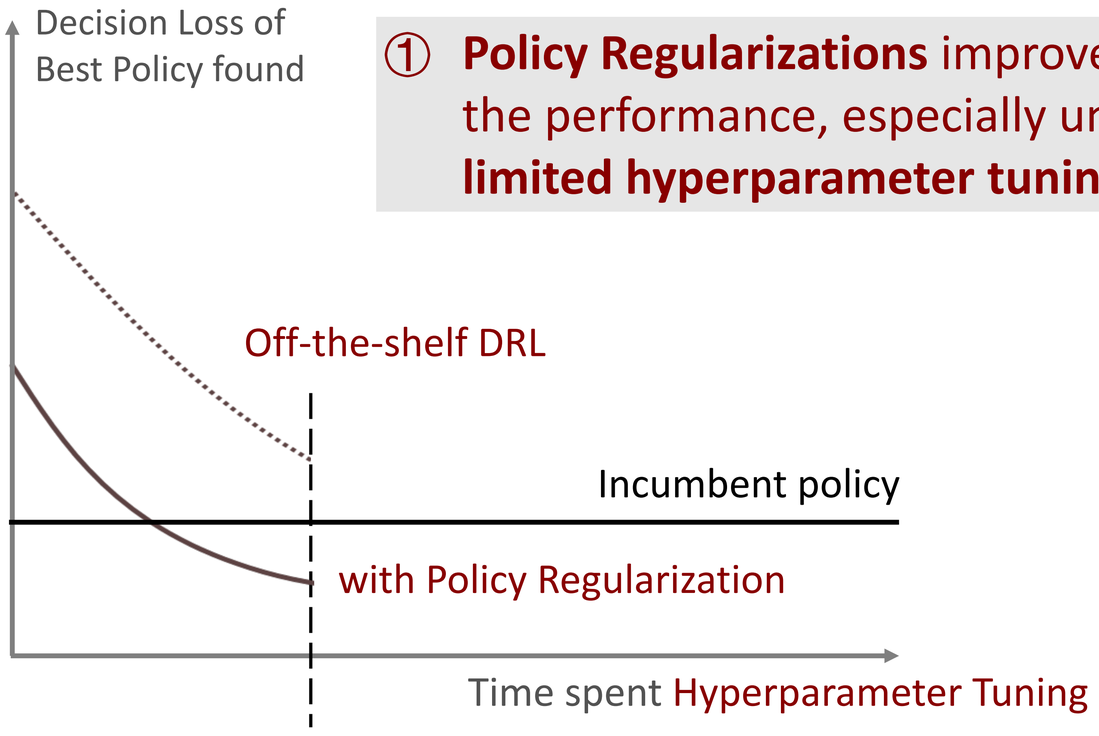
DeepStock: Reinforcement Learning with Policy Regularizations for Inventory Management

Yaqi Xie[1], Xinru Hao[2], Jiaxi Liu[3], Will Ma[4], Linwei Xin[5], Lei Cao[2], Yidong Zhang[2]

1 University of Chicago, USA; 2 Taobao & Tmall Group, China; 3 Sichuan University, China;
4 Columbia University, USA; 5 Cornell University, USA

Deep Reinforcement Learning (DRL) for Inventory

Inventory State, Feature Vector → Neural Network (NN)-based Policy → Order Quantity



① **Policy Regularizations** improve the performance, especially under **limited hyperparameter tuning**

≈ 10 days, realistic for Alibaba (each training run takes > 1 day)

Policy Regularizations

- **"BASE"**: NN outputs **Basestock level**
  Order quantity = max{basestock level − on-hand inventory position, 0}
- **"COEFF"**: NN outputs **Coefficient vector**
  Order quantity = coefficient vector^T selected features
  e.g., selected features = recent/forecasted demands
- **"BOTH":** max{coefficient vector^T selected features − on-hand inventory position, 0}

Alibaba's Setting

- Tmall: 100,000+ SKU's B2C across ≈20 warehouses

Stockout Rate: % of days the SKU ran out of inventory	Performance metrics	**Turnover Time**: average duration the SKU stayed

- Metrics: take weighted average across all SKU-warehouse inventories

② **One unified** policy to manage **all** 1,000,000+ inventories at Alibaba Tmall

Alibaba.com 天猫 TMALL.COM

Compared to July-August 2024:
no change in Stockout Rate
−20% in Turnover Time (Tmall global)

Project start June 2023 → Pilot Test July 2024 → Major Rollout April 2025 → 100% Adoption as of October 2025

Difference-in Differences:
-0.83 % in Stockout Rate
-9.53 days in Turnover Time

no change in Stockout Rate
−1 day in Turnover Time (Tmall global)
−2 days in Turnover Time (Tmall)

Offline Comparison in Alibaba's setting

- Train/Validate on 55,000 SKU-warehouse combinations
- Test on chronologically-later days through simulation

Compared to DDPG BOTH; best results found after searching 10 hyperparameter configurations:

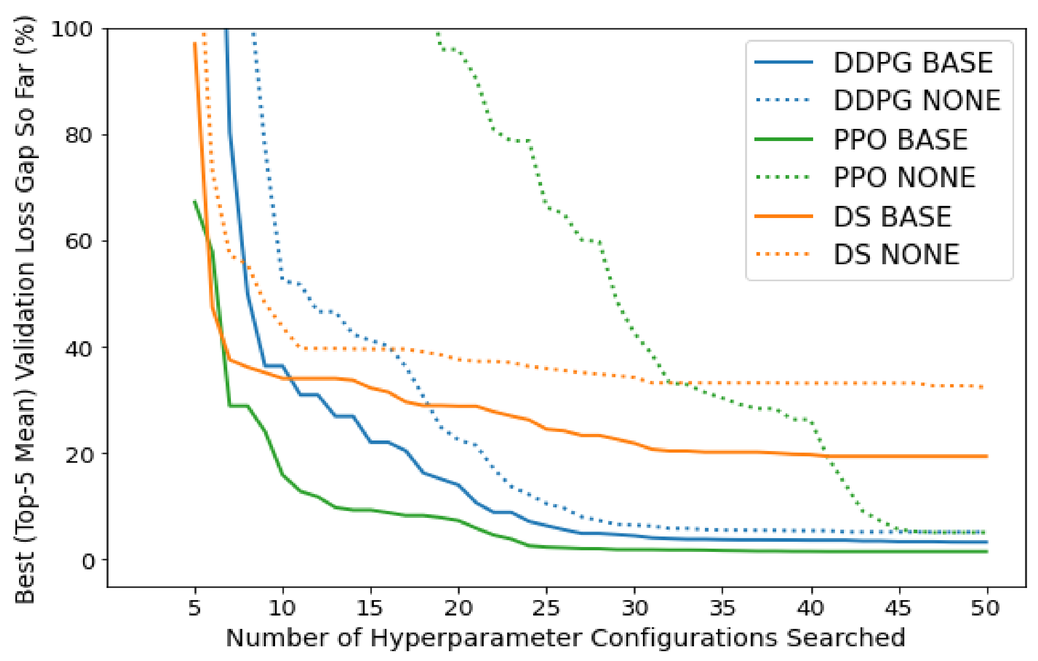
Deep Deterministic Policy Gradient (DDPG)				
Policy Regularization	NONE	BASE	COEFF	BOTH
Stockout Rate (%)	+10.10	+6.03	+4.41	-
Turnover Time (days)	+6.13	+6.46	-0.41	-

Differentiable Simulator (DS)				
Policy Regularization	NONE	BASE	COEFF	BOTH
Stockout Rate (%)	+2.10	+2.18	+1.74	+1.91
Turnover Time (days)	-1.25	-2.81	+3.80	+0.23

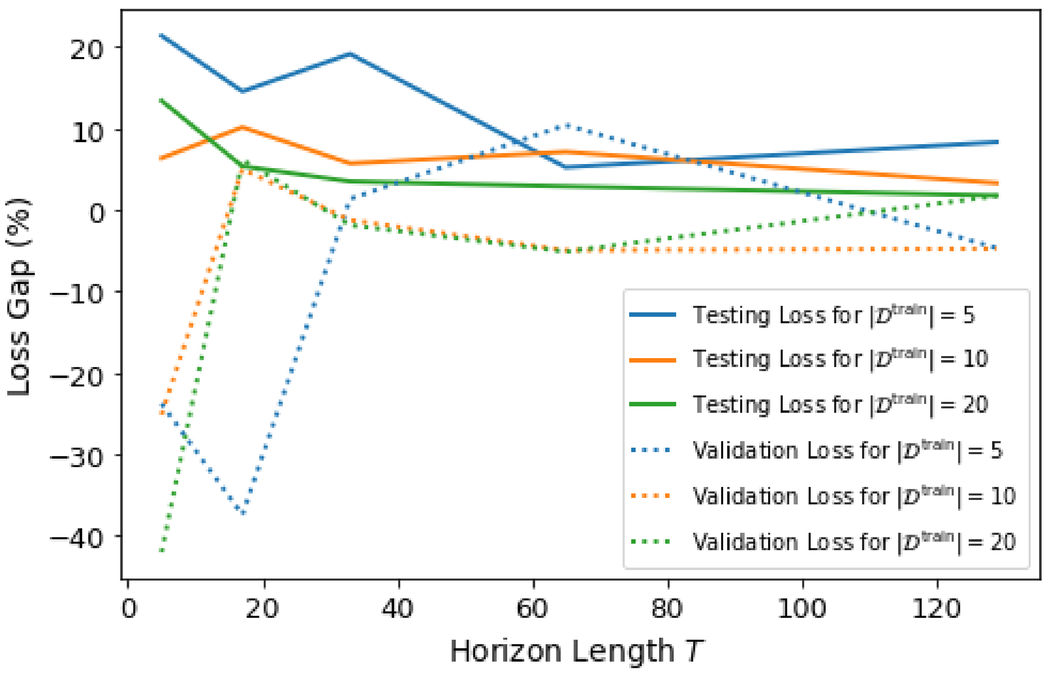
- New DRL method "DS", not sensitive to hyperparameters [Madeka et al. 22; Alvo et al. 25]

Synthetic Data

- Independent demands over 29 days
- 1-dimensional feature; train/validate on 20 trajectories



③ **Reshape the narrative** on the best DRL method among DS and traditional DDPG & PPO for inventory



Why do Policy Regularizations help DRL training?

- Stabilize learning of Q-function when minimizing the Bellman error

Why is the final performance of DS worse?

- Not cross-learning over time; overfitting to idiosyncrasies in trajectories