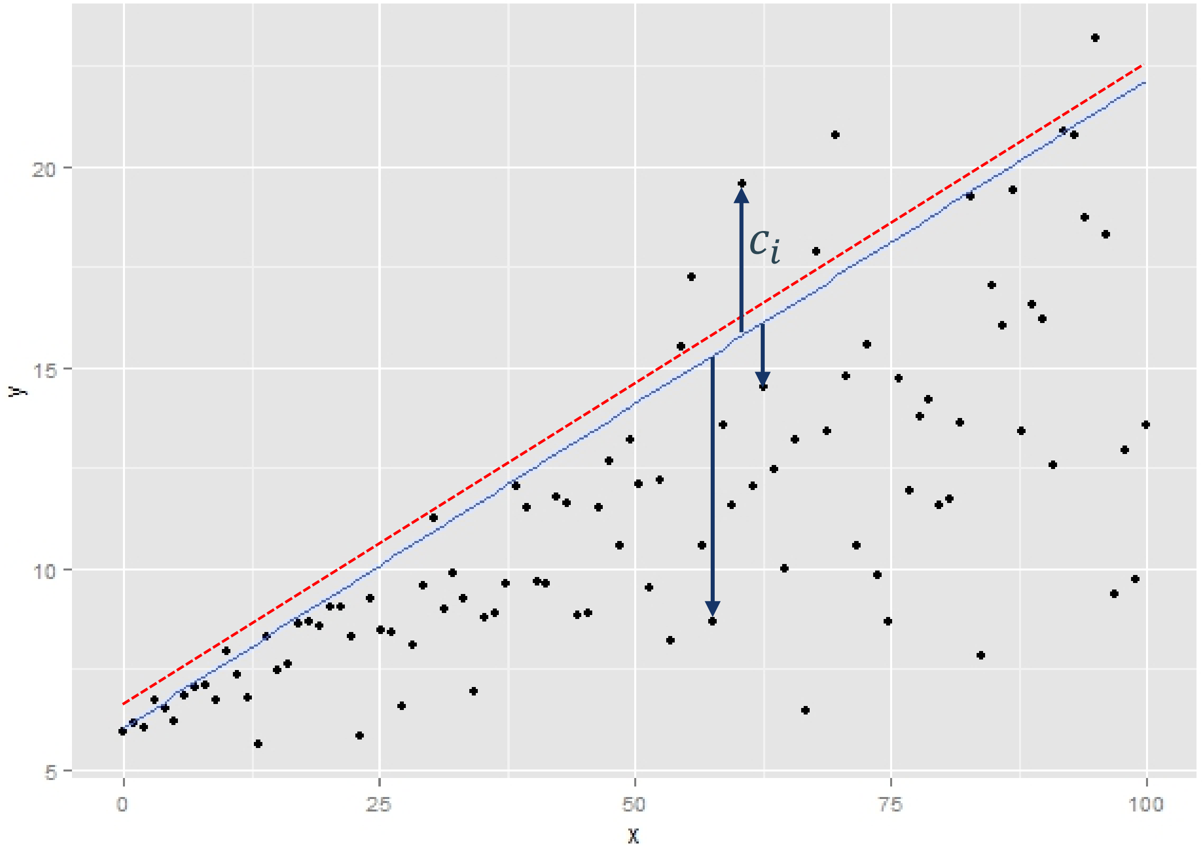
- $P(y|x)$ depends arbitrarily on x
- $F(y|x)$
  - cumulative distribution function of y at x
- $F^{-1}(q|x)$
  - the value of y such that $F(y|x) = q$
- Many algorithms for quantile regression
- We employ Quantile Random Forests (Meinshausen, 2006) to compute the $\delta/2$ and $1\ - \delta/2$ quantiles



https://data.library.virginia.edu/files/qreg_fig_5.jpeg

Quantile Regression with Guarantees

- Romano, Patterson & Candes (NeurIPS 2019) Conformalized Quantile Regression
- Idea: Compute the "error" between the observed values y_i and the predicted quantile $F^{-1}(x_i; q)$ and conformalize to get a "correction"
- Two data sets:
  - D_1: used for fitting $F^{-1}(x; q)$
  - D_2: used for conformalization
- For $(x_i, y_i) \in D_2; \quad i = 1, \dots, n$
  - $c_i \coloneqq y_i - F^{-1}(x_i; q)$
- Sort to obtain $c_{(1)}, \dots, c_{(n)}$
- Bound: $hi(x) \coloneqq F^{-1}(x; q) + c_{(\lceil(1-\delta)(n+1)\rceil)}$
- Let (x_{n+1}, y_{n+1}) be a new data point
  - $c_{n+1} := y_{n+1} - F^{-1}(x, q)$
- Claim: The c_i values are exchangeable ➔ rank of c_{n+1} will be uniformly distributed in $c_{(1)}, \dots, c_{(n+1)}$
- Therefore, $P[c_{n+1} \leq hi(x_{n+1})] \geq 1 - \delta$

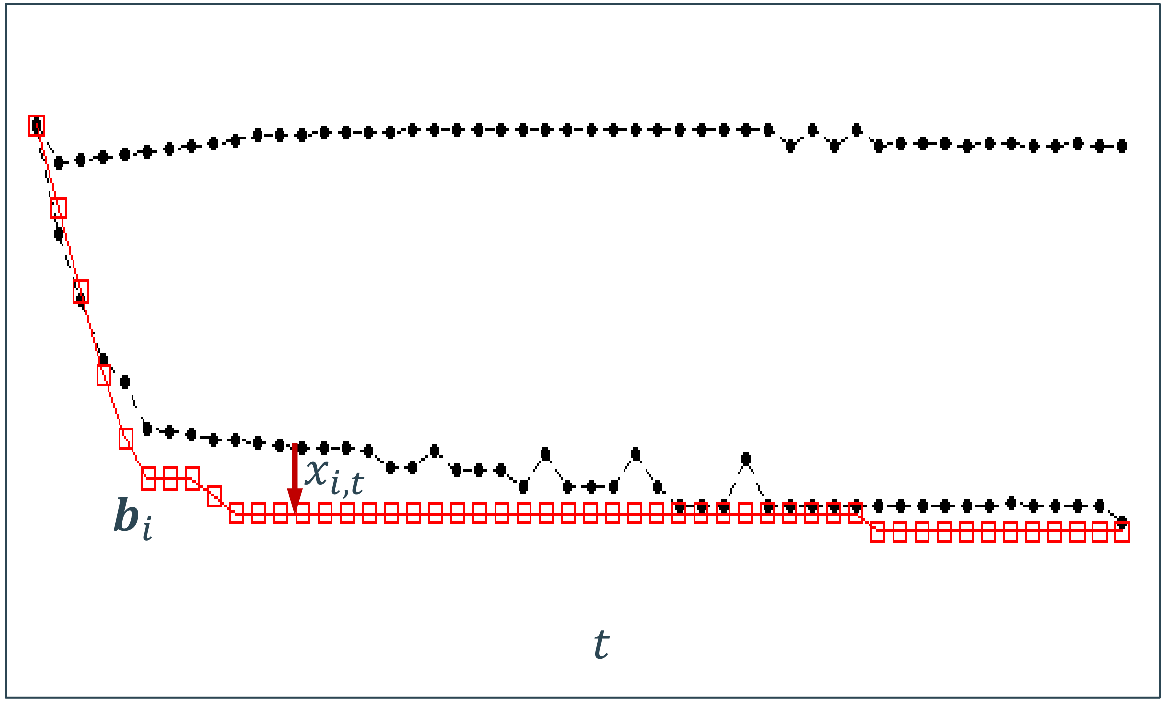


Idea: Extend Conformalized Quantile Regression to Multiple Dimensions

- Three data sets
  - D_1: behavior vectors for quantile regression
  - D_2: behavior vectors for computing $\hat{\sigma}_t$
  - D_3: behavior vectors for computing $\boldsymbol{lo}, \boldsymbol{hi}$
- Plan:
  - Fit quantile regression models $F_t^{-1}(s_0; \delta/2)$ and $F_t^{-1}(s_0; 1 - \delta/2)$ to D_1 for each time step t
  - Compute "exceedances": the amount that each trajectory goes outside the quantile regression prediction
  - Compute $\hat{\sigma}_t$ for the exceedances at time t using D_2
  - Use $\hat{\sigma}_t$ to standardize the exceedances of D_3
  - Compute conformal prediction intervals on the exceedances

Compute “exceedances”

- $x_{i,t} = \max\left(0,\ F_t^{-1}\left(s_0(\tau_i), \frac{\delta}{2}\right) - b_{i,t}, b_{i,t} - F_t^{-1}\left(s_0(\tau_i), 1 - \frac{\delta}{2}\right)\right)$



$F_t^{-1}\left(s_0(\tau_i), 1 - \frac{\delta}{2}\right)$

$F_t^{-1}\left(s_0(\tau_i), \frac{\delta}{2}\right)$

Conformalized Quantile Regression: SDScaledQuantiles

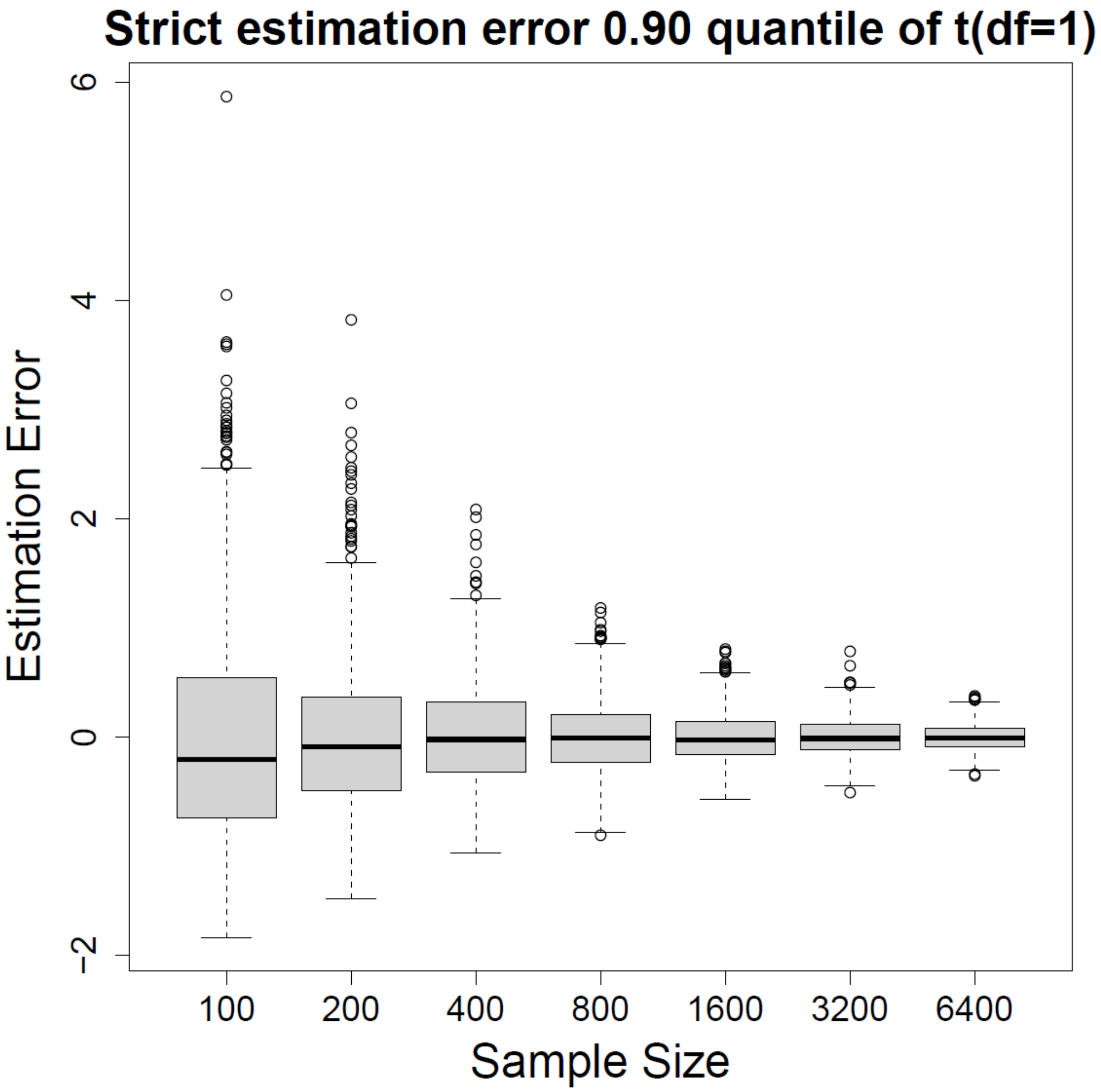
- Given:
  - D_1: behavior vectors for quantile regression
  - D_2: behavior vectors for computing $\hat{\sigma}_t$
  - D_3: behavior vectors for computing $\boldsymbol{lo}, \boldsymbol{hi}$
- Fit quantile regression models $F_t^{-1}(s_0; \delta/2)$ and $F_t^{-1}(s_0; 1 - \delta/2)$ to D_1 for each time step t
- Compute "exceedances": $x_{i,t}$ the amount that trajectory i goes outside the quantile regression prediction at time t
- Compute $\hat{\sigma}_t$ of the exceedances $x_{i,t}$ at time t using D_2
- Rescale exceedances: $x'_{i,t} \coloneqq \frac{x_{i,t}}{\hat{\sigma}_t}$
- Compute c_i for each trajectory in D_3
  - $c_i \coloneqq \max_t x'_{i,t}$
- Compute order statistics $c_{(1)}, \ldots, c_{(n)}$
- $\beta \coloneqq (1 - \delta)\left(\frac{n+1}{n}\right)$ quantile of the c values
- $\boldsymbol{lo}_t \coloneqq F_t^{-1}(s_0; \delta/2) - \beta\hat{\sigma}_{\mathrm{t}}$
- $\boldsymbol{hi}_t \coloneqq F_t^{-1}(s_0; 1 - \delta/2) + \beta\hat{\sigma}_{\mathrm{t}}$

Theorem 2. The behavior vector $\boldsymbol{b}(\tau_{n+1})$ will fall within the prediction interval $[\boldsymbol{lo}, \boldsymbol{hi}]$ returned by SDSCALEDQUANTILES with probability $1 - \delta$

See also: Lei, Rinaldo & Wasserman (2013). Related result for general functional data

Theory and Practice

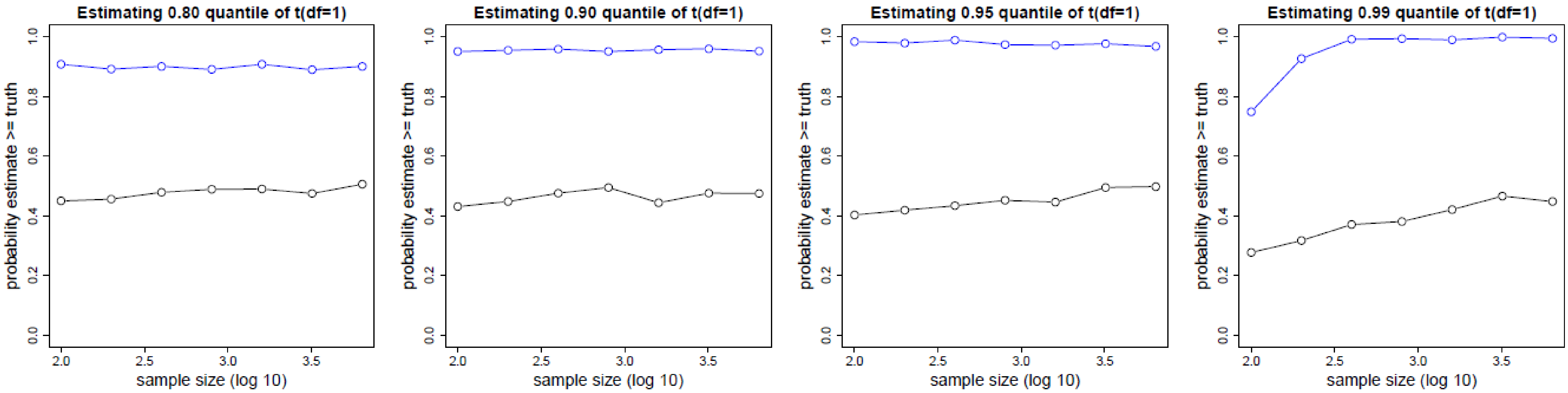
- The $c_{(\lceil (1-\delta)(n+1) \rceil)}$ estimate of the $(1-\delta)\frac{n+1}{n}$ quantile is unbiased but often low for small samples
- We want the estimation error to be ≥ 0 with probability $1-\delta$
- Solution: Use the $1-\delta$ upper bound confidence interval on the target quantile (heuristic)



SRI International

Oregon State University

Quantile Estimation: Strict vs. CI



- Fraction of 1000 trials in which Strict and CI methods exceeded the true target quantile
- CI computed according to Nyblom (1992)

So many quantiles…

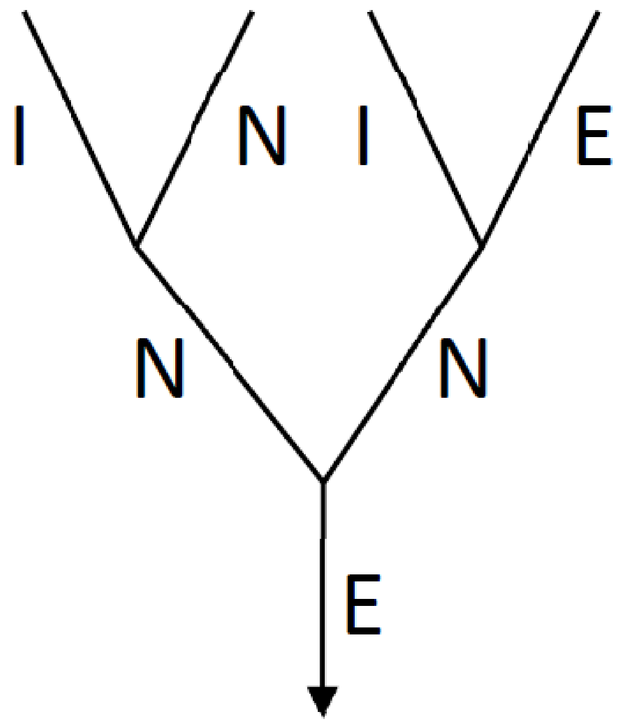
1. Quantile regression
2. $(1-\delta)\frac{n+1}{n}$ quantile of c from conformalization
3. $(1-\delta)$ upper confidence bound on #2

Outline

- Background: Conformal Prediction Intervals
- Core Problem: Multivariate Prediction Interval
- Prediction Intervals for MDP Trajectories
- Experimental Results
  - Tamarisk invasions
  - StarCraft battles
- Assessment

Problem 1: Tamarisk Invasions in River Networks

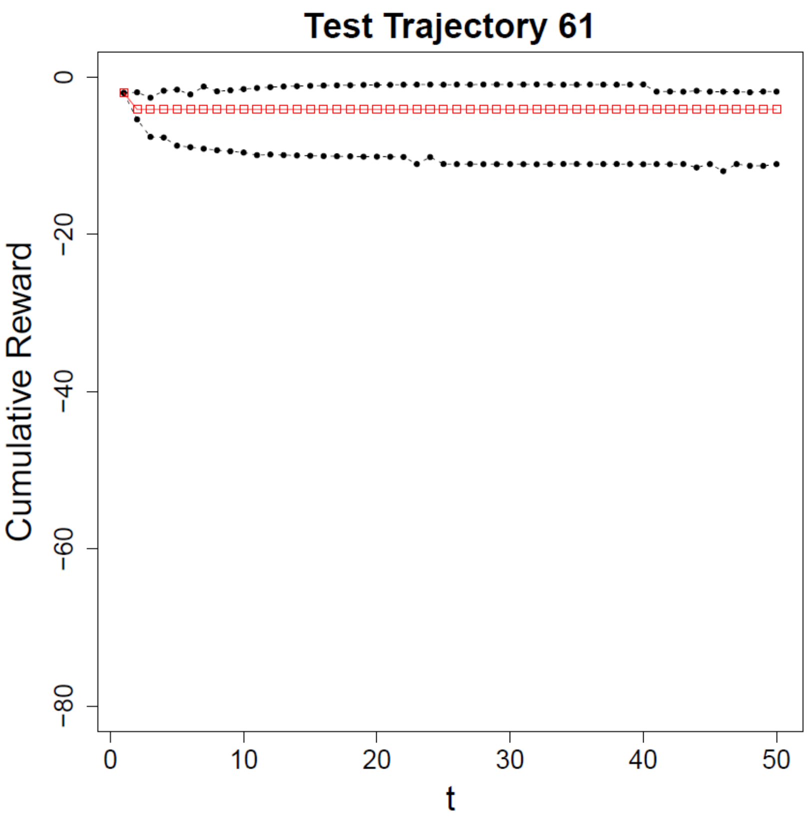
- States:
  - 7 edge river network
  - edge can be
    - I: invaded with tamarisk tree
    - N: occupied by native tree
    - E: empty
- Actions:
  - Plant native
  - Eradicate tamarisk
  - Eradicate + Plant
  - No-Op
- Budget restricts us to one action on one edge per time step
- See Hall, Albers, Alkaee-Taleghan, Dietterich (2018)



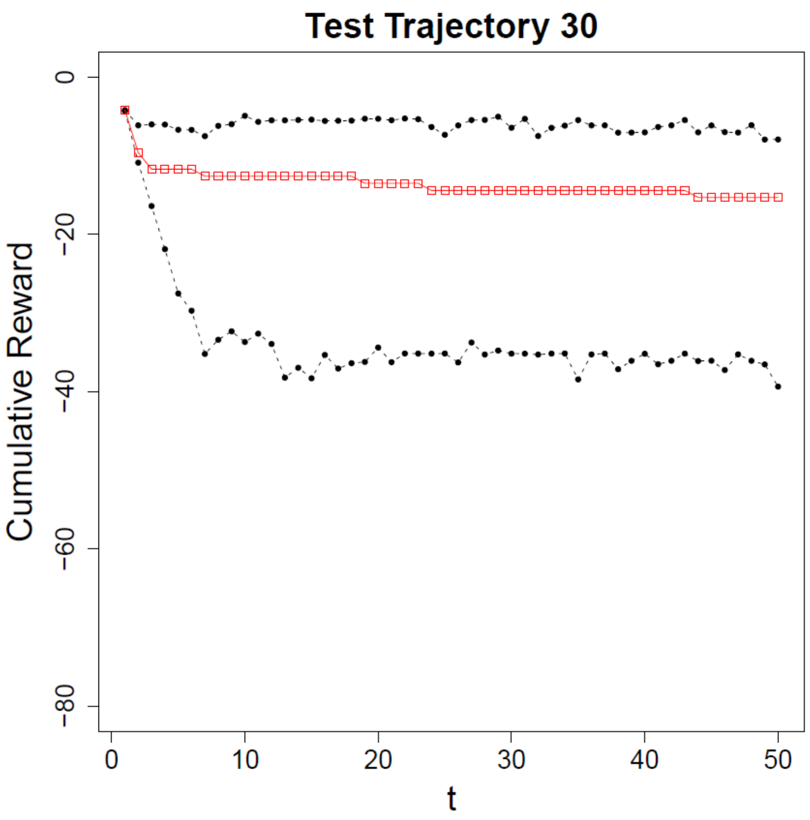
SRI International

Oregon State University

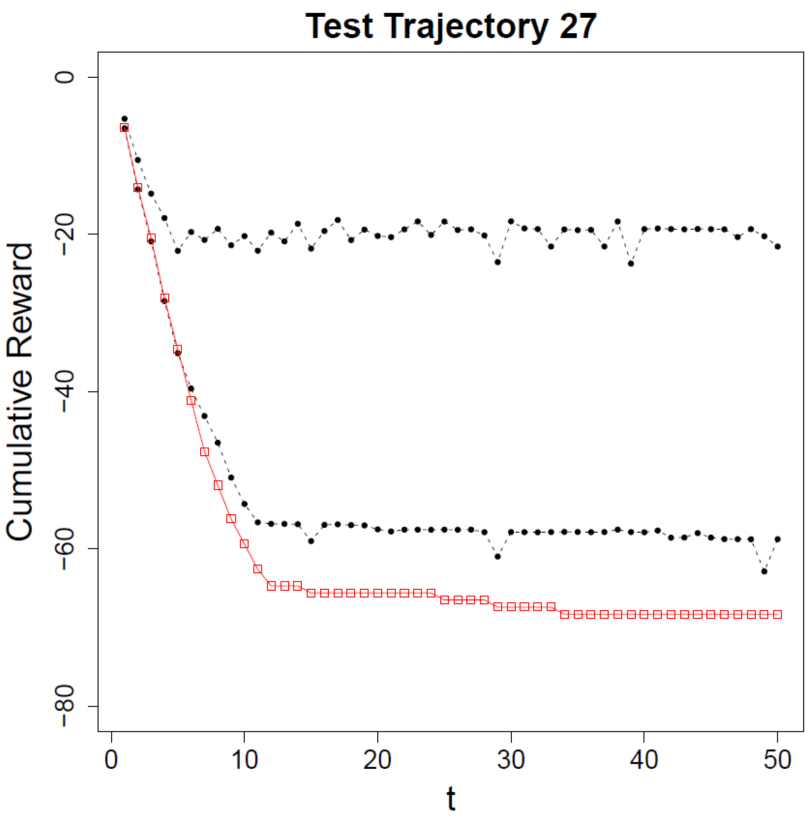
Example Prospective Intervals and Actual Trajectories



$s_0 =$ EEENENT



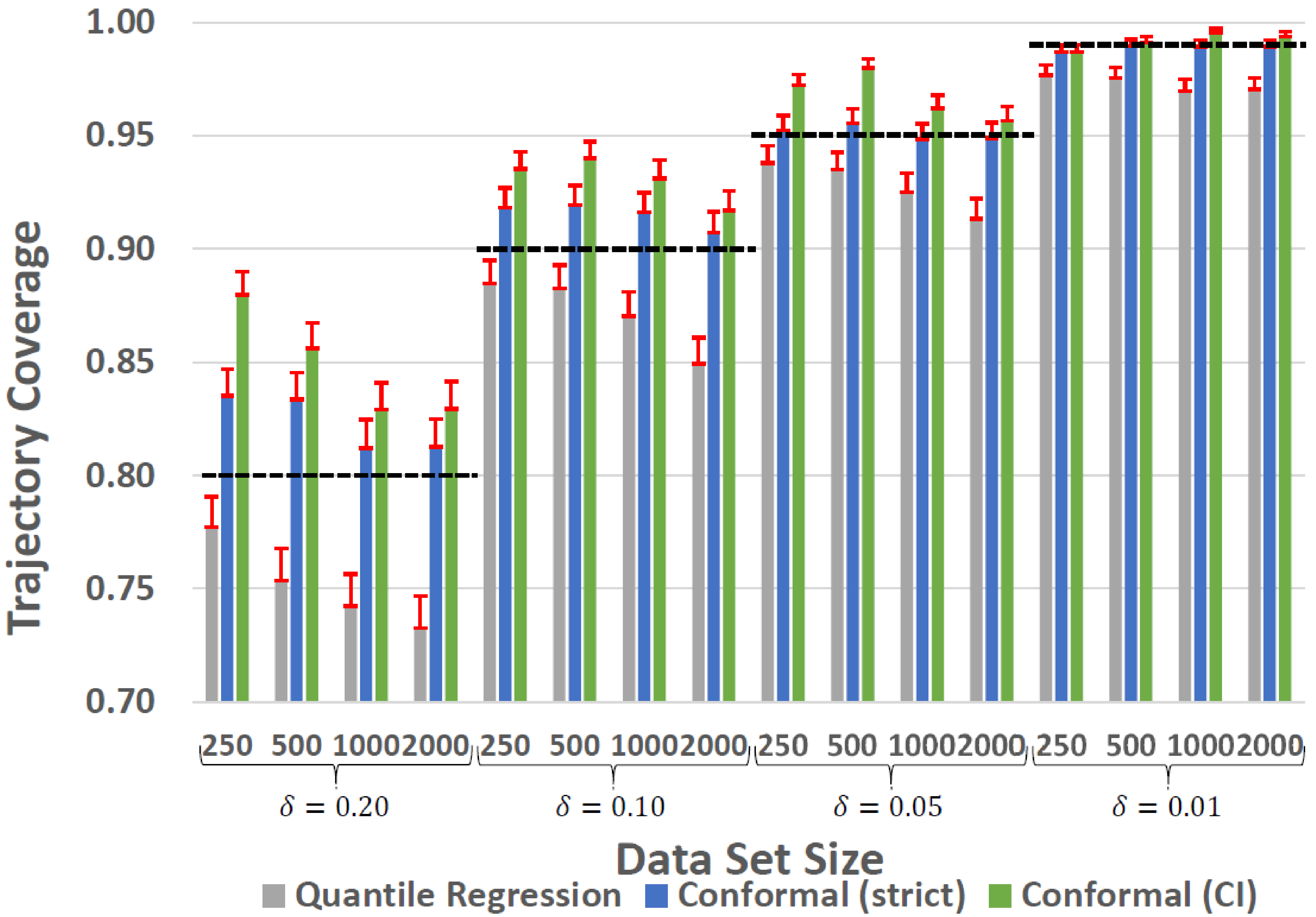
$s_0 =$ ETETNTE



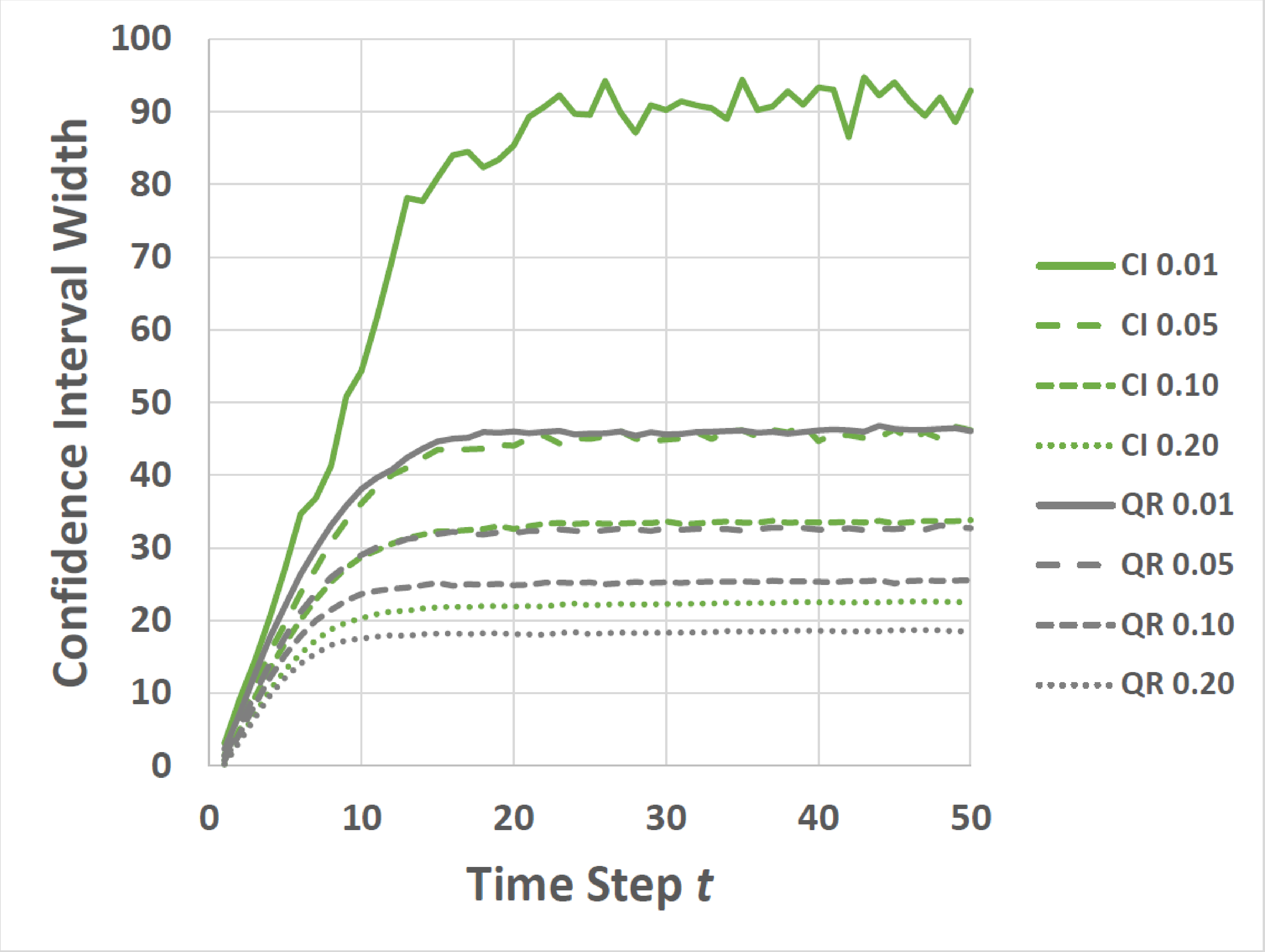
$s_0 =$ TETTTET

Tamarisk Prediction Interval Coverage

Raw QR: 0/16
Strict: 16/16
CI: 16/16



Prediction Interval Widths



MDP 2: Starcraft Battles

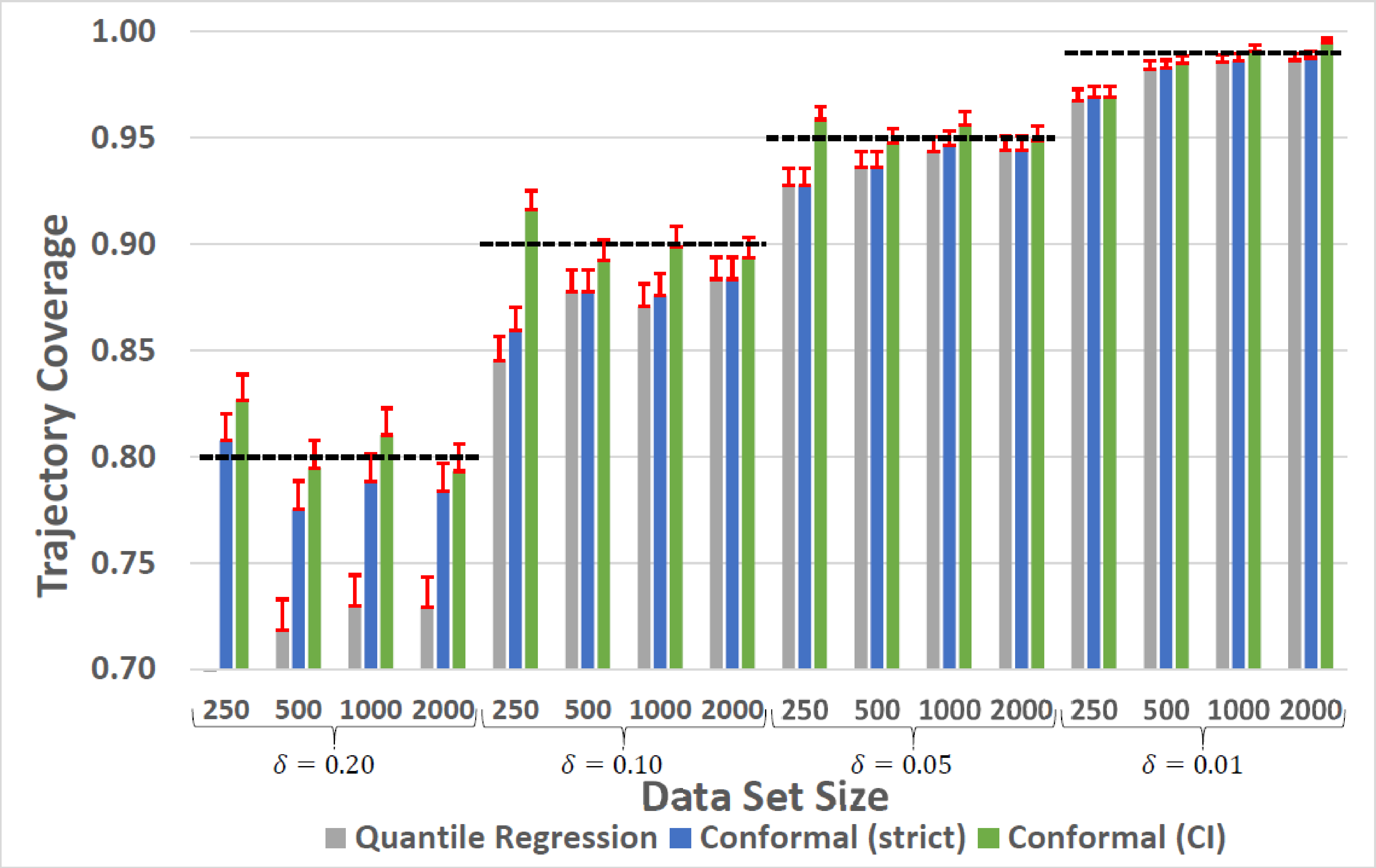
- Blue units: $\text{unif}(5, 20)$
- Red units: $\text{unif}(5, 10)$
- At time 14, Red receives reinforcements $\text{unif}(0, N)$ where $N \sim \text{unif}(0,15)$

SRI International®

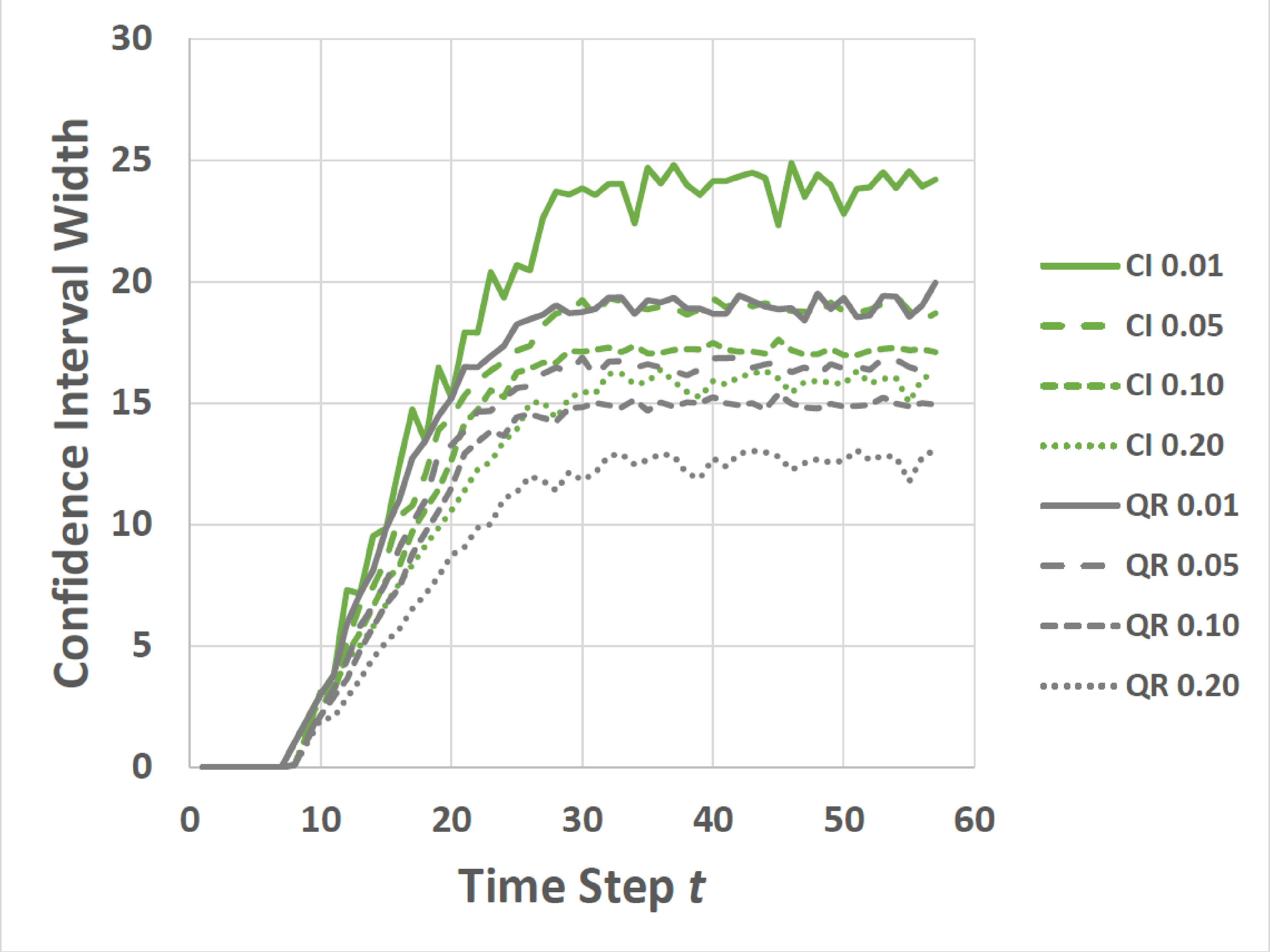
Oregon State University

Starcraft Prediction Interval Coverage

Raw QR: 2/16
Strict: 5/16
CI: 14/16



Starcraft Prediction Interval Widths



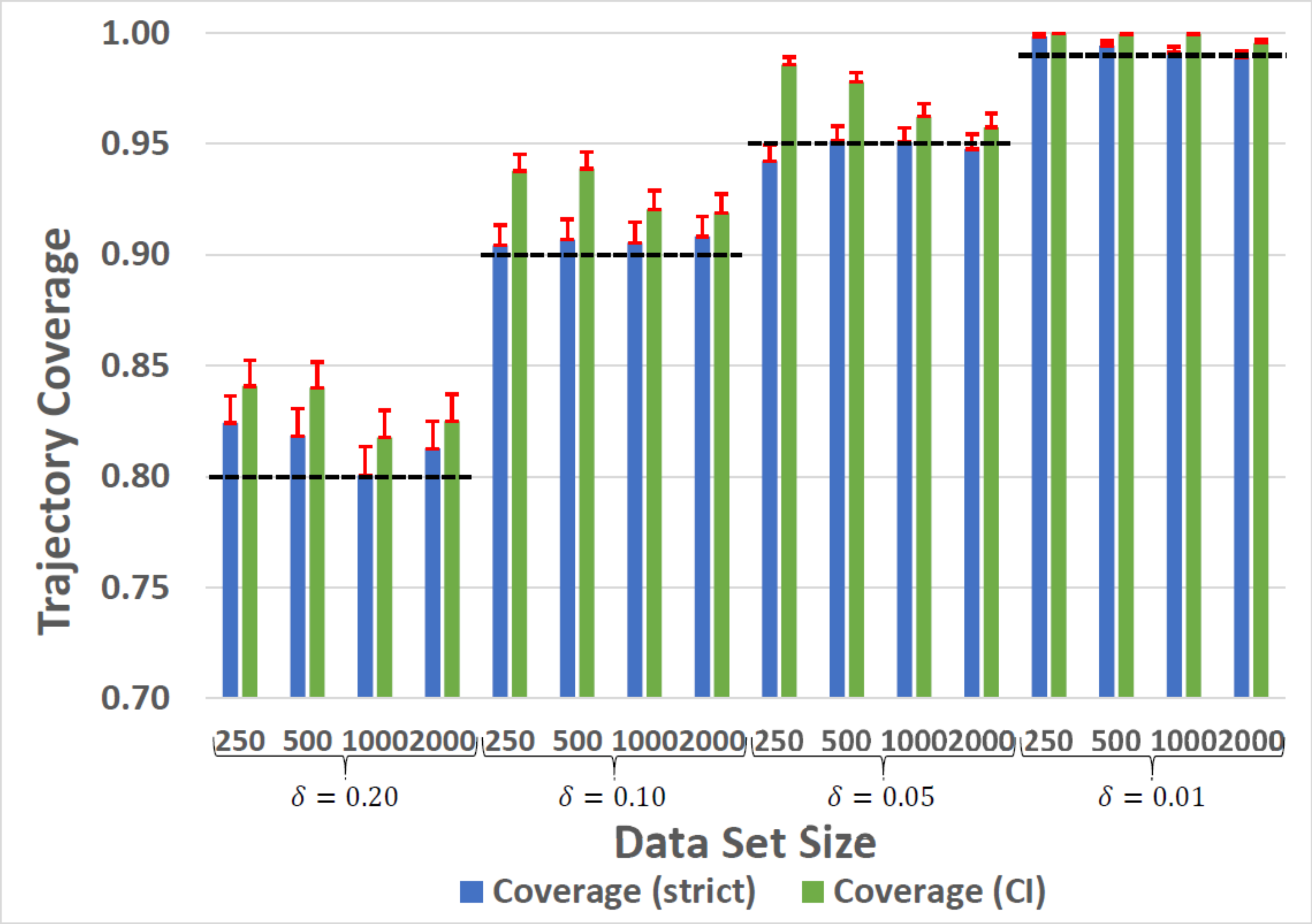
SRI International
Oregon State University

An Alternative: Total Exceedance Bound

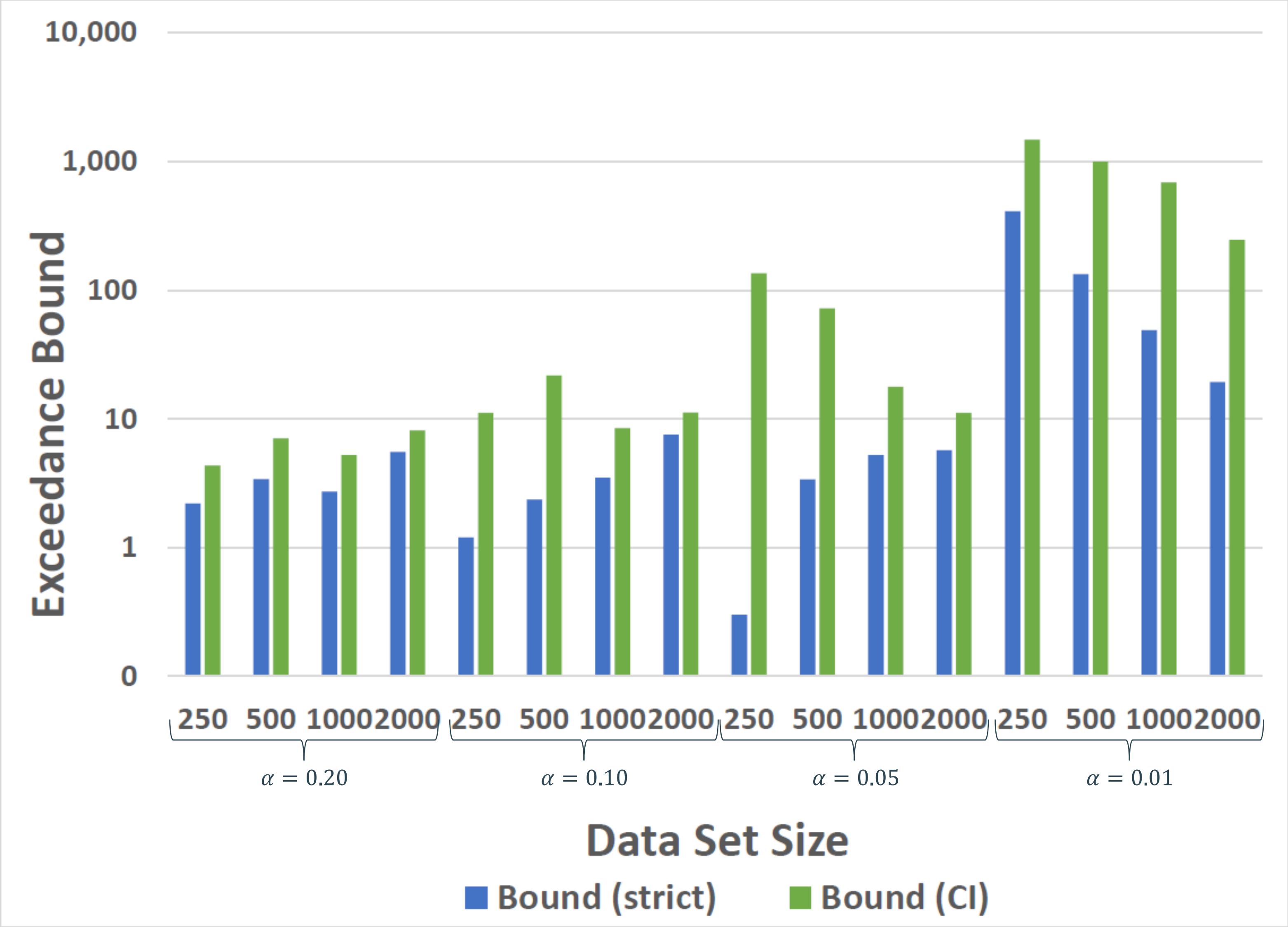
- Guarantee: With probability $1 - \delta$ the total exceedance along the trajectory will be T
- Compute the Quantile Regression predictions
- Let $c_i =$ the *total exceedance* of each trajectory in D_3
- Sort to obtain $c_{(1)}, \ldots, c_{(n)}$
- Compute the $\left\lceil (1 - \delta) \left(\frac{n+1}{n}\right) \right\rceil$ quantile of these as the upper bound

Tamarisk Total Exceedance Coverage

Both methods achieve target coverage in all cases



Tamarisk Total Exceedance Bounds

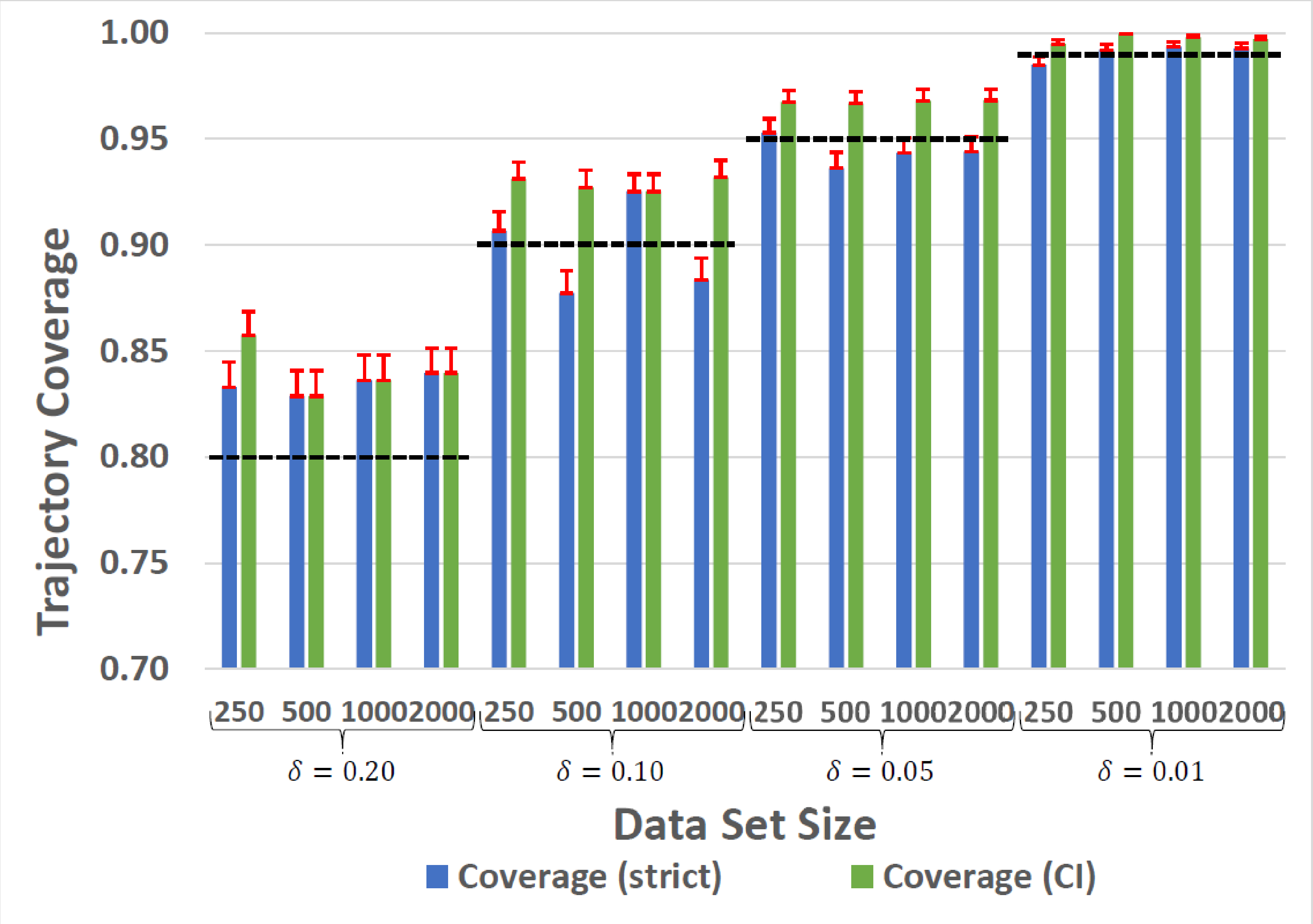


SRI International

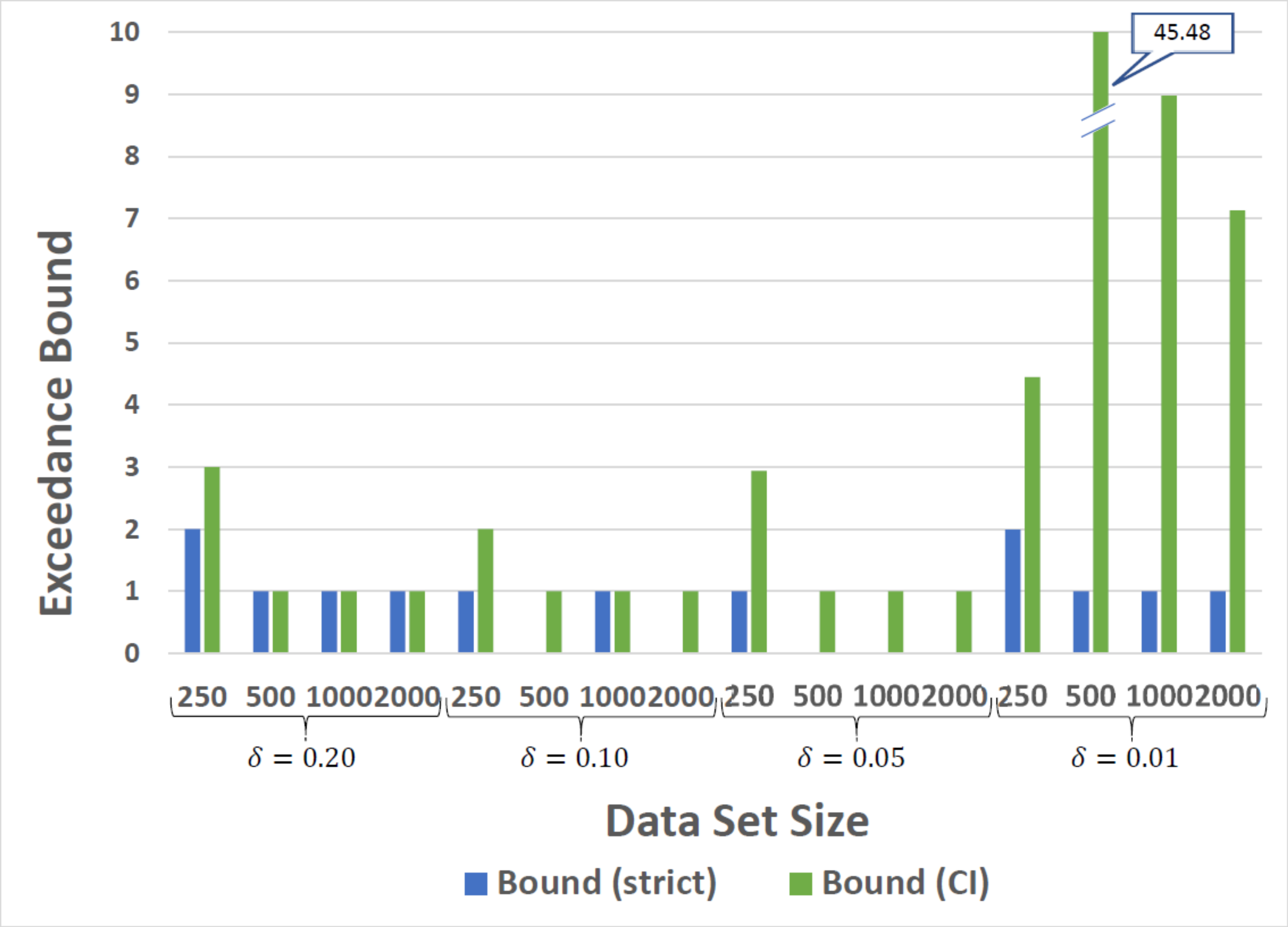
Oregon State University

Starcraft Total Exceedance Coverage

Strict method fails 3 times
CI method covers all 16 cases



Starcraft Total Exceedance Bounds



Summary

- Conformal Prediction Intervals for d-dimensional data
- Trajectory-wise Prediction Intervals
- Excellent performance on two MDPs
- Alternative of Total Exceedance Bounds is ok for Starcraft but not for Tamarisk

Assessment

- The guarantees are semi-conditional
  - The quantile regressions are conditioned on s_0
  - The conformal corrections are unconditional ("marginal") and are taken over P_0
- If the failures are scattered throughout the state space, this is not a serious issue
- But if the failures are concentrated in one region, then the claim is misleading
- Additional techniques are required to address this shortcoming

Acknowledgements

- Funding:
  - Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001119C0112
  - US National Science Foundation under Grant Nos. 0832804, 1331932, and 1521687
- Collaboration and feedback
  - Kiri Wagstaff, Si Liu, Kim Meyer-Hall, Majid Alkaee-Taleghan, H. Jo Albers
- Disclaimer:
  - This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001119C0112. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA

References

- Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2019). The limits of distribution-free conditional predictive inference. *ArXiv*, *1903.04684*, 1–34. http://arxiv.org/abs/1903.04684
- Hall, K. M., Albers, H. J., Alkaee Taleghan, M., & Dietterich, T. G. (2018). Optimal Spatial-Dynamic Management of Stochastic Species Invasions. *Environmental and Resource Economics*, *70*(2), 403–427. https://doi.org/10.1007/s10640-017-0127-6
- Lei, J., Rinaldo, A., & Wasserman, L. (2013). A Conformal Prediction Approach to Explore Functional Data. Annals of Mathematics and Artificial Intelligence, 74(1), 23–43. http://arxiv.org/abs/1302.6452
- Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*, 983–999.
- Nyblom, J. (1992). Note on interpolated order statistics. *Statistics and Probability Letters*, *14*, 129–131.
- Romano, Y., Patterson, E., & Candès, E. J. (2019). Conformalized Quantile Regression. http://arxiv.org/abs/1905.03222; NeurIPS 2019
- Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko, Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. Starcraft II: A new challenge for reinforcement learning. arXiv, 1708.04782, 2017.
- Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.