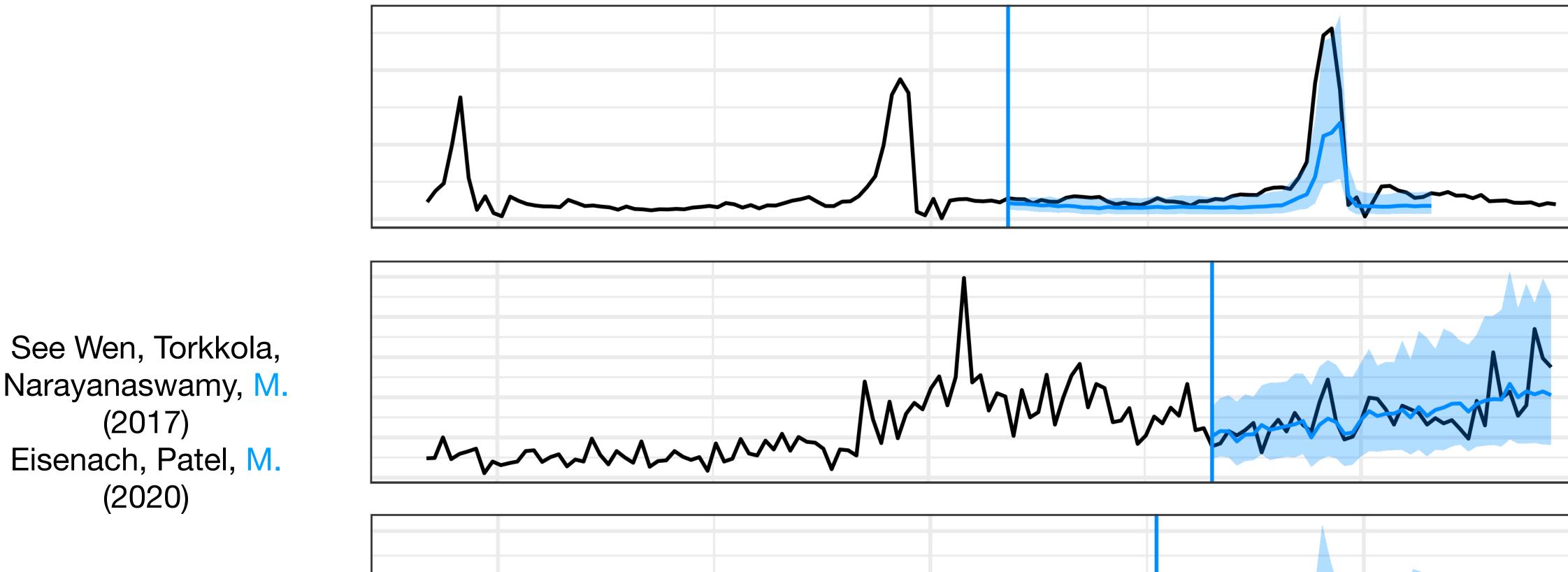
Reinforcement Learning for Supply Chains

Dhruv Madeka*, Dean Foster*, Sham M. Kakade^* *Amazon **^Harvard University**



First, let's look at supervised learning at Amazon.

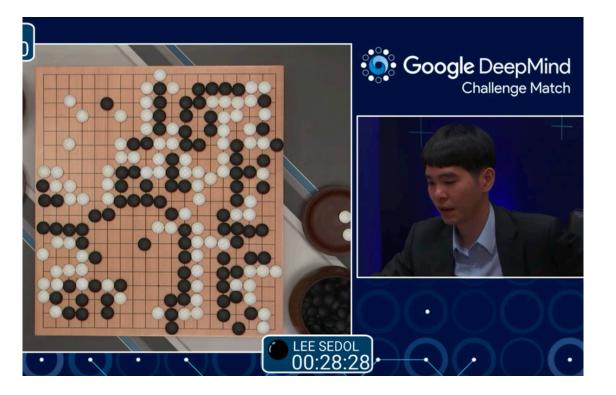


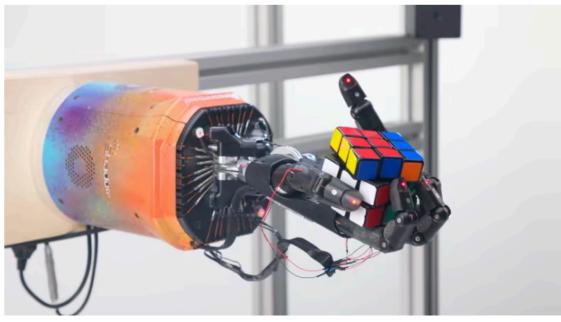
- Finite hypothesis class: need $\mathcal{O}(\log | \Theta)$ Supervised Learning: We can generalize from iid data
- Data reuse: We can compute the loss of every function in a hypothesis class

_	
	ł
	I
	4
	1

_	







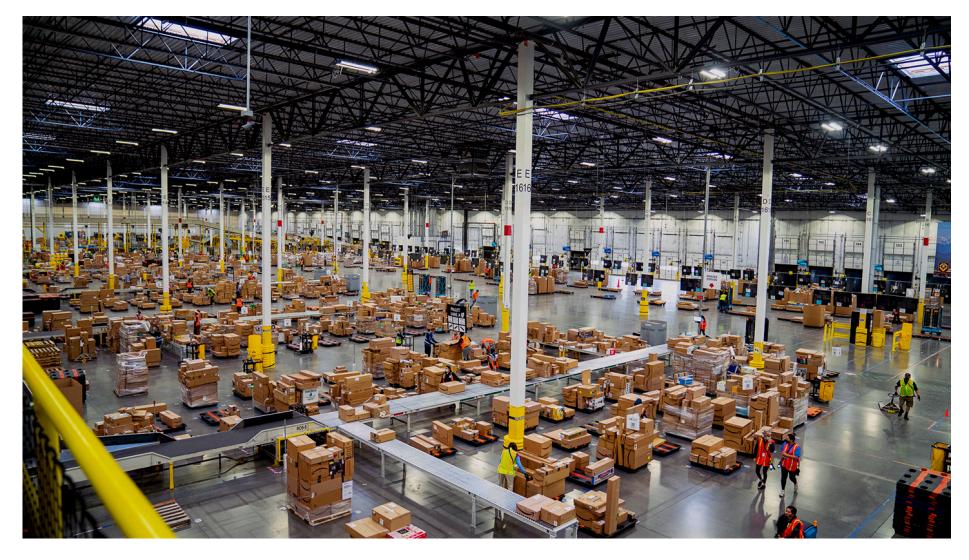




The core challenges Amazon faces are sequential decision making problems.

Can RL help in this space?

Real-world RL is hard.



Deliver to Sham prime © Boston 02118	Rello, Sham ■ EN → Account & Lis	Re sts - &
🗮 All Early Black Friday Deals Holiday Gift Guide Clinic Ar	mazon Basics Customer Service Best Sellers Buy Again Prime Video Pet Supplies	Sho
All Electronics Deals Best Sellers TV & Video Audio & Home Th	neater Computers Camera & Photo Wearable Technology Car Electronics & GPS Portable Audio	Cell Phone
Hisense Roku TV	Hisense 40-Inch Class H4 Series LED Roku Smart TV with 🔶 🚖 🚖 🏫 🏌 8,970	
Electronics > Television & Video > Televisions > OLED TVs		
Share	Edition Smart TV OLED77C2PUA, 2022 - Al-	12 monthl \$208.09/n (\$2,496.99
	Powered 4K TV, Alexa Built-in Visit the LG Store	One-time
	$\Rightarrow \Rightarrow $	^{\$} 2,49 <mark>√prime</mark> s
	-5% \$ 2,496 ⁹⁹	FREE Insid delivery as Saturday, 9 AM - 12
	or 12 monthly payments ~ of \$208.09 i Get 10% back on amount charged to an Amazon Prime credit card. Learn	Ships f more
6 VIDEOS	Size: 77 inch	02118
Roll over image to zoom in	42 inch 48 inch 55 inch 65 inch 77 inch 83 inch	In Stock
	Style: TV Only	Qty: 1 🗸
Product Energy Guide	TV + S65Q TV + S75Q TV + S80QY TV Only TV + S90QY	А
	TV wall mounting options: Get expert TV wall mounting Details	
	Without expert wall mounting +\$200.00 per unit Expert wall mounting +\$200.00 per unit Image: Comparison of the second se	Secure
	~ What's included	Sold by





99 / 12 mo.) e paymer **)6**99 le Entrywa November 26 from nearby Lea r to Sham - Boston Add to Cart Buy Now re transaction

Amazon.com Amazon.com Packaging Shows what's insi..

RL is hard!

• Sample complexity can be as large as min($|\Theta|, 2^T$)

Large state/action spaces

Exploration

Credit assignment problem

Dexterous Robotic Hand Manipulation <u>OpenAl, '19</u>



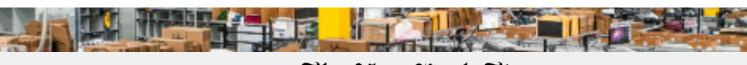
The Supply Chain Problem

- Supply Chain is about buying, storing, and transporting goods.
- Amazon has been running it's Supply Chain for decades now
 - There is a lot of historical "off-policy" data
 - How do we use it?
 - Concern: counterfactual issue?
- This talk: how can we use this data to solve the inventory management problem?

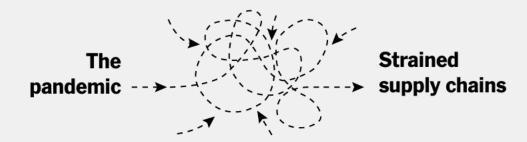
The New York Times

Supply Chain Hurdles Will Outlast Pandemic, White House Says

The administration's economic advisers see climate change and other factors complicating global trade patterns for years to come.



The New York Times



How the Supply Chain Crisis Unfolded



Can we use historical data to solve inventory management problems in supply chain?

- Part I: Utilizing Historical Data
- Part II: Moving to real-world inventory management problems
- Part III: Real World Results

Outline

Deep Inventory Management

Dhruv Madeka Amazon, maded@amazon.com

Kari Torkkola Amazon, karito@amazon.com

Carson Eisenach Amazon, ceisen@amazon.com

Anna Luo Pinterest*, annaluo676@gmail.com

Dean P. Foster Amazon, foster@amazon.com

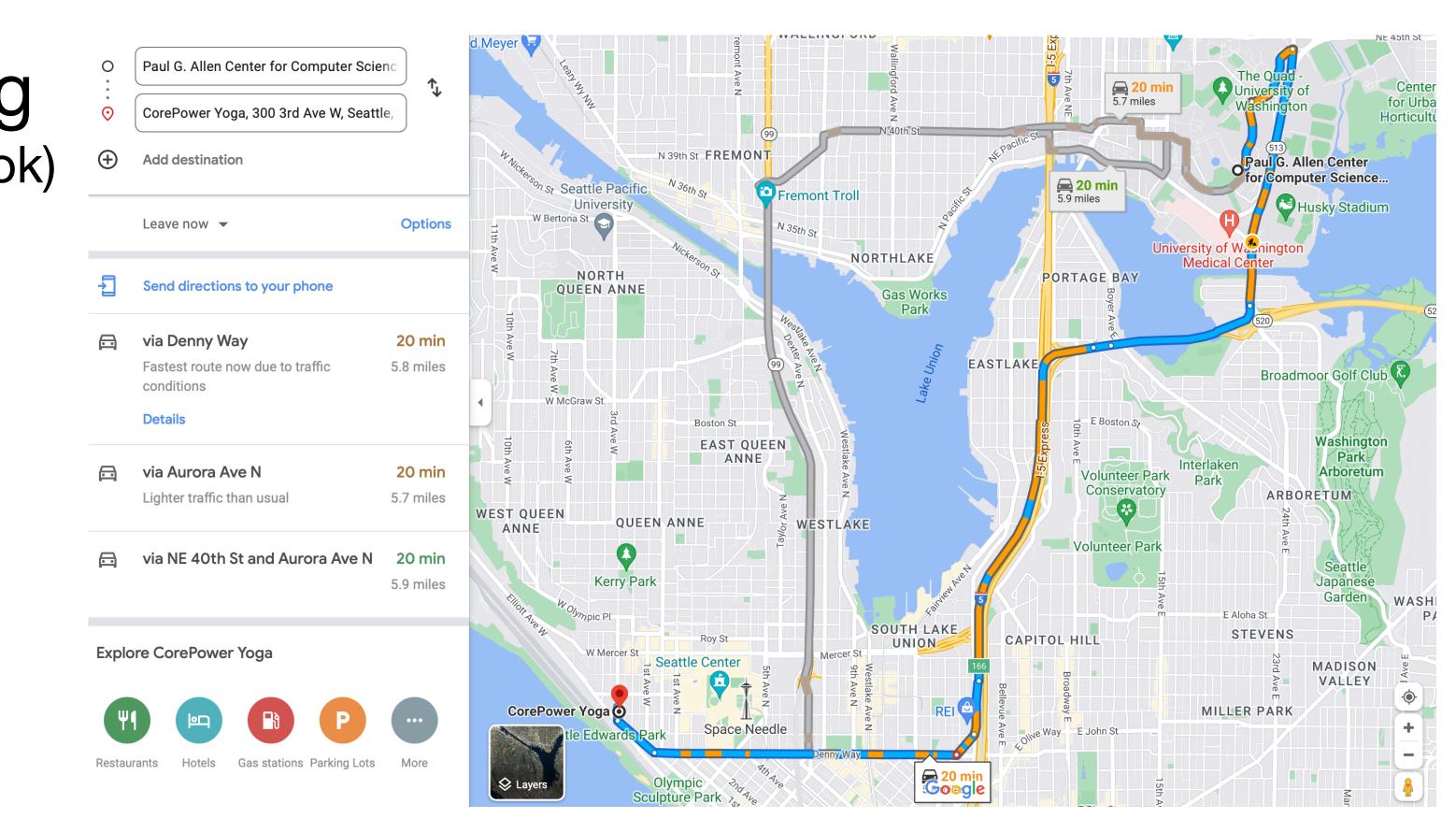
Sham M. Kakade Amazon, Harvard University, shamisme@amazon.com

Largely based on this paper: <u>arxiv/2210.03137</u>



I: Utilizing historical data

- Warm up: Vehicle Routing (when using historical data might be ok)
- We want a good policy for routing a single car.
- Policy π : features -> directions features: time of day, holiday indicators, current traffic, sports games, accidents, location, weather,



Historical Data:

suppose we have logged historical data of features

- Backtesting policies:
 - Key idea: a single route minimally affects traffic
 - Counterfactual: with the historical data, we can see what would have happened with ulletanother policy.

Warm up 2: Fleet Routing

- We want to route a whole fleet of self-driving taxis.
- Policy π : features -> directions
 - features: customer demand, time of day, holiday indicators, current traffic, sports games, accidents, location, weather...
- Historical Data:

suppose we have logged historical data of features

- Backtesting policies:
 - Key idea: a small fleet route may have small affects on traffic. Counterfactual: with the historical data, we can see what would have happened with
 - another policy.



Supply Chain Data

Time	Inventory	Inventory Demand		Revenue
0	100	20	_	40
0	80	_	10	-10
1	90	20	_	40
1	70	_	50	-50
2	120	60	_	120
2	60		10	-10

Price= \$2 Cost= \$1

Backtesting a policy

Time	Inventory	Demand	Order	Revenue
0	100	20	_	40
0	80	Ι	10-40	<u> </u>
1	<u> </u>	20	_	40
1	-70- <i>100</i>	_	<u>-50</u> -20	<u> </u>
2	120	60		120
2	60	-	10	-10

Price =Cost= \$1

• Current order doesn't impact future demand.

- This allows us to backtest!
- Empirically, backlog due to unmet demand does not look significant.¹



Formalization of the Supply Chain Problem

- process [Efroni et al 2021, Sinclair et al 2022]
- A formalization of the model:
 - Action a_{f} : how much you buy ullet
 - Exogenous random variables: evolving under Pr and not dependent on our actions $(Demand_t, Price_t, Cost_t, Lead Time_t, Covariates_t) := s_t$
 - Controllable part (inventory) I_t : evolution is dependent on our action.
 - $I_t = \max(I_{t-1} + a_{t-1} D_t, 0)$ (and suppose we start at I_0).
 - Reward is just the sum of profits: $r(s_t, I_t, a_t) := \text{Price}_t \times \min(\text{Demand}_t, I_t) \text{Cost}_t \times a_t$ ullet
- Learning setting:
 - Data collection: We observe N historical trajectories, where each sequence is sampled $s_1, \ldots, s_T \sim \Pr$ Goal: maximize our rewards cumulative reward over T periods
 - lacksquare

 $V_T(\pi) = E$

Growing literature around a class of MDPs where a large part of the state is driven by an exogenous noise

$$E_{\pi}\left[\sum_{t=1}^{T} \gamma^{t} r(s_{t}, I_{t}, a_{t})\right]$$





Why is it an interesting RL problem?

- Lots of time dependence!
 - If you buy too much, you're left with the inventory for months!
 - Your actions (orders) affect the state at a random time later
 - Tons of correlation across time (Demand, Price, Cost)
- We can explore (linear risk in every product)

Theorem: Backtesting in ExoMDPs

Theorem [M., Torkkola, Eisenach, Luo, Foster, Kakade '22]: Suppose we have a set of K policies $\Pi = \{\pi_1, \dots, \pi_K\}$, and we have N sampled exogenous paths. Then we can accurately backtest up to nearly $K \approx 2^N$ policies. Formally, for any $\delta \in (0,1)$, with probability greater than $1 - \delta$ - we have that for all $\pi \in \Pi$: $|V_T(\pi) - \hat{V}_T(\pi)|$ (assuming the reward r_t is bounded by 1).

- Implications:
 - We can optimize a neural policy on the past data.
 - $min\{2^T, K\}$, using historical data due to the counterfactual issue.

$$|_{T}(\pi)| \leq T \sqrt{\frac{\log(K/\delta)}{N}}$$

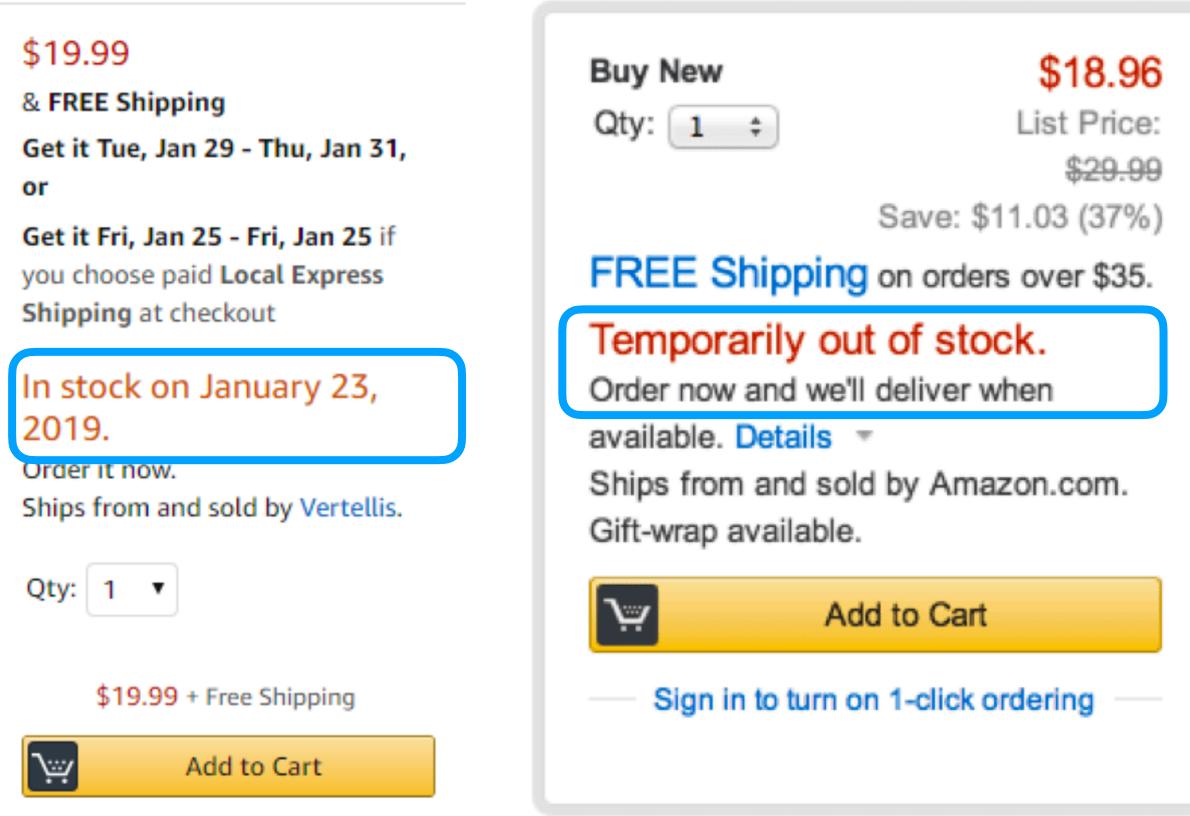
In the usual RL setting (not exogenous), we would have an amplification factor of (at least)



II: Real World Inventory Management Problems

Real-world Issue: Censored Demand

• When demand \geq inventory, what customers see:



We only observe sales not the demand: **Sales** := min(**Demand**, **Inventory**)

Can we still backtest?



Our historical data is then censored....

Sales := min(Demand, Inventory)

Time	Inventory	True Demand	Sales	Order	Revenue
Т	10	??	10	_	20
•	•	\$19.99 & FREE Shipping Get it Tue, Jan 29 - 1	Thu, Jan 31,	Buy New Qty: 1 ‡	\$18.96 List Price:
•	•	or Get it Fri, Jan 25 - Fr you choose paid Loc Shipping at checkou	ri, Jan 25 if al Express		\$29.99 Save: \$11.03 (37%) on orders over \$35.
•	•	In stock on Janu 2019. Order it now. Ships from and sold		Temporarily ou Order now and we'l available. Details Ships from and sol	d by Amazon.com.
	•	Qty: 1			dd to Cart
		\$19.99 + Free 9		Sign in to turn o	on 1-click ordering

Price= \$2 Cost= \$1

If we could fill in the missing demand, then we could still backtest!



We have many observed historical covariates

- Covariates: Sales, Web Site, Glance Views, Product Text, Reviews
- Example: the #times customers look at an item gives us info about the unobserved demand.

 Let's forecast the missing variables from the observed covariates! $\hat{\mathbb{P}}(\text{Missing Data} | \text{Observed Data})$

Buy New Qty: 1 ‡	\$18.96 List Price: \$29.99 Save: \$11.03 (37%)		
FREE Shippin	g on orders over \$35.		
Temporarily out of stock. Order now and we'll deliver when available. Details Ships from and sold by Amazon.com. Gift-wrap available.			
ب ې ۸	dd to Cart		
Sign in to turn	on 1-click ordering		



Uncensoring the data...

Sales := min(Demand, Inventory)

Time	Inventory	True Demand	Sales	Order	Revenue
Т	10	40	10	_	20
•		•	•	Buy New Qty: 1 ‡	\$18.96 List Price:
•		•	:	FREE Shipping	\$29.99 Save: \$11.03 (37%) on orders over \$35.
•		•	:	Temporarily ou Order now and we'l available. Details	I deliver when
•				Ships from and sol Gift-wrap available.	-
			•		on 1-click ordering



Price = \$2Cost= \$1

Key idea: **Use covariates** (e.g. glance views) to forecast missing demand, vendor lead times, etc



Theorem: Backtesting in Uncensored ExoMDPs

Theorem [M., Torkkola, Eisenach, Luo, Foster, Kakade 22]: If we can forecast the missing variables accurately (in a total variation sense), then we can backtest accurately. More formally,

Setting: we have N sampled sequences $\{s_1^i, s_2^i, \dots, s_T^i\}_{i=1}^N$, where M_i and O_i are the missing and observed exogenous variables in sequence *i*.

Forecast: $\widehat{\mathbb{P}}^{i} = \widehat{\Pr}(M_{i} | O_{i})$ is our forecast of $\mathbb{P}^{i} =$

Assume: With pr. 1, forecasting has low error:

Guarantee: For any $\delta \in (0,1)$, with pr. greater than

 $|V_T(\pi) - \hat{V}_T(\pi)|$

Key idea: We can backtest even in the censored setting!

$$\begin{aligned} &\Pr(M_i \mid O_i). \\ & \frac{1}{N} \sum_{i=1}^{N} \operatorname{TotalVar} \left(\mathbb{P}^i, \ \widehat{\mathbb{P}}^i \right) \leq \epsilon_{\sup} \\ & \text{for all } \pi \in \Pi: \\ & \leq T \left(\epsilon_{\sup} + \sqrt{\frac{\log(K/\delta)}{N}} \right) \end{aligned}$$



III: Training Policies & Empirical Results

The Simulator

- Collection of historical trajectories:
 - 1 million products
 - 104 weeks of data per product
- Uncensoring:
 - Demand
 - Vendor Lead Times
- Policy gradient methods in a "gym":
 - "gym" \leftrightarrow backtesting \leftrightarrow simulator (note the "simulator" isn't a good world model).
 - The policy can depend on many features. (seasonality, holiday indicators, demand history, ASIN, text features)







Differentiable Control Problem

• So, we can take gradients directly from our Reward through our policy

• This is our current production policy, called *DirectBackprop*

• Similar in spirit to Perturbation Analysis (Glasserman et al 1995), except it uses a neural policy

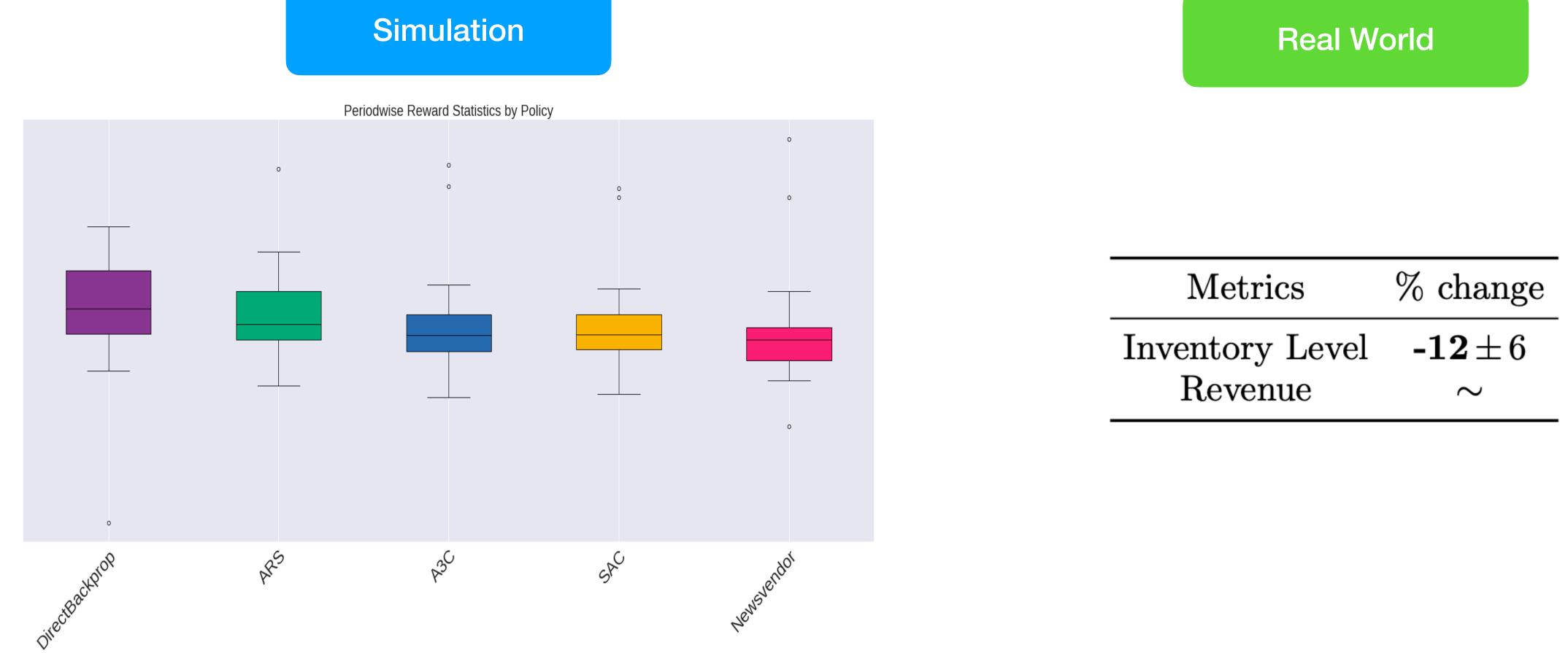
Note that each term of our state evolution is a differentiable function of previous actions



Sim to Real Transfer

Reward

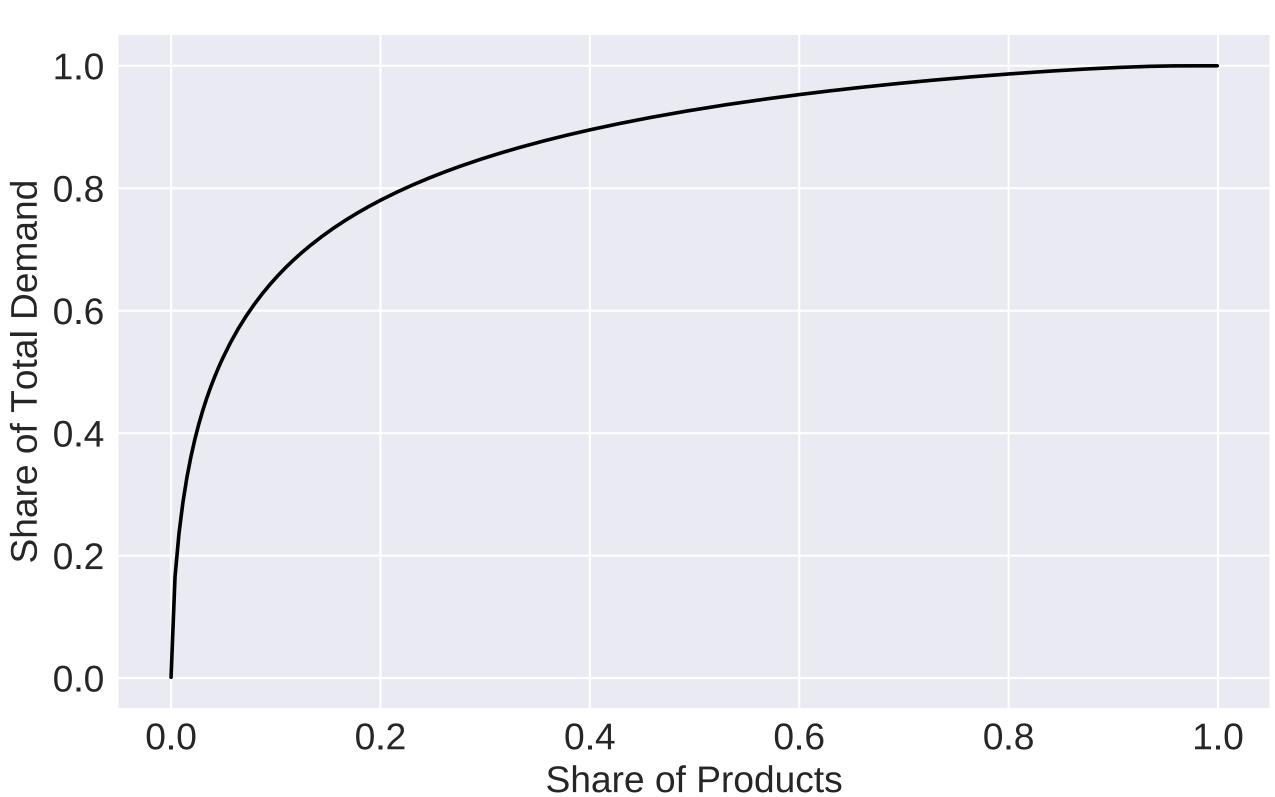
- Sim: the backtest of DirectBackprop improves on Newsvendor.
- total revenue.



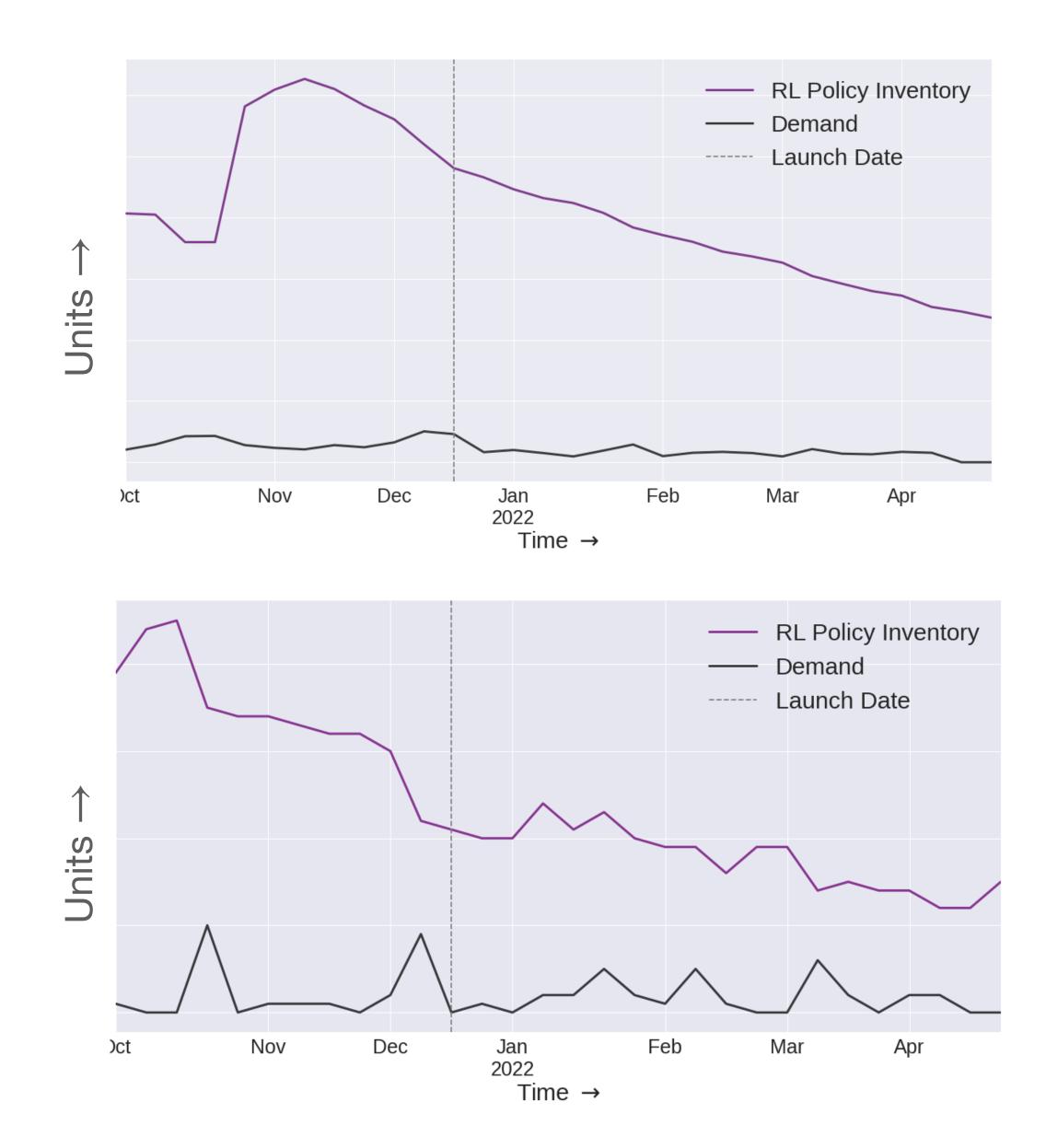
Real: DirectBackprop significantly reduces inventory without significantly reducing

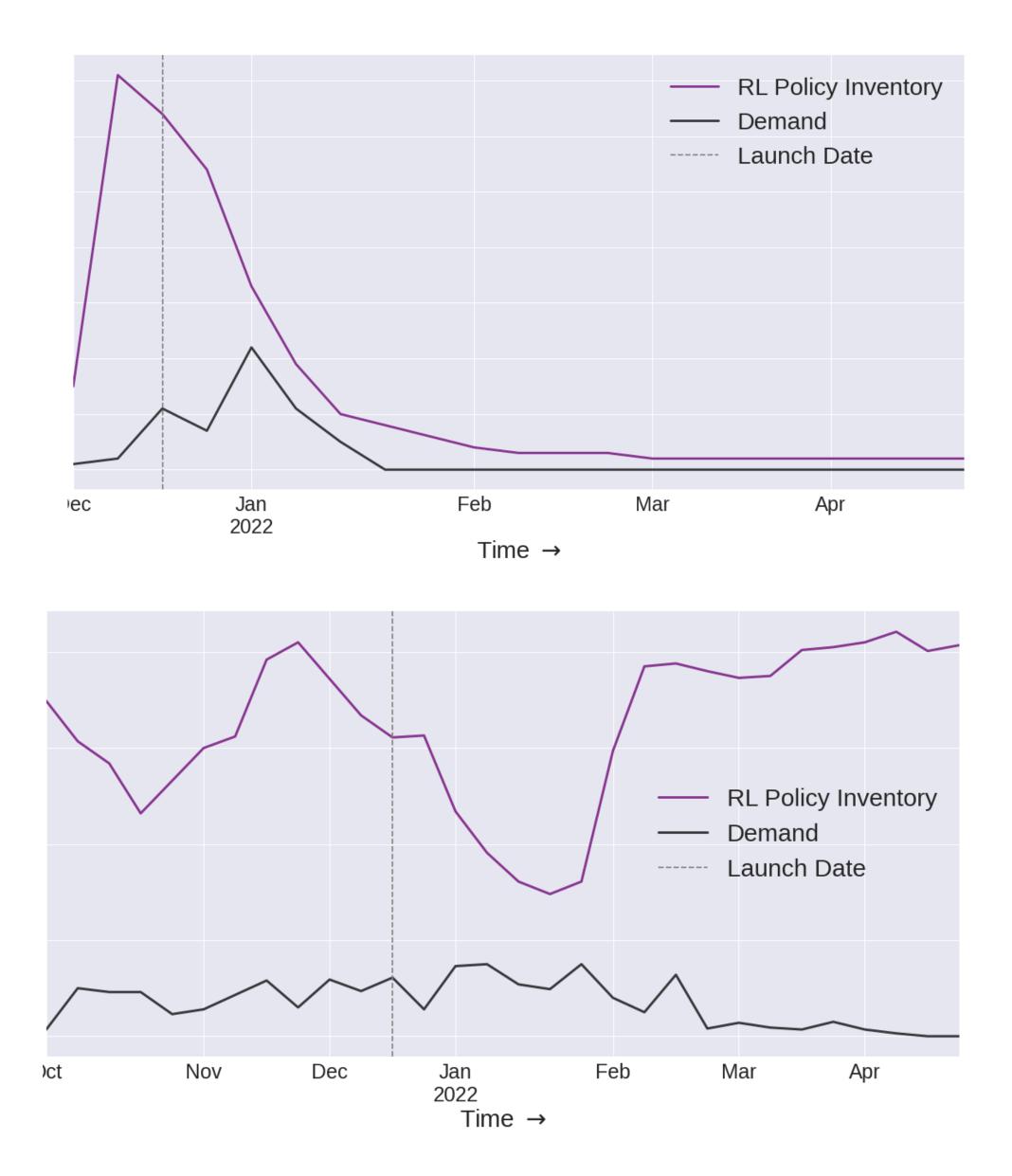
What about in the real world?

- Really hard to measure! (Tripuraneni, M et. al 2021)
- Heavy tailed data:
 - A few products contribute to most of the reward



Anecdotally, RL has reasonable strategies in the real world...





Real World RL Challenges

Cross product constraints are computationally intensive

Not every Supply Chain problem can be written in this framework

World is not perfectly exogenous (some terms may depend on our actions)

Conclusion

- There are a class of RL Problems that work in the real world!
- The exogenous assumption allows us to backtest any policy on historical data
- A large number of classical Operations Research problems fall into this class of Interactive Decision-Making problems



Carson

Kari







Anna

Dhruv

Sham

Conclusion

- There are a class of RL Problems that work in the real world!
- The exogenous assumption allows us to backtest any policy on historical data
- A large number of classical Operations Research problems fall into this class of Interactive Decision-Making problems





Kari

Dhruv







Anna

Dean

Sham