Perils and Benefits of Free Trials in Large Scale Service Systems: An Operational Perspective

Yaşar Levent Koçağa

Sy Syms School of Business, Yeshiva University

Free trials in Service Systems

- Free trials in *congestion-prone service* systems:
  - Many service systems offer a free trial service to customers (e.g. Netflix, Hulu).
  - Customers enjoy the service for free for a given duration (e.g. 15 or 30 days).
  - This allows the customers to experience the service without commitment (e.g. customers learn their valuations).
  - After the free trial period a portion of customers convert to paying customers (conversion rate).

Free trials in Service Systems

- Free trials in *congestion-prone service* systems:
  - The idea is to tap into a market of customers who must experience the service before paying for the service.
  - Similar to free samples, both aim to generate more demand.
  - Different than free samples, capacity is shared.
  - Essentially we are giving up capacity to generate future demand (double-edged sword phenomenon).

Research Questions

We aim to answer the following questions:

1. What are the implications of offering free trials for a congestion-prone service firm which maximizes capacity and price are ex ante (before offering free trials)? Is offering free trials really beneficial or does it lead to a double-edged sword?
2. Can the firm do better by leveraging some form of control such as capacity adjustment?
3. Under what regime does a firm operate without free trials?

Service System: Queueing Dynamics

- Regular customers arrive according to a Poisson process with rate Λ_1 per day (regular demand).
- Service requirements i.i.d exponential with mean $1/\mu$.
- C units of capacity
- $M/M/C$ (Erlang C) queue

Service System: Customer Choice Model

- Each customer is endowed with a service valuation v.
- v is IGFR with finite mean and continuously diff cdf $F(\cdot)$.
- An arriving customer joins the systems if his valuation exceeds expected steady state full price of service.
- Full price includes price p and cost q for equilibrium expected delay $d := E[D]$ (unobservable queue).

Service System with Free Trials

- Free trial customers arrive according to a Poisson process with rate $\Lambda_2/\tau < C\mu$ per day.
- Free trial period is fixed at τ days.
- Service requirements identical to regular customers.
- Free trial customers joining on day t convert to regular customers at rate δ_t.
- Valuations identical after free trial period.

Timeline

Service Systems with Free Trials: Equilibrium Demand

- We assume the system reached equilibrium within the day.
- Then the equilibrium arrival rate in free trial cycle t is given by

$$\lambda_t(C,p) = \left(\Lambda_1 + \frac{\Lambda_2}{\tau}\sum_{j=1}^{t-\tau}\delta_t\right) P(v \geq p + qE\left[D_t\right]) + \Lambda_2$$
$$= \left(\Lambda_1 + \Lambda_2\sigma_t\right)\bar{F}(p + qd_t) + \Lambda_2,$$

where $\sigma_t := \frac{1}{\tau}\sum_{j=1}^{t-\tau}\delta_t$.

Service System with Free Trials: Formulations

We assume capacity and price is ex ante set optimally at $C_0^\star$ and $p_0^\star$. Then we consider the following two scenarios with free trials:

1. No Control with free trials
   - The total revenue $R(C_0^\star, p_0^\star)$ for T days with free trials

$$R(C_0^\star, p_0^\star) = \sum_{t=1}^{T} R_t(C_0^\star, p_0^\star) = \sum_{t=1}^{T} (\Lambda_1 + \Lambda_2 \sigma_t)\, p_0^\star \bar{F}(p_0^\star + q d_t)$$

2. Capacity optimization with free trials
   - The objective is to maximize total equilibrium profit when price is kept fixed at $p_0^\star$

$$\max_{C_t > 0} \sum_{t=1}^{T} P_t(C_t) = \max_{C_t > 0} \sum_{t=1}^{T} R_t(C_t, p_0^\star) - w\mu C_t$$

Related Literature

- **Free trials in service systems:** Zhou et al (EJOR 14), Datta et al (JMR 15), Foubert & Gijsbrechts (Mar Sci 16), Lian et al (EJOR 16).
- **Economics of Queues:** Naor (Econometrica 69), Mendelson (ACM 85), Mendelson & Whang (MS 90), Stidham (IEEE 85), **Maglaras & Zeevi (MS 03)**, Whitt (OR 03), Kumar & Randhawa (MSOM 2010), Randhawa (ORL 2013), Maglaras et al (MS 17).
- **QED heavy traffic:** **Halfin & Whitt (OR 81)**, Whitt (MS 92), Garnet et al (MSOM 02), Whitt (OR 03), Armony & Maglaras (OR 04), **Borst et al (OR 04)**.

Exact Solution

Proposition

For each capacity $C > 0$ and price $p > 0$, there exists a unique equilibrium arrival rate $\lambda_t(C, p)$ and equilibrium delay d_t.

- We can *numerically* compute $\lambda_t(p)$, and thus $R_t(p)$.
- However, not much **insight** can be gleaned from this approach.
- This is important because we want to understand the implications of offering free trials.

Asymptotic Approach

- So, we consider a sequence of systems where market size(s) and capacity grow large.
- First, we consider a system w/o free trials where capacity and price is optimized.
- Then, we incorporate free trials on with no further control.
- Eventually we optimize capacity with free trials.

Scaling of Service System

- Suppose capacity C^n, regular market size Λ_1^n, and free trial market size Λ_2^n grow large proportionally such that

$$C^n = n, \ \Lambda_1^n = \kappa_1 \mu n, \ \text{and} \ \Lambda_2^n = \kappa_2 \mu n$$

(where $\kappa_1 := \frac{\Lambda_1}{C\mu} > 0$ and $0 \leq \kappa_2 =: \frac{\Lambda_2}{C\mu} < 1$)

- Then, define $\bar{p}_t$ such that

$$(\Lambda_1^n + \sigma_t \Lambda_2^n)\bar{F}(\bar{p}_t) + \Lambda_2^n = C^n \mu$$

$$\bar{p}_t = \bar{F}^{-1}\left(\frac{C^n\mu - \Lambda_2^n}{\Lambda_1^n + \sigma_t \Lambda_2^n}\right) = \bar{F}^{-1}\left(\frac{C\mu - \Lambda_2}{\Lambda_1 + \sigma_t \Lambda_2}\right) = \bar{F}^{-1}(\kappa_t)$$

where $\kappa_t := \frac{1-\kappa_2}{\kappa_1 + \sigma_t \kappa_2}$ (so $\bar{p}_t$ is constant along the sequence).

Price and Capacity Optimization w/o Free Trials

Theorem

The optimal system operates in the QED heavy traffic regime and

$$\lambda_0^\star := \lambda(C_0^\star, p_0^\star) \approx C_0^\star \mu - \beta^\star \sqrt{C_0^\star \mu} \text{ and } p_0^\star \approx \bar{p}_0(C_0^\star) + \frac{\pi^\star}{\sqrt{n}}$$

where $\bar{p}_0 = \bar{p}(C_0^\star)$ *is such that* $\Lambda_1 \bar{F}(\bar{p}_0) = C_0^\star \mu$*, and the optimal profit scales as*

$$P_0^\star = C_0^\star \bar{p}(C_0^\star) - \sqrt{C_0^\star \mu}\,(\bar{p}(C_0^\star)\beta^\star - \pi^\star) - C_0^\star \mu w$$

Outline of Proof of Optimality of QED

- We first fix the capacity and consider the pricing problem under two separate scenarios in which the a so-called elasticity condition (EC) either holds or doesn't hold
- The former case yields the Halfin-Whitt aka QED HT regime, the latter the QD HT regime as the optimal regime
- Then, we go back to the joint problem and show that the (EC) holds once system scale is large enough

Elasticity Condition

Condition

Define the demand function $\Lambda_t(p) := (\Lambda_1 + \sigma_t \Lambda_2)\bar{F}(p)$. *We assume that the first order price* $\bar{p}_t$ *is well-defined and that* $\Lambda_t(p)$ *is elastic*[1] *for all* $p \geq \bar{p}_t$, *or equivalently*

$$\begin{aligned} \max_{p \geq 0} \quad & (\Lambda_1 + \sigma_t \Lambda_2) p \bar{F}(p) \\ s.t. \quad & (\Lambda_1 + \sigma_t \Lambda_2)\bar{F}(p) + \Lambda_2 \leq C\mu \end{aligned}$$

is maximized at $\bar{p}_t > \tilde{p} = \arg\max p\bar{F}(p)$.

[1] $\lambda(p)$ is elastic at price p if $\varepsilon(p) > 1$ where $\varepsilon(p) := -\frac{d\lambda(p)}{dp}\frac{p}{\lambda(p)}$

Pricing Problem When (EC) Holds

Theorem

(Optimality of QED) *Suppose (EC) holds and that capacity scales proportionally to demand. Then, the optimal price scales as*

$$p^{\star,n} = \bar{p} + \frac{\pi^\star}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)$$

and the optimal revenue scales as

$$R^{\star,n} = (n\mu - \Lambda_2)\,\bar{p} - \sqrt{n}\mu\left(\bar{p}\beta\left(\pi^\star\right) - \frac{\kappa_2 - 1}{\kappa_2}\pi^\star\right) + o\left(\sqrt{n}\right)$$

where $\pi^\star := \arg\min_{\pi \in R}\{\bar{p}\beta\left(\pi\right) - \frac{\kappa_2 - 1}{\kappa_2}\pi\}$.

Optimality of QD When (EC) Doesn't Hold

Theorem

(Optimality of QD) *Suppose (EC) does not hold and that capacity scales proportionally to demand. Then, under the optimal pricing policy, the system operates in the QD heavy traffic regime. Furthermore, the optimal price satisfies*

$$p_t^{\star,n} \to \tilde{p}_t \ \text{ as } \ n \to \infty,$$

and the revenue scales as

$$R_t^{\star}(\Lambda_2) \approx \bar{R}_t(\Lambda_2) := (\Lambda_1 + \Lambda_{2,t})\,\tilde{p}_t \bar{F}(\tilde{p}_t).$$

No Control with Free Trials

Theorem

Suppose regular market size and free trial market scale proportionally to the ex ante optimal capacity. Then, with free trials the system moves into the Efficiency-Driven (ED) heavy traffic regime.

$$d_t \approx \bar{d}_t := \frac{\bar{p}_t - \bar{p}_0}{q}$$

where $\bar{p}_t$ *satisfies* $(\Lambda_1 + \sigma_t \Lambda_2)\bar{F}(\bar{p}_t) + \Lambda_2 = C_0^\star \mu$. *Furthermore, optimal revenue scales as*

$$R_t(C_0^\star, p_0^\star) \approx \left[(C_0^\star - \Lambda_2) - \frac{1}{\bar{d}_t}\right] p_0^\star \approx \left[(C_0^\star - \Lambda_2) - \frac{1}{\bar{d}_t}\right] \bar{p}_0$$

On Emergence of ED

- Price cannot increase to thin the arrival rate, so expected delay does:

$$d_t \approx \bar{d}_t := \frac{\bar{p}_t - \bar{p}_0}{q}$$

- Where is the effect of conversion rates?

$$R_t(C_0^\star, p_0^\star) \approx \left[(C_0^\star - \Lambda_2) - \frac{1}{\bar{d}_t}\right] p_0^\star \approx \left[(C_0^\star - \Lambda_2) - \frac{1}{\bar{d}_t}\right] \bar{p}_0$$

- Whitt (03 OR), also observes that ED emerges under congestion sensitive demand.

Effect of Offering Free trials: What to expect

- Recall w/o free trials we have:

$$R_0 \approx \left[C_0^\star \mu - \beta^\star \sqrt{C_0^\star \mu} \right] p_0^\star$$

  - Ample capacity is of order of square root of capacity.
- And with free trials we have:

$$R_t \approx \left[(C_0^\star \mu - \Lambda_2) - \frac{1}{\bar{d}_t} \right] p_0^\star$$

  - When the free trial market size is large, offering free trials will decrease revenue rate.
  - When the free trial market size is small, offering free trials might yield small benefits that are of order of square root of capacity.

.

Effect of Offering Free Trial: What really happens

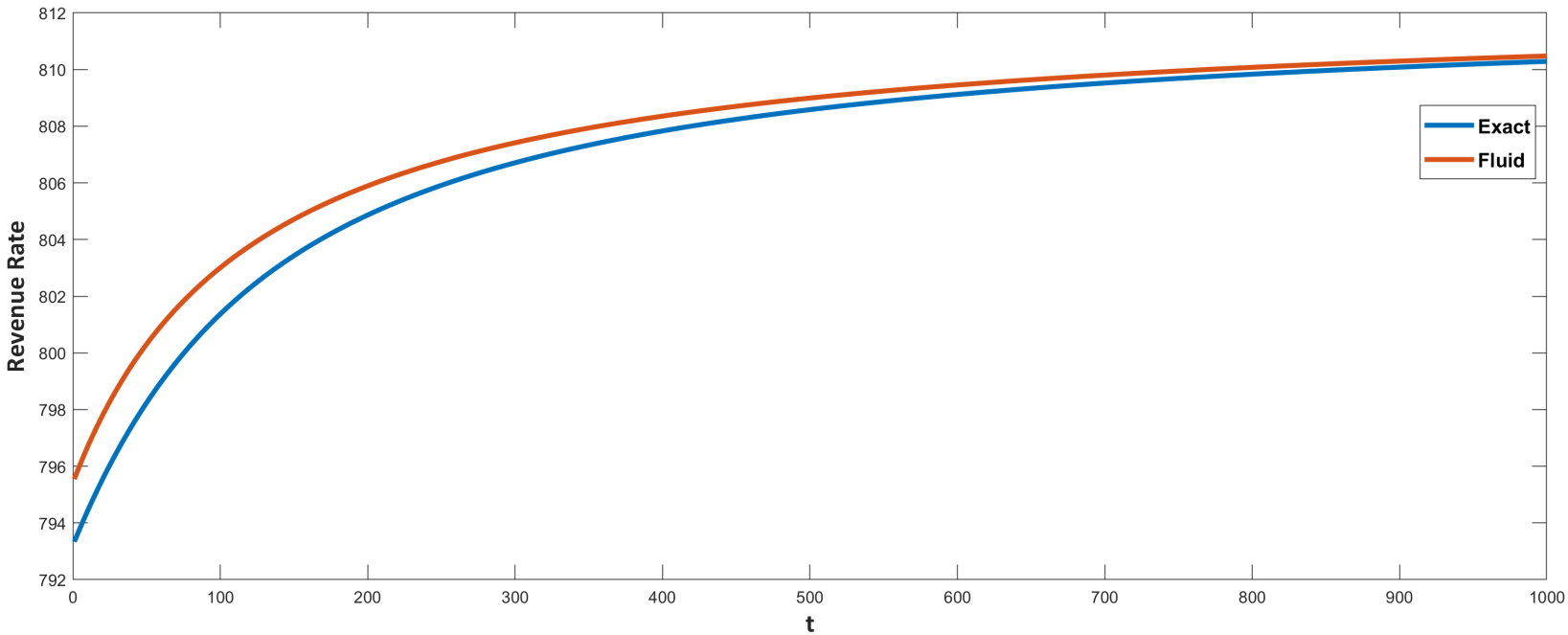

Figure 1: Exact vs Approximate Revenue Rate [$C = 400$, $\Lambda_1 = 825$, $\Lambda_2 = 25$, $v \sim U(0, 4)$, $\mu = 1$, $q = 1$, $\delta_t = 1$, $\tau = 30$]

Effect of Offering Free Trial: What really happens

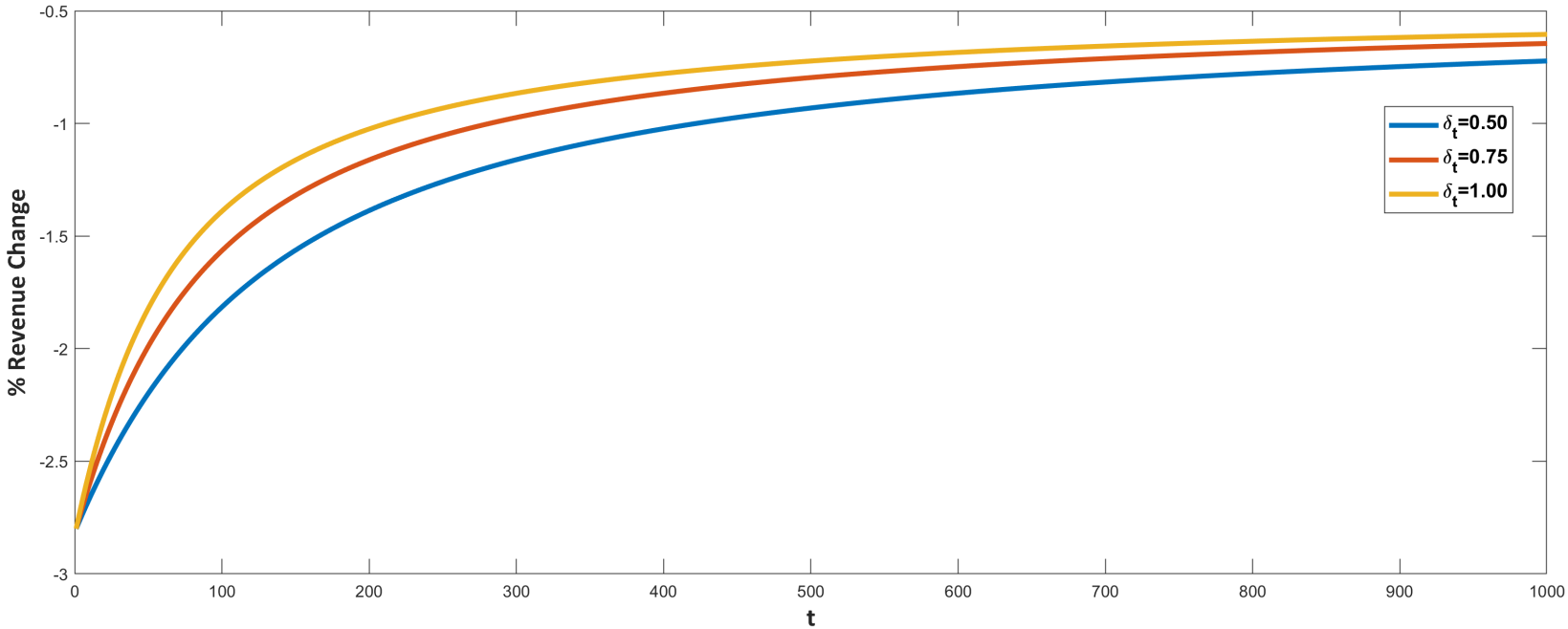

Figure 2: Varying Conversion Rates [$C = 400$, $\Lambda_1 = 825$, $\Lambda_2 = 25$, $v \sim U(0, 4)$, $\mu = 1$, $q = 1$, $\tau = 30$]

Effect of Offering Free Trial: What really happens

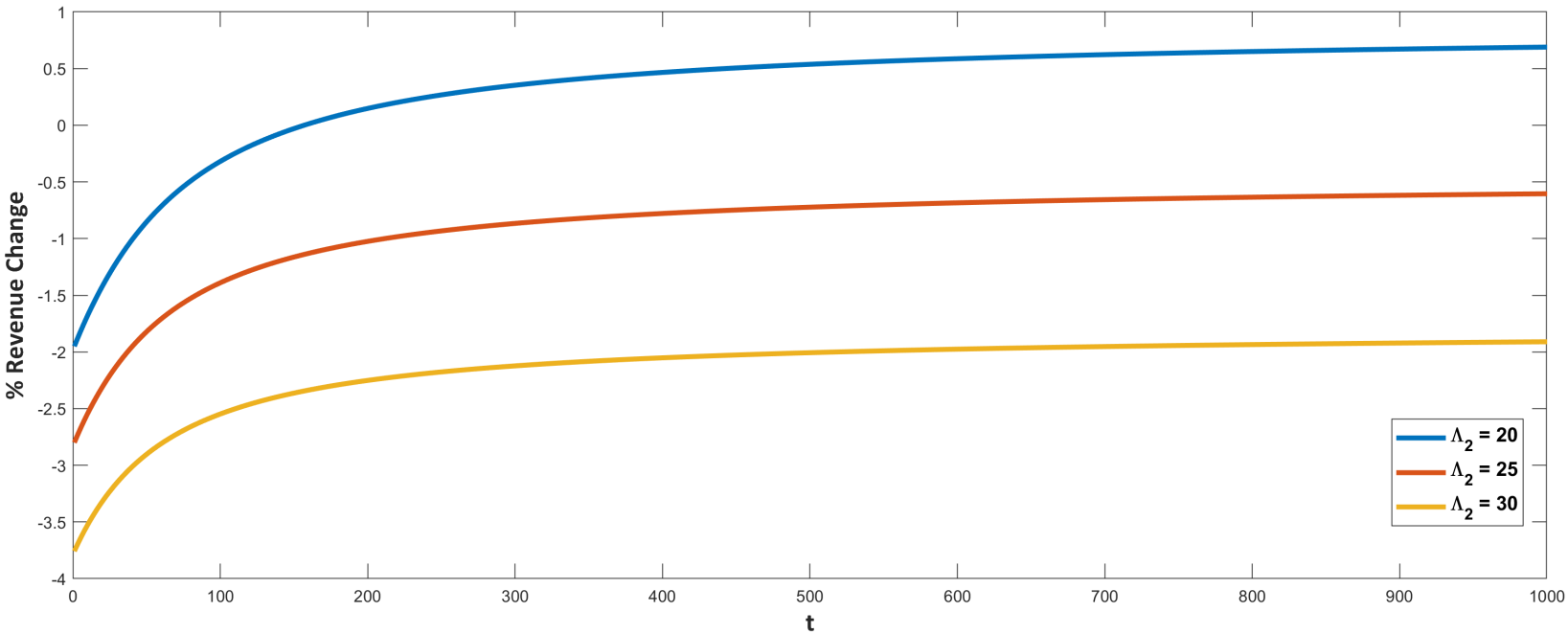

Figure 3: Varying Free Trial Market Size [$C = 400$, $\Lambda_1 = 825$, $v \sim U(0, 4)$, $\mu = 1$, $q = 1$, $\delta_t = 1$, $\tau = 30$]

Effect of Offering Free trials: Practical Implications

- The question is can we convert free trials to paying throughput?
- For a system with capacity and price optimized ex ante, the answer is only if the free trial market size is of order of the square root of the capacity.
- Even so, the benefit is of order of the square root of the capacity and thus limited.
- A smaller free trial market size is preferable increasing the free trial market size to compensate for small conversion rates can be detrimental.
- Trying to increase the conversion rate wouldn't improve the system in the long rung.

Capacity Optimization with Free Trials

- Capacity Sizing: Optimize the first-order approximation for optimal profit given by

$$\max_{C \geq 0} \hat{P}(C) := (C\mu - \Lambda_2)\,\bar{p}\,(C) - w\mu C,$$

  and let $\hat{C}$ be the corresponding maximizer. Also, set the first order price term $\hat{\bar{p}} := \bar{p}(\hat{C})$.
- This is justified because $\hat{\bar{p}} := \bar{p}(\hat{C}) \approx \bar{p}(C^\star)$, and we can show the system still operates under QED with capacity $\hat{C}$.

Capacity Optimization with Free Trials

Theorem

Suppose the regular and free market sizes scale proportionally. Then, the first order optimal capacity $\hat{C}^n$ is asymptotically optimal, i.e,

$$\frac{P^n(\hat{C}^n, p_0^{\star,n})}{P^n(C^{\star,n}, p_0^{\star,n})} \to 1 \text{ as } n \to \infty.$$

Furthermore, the optimal system operates in the QED regime so thatdelay is negligible for large scale systems, i.e.,

$$d^{\star,n} \to 0 \text{ as } n \to \infty.$$

Effect of Offering Free trials with Capacity Optimization: What to expect

$$P(C^\star) \approx P(\hat{C}) = \Lambda_1 \left[\hat{\bar{p}} \bar{F}\left(\hat{\bar{p}}\right) - w\right] + \Lambda_2 \left[(\hat{\bar{p}} - w)\sigma_t \bar{F}\left(\hat{\bar{p}}\right) - w\right]$$

Then, offering free trials is beneficial on day t if and only if

$$(\hat{\bar{p}} - w)\sigma_t \bar{F}\left(\hat{\bar{p}}\right) - w > 0$$

- Whether offering free trials is beneficial depends on F, w, and σ_t; furthermore can be explicitly determined:
  - Under $U[0, \theta]$ valuations, we require $\theta > (3 + 2\sqrt{2})w$ when $\sigma_t = 1$.
- Benefit is (approximately) linear in t if δ_t is constant.

Effect of Offering Free trials with Capacity Optimization: What really happens

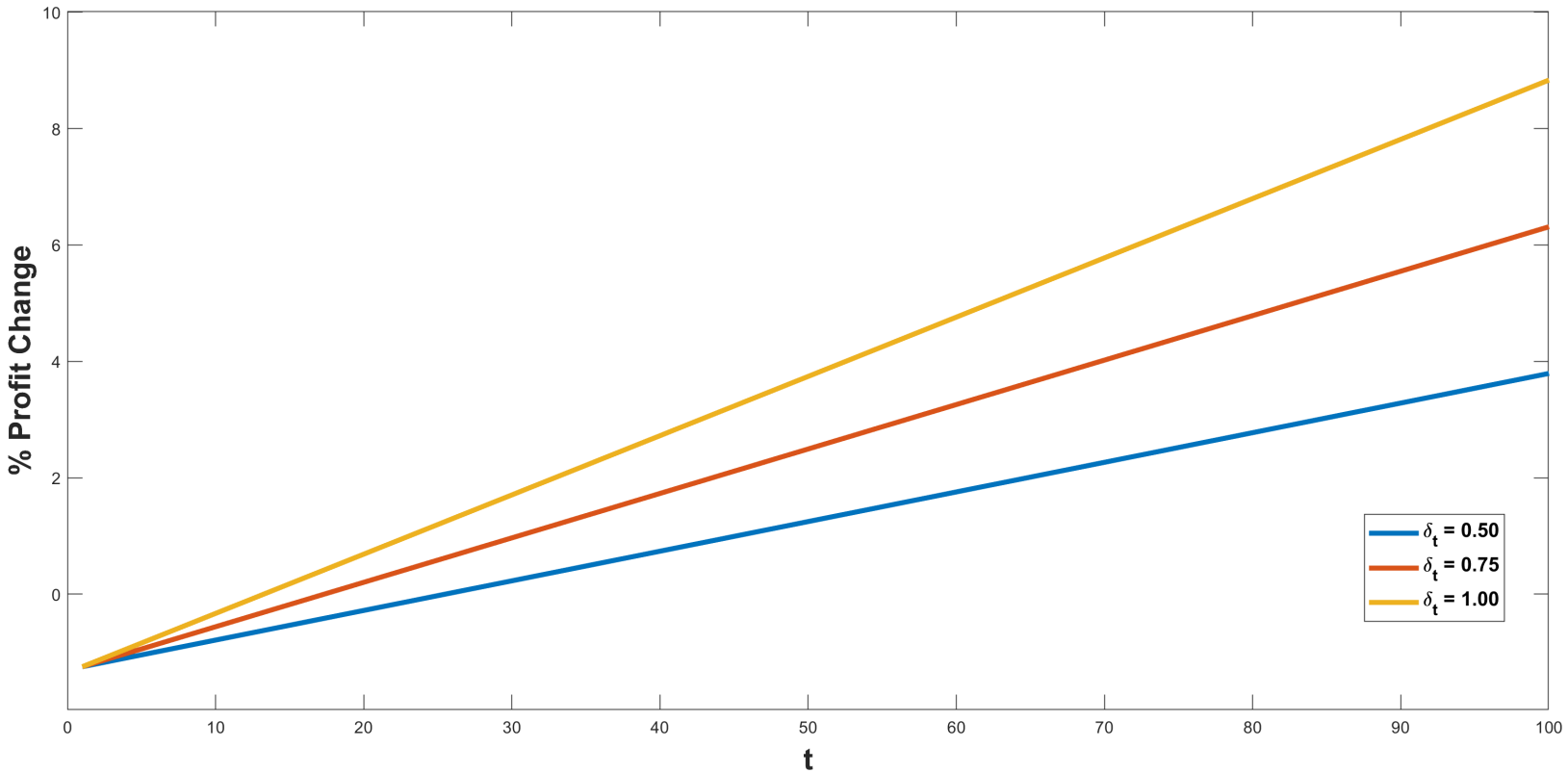

Figure 4: % Effect of Offering Free Trials under Various Conversion Rates

Some Extensions

- QoS dependent conversion rates
  - Small expected delays
  - Replace δ_t with $\delta_t(\mathbf{d_t})$, where $\mathbf{d_t} := [d_t, d_{t+1}, \cdots, d_{t+\tau}]$
  - Previous results equivalent to assuming $\delta_t(\mathbf{0})$
  - Even smaller benefit without control
  - Dual role of QED with capacity optimization
- Heterogeneous Valuations
  - Suppose free trial customers have stochastically smaller valuations
  - Then most results will still hold
- Time-dependent free trial market size
- Subscriptions

Takeaways

- Large scale service systems where capacity and price is jointly optimized operate under QED
  - Small expected delays
  - High utilization with ample capacity of order of square root
- The latter implies that free trials can only be beneficial up to the same order and as such the benefit is relatively limited, if beneficial at all.
- The benefit will even be further limited if conversion rates are QoS dependent
- Increasing conversion rates doesn't benefit the system much
- The service firm should opt to increase capacity to turn the converted customers into throughput and to fully unlock the benefits of free trials.