QUEUING WITH FUTURE INFORMATION

Joel Spencer, Courant Institute of Mathematical Sciences, New York University, spencer@courant.nyu.edu
Madhu Sudan, Microsoft Research New England, madhu@mit.edu
Kuang Xu, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, kuangxu@mit.edu

We study an admissions control problem, where a queue with service rate $1-p$ receives incoming jobs at rate $\lambda \in (1-p, 1)$, and the decision maker is allowed to redirect away jobs up to a rate of p, with the objective of minimizing the time-average queue length.

We show that the amount of *information about the future* has a significant impact on system performance, in the heavy-traffic regime. When the future is unknown, the optimal average queue length diverges at rate $\sim \log_{\frac{1}{1-p}} \frac{1}{1-\lambda}$, as $\lambda \to 1$. In sharp contrast, when all future arrival and service times are revealed beforehand, the optimal average queue length converges to a finite constant, $(1-p)/p$, as $\lambda \to 1$. We further show that the finite limit of $(1-p)/p$ can be achieved using only a *finite* lookahead window starting from the current time frame, whose length scales as $\mathcal{O}\left(\log \frac{1}{1-\lambda}\right)$, as $\lambda \to 1$. This leads to the conjecture of an interesting duality between queuing delay and the amount of information about the future.