# IE Decision Systems Engineering
# Spring '21 Seminar Series

**Friday, March 5, 12-1 p.m.**

**Zoom https://asu.zoom.us/j/81413425044**

This talk will be recorded and will have a Q&A at the end.

# "Learning in structured MDPs with convex cost functions: Improved regret bounds for inventory management"


## Shipra Agrawal

Assistant Professor

Department of Industrial Engineering and Operations Research, Columbia University


## Bio

Shipra Agrawal is Cyrus Derman Assistant Professor of the Department of Industrial Engineering and Operations Research. She is also affiliated with the Department of Computer Science and the Data Science Institute, at Columbia University. Her research spans several areas of optimization and machine learning, including online optimization, multi-armed bandits, online learning, and reinforcement learning. Shipra serves as an associate editor for Management Science, Mathematics of Operations Research, and INFORMS Journal on Optimization. Her research is supported by an NSF CAREER award and faculty research awards from Google and Amazon.


## Abstract

The stochastic inventory control problem under censored demands is a fundamental problem in revenue and supply chain management. A simple class of policies called "base-stock policies" is known to be asymptotically optimal for this problem in certain settings, and further, the convexity of long-run average-cost under such policies has been established. In this work, we present a learning algorithm for the stochastic inventory control problem under lost sales penalty and positive lead times, when the demand distribution is a priori unknown. Our main result is a bound of O(L\sqrt{T}+D) on the regret against the best base-stock policy. Here T is the time horizon, L is the fixed and known lead time, and D is an unknown parameter of the demand distribution described roughly as the number of time steps needed to generate enough demand for depleting one unit of inventory. Our results significantly improve the existing regret bounds for this problem. Notably, even though the state space of the underlying Markov Decision Process (MDP) in this problem is continuous and L-dimensional, our regret bounds depend linearly on L. Our techniques utilize convexity of the long-run average cost and a newly derived bound on `bias' of base-stock policies, to establish an almost blackbox connection between the problem of learning and optimization in such MDPs and stochastic convex bandit optimization. The techniques presented here may be of independent interest for other settings that involve large structured MDPs but with convex cost functions.

This talk is based on joint work with Randy Jia.


Ira A. Fulton Schools of Engineering
Arizona State University

**School of Computing, Informatics, and Decision Systems Engineering**

Hosted by: Giulia Pedrielli