Online Appendix to "Courier Dispatch in On-Demand Delivery": Supplementary Derivations and Proofs

A. Major Proofs

Proof of Propositions 1 and 2. We first show the result of Proposition 1. First, note that we have $\lim_{\lambda\to 0} W_B(\lambda,\mu_B,C_B)=\infty$ and $\lim_{\lambda\to\mu_D} W_D(\lambda,\mu_D,C_D)=\infty$, according to (4) and (12). Thus, we only need to show that $W_D(\lambda,\mu_D,C_D)-W_B(\lambda,\mu_B,C_B)=0$ has a unique solution in λ to established the desired threshold result in λ.

Consider function

$$f(a):=\frac{1}{\lambda}\left[\frac{a^2C_1}{2T_1(T_1-a)}-\frac{1}{2}-\frac{a^2C_2}{T_2(T_2-a)}-\frac{a}{3}\right],\quad \forall a>0, \tag{A.1}$$

where

$$T_1=\frac{3}{4},\ T_2=\frac{45\pi}{2(32+15\pi)},\ C_1=\frac{9}{8},\ C_2=0.583. \tag{A.2}$$

Then, according to (4) and (12), we have $f(\lambda r)=W_D(\lambda,\mu_D,C_D)-W_B(\lambda,\mu_B,C_B)$. Note that $f(a)=0$ is a single variable cubic equation, which can be solved using standard methods. In particular, equation $f(a)=0$ only has two positive solutions: $a_1\approx 0.5689$ and $a_2\approx 1.150$ (the exact symbolic solutions with T_1, T_2, C_1, C_2 are cumbersome, thus omitted). Thus, we have $0\ll a_1\ll T_1=\mu_D r$ and $T_1\ll a_2$, which implies that $0=W_D(\lambda,\mu_D,C_D)-W_B(\lambda,\mu_B,C_B)$ has a unique solution $\lambda_{ex}=a_1/r$ on $(0,\mu_D)$.

We use the same technique to show Proposition 2. According to (4) and (12), with slight abuse of notation we write $W_B(\lambda,\mu_B,C_B)$ and $W_D(\lambda,\mu_D,C_D)$ as $W_B(r,\lambda,\mu_B,C_B)$ and $W_D(r,\lambda,\mu_D,C_D)$, respectively, to emphasize their dependence on r. have

$$\lim_{r\to 0} W_B(r,\lambda,\mu_B,C_B)=\frac{1}{2\lambda}>0=\lim_{r\to 0} W_D(r,\lambda,\mu_D,C_D),$$

and $\lim_{r\to T_1} W_D(r,\lambda,\mu_D,C_D)=\infty$. Next, since a_1 is the unique solution to $f(a)=0$ on $a\in(0,T_1)$, there is a unique $r_{ex}=a_1/\lambda$ such that $W_D(r,\lambda,\mu_D,C_D)=W_B(r,\lambda,\mu_B,C_B)$. This completes the proof. □

Proof of Corollary 1. (i) This part follows Proposition 1 and 2 directly, where we have shown that serving dedicated leads to shorter wait time comparing to serving batch when either the demand is low or the radius is small, and vice versa. Thus, using the expression for the price in (1), we reach the desired result.

(ii) We use the exactly same proof technique as in the proof of Proposition 1 and 2. We only need to modify the definition of function $f(\cdot)$ is in (A.1) to

$$f(a) := \frac{1}{\lambda}\left[\frac{a^2 C_1}{2T_1(T_1-a)} - \frac{1}{4} - \frac{a^2 C_2}{2T_2(T_2-a)-\frac{a}{6}}\right], \quad \forall a > 0,$$

where parameters T_1, T_2, C_1, and C_2 are defined in (A.2). Then, we have $f(\lambda r) = W_D(\lambda, \mu_D, C_D) - W_B(\lambda, \mu_B, C_B)/2$. Again, function $f(a) = 0$ is a simple cubic equation, having a unique solution on $a \approx 0.5087 \in (0, T1]$. We omit the details to avoid repetition. □

Before proving Propositions 3 and 4, we first provide some properties of the optimal demand rate. For notational convenience, denote

$$V_D^\infty(\lambda) := V_\infty(\lambda, W_D(\lambda, \mu_D, C_D)) \text{ and } V_B^\infty(\lambda) := V_\infty(\lambda, W_B(\lambda, \mu_B, C_B)),$$

and the optimal solutions to the optimization problems

$$\max_{\lambda \in [0, \mu_D)} V_D^\infty(\lambda), \text{ and } \max_{\lambda \in [0, 2\mu_B)} V_B^\infty(\lambda), \tag{A.3}$$

as λ_F^∞ and λ_B^∞, respectively. Further, denote $T_B = 2\mu_B r = \dfrac{45\pi}{2(32+15\pi)}$ and $T_D = \mu_D r = \dfrac{3}{4}$.

Next, we present the optimal solutions and objective values to the optimization problems in (A.3). Note that the objective functions in (A.3) are strictly concave in λ, since

$$\frac{d^2 V_D^\infty(\lambda)}{d\lambda^2} = \frac{cC_D T_D}{r(\lambda - \frac{T_D}{r})^3} < 0, \text{ and } \frac{d^2 V_B^\infty(\lambda)}{d\lambda^2} = \frac{2cC_B T_B}{r(\lambda - \frac{T_B}{r})^3} < 0,$$

respectively. Furthermore, by solving the first order conditions when $cr < T_B < 2T_D$, we have

$$\frac{dV_D^\infty(\lambda)}{d\lambda} = 1 + \frac{crC_D\left(1 - \frac{T_D^2}{(T_D - \lambda r)^2}\right)}{2T_D} = 0, \text{ and } \frac{dV_B^\infty(\lambda)}{d\lambda} = 1 + \frac{crC_B\left(1 - \frac{T_B^2}{(T_B - \lambda r)^2}\right)}{T_B} - \frac{cr}{3} = 0,$$

respectively, which imply

$$\lambda_F^\infty = \frac{T_D}{r}\left(1 - \frac{crC_D}{\sqrt{crC_D[2T_D + crC_D]}}\right), \lambda_B^\infty = \frac{T_B}{r}\left(1 - \frac{\sqrt{3}crC_B}{\sqrt{crC_B[3T_B + cr(3C_B - T_B)]}}\right), \tag{A.4}$$

respectively, and

$$V_D^\infty(\lambda_F^\infty) = cC_D + \frac{T_D}{r} - \frac{1}{r}\sqrt{crC_D[2T_D + crC_D]}, \tag{A.5}$$

$$V_B^\infty(\lambda_B^\infty) = c\left(2C_B - \frac{1}{2} - \frac{1}{3}T_B\right) + \frac{T_B}{r} - \frac{4}{r}\sqrt{3crC_B[3T_B + cr(3_B - T_B)]}, \tag{A.6}$$

where we have omitted the solutions that are outside the feasible regions.

Proof of Proposition 3. Fix $r>0$ and define function

$$g(\alpha) := r\left(V_B^\infty(\lambda_B^\infty) - V_D^\infty(\lambda_F^\infty)\right) = \alpha\left(2C_B - \frac{1}{2} - \frac{T_B}{3}\right) + T_B - 4\sqrt{3\alpha C_B[3T_B + \alpha(3C_B - T_B)]} - \left[\alpha C_D + T_D - \sqrt{\alpha C_D[2T_D + \alpha C_D]}\right], \tag{A.7}$$

where $\alpha := cr$. By plugging in $C_D = 9/8$, $C_B = 0.583$, $T_D = 3/4$, $T_B = 45\pi/(2(32+15\pi))$ and solving $g(\alpha) = 0$, we obtain a unique solution $\alpha^* \approx 0.1355$, which implies that function $g(\cdot)$ only "crosses" 0 once. Finally, one can easily verify that $\lim_{\alpha\to 0} g(\alpha) > 0$. Thus, we have $g(\alpha) > 0$ for all $\alpha \in (0, \alpha^*)$ and $g(\alpha) \le 0$ when $\alpha \ge \alpha^*$. Therefore, it is better for the vendor to serve dedicated if $c \ge \alpha^*/r$ and to operate batch otherwise, which implies the first statements in Proposition 3.

Next, we show the second statements in Proposition 3. Recall that the threshold on the wait cost c can be expressed as $\alpha = cr$ and the vendor switches from serving batch to dedicated when $\alpha = \alpha^*$. Thus, we only need to show that $\lambda_B^\infty > \lambda_F^\infty$ when $cr = \alpha = \alpha^*$, where the optimal non-zero demands λ_F^∞ and λ_B^∞ are defined in (A.4).

Note that for fixed r and $cr = \alpha^* \approx 0.1355$, we have

$$\lambda_B^\infty - \lambda_F^\infty = \frac{1}{r}\left(T_B - T_D + \frac{\alpha C_D T_D}{\sqrt{\alpha C_D[2T_D + \alpha C_D]}} - \frac{\sqrt{3}\alpha C_B}{\sqrt{\alpha C_B T_B[3T_B + \alpha(3C_B - T_B)]}}\right) > 0,$$

where we plug in the value of α^*, C_D, C_B, T_D, and T_B to reach the inequity. Thus, we conclude that when switching from serving batch to dedicated, the optimal demand rate decreases. Finally, recall the revenue function in (2) equals to demand times price, i.e., λp. When the optimal demand switches from λ_B to $\lambda_F < \lambda_B$ at $cr = \alpha^*$, the corresponding optimal price surges upwards accordingly. This completes the proof. □

Proof of Proposition 4. This proof follows from the same steps as the proof of Proposition 3. Fix $c>0$ and define function

$$h(\alpha) := \frac{1}{c}\left(V_B^\infty(\lambda_B^\infty) - V_D^\infty(\lambda_F^\infty)\right) = \left(2C_B - \frac{1}{2} - \frac{T_B}{3}\right) + \frac{T_B}{\alpha} - \frac{4\sqrt{3\alpha C_B[3T_B + \alpha(3C_B - T_B)]}}{\alpha} - \left[C_D + \frac{T_D}{\alpha} - \frac{\sqrt{\alpha C_D[2T_D + \alpha C_D]}}{\alpha}\right], \tag{A.8}$$

where $\alpha = cr$. By solving $h(\alpha) = 0$, we obtain the same unique solution $\alpha^* \approx 0.1355$. We omit the details for the rest of the proof to avoid repetition. □

Proof of Proposition 5. To show the first statement, we show that the social welfare function in (18) is an non-increasing function *w.r.t.* w by taking the first order derivative:

$$\frac{\partial SW(w)}{\partial w} = -c\int_{F^{-1}\left(1-\frac{\lambda}{\Lambda r^2}\right)}^{1} dF(v) = -c\left(1 - \frac{\lambda}{\Lambda r^2}\right) \le 0.$$

Therefore, when fixing λ, the smaller expected wait time translates to higher social welfare. Therefore, the first statement in Proposition 5 simply follows Propositions 1 and (2).

In order to show the second statement, we consider the large market regime and redefine the social welfare rate as

$$SW_n(\lambda, w) = \Lambda n r^2 \int_{F^{-1}\left(1-\frac{\lambda}{\Lambda n r^2}\right)}^{1} (v - cw) dF(v). \tag{A.9}$$

First, note that we have

$$\begin{aligned}
\lim_{n\to\infty} SW_n(\lambda, w) &= \lim_{n\to\infty} \Lambda n r^2 \int_{F^{-1}\left(1-\frac{\lambda}{\Lambda n r^2}\right)}^{1} (v - cw) dF(v) && \text{(A.10)}\\
&= \lambda \frac{\int_{F^{-1}\left(1-\frac{\lambda}{\Lambda n r^2}\right)}^{1} (v - cw) dF(v)}{\int_{F^{-1}\left(1-\frac{\lambda}{\Lambda n r^2}\right)}^{1} dF(v)} && \text{(A.11)}\\
&= \lambda \lim_{n\to\infty} \frac{-\frac{dF^{-1}\left(1-\frac{\lambda}{\Lambda c r^2}\right)}{dn} \left(F^{-1}\left(1-\frac{\lambda}{\Lambda c r^2}\right) - cw\right) f\left(F^{-1}\left(1-\frac{\lambda}{\Lambda c r^2}\right)\right)}{-\frac{dF^{-1}\left(1-\frac{\lambda}{\Lambda c r^2}\right)}{dn} f\left(F^{-1}\left(1-\frac{\lambda}{\Lambda c r^2}\right)\right)} && \text{(A.12)}\\
&= \lambda(1 - cw) := SW_{\inf}(\lambda, w), && \text{(A.13)}
\end{aligned}$$

where the first equality follows (1); the second equality follows L'Hopital rule and Leibniz rule.

Finally, note that we have both $\max\{SW_n(\lambda, W_D(\lambda, \mu_D, C_D)), 0\}$ and $\max\{SW_n(\lambda, W_B(\lambda, \mu_B, C_B)), 0\}$ are Lipschitz continuous. To see this, take

$$\max\{SW_n(\lambda, W_D(\lambda, \mu_D, C_D)), 0\} = \max\left\{-\Lambda n r^2 \int_{F^{-1}\left(1-\frac{\lambda}{\Lambda n r^2}\right)}^{1} v\, dF(v) + \lambda c W_D(\lambda, \mu_D, C_D), 0\right\},$$

as the example. It is easy to verify there exists a $\hat{\lambda}$ such that $\max\{SW_n(\lambda, W_D(\lambda, \mu_D, C_D)), 0\} = 0$ for all $\lambda \in [\hat{\lambda}, \mu_D)$ since $W_D(\lambda, \mu_D, C_D)$ is convex, increasing and approaching infinity as $\lambda \to \mu_D$. Thus, we only need to focus on $\lambda \in [0, \hat{\lambda}]$. Note that the derivative of $-\Lambda n r^2 \int_{F^{-1}\left(1-\frac{\lambda}{\Lambda n r^2}\right)}^{1} v\, dF(v)$ is

$$\Lambda n r^2 \frac{\partial F^{-1}\left(1-\frac{\lambda}{\Lambda n r^2}\right)}{\partial \lambda} F^{-1}\left(1 - \frac{\lambda}{\Lambda n r^2}\right) f\left(F^{-1}\left(1 - \frac{\lambda}{\Lambda n r^2}\right)\right) < \infty,$$

since $F^{-1}(\cdot)$ is a Lipschitz continuous function and thus $\frac{\partial F^{-1}\left(1-\frac{\lambda}{\Lambda n r^2}\right)}{\partial \lambda}$ is finite. Therefore, the term $-\Lambda n r^2 \int_{F^{-1}\left(1-\frac{\lambda}{\Lambda n r^2}\right)}^{1} v\, dF(v)$ is also Lipschitz continuous. It is also straight forward to verify that $\lambda c W_D(\lambda, \mu_D, C_D)$ is Lipschitz continuous (see the proof of Lemma 1). Thus, $\max\{SW_n(\lambda, W_D(\lambda, \mu_D, C_D)), 0\}$ is a Lipschitz continuous function. Then one can show that

$$\lim_{n\to 0} \max_{\lambda\in[0,\mu_D)} SW_n(\lambda, W_D(\lambda, \mu_D, C_D)) = \max_{\lambda\in[0,\mu_D)} \lim_{n\to 0} SW_n(\lambda, W_D(\lambda, \mu_D, C_D)) = \max_{\lambda\in[0,\mu_D)} \lambda\left(1 - W_D(\lambda, \mu_D, C_D)\right)$$

$$\lim_{n\to 0}\max_{\lambda\in[0,2\mu_B)} SW_n(\lambda, W_B(\lambda,\mu_B,C_B)) = \max_{\lambda\in[0,2\mu_B)}\lim_{n\to 0} SW_n(\lambda, W_B(\lambda,\mu_B,C_B)) = \max_{\lambda\in[0,2\mu_B)} \lambda\left(1-W_B(\lambda,\mu_B,C_B)\right),$$

following the exactly same proof techniques in the one of Lemma 1. We omit the details to avoid repetition.

Thus, under a crowded market, social welfare maximization over the demand rate is equivalent to revenue maximization in Section 5, thus, producing the same results. □

Proof of Proposition 6. We prove this result using a coupling argument. First, we note that the underlying stochasticity comes from the Poisson arrival of orders with the rate λ and their locations, which are independent and uniformly distributed on a disk. Furthermore, in all dispatch policies, since we normalized the courier's speed to 1, the service time is simply the travel distance.

Denote by L_D and L_C the number of unfilled orders under dedicated and contingent policies, respectively. We show path-wise dominance: $L_D \geq L_C$ when the underlying random events are coupled.

Consider a hypothetical *alternative policy* (short for "alternative contingent policy"), which mimics the contingent policy and follows its delivery decision for each order. More specifically, if two arrivals under the contingent policy are served in a batch, in the alternative policy, they are also served in the same batch. However, unlike the original contingent policy, the courier does not serve both orders in a single trip. Instead, the courier travels to the first location but then travels back to the hub before heading to the second location. That is, although two orders are leaving the hub together, the delivery routing is the same as that of dedicated strategy. Thus, the alternative policy's total travel distance/service time should be the same as the one under a dedicated policy, not taking advantage of spatial pooling.

As mentioned, to conduct this alternative policy, we need to mimic the contingent policy. More specifically, we need to know which orders need to be served in batches and which do not. We first argue that the alternative policy is feasible at any time t. That is, to execute this alternative policy, we only need information up to time t of the contingent policy and do not require clairvoyant information. Note that systems under the alternative policy and the contingent policy behave the same until the first time two orders need to be batched. For any two orders with locations $\mathbf{x}_1$ and $\mathbf{x}_2$, the travel distance under the contingent policy is always no greater than that of the alternative policy due to triangular inequality (i.e. $|\mathbf{x}_1| + |\mathbf{x}_2| \geq |\mathbf{x}_1 - \mathbf{x}_2|$). So after the first batch delivery occurs, the alternative policy induces longer service time and "lags" behind the contingent policy regarding served orders. Thus, by an induction argument, any subsequent delivery decision under the contingent policy (which will be mimicked by the alternative policy) is always made no later than the time the courier becomes idle or the arrival time of the next order in the alternative policy,

requiring no clairvoyant information for the alternative policy to be implemented. Next, denote by L_A the number of unfilled orders under the alternative contingent policy. Since the decisions of which orders are served in batches are the same between the two systems, and the alternative policy leads to no shorter service time, we have $L_A \geq L_C$ path-wise.

Next, note that the alternative policy is identical to the dedicated policy in terms of the total service time when coupled since the two policies induce the same routing policy. The only difference between the two policies is that two orders leave the queue whenever a batch decision is made in the alternative policy. However, only one order leaves the system at a time in the corresponding dedicated system. Thus we have $L_D \geq L_A$. Therefore, we have reached $L_D \geq L_A \geq L_C$. Finally, the result on the expected wait time holds by Little's Law.

We prove the second statement by considering extreme cases. When the demand rate is close to zero, the contingent policy leads to a shorter expected wait time than batch delivery since the time to accumulate orders in batches goes to infinity. On the other hand, when the demand rate is very large, which leads to a shorter order accumulation time than the travel time for a single order ($1/\lambda \ll 2r$), it is beneficial for the courier to serve in batches as opposed to adopting the contingent policy. This completes the proof. □

Before proving the rest of Propositions in Section 6, we first present a lemma on the properties of the revenue function of serving batch, when customer valuations follow a standard uniform distribution without the large market assumption.

LEMMA 2. *Consider function*

$$g(\lambda, r, c, \Lambda) := \lambda \left\{ 1 - \frac{\lambda}{\Lambda r^2} - c \left[\frac{1}{2\lambda} + \frac{\lambda r^2 C}{T(T - \lambda r)} + \frac{r}{3} \right] \right\}, \tag{A.14}$$

with $C, T > 0$.

(i) Fix $\Lambda, r > 0$. *Function* g *is submodular in* $(\lambda; c)$ *for* $c > 0$ *and* $\lambda \in (0, T/r)$.

(ii) Fix $c, \Lambda > 0$ *and* $C < 1$. *Function* g *is submodular in* $(\lambda; r)$ *for* $r > \left(\frac{T^2}{cC\Lambda} \right)^{\frac{1}{4}}$ *and* $\lambda \in (0, T/r)$.

(iii) Fix $c, r > 0$. *Function* g *is supermodular in* $(\lambda; \Lambda)$.

For the proofs of Propositions 7 and 8, denote by λ_D^* and λ_B^* the optimal solutions to (19) and (20), respectively.

Proof of Proposition 7. We show that, when fixing $r, \Lambda > 0$, there exist some c_{en} such that we have $V_B(\lambda_B^*, W_B(\lambda_B^*, \mu_B, C_B)) \leq V_D(\lambda_D^*, W_D(\lambda_D^*, \mu_D, C_D))$ if $c \geq c_{en}$.

With slight abuse notation, denote the optimal solution to

$$\max_{\lambda \in [0, 2\mu_B)} \lambda \left[1 - \frac{\lambda}{\Lambda r^2} - c W_B(\lambda, \mu_B, C_B) \right], \tag{A.15}$$

by $\lambda_B^*(c)$ for every $c>0$. Also recall Proposition 1 and denote by λ_{ex} the unique solution to $f(\lambda,r)=0$ for $r>0$, which is independent of c.

Consider $T=\dfrac{45\pi}{2(32+15\pi)}$ and $C=C_B$ for function g defined in (A.14). Recall $C_B\approx 0.583<1$ and we recognize function g is the revenue function in (A.15) with $T=T_B$ and $C=C_B$, so we can apply the result of Lemma 2 (i) directly. Thus, we have $\lambda_B^*(c)$ is decreasing when c is increasing using Topkis's Theorem (see, e.g., Topkis 1978) since the objective function is submodular in $(\lambda;c)$.

It is obvious that when c is large enough, we must have $\lambda_B^*(c)=0$. Thus, as $\lambda_{ex}>0$ is independent *w.r.t.* c where λ_{ex} is the threshold on which the wait times are equal under dedicated and batch in Proposition 1, there exists some $c_{en}\in(0,2\mu_B)$ such that for all $c>c_{en}$, we have $\lambda_{ex}>\lambda_B^*(c)$ since $\lambda_B^*(c)$ is decreasing *w.r.t.* c. Therefore, when $c\geq c_{en}$, we have

$$V_B(\lambda_B^*(c),W_B(\lambda_B^*(c),\mu_B,C_B))\leq \lambda_B^*(c)\left[1-\frac{\lambda_B^*(c)}{\Lambda r^2}-cW_D(\lambda_B^*(c),\mu_D,C_D)\right]\leq V_D(\lambda_D^*,W_D(\lambda_D^*,\mu_D,C_D)),$$

where the first inequality follows from Proposition 1 since $\lambda_{ex}>\lambda_B^*(c)$ and the second inequality follows from the definition of λ_D^* as the optimal solution. □

Proof of Proposition 8. We show that, when considering $\Lambda,c>0$ such that $\dfrac{\Lambda}{c^3}>L:=\dfrac{1+8(\sqrt{2C_B}+6C_B+8\sqrt{2C_B^3}+8C_B^2)}{16C_B}\left(\dfrac{2(32+16\pi)}{45\pi}\right)^2\approx 13.39$, there exists a threshold, r_{en}, on the service radius r, such that we have $V_B(\lambda_B^*,W_B(\lambda_B^*,\mu_B,C_B))\leq V_D(\lambda_D^*,W_D(\lambda_D^*,\mu_D,C_D))$ if $r\geq r_{en}$.

Denote by $\lambda_B^*(r)$ the optimal solution to

$$\max_{\lambda\in[0,1/(2\mu_B))}\lambda\left[1-\frac{\lambda}{\Lambda r^2}-cW_B(\lambda,\mu_B,C_B)\right], \tag{A.16}$$

when fixing $r>\underline{r}:=3\sqrt{\dfrac{5\pi}{32+15\pi}}\dfrac{1}{(0.583\Lambda c)^{\frac{1}{4}}}$. Note that we only need to consider the case where $\lambda_B^*(r)$ solves the first order condition since any $\lambda_B^*(r)$ that approaches the boundaries leads to a negative objective value, which implies $V_D(\lambda_D^*,W_D(\lambda_D^*,\mu_D,C_D))\geq V_B(\lambda_B^*,W_B(\lambda_B^*,\mu_B,C_B))$ immediately as $V_D(\lambda_D^*,W_D(\lambda_D^*,\mu_D,C_D))\geq 0$. Furthermore, note that by plugging in $T=\dfrac{2\mu_B}{r}=\dfrac{45\pi}{2(32+15\pi)}$ and $C=C_B$, function g in (A.14) is the revenue function of serving batch in (20). Since we have

$$C_B\approx 0.583<1,\ \text{and}\ r>\underline{r}=3\sqrt{\frac{5\pi}{32+15\pi}}\frac{1}{(0.583\Lambda c)^{\frac{1}{4}}}=\left(\frac{T^2}{\Lambda c C_B}\right)^{\frac{1}{4}},$$

we can apply Lemma 2(ii) directly and conclude that $\lambda_B^*(r)$ is decreasing *w.r.t* r when $r>\underline{r}$.

Denote $\bar{r}=45\pi/(c(32+15\pi)(1+2\sqrt{1.166}))$. When $r>\bar{r}$ have

$$\lambda\left[1-\frac{\lambda}{\Lambda r^2}-cW_B(\lambda,\mu_B,C_B)\right]=\lambda\left\{1-\frac{\lambda}{\Lambda r^2}-c\left[\frac{1}{2\lambda}+\frac{\lambda r^2 C_B}{T_B(T_B-\lambda r)}\right]\right\}$$

$$\leq \lambda\left\{1-c\left[\frac{1}{2\lambda}+\frac{\lambda r^2 C_B}{T(T-\lambda r)}\right]\right\}$$
$$\leq \lambda\left\{1-cr\frac{1+2\sqrt{2C_B}}{2T}\right\}\leq 0,$$

where the second inequality follows the fact that the term $\frac{1}{2\lambda}+\frac{\lambda r^2 C_B}{T(T-\lambda r)}$ is convex in λ and attains its minimum $r\frac{1+2\sqrt{2C_B}}{2T}$ when $\lambda=\frac{T(\sqrt{2C_B}-1)}{r(2C_B-1)}$, and the second inequality follows $r>\bar{r}$.

Note that there is

$$\bar{r}=\frac{45\pi}{c(32+15\pi)(1+2\sqrt{1.166})}=\frac{2T}{c(1+2\sqrt{2C_B})}>\left(\frac{T^2}{\Lambda c C_B}\right)^{\frac{1}{4}}=\underline{r},$$

where the inequality follows from $\frac{\Lambda}{c^3}>L$. As $\lambda_B^*(r)$ is decreasing *w.r.t.* r, we have $\lim_{r\to\bar{r}}\lambda_B^*(r)=0$. Recall that for a finite r, we have λ_{ex}, the solution to $W_D(\lambda,\mu_D,C_D)=W_B(\lambda,\mu_B,C_B)$, never equals to 0. Thus, denote $\ell=\min_{r\in(\underline{r},\bar{r})}\lambda_{ex}$, which is strictly greater than 0. Then by definition of the limit, there exists some $r_{en}\in(\underline{r},\bar{r})$ such that $\lambda_B^*(r)\leq\ell\leq\lambda_{ex}$ for all $r\geq r_{en}$.

Finally, consider $r\geq r_{en}$. We have

$$V_B(\lambda_B^*,W_B(\lambda_B^*(r),\mu_B,C_B))\leq\lambda_B^*(r)\left[1-\frac{\lambda_B^*(r)}{\Lambda r}-cW_D(\lambda_B^*(r),\mu_D,C_D)\right]\leq V_D(\lambda_D^*,W_D(\lambda_D^*,\mu_D,C_D)),$$

where the first inequality follows from Proposition 1 since we have $\lambda_B^*(r)\leq\lambda_{ex}$ for all $r\geq r_{en}$ and the second inequality follows from that λ_D^* is the optimal solution. This completes the proof. □

Proof of Proposition 9. The proof of the first statement follows from the exactly same steps as the proof of Propositions 3 and 4. We only need to substitute the values of constants by $T_B=2\mu_{B,C}r=\frac{4}{4+\pi}$, $T_D=2\mu_{D,C}r=\frac{1}{2}$, $C_{B,C}=\frac{1}{2}+\frac{\pi^2}{3(4+\pi)^2}$, and $C_{D,C}=1$. We still denote $\alpha=cr$ and let $\hat{\alpha}^*$ be the new threshold in cr when serving a circular region. We have $\hat{\alpha}^*\approx 0.057\ll 0.1809\approx\alpha^*$, where α^* is the optimal value of alpha when the service region is a disk. That is, when $cr\geq\hat{\alpha}^*$, serving dedicated is better than batch, and vice versa. We omit the details to avoid repetition. □

Proof of Proposition 10. To show the first statement, we prove that for any radii $r_1\geq r_2$, we have $\lambda_D^*(r_1)\leq\lambda_D^*(r_2)$ and $\lambda_B^*(r_1)\leq\lambda_B^*(r_2)$. Since the vendor uses a single fixed price, denote by p_D^* and p_B^* the optimal prices under dedicated and batch services. Then, for a disk with radius r, we must have

$$p_D^*=F^{-1}\left(1-\frac{2\lambda_D^*(s)}{\Lambda r^2}\right)-cW_D(s)\text{ and }p_B^*=F^{-1}\left(1-\frac{2\lambda_B^*(s)}{\Lambda r^2}\right)-cW_B(s),\tag{A.17}$$

whenever $\lambda_D^*(s)$, $\lambda_B^*(s)>0$, $s\in[0,1]$, where $W_D(\cdot)$ and $W_B(\cdot)$ are expected wait times for customers at different locations under dedicated and batch services, respectively. Whenever $\lambda_D^*(s)=\lambda_B^*(s)=0$ for some $s\in[0,1]$, we have

$$p_D^*\geq F^{-1}\left(1-\frac{2\lambda_D^*(s)}{\Lambda r^2}\right)-cW_D(s)\text{ and }p_B^*\geq F^{-1}\left(1-\frac{2\lambda_B^*(s)}{\Lambda r^2}\right)-cW_B(s).\tag{A.18}$$

We show the desired results by contradiction. Suppose for $r_1 \geq r_2$, we have $\lambda_D^*(r_1) > \lambda_D^*(r_2) \geq 0$ and $\lambda_B^*(r_1) > \lambda_B^*(r_2) \geq 0$, respectively. Since F^{-1} is a non-decreasing function, we must have

$$F^{-1}\left(1 - \frac{2\lambda_D^*(r_1)}{\Lambda r^2}\right) \leq F^{-1}\left(1 - \frac{2\lambda_D^*(r_2)}{\Lambda r^2}\right), \text{ and } F^{-1}\left(1 - \frac{2\lambda_B^*(r_1)}{\Lambda r^2}\right) \leq F^{-1}\left(1 - \frac{2\lambda_B^*(r_2)}{\Lambda r^2}\right). \tag{A.19}$$

Furthermore, note that in this setting of distance-dependent wait time, the expected wait time consists of two parts: in-line delay and the expected travel time of the courier. Although the in-line delay is the same for all customers despite different locations, the expected travel time for the courier is increasing *w.r.t.* the distance from the vendor. Thus, we have $W_D(r_1) > W_D(r_2)$ and $W_B(r_1) > W_B(r_2)$. Combining this with (A.19), we have

$$\begin{aligned} p_D^* &= F^{-1}\left(1 - \frac{2\lambda_D^*(r_1)}{\Lambda r^2}\right) - cW_D(r_1) < F^{-1}\left(1 - \frac{2\lambda_D^*(r_2)}{\Lambda r^2}\right) - cW_D(r_2) \leq p_D^*, \text{ and} \\ p_B^* &= F^{-1}\left(1 - \frac{2\lambda_B^*(r_1)}{\Lambda r^2}\right) - cW_B(r_1) < F^{-1}\left(1 - \frac{2\lambda_B^*(r_2)}{\Lambda r^2}\right) - cW_B(r_2) \leq p_B^*, \text{ respectively,} \end{aligned}$$

which follows (A.17) and (A.18). Thus, we have reached the contraction and this completes the proof for the desired statement.

The second statement in Proposition 10 follows directly from the first statement. The arguments for dedicated and batch service are the same so we also show the result for a dedicated delivery system. Suppose the optimal price leads to that $\lambda_D(s)$ equals to zero. Then, for any $s' \in [s, r]$, we have $\lambda_D(s') = 0$. Thus, the service region is shrunk to a smaller disk. This completes the proof. □

B. Supplementary Proofs in Section 5

Proof of Lemma 1. We only show the argument for function $V_n(\lambda, W_D)$ since the same steps can be applied to function $V_n(\lambda, W_B)$.

Denote function $J_n := \min\{-V_n(\lambda, W_D), 0\}$. Thus, maximizing $V_n(\lambda, W_D)$ is equivalent to minimizing J_n. Furthermore, by equation (15), denote function

$$J := \lim_{n\to\infty} J_n = \min\{-\lambda(1 - cW_D(\lambda, \mu_D, C_D)), 0\}.$$

The rest of this proof is broken into three steps:

1. We show that for each $\lambda \in [0, \mu_D)$, there exists a sequence $\{\lambda_n'\}$ converging to λ such that

$$\lim_{n\to\infty} J_n(\lambda_n') = J(\lambda). \tag{B.1}$$

2. We show that for every $\lambda \in [0, \mu_D)$ and for every sequence $\{\lambda_n'\}$ converging to λ, there is

$$\liminf_{n\to\infty} J_n(\lambda_n') = J(\lambda). \tag{B.2}$$

3. Once conditions in Steps 1 and 2 are satisfied, we have that function J_n $\Gamma-$converges (see, e.g., Dal Maso 1993) to function J, which implies that

$$\lim_{n\to\infty} \min J_n(\lambda) = \min \lim_{n\to\infty} J_n(\lambda).$$

Thus, the desired result can be obtained.

Step 1: We show a stronger result here: for every $\lambda \in [0, \mu_D)$ and for every sequence $\{\lambda_n\}$ converging to λ, the equation in (B.1) holds. This stronger result also helps us to show the statement in Step 2.

We begin by showing that function J_n is Lipschitz continuous for every $n \in \mathbb{N}$. First, note that $0 \le F^{-1}\left(1 - \frac{\lambda}{\Lambda n r^2}\right) \le 1$, since w.l.o.g., the bounded support of the valuation distribution function F is normalized to $[0,1]$. From the proof of Proposition 1, we know that the wait time function $W_D(\lambda, \mu_D, C_D)$ is strictly convex and increasing in λ with $\lim_{\lambda\to\mu_D} W_D(\lambda, \mu_D, C_D) = \infty$. Therefore, there exists some $\hat{\lambda} < \mu_D$, such that

$$F^{-1}\left(1 - \frac{\lambda}{\Lambda n r^2}\right) - cW_D(\lambda, \mu_D, C_D) \le 0, \text{ so that } J_n(\lambda) = 0, \quad \forall \lambda > \hat{\lambda}.$$

Furthermore, denote K_F as the Lipschitz constant for function F^{-1} and

$$K_W = \frac{\partial W_D(\lambda, \mu_D, C_D)}{\partial \lambda}\Big|_{\lambda=\hat{\lambda}}. \tag{B.3}$$

Then we have that function W_D is Lipschitz continuous when $\lambda \in (0, \hat{\lambda})$ with constant K_W since function W_D is strictly convex from the proof of Proposition 1. Furthermore, we have function

$$g_n(\lambda) := -\lambda\left[F^{-1}\left(1 - \frac{\lambda}{\Lambda n r^2}\right) - cW_D(\lambda, \mu_D, C_D)\right],$$

is also Lipschitz continuous with constant

$$K_n = \mu_D \max\left\{\frac{K_F}{\Lambda n r^2}, cK_W\right\}, \tag{B.4}$$

since function $g_n(\lambda)/\lambda$ is the difference between two Lipschitz continuous functions and $\lambda < \mu_D$. Since there is $J_n = \min\{0, g_n\}$, we conclude that function J_n is Lipschitz with factor K_n.

Next, we have

$$\begin{aligned} |J_n(\lambda) - J(\lambda)| &\le |g_n(\lambda) + \lambda(1 - cW_D(\lambda, \mu_D, C_D))| \\ &= \lambda\left(1 - F^{-1}\left(1 - \frac{\lambda}{\Lambda n r^2}\right)\right) \\ &< \mu_D\left(1 - F^{-1}\left(1 - \frac{\lambda}{\Lambda n r^2}\right)\right), \end{aligned} \tag{B.5}$$

according to the definitions of functions J_n and J, and $\lambda < \mu_D$.

Consider any $\lambda \in [0, \mu_D)$ and any sequence $\{\lambda'_n\}$ converging to λ. By the definition of convergence, consider $\epsilon > 0$ and we can find N_1 such that there is

$$|\lambda'_n - \lambda| < \frac{\epsilon}{2\bar{K}}, \quad \forall n \geq N_1, \tag{B.6}$$

where $\bar{K} = \max\{K_n | n \geq N_1\}$. Furthermore, since F^{-1} is Lipschitz continuous and monotone increasing, fixing $\epsilon > 0$, we can find N_2 such that

$$1 - F^{-1}\left(1 - \frac{\lambda}{\Lambda n r^2}\right) \leq \frac{\epsilon}{2\mu_D}, \quad \forall n \geq N_2. \tag{B.7}$$

Fix $n > N_\epsilon := \max\{N_1, N_2\}$, we have

$$\begin{aligned}
|J_n(\lambda_n) - J(\lambda)| &\leq |J_n(\lambda_n) - J_n(\lambda)| + |J_n(\lambda) - J(\lambda)| \\
&\leq \bar{K}|\lambda_n - \lambda| + \mu_D \left(1 - F^{-1}\left(1 - \frac{\lambda}{\Lambda n r^2}\right)\right) \\
&\leq \bar{K}|\lambda_n - \lambda| + \frac{\epsilon}{2} \\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,
\end{aligned}$$

where the first inequality follows from the triangular inequality; the second inequality follows from (B.5) and that function J_n is Lipschitz with constant $K_n \leq \bar{K}$; the third inequity follows from (B.7); and the last inequality follows from (B.6). Therefore, we conclude that (B.1) holds for every $\lambda \in [0, \mu_D)$ and for every sequence $\{\lambda_n\}$ converging to λ.

Step 2: The stronger statement we have shown in step 1 implies the desired result in this step. Consider any $\lambda \in [0, \mu_D)$ and any sequence $\{\lambda_n\}$ converging to λ. Fix $n > N_\epsilon = \max\{N_1, N_2\}$ and denote $m^* \in \arg\inf_{m \geq n} J_m(x_m)$. We have

$$\begin{aligned}
|J_{m^*}(\lambda^*_m) - J(\lambda)| &\leq |J_{m^*}(\lambda^*_m) - J_{m^*}(\lambda)| + |J_{m^*}(\lambda) - J(\lambda)| \\
&\leq \bar{K}|\lambda^*_m - \lambda| + \mu_D \left(1 - F^{-1}\left(1 - \frac{\lambda}{\Lambda n r^2}\right)\right) \leq \epsilon.
\end{aligned}$$

Step 3: Now we can conclude that function J_n Γ-Converges to function J as n approaches infinity, since we have: 1) for each $\lambda \in [0, \mu_D)$, there exists a sequence $\{\lambda_n\}$ converging to λ such that equation (B.1) holds from step 1; 2) equation (B.2) holds for every $\lambda \in [0, \mu_D)$ and for every sequence $\{\lambda_n\}$ converging to λ from step 2. Thus, by the property of Γ-Convergence, we have $\lim_{n\to\infty} \max_{\lambda \in [0, \mu_D)} V_D^n(\lambda) = \max_{\lambda \in [0, \mu_D)} \lim_{n\to\infty} V_D^n(\lambda)$, and this completes the proof. □

C. Supplementary Results and Proofs in Section 6

C.1. General Arrival Rate

Proof of Lemma 2. Noting that the function g is continuous and twice differentiable *w.r.t.* each variable, we can verify the statements by taking the mixed second derivatives.

(i) Fix $\Lambda, r > 0$ and we have

$$\frac{\partial^2 g(\lambda, r, c, \Lambda)}{\partial \lambda \partial c} = Cr\left(\frac{1}{T} - \frac{T}{(T-\lambda r)^2}\right) - \frac{r}{3} = \frac{C\lambda r^2(\lambda r - 2T)}{T(T-\lambda r)^2} - \frac{r}{3} < 0,$$

where the inequality follows from $\lambda r < T$.

(ii) Fix $c, \Lambda > 0$ and there is

$$\frac{\partial^2 g(\lambda, r, c, \Lambda)}{\partial \lambda \partial r} = \frac{4\lambda}{\Lambda r^3} + \frac{cC}{T}\left(1 - \frac{T^2(T+\lambda r)}{(T-\lambda r)^3}\right) - \frac{c}{3} \tag{C.1}$$

We verify the right-hand-side of (C.1) is decreasing in λ by taking its derivative *w.r.t.* λ:

$$\frac{4}{\Lambda r^3} - \frac{2crCT(2T+\lambda r)}{(T-\lambda r)^4} < \frac{4}{\Lambda r^3} - \frac{4crC}{T^2} < 0,$$

where the first inequality follows that the term $\dfrac{2crCT(2T+\lambda r)}{(T-\lambda r)^4}$ is increasing in λ, thus, letting $\lambda = 0$, and the second inequality follows $r > \left(\dfrac{T^2}{cC\Lambda}\right)^{\frac{1}{4}}$. Therefore, the expression in (C.1) reaches its maximum $-c/3$ when $\lambda = 0$, suggesting submodularity.

(iii) Fix $c, r > 0$ and we have

$$\frac{\partial^2 g(\lambda, r, c, \Lambda)}{\partial \lambda \partial \Lambda} = \frac{2\lambda}{\Lambda^2 r^2} > 0,$$

and this completes the proof. □

C.2. Discussion of the Uniform Distribution Assumption on Customers' Valuation

As mentioned in Section 6, the assumption that function $F(v) = v$ for $v \in [0,1]$ and $F(v) = 0$ is not restrictive. According to the proof of Proposition 7 and 8, all we need is that the revenue function when serving batch V_B is submodular in $(\lambda; c)$ and $(\lambda; r)$, respectively. Thus, as long as the distribution function F of customer valuations induces a inverse function F^{-1} leading to submodularity, similar to Lemma 2 (i) and (ii), we can still find thresholds in wait cost c and radius r above which, serving dedicated is optimal.

In particular, for any continuous and twice differentiable function F^{-1}, the result in Proposition 7 still holds. To see this, note that the base price $F^{-1}\left(1 - \dfrac{\lambda}{\Lambda r^2}\right)$ is not a function of c, so it does not affect submodularity of the revenue function. Thus, the result in lemma 2 still holds.

Similarly, we can extend Proposition 8 under any continuous and twice differentiable function F^{-1} such that

$$\lambda \frac{\partial^2 F^{-1}\left(1 - \frac{\lambda}{\Lambda r^2}\right)}{\partial \lambda \partial r} + \frac{\partial F^{-1}\left(1 - \frac{\lambda}{\Lambda r^2}\right)}{\partial r} + \frac{cC}{T}\left(2 - \frac{T^2(T+\lambda r)}{(T-\lambda r)^3}\right) - \frac{c}{3} < 0,$$

where $T = 2\mu_B/r$ and $C = C_B$.

C.3. Circular City

PROPOSITION C.1. *Consider a circular service area with radius r. When the demand rate $\lambda > 0$ is exogenous:*

(i) There exists a threshold on the demand rate, below which serving dedicated leads to a shorter wait time and thus a higher revenue, and above which serving batch is optimal.

(ii) There exists a threshold on the service radius, below which serving dedicated leads to a shorter wait time and thus a higher revenue, and above which serving batch is optimal.

Without the large market assumption, when the demand rate λ is endogenously determined by the vendor and customer valuations follow a standard uniform distribution:

(iii) There exists a threshold on the customers' wait cost parameter, above which it is optimal to serve dedicated.

(iv) There exists a threshold on the service radius, above which it is optimal to serve dedicated.

Proposition C.1 confirms that all the major results in Sections 4 and 5 still hold even if we change the service area from a disk to a circle.

Proof of Proposition C.1. We only present a proof sketch for each statement as the details greatly resemble the previous proofs by substituting in $C_{D,C}$, and $C_{B,C}$ for C_D, and C_B:

Statements in (i) and (ii) follow from the proof of Propositions 1 and 2, with $T_1 = \mu_{D,C} r$, $T_2 = 2\mu_{B,C} r$, $C_1 = 1$ and $C_2 = \frac{1}{2} + \frac{\pi^2}{3(4+\pi)^2}$. Then we verify that the cubic equation $f(a) = 0$ has a unique solution on $a \in (0, T_1)$, which completes the proof.

Statement (iii) follows from the proof of Proposition 7, and statement (iv) follows from the proof of Proposition 8. We omit the details. □

C.4. Multiple Couriers

PROPOSITION C.2. *Consider the vendor hires $k \geq 2$ couriers covering the service area.*

(i) When the demand rate $\lambda > 0$ is exogenous, there exists a threshold on the demand rate, below which serving dedicated leads to a shorter wait time and thus a higher revenue, and above which serving batch is optimal.

(ii) When the demand rate λ is endogenously determined by the vendor and customer valuations follow a standard uniform distribution, there exists a threshold on the customers' wait cost parameter, above which it is optimal to serve dedicated.

As Proposition C.2 suggests, there is still a threshold on the exogenous demand rate, below which serving dedicated leads to a smaller expected wait time and above which serving batch has the edge when there are k couriers. Furthermore, when the wait cost parameter is relatively large, we still have that serving dedicated dominates serving batch.

Proof of Proposition C.2. (i) For notational convenience, we denote $\mu_1 := 2\mu_{B,k}$, $\mu_2 := \mu_{F,k}$ and $a := \sqrt{2(k+1)}$. By definition, we have $\mu_2 > \mu_1$. Furthermore, define function

$$\begin{aligned} f(\lambda) &:= W_{F,k} - W_{B,k} \\ &= \frac{C_D}{2(k\mu_1-\lambda)}\left(\frac{\lambda}{k\mu_1}\right)^{a-1} - \left(\frac{1}{2\lambda} + \frac{C_B}{2(k\mu_2-\lambda)}\left(\frac{\lambda}{k\mu_2}\right)^{a-1} + \frac{r}{3}\right), \lambda \in (0,\mu_1). \end{aligned} \tag{C.2}$$

Note that function f represents the difference in wait times when serving dedicated and batch. First, we show that function f is strictly increasing *w.r.t.* λ by taking the first order derivative. We have

$$\frac{\partial f(\lambda)}{\partial \lambda} = \frac{1}{2\lambda^2}\left(1 + C_D h(\mu_1) - C_B h(\mu_2)\right),$$

where

$$h(\mu,\lambda) = \frac{k\mu\left(\frac{\lambda}{k\mu}\right)^a (k\mu(a-1) - \lambda(a-2))}{(k\mu-\lambda)^2}, \mu \in (\lambda/k, \infty). \tag{C.3}$$

Note that function h is non-increasing *w.r.t.* μ since by taking the first order derivative, we have

$$\begin{aligned} \frac{\partial h(\mu,\lambda)}{\partial \mu} &= \frac{k\left(\frac{\lambda}{k\mu}\right)^a}{(k\mu-\lambda)^3}\left[(a-1)ak^2\mu^2 - 2(a-2)ak\lambda\mu + (a-1)(a-2)\lambda^2\right] \\ &= \frac{k\left(\frac{\lambda}{k\mu}\right)^a}{(k\mu-\lambda)^3}\left\{a(k\mu-\lambda)[(a-1)k\mu - (a-2)\lambda] + \lambda[ak\mu - (a-2)\lambda]\right\} \geq 0, \end{aligned}$$

where the inequality follows from that $k\mu > \lambda$. Thus, we have

$$\frac{\partial f(\lambda)}{\partial \lambda} \geq \frac{1}{2\lambda^2}\left(1 + C_D h(\mu_1,\lambda) - C_B h(\mu_2,\lambda)\right) > 0,$$

where the first inequality follows from that function h in non-increasing in μ together with $\mu_2 > \mu_1$ and the second inequality follows from $C_D > C_B$.

Next, by acknowledging function f goes to negative infinity and positive infinity when λ approaches 0 and μ_1, respectively, we can reach the first statement in Proposition C.2, since $f(\lambda) = 0$ has a unique solution.

(ii) We only need to verify that the revenue function of serving batch,

$$U(\lambda, c) := \lambda\left\{1 - \frac{\lambda}{\Lambda r^2} - c\left[\frac{1}{2\lambda} + \frac{C_B}{2(2k\mu_1-\lambda)}\left(\frac{\lambda}{2k\mu_1}\right)^{a-1} + \frac{r}{3}\right]\right\}, \tag{C.4}$$

is submodular in (λ, c). By taking its derivatives *w.r.t.* λ and c, we have

$$\frac{\partial^2 U(\lambda,c)}{\partial\lambda\partial c} = -\frac{C_B\left(\frac{\lambda}{k\mu_1}\right)^{a-1}}{2(k\mu_1-\lambda)^2}(ak\mu - (a-1)\lambda) - \frac{r}{3} < 0,$$

where the inequality follows from $k\mu - \lambda > 0$.

The rest of the proof follows from the exactly same steps as in the proof of Proposition 7. Thus, we omit the details. □

D. Distance-Dependent Wait Time

We dedicate this section to further elaborate the model where customers are sensitive to the courier's travel time, besides in-line delay, in Section 6.6.

Unfortunately, it is difficult even to conduct numerical analysis or simulations based on a disk-shaped area since we do not have closed-form expressions on the demand distribution with respect to customers' locations. The cumulative demand on the entire disk is an equilibrium outcome as customers need to use it when deciding whether to place an order or not. To make the analysis tractable, we consider a simplified city structure where the disk is stripped down to two rings, with radii $\bar{r} > \underline{r}$, respectively. Orders only come from locations on the two rings. Since customers are also sensitive to the travel time, we can expect that the customers on the outer ring need to wait longer on average compared to those on the inner ring. As a result, the demand for the outer ring can be different from that of the inner ring. We also assume that the underlying arrival rate of customers is $\Lambda\bar{r}$ and $\Lambda\underline{r}$, respectively, proportional to the circumference of rings.

We consider dedicated delivery first. Since orders are coming from somewhere either on the outer ring or on the inner ring, then w.l.o.g., denote by $\lambda_{D,\bar{r}}$ and $\lambda_{D,\underline{r}}$ the demand rate at these locations, respectively. Then, given wait times $w_{\bar{r}}$ and $w_{\underline{r}}$ on the two rings, respectively, and price p, we have

$$\frac{\lambda_{D,\bar{r}}}{\Lambda\bar{r}} = \mathbb{P}(v \geq cw_{\bar{r}} + p), \text{ and } \frac{\lambda_{D,\underline{r}}}{\Lambda\underline{r}} = \mathbb{P}(v \geq cw_{\underline{r}} + p). \tag{D.1}$$

As a result, whenever an order arrives, with probability $\dfrac{\lambda_{D,\underline{r}}}{\lambda_{D,\underline{r}} + \lambda_{D,\bar{r}}}$ the order is on the inner ring and with probability $\dfrac{\lambda_{D,\bar{r}}}{\lambda_{D,\underline{r}} + \lambda_{D,\bar{r}}}$, it comes from the outer ring. Thus, the courier's total travel distance per delivery trip is either $2\bar{r}$ if the order is on the outer ring, or $2\underline{r}$ if the order is on the inner ring. Thus, we use a Bernoulli random variable $\tilde{X}_D$ to denote the courier's travel distance, where

$$\mathbb{E}[\tilde{X}_D] = 2\bar{r}\frac{\lambda_{D,\bar{r}}}{\lambda_{D,\underline{r}} + \lambda_{D,\bar{r}}} + 2\underline{r}\frac{\lambda_{D,\underline{r}}}{\lambda_{D,\underline{r}} + \lambda_{D,\bar{r}}}, \text{ and } \mathbb{E}[\tilde{X}_D^2] = 4\bar{r}^2\frac{\lambda_{D,\bar{r}}}{\lambda_{D,\underline{r}} + \lambda_{D,\bar{r}}} + 4\underline{r}^2\frac{\lambda_{D,\underline{r}}}{\lambda_{D,\underline{r}} + \lambda_{D,\bar{r}}}. \tag{D.2}$$

Just like the base model, we can treat this system as an M/G/1 queue with

$$\tilde{\mu}_D = \frac{1}{\mathbb{E}[\tilde{X}_D]} = \frac{\lambda_{D,\underline{r}} + \lambda_{D,\bar{r}}}{2(\underline{r}\lambda_{D,\underline{r}} + \bar{r}\lambda_{D,\bar{r}})}, \text{ and } \tilde{\rho}_D = \frac{\lambda_{D,\underline{r}} + \lambda_{D,\bar{r}}}{\tilde{\mu}_D} = 2(\underline{r}\lambda_{D,\underline{r}} + \bar{r}\lambda_{D,\bar{r}}). \tag{D.3}$$

We then derive the coefficient of variation of the arrival and service processes as

$$\tilde{C}_D = 1 + \frac{\mathbb{E}[\tilde{X}_D^2] - (\mathbb{E}[\tilde{X}_D])^2}{(\mathbb{E}[\tilde{X}_D])^2} = \frac{(\lambda_{D,\underline{r}} + \lambda_{D,\bar{r}})(\underline{r}^2\lambda_{D,\underline{r}} + \bar{r}^2\lambda_{D,\bar{r}})}{(\underline{r}\lambda_{D,\underline{r}} + \bar{r}\lambda_{D,\bar{r}})^2}. \tag{D.4}$$

Finally, we define the vendor's revenue as

$$\tilde{V}_D = (\lambda_{D,\underline{r}} + \lambda_{D,\bar{r}})\, p$$

$$
\begin{aligned}
&= \lambda_{D,\underline{r}}\left[F^{-1}\left(1-\frac{\lambda_{D,\underline{r}}}{\Lambda \underline{r}}\right) - c(W_D(\lambda_{D,\underline{r}}+\lambda_{D,\bar{r}},\tilde{\mu}_D,\tilde{C}_D)+\underline{r})\right] \\
&\quad +\lambda_{D,\bar{r}}\left[F^{-1}\left(1-\frac{\lambda_{D,\bar{r}}}{\Lambda \bar{r}}\right) - c(W_D(\lambda_{D,\underline{r}}+\lambda_{D,\bar{r}},\tilde{\mu}_D,\tilde{C}_D)+\bar{r})\right] \\
&= \lambda_{D,\underline{r}}\left[F^{-1}\left(1-\frac{\lambda_{D,\underline{r}}}{\Lambda \underline{r}}\right) - c\left(\frac{2(\underline{r}^2\lambda_{D,\underline{r}}+\bar{r}^2\lambda_{D,\bar{r}})}{1-2(\underline{r}\lambda_{D,\underline{r}}+\bar{r}\lambda_{D,\bar{r}})}+\underline{r}\right)\right] \\
&\quad +\lambda_{D,\bar{r}}\left[F^{-1}\left(1-\frac{\lambda_{D,\bar{r}}}{\Lambda \bar{r}}\right) - c\left(\frac{2(\underline{r}^2\lambda_{D,\underline{r}}+\bar{r}^2\lambda_{D,\bar{r}})}{1-2(\underline{r}\lambda_{D,\underline{r}}+\bar{r}\lambda_{D,\bar{r}})}+\bar{r}\right)\right]. \qquad \text{(D.5)}
\end{aligned}
$$

Next, we consider batch delivery. Denote the demand rates on the outer and inner rings by $\lambda_{B,\bar{r}}$ and $\lambda_{B,\underline{r}}$, respectively. Denote by $\tilde{X}_B$ the courier's travel distance when serving batch delivery. Then we have

$$
\begin{aligned}
\mathbb{E}[\tilde{X}_B] &= 2\frac{\lambda_{B,\bar{r}}}{\lambda_{B,\bar{r}}+\lambda_{B,\underline{r}}}\frac{\lambda_{B,\underline{r}}}{\lambda_{B,\bar{r}}+\lambda_{B,\underline{r}}}\left(\underline{r}+\bar{r}+\frac{1}{\pi}\int_0^{\pi}\sqrt{\bar{r}^2+\underline{r}^2-2\bar{r}\underline{r}\cos(\theta)}\,d\theta\right) \\
&\quad +\left(2\bar{r}+\frac{\pi\bar{r}}{2}\right)\left(\frac{\lambda_{B,\bar{r}}}{\lambda_{B,\bar{r}}+\lambda_{B,\underline{r}}}\right)^2+\left(2\underline{r}+\frac{\pi\underline{r}}{2}\right)\left(\frac{\lambda_{B,\underline{r}}}{\lambda_{B,\underline{r}}+\lambda_{B,\underline{r}}}\right)^2 \\
\mathbb{E}[\tilde{X}_B^2] &= 2\frac{\lambda_{B,\bar{r}}}{\lambda_{B,\bar{r}}+\lambda_{B,\underline{r}}}\frac{\lambda_{B,\underline{r}}}{\lambda_{B,\bar{r}}+\lambda_{B,\underline{r}}}\frac{1}{\pi}\int_0^{\pi}\left(\underline{r}+\bar{r}+\sqrt{\bar{r}^2+\underline{r}^2-2\bar{r}\underline{r}\cos(\theta)}\right)^2 d\theta \\
&\quad +\left(\frac{\lambda_{B,\bar{r}}}{\lambda_{B,\bar{r}}+\lambda_{B,\underline{r}}}\right)^2\int_0^{\pi\bar{r}}(2\bar{r}+u)^2\frac{1}{\pi\bar{r}}\,du+\left(\frac{\lambda_{B,\underline{r}}}{\lambda_{B,\bar{r}}+\lambda_{B,\underline{r}}}\right)^2\int_0^{\pi\underline{r}}(2\underline{r}+u)^2\frac{1}{\pi\underline{r}}\,du. \qquad \text{(D.6)}
\end{aligned}
$$

We can also write the service rate as $\tilde{\mu}_B = 1/\mathbb{E}[\tilde{X}_B]$.

Just like the base model, we can still treat the arrival process as a Erlang-2 process with arrival rate $\frac{1}{2}(\lambda_{B,\underline{r}}+\lambda_{B,\bar{r}})$. Thus, this is again a $E_2/G/1$ queue. Finally, the coefficient of variation is

$$
\tilde{C}_B = \frac{1}{2}+\frac{\mathbb{E}[\tilde{X}_B^2]-(\mathbb{E}[\tilde{X}_B])^2}{(\mathbb{E}[\tilde{X}_B])^2}. \qquad \text{(D.7)}
$$

Using the same approximation in the base model, namely w_q as in (11), a customer on the inner ring has the expected wait time

$$
\frac{1}{2(\lambda_{B,\underline{r}}+\lambda_{B,\bar{r}})}+w_q+\underline{r} \approx \frac{1}{2(\lambda_{B,\underline{r}}+\lambda_{B,\bar{r}})}+\frac{\tilde{C}_B}{2}\frac{(\lambda_{B,\underline{r}}+\lambda_{B,\bar{r}})}{\tilde{\mu}_B(2\tilde{\mu}_B-(\lambda_{B,\underline{r}}+\lambda_{B,\bar{r}}))}+\underline{r}, \qquad \text{(D.8)}
$$

and a customer on the outer ring has the expected wait time

$$
\begin{aligned}
&\frac{1}{2(\lambda_{B,\underline{r}}+\lambda_{B,\bar{r}})}+w_q+\underline{r}+\frac{1}{\pi}\int_0^{\pi}\sqrt{\bar{r}^2+\underline{r}^2-2\bar{r}\underline{r}\cos(\theta)}d\theta \\
\approx\ &\frac{1}{2(\lambda_{B,\underline{r}}+\lambda_{B,\bar{r}})}+\frac{\tilde{C}_B}{2}\frac{(\lambda_{B,\underline{r}}+\lambda_{B,\bar{r}})}{\tilde{\mu}_B(2\tilde{\mu}_B-(\lambda_{B,\underline{r}}+\lambda_{B,\bar{r}}))}+\underline{r}+\frac{1}{\pi}\int_0^{\pi}\sqrt{\bar{r}^2+\underline{r}^2-2\bar{r}\underline{r}\cos(\theta)}d\theta, \qquad \text{(D.9)}
\end{aligned}
$$

where we assume the courier always delivers to the inner ring before delivering to the outer ring when serving in batches[8]. Thus, we can write the vendor's revenue function when serving batch as

$$
\tilde{V}_B = (\lambda_{B,\underline{r}}+\lambda_{B,\bar{r}})p
$$

[8] Our numerical studies in this section can be easily generalized to the case where the fulfillment order prioritizes the outer ring or is in a random order.

$$= \lambda_{B,\bar{r}}\left[F^{-1}\left(1-\frac{\lambda_{B,\bar{r}}}{\Lambda\bar{r}}\right)-c\left(\frac{1}{2(\lambda_{B,\underline{r}}+\lambda_{B,\bar{r}})}+w_q+\underline{r}+\frac{1}{\pi}\int_0^{\pi}\sqrt{\bar{r}^2+\underline{r}^2-2\bar{r}\underline{r}\cos(\theta)}d\theta\right)\right]$$
$$+\lambda_{B,\underline{r}}\left[F^{-1}\left(1-\frac{\lambda_{B,\underline{r}}}{\Lambda\underline{r}}\right)-c\left(\frac{1}{2(\lambda_{B,\underline{r}}+\lambda_{B,\bar{r}})}+w_q+\underline{r}\right)\right]. \tag{D.10}$$

When the demand is exogenous, our major insights still hold. We summarize the results in the next corollary.

COROLLARY 2. *Suppose the demand is exogenous with $\lambda_{D,\bar{r}}=\lambda_{D,\underline{r}}=\lambda_{B,\bar{r}}=\lambda_{B,\underline{r}}$. Then we have:*

(i) Dedicated delivery leads to higher revenue when the exogenous demand rate is very low and batch delivery leads to higher revenue when the demand rate is very high.

(ii) Furthermore ,dedicated delivery leads to higher revenue when the serve region is very small, i.e., $\underline{r}$ is small enough, and batch delivery leads to higher revenue when the service region is very large, i.e., $\underline{r}$ is large enough.

Proof of Corollary 2. Denote the exogenous demand rate by λ. Note that when the demand is exogenous and the same among the two rings, shorting service time leads to higher revenue, just like in our base model.

We first show that $\mathbb{E}[\tilde{X}_B]\le 2\mathbb{E}[\tilde{X}_D]$ so that the system with batch service has a larger load factor compared to the system with dedicated service. In other words, the batch system can handle a larger exogenous demand rate. Using the expressions in (D.2) and (D.6), when $\lambda_{D,\bar{r}}=\lambda_{D,\underline{r}}=\lambda_{B,\bar{r}}=\lambda_{B,\underline{r}}$, we have

$$\begin{aligned}
&2\mathbb{E}[\tilde{X}_D]-\mathbb{E}[\tilde{X}_B]\\
&=2\bar{r}+2\underline{r}-\left(\frac{1}{2}\left(\bar{r}+\underline{r}+\frac{1}{\pi}\int_0^{\pi}\sqrt{\bar{r}^2+\underline{r}^2-2\bar{r}\underline{r}\cos(\theta)}\,d\theta\right)+\frac{1}{4}\left(2\bar{r}+\frac{\pi\bar{r}}{2}\right)+\frac{1}{4}\left(2\underline{r}+\frac{\pi\underline{r}}{2}\right)\right)\\
&=\frac{1}{2}\left(\bar{r}+\underline{r}-\frac{1}{\pi}\int_0^{\pi}\sqrt{\bar{r}^2+\underline{r}^2-2\bar{r}\underline{r}\cos(\theta)}\,d\theta\right)+\frac{1}{4}\left(2\bar{r}-\frac{\pi\bar{r}}{2}\right)+\frac{1}{4}\left(2\underline{r}-\frac{\pi\underline{r}}{2}\right)\ge 0,
\end{aligned}$$

where the inequality follows the triangle inequality and the fact that $\pi<4$. The rest of the proof then follows the logic of Propositions 1 and 2.

When λ approaches 0, the wait times on both rings under batch go to infinity, according to (D.8) and (D.9), since the term $1/(2\lambda_{B,\bar{r}}+2\lambda_{B,\underline{r}})=1/(4\lambda)$ goes to infinity in both expressions. On the other hand, the wait time under dedicated delivery is constant even if λ approaches 0, since W_D goes to zero. When λ approaches $\tilde{\mu}_D=1/\mathbb{E}[\tilde{X}_D]=\bar{r}+\underline{r}$, which is the upper bound on the demand rate that can be handled by the dedicated system, the wait time of a dedicated system goes to infinity. However, the wait time of a batch system is still finite since we have shown that it has a larger load factor. This shows the first statement.

We follow the same logic to show the second statement. Suppose $\bar{r} = a\underline{r}$ for some $a > 1$. When $\underline{r}$ approaches 0, we have that $\bar{r} = a\underline{r}$ also goes to 0 for any $a > 0$. In this case, the wait time of dedicated delivery approaches zero since $W_D(2\lambda, \tilde{\mu}_D, \tilde{C}_D) = 2(\bar{r}^2 + \underline{r}^2)\lambda/(1 - 2\lambda(\bar{r} + \underline{r}))$ goes to zero. However, the wait times on both rings under batch delivery are positive according to Propositions 1 and 2, due to the order accumulation time. When $\underline{r}$ approaches $1/(\underline{r} + a\underline{r})$, the load under dedicated service approaches 1, leading to infinite wait time. However, batch service still has finite wait time due to a larger load factor. This completes the proof. □

As we can see, it is very challenging to conduct further analysis using the expressions derived above, especially for batch delivery, due to the cumbersome expressions. Given the scope of this paper, we decide to conduct numerical calculations when the vendor endogenizes the demand rate and customers have uniformly distributed valuations. Furthermore, we assume that the vendor can only offer one price for customers on the inner and outer rings when endogenizing demand. Our base model's insights still carry over to this setting, as shown in Figure 7. We leave the analytical exploration of distance-dependent wait time as a future research direction.

E. Distance-Dependent Delivery Policy

In this section, we consider a distance-dependent delivery policy. For tractability and coherence, we consider *two* couriers on a two-rings-structured service area, introduced in Appendix D. A distance-dependent delivery policy asks one courier to only focus on the orders on the inner ring with radius $\underline{r}$ and serve in Batch. On the other hand, it asks the other courier to only serve orders on the outer ring with radius $\bar{r}$ with dedicated service. We shall compare this distance-dependent delivery policy with two benchmarks, serving only dedicated or batch.

For the first benchmark, the two couriers only serve the two rings with dedicated service. We combine the derivations from Section 6.7 and Appendix D together. From Appendix D, when given the wait time and price, the service level on the two rings follows (D.1). Moreover, the first and second moments of the travel distances follows (D.2); the service rate $\tilde{\mu}_D$ follows (D.3); and the coefficient of variations $\tilde{C}_D$ follows (D.4). From Section 6.7, we have the expected wait time for a customer as

$$\tilde{W}_{D,2}(\lambda, d) \approx \frac{\tilde{C}_D}{2(2\tilde{\mu}_D - \lambda)} \left(\frac{\lambda}{2\tilde{\mu}_D}\right)^{\sqrt{6}-1} + d, \tag{E.1}$$

where the first term is in-line delay and the second term is en-route delay (travel time). Thus, the vendor's revenue is

$$\begin{aligned} \tilde{V}_{D,2} &= (\lambda_{D,\underline{r}} + \lambda_{D,\bar{r}})\, p \\ &= \lambda_{D,\underline{r}} \left[F^{-1}\left(1 - \frac{\lambda_{D,\underline{r}}}{\Lambda \underline{r}}\right) - c\tilde{W}_{D,2}(\lambda_{D,\underline{r}} + \lambda_{D,\bar{r}}, \underline{r}) \right] \end{aligned}$$

$$+\lambda_{D,\bar{r}}\left[F^{-1}\left(1-\frac{\lambda_{D,\bar{r}}}{\Lambda\bar{r}}\right)-c\tilde{W}_{D,2}(\lambda_{D,\underline{r}}+\lambda_{D,\bar{r}},\bar{r})\right], \tag{E.2}$$

such that

$$F^{-1}\left(1-\frac{\lambda_{D,\underline{r}}}{\Lambda\underline{r}}\right)-c\tilde{W}_{D,2}(\lambda_{D,\underline{r}}+\lambda_{D,\bar{r}},\underline{r})=F^{-1}\left(1-\frac{\lambda_{D,\bar{r}}}{\Lambda\bar{r}}\right)-c\tilde{W}_{D,2}(\lambda_{D,\underline{r}}+\lambda_{D,\bar{r}},\bar{r}). \tag{E.3}$$

For the second benchmark, we consider that the two couriers only serve the two rings in batches. From Appendix D, we have that the service level still follows (D.1); the first and second moments of the travel distances follow (D.6); the service rate follows $\tilde{\mu}_D = 1/\mathbb{E}[\tilde{X}_B]$; and the coefficient of variations $\tilde{C}_B$ follows (D.7). Following the base model, assume that the inner-ring orders are fulfilled first. Thus, using the expression of (26), a customer has the expected wait time of

$$\tilde{W}_{B,2}(\lambda,d)=\frac{1}{2\lambda}+\frac{\tilde{C}_B}{2(4\tilde{\mu}_B-\lambda)}\left(\frac{\lambda}{4\tilde{\mu}_B}\right)^{\sqrt{6}-1}+d, \tag{E.4}$$

where the first term is the time it takes to cumulate a batch; the second term is the in-line delay; and the third term is the en-route delay.

Thus, the vendor's revenue is

$$\begin{aligned}\tilde{V}_{B,2} &= (\lambda_{B,\underline{r}}+\lambda_{B,\bar{r}})\,p\\ &= \lambda_{B,\underline{r}}\left[F^{-1}\left(1-\frac{\lambda_{B,\underline{r}}}{\Lambda\underline{r}}\right)-c\tilde{W}_{B,2}(\lambda_{B,\underline{r}}+\lambda_{B,\bar{r}},\underline{r})\right]\\ &\quad+\lambda_{B,\bar{r}}\left[F^{-1}\left(1-\frac{\lambda_{B,\bar{r}}}{\Lambda\bar{r}}\right)-c\tilde{W}_{B,2}\left(\lambda_{B,\underline{r}}+\lambda_{B,\bar{r}},\underline{r}+\frac{1}{\pi}\int_0^{\pi}\sqrt{\bar{r}^2+\underline{r}^2-2\bar{r}\underline{r}\cos(\theta)}d\theta\right)\right],\end{aligned} \tag{E.5}$$

such that

$$\begin{aligned}&F^{-1}\left(1-\frac{\lambda_{B,\underline{r}}}{\Lambda\underline{r}}\right)-c\tilde{W}_{B,2}(\lambda_{B,\underline{r}}+\lambda_{B,\bar{r}},\underline{r})\\ =\;&F^{-1}\left(1-\frac{\lambda_{B,\bar{r}}}{\Lambda\bar{r}}\right)-c\tilde{W}_{B,2}\left(\lambda_{B,\underline{r}}+\lambda_{B,\bar{r}},\underline{r}+\frac{1}{\pi}\int_0^{\pi}\sqrt{\bar{r}^2+\underline{r}^2-2\bar{r}\underline{r}\cos(\theta)}d\theta\right).\end{aligned} \tag{E.6}$$

Finally, we characterize the distance-dependent delivery policy (which we refer to as a hybrid policy denoted by subscript $_H$). Under this policy, since each courier only focuses on one ring, we can decompose this service system into two sub-systems. More specifically, we can use the derivations from Sections 3 and 6.5 to treat the system as a two single-courier service system. To make a proper comparison, we still define the service level using (D.1), where $\lambda_{H,\bar{r}}$ and $\lambda_{H,\underline{r}}$ are the demand rates on the outer and inner rings, respectively. Then the vendor's revenue is

$$\begin{aligned}\tilde{V}_{H,2} &= (\lambda_{H,\underline{r}}+\lambda_{H,\bar{r}})\,p\\ &= \lambda_{D,\underline{r}}\left[F^{-1}\left(1-\frac{\lambda_{H,\underline{r}}}{\Lambda\underline{r}}\right)-c\left(\frac{1}{2\lambda_{H,\underline{r}}}+\frac{\lambda_{H,\underline{r}}C_{B,C}}{2\underline{\mu}_{B,C}(2\underline{\mu}_{B,C}-\lambda_{H,\underline{r}})}+\underline{r}+\frac{1}{2\mu_{B,C}}\right)\right]\end{aligned}$$

$$+\lambda_{D,\bar{r}}\left[F^{-1}\left(1-\frac{\lambda_{H,\bar{r}}}{\Lambda\bar{r}}\right)-c\left(\frac{\lambda_{H,\bar{r}}C_{D,C}}{2\bar{\mu}_{D,C}(\bar{\mu}_{D,C}-\lambda_{H,\bar{r}})}+\bar{r}\right)\right], \tag{E.7}$$

such that

$$\begin{aligned} &F^{-1}\left(1-\frac{\lambda_{H,\underline{r}}}{\Lambda\underline{r}}\right)-c\left(\frac{1}{2\lambda_{H,\underline{r}}}+\frac{\lambda_{H,\underline{r}}C_{B,C}}{2\underline{\mu}_{B,C}(2\underline{\mu}_{B,C}-\lambda_{H,\underline{r}})}+\underline{r}+\frac{1}{2\mu_{B,C}}\right)\\ =\ &F^{-1}\left(1-\frac{\lambda_{H,\bar{r}}}{\Lambda\bar{r}}\right)-c\left(\frac{\lambda_{H,\bar{r}}C_{D,C}}{2\bar{\mu}_{D,C}(\bar{\mu}_{D,C}-\lambda_{H,\bar{r}})}+\bar{r}\right), \end{aligned} \tag{E.8}$$

where

$$\bar{\mu}_{D,C}=\frac{1}{2\bar{r}},\ \text{and}\ \underline{\mu}_{B,C}=\frac{1}{2\underline{r}+\frac{\pi\underline{r}}{2}}, \tag{E.9}$$

where the service rates and constants $C_{B,C}, C_{D,C}$ are all defined in Section 6.5.

Indeed, the hybrid policy mentioned above can take advantage of the en-route efficiency of batch and the lack of order aggregation time of dedicated delivery. However, it does not always dominate either the dedicated or batch delivery. In extreme cases in which the two radii are extremely close to each other, the two rings shrink to one, and then one of the rings shall lose its edge. For example, with two rings very close to each other and the radii are large, the inner-ring service is inefficient due to serving batch with a large radius and thus a low sustainable demand rate. Similarly, two very small rings shall induce an inefficient outer ring, since the outer ring does not take advantage of a higher sustainable demand rate but serves dedicated delivery.

Even in non-extreme cases, based on the current hybrid policy, we can still find numerical examples such that it is dominated by dedicated and batch delivery in some parameter regimes. Another drawback of the current hybrid policy is that it enforces the same price on the two rings. As a result, services on the two rings need to make compromises with each other can create sustainable demand levels on each ring. Thus, the resulting demand rates may not be favorable to either ring, but just rates that keep the service systems on the two rings feasible. Figure E.1 provides an example of such rare scenarios that the hybrid system can be dominated by both dedicated and batch delivery.

Our analysis in this section can shed some light and inspire more research on distance-based delivery policies, since it is a very important and relevant issue that is worth careful consideration in practice. In our setup, we focus on the case with two couriers, each serving a single ring in the hybrid system. We point out that it is very difficult to characterize the resulting queueing system if there is only a single courier who decides to serve either batch or dedicated delivery based on the order location. Our work may inspire more papers to investigate distance-based service policies in on-demand delivery.

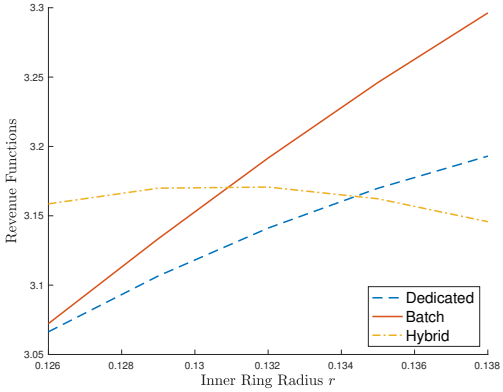


Figure E.1 Revenue functions under hybrid, dedicated, and batch services with two couriers. $c = 0.1$, $\Lambda = 50$, $\bar{r}/\underline{r} = 1.1$

F. Simulation Results

In this section, we present the simulation results on the accuracy of several approximations on expected wait times in this paper.

First, we briefly state our simulation methods. In each sample, consider a unit disk (with radius $r = 1$), and generate 1,000 arrivals uniformly distributed inside the disk using a fixed arrival rate λ. We calculate the average wait time (sample mean) for each sample. When calculating the sample mean, we exclude the first 200 arrivals to ensure the system has reached a steady state. In total, we simulate 100 samples for each arrival rate and take the average of sample means as the simulated expected wait time.

Next, we present the results by comparing the simulated expected wait times versus the approximated wait times. Table F.1 demonstrates that our approximation for the expected wait time in an $E_2/G/1$ queue using Kingman's formula in (12) is reasonably good. The percentage difference is calculated as

$$\text{Difference}\,(\%) := \frac{|\text{Simulated Time} - \text{Approx. Time}|}{\text{Approx. Time}}.$$

Lastly, we choose parameters $\lambda \leq 2\mu_B$ so that the approximated system is finite and the comparisons are meaningful.

In Table F.3, we present the expected wait time with the approximation in (25) and (26), which are compared to their simulation counterparts, respectively. We choose courier number $k = 3$.

As we can see from Tables F.2 and F.3, our approximation is relatively accurate except for light traffic dedicated systems or nearly overloaded batch systems. In the next two figures, we show that even with approximation errors, our insights on the thresholds in wait cost c and service radius r still hold with simulation results.

Table F.1 Performance of approximated expected wait time with Kingman's formula on $E_2/G/1$ queue

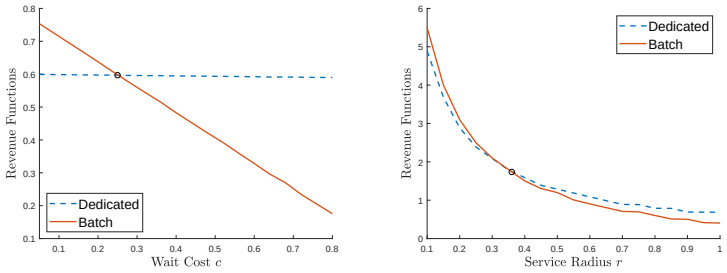
$\lambda(\leq 2\mu_B)$	utilization ρ	Simulated Time	Approx. Time	Difference (%)
0.2	0.22	2.92	3.02	3.47%
0.3	0.34	2.20	2.33	5.49%
0.4	0.45	1.95	2.11	7.71%
0.5	0.56	1.97	2.16	8.80%
0.6	0.67	2.30	2.50	8.23%
0.7	0.78	3.21	3.41	5.90%
0.8	0.90	6.13	6.55	6.45%

Table F.2 Performance of approximated expected wait time with dedicated delivery and $k=3$ couriers

$\lambda(\leq k\mu_D)$	utilization ρ	Simulated Time	Approx. Time	Difference (%)
0.3	0.13	0.003	0.007	57.48%
0.6	0.27	0.02	0.03	32.91%
0.9	0.40	0.07	0.08	10.18%
1.2	0.53	0.33	0.36	5.38%
1.5	0.67	0.43	0.42	8.35%
1.8	0.80	0.84	0.83	0.48%
2.1	0.93	3.10	3.31	6.12%

Table F.3 Performance of approximated expected wait time with batch delivery and $k=3$ couriers

$\lambda(\leq 2k\mu_B)$	utilization ρ	Simulated Time	Approx. Time	Difference (%)
0.3	0.11	2.00	2.00	0.47%
0.6	0.22	1.17	1.18	1.06%
0.9	0.34	0.91	0.91	0.39%
1.2	0.45	0.79	0.80	0.70%
1.5	0.56	0.77	0.75	1.76%
1.8	0.67	0.84	0.77	9.51%
2.1	0.78	1.14	0.89	27.3%



(a) (b)

Figure F.1 Revenue functions when serving dedicated versus batch under a large market. (a) $r=1$ $k=3$, (b) $c=0.5$, $k=3$

G. Other Delivery Delay

In our base model, we consider that the courier does not stay for any positive amount of time at each order location, nor has any delay caused by other factors. In this section, we discuss how to extend our base model to incorporate these features into the model.

Take delays that occurred at the delivery locations as an example. Suppose each order incurs an independently and identically distributed extra delay denoted as d at each order location, where d has mean μ_d and variance σ_d^2, following a known distribution. Further assume that this extra delay is also independent of the courier's travel time (the time to cover the distance between two locations). In reality, there is at most a negligible correlation between the actual travel time and the time finding a parking space at the delivery location.

Using this setup, we can rewrite the service rate and coefficient of variation as

$$\hat{\mu}_D = \frac{1}{\mathbb{E}[X_D] + \mu_d} = \frac{1}{\frac{4}{3}r + \mu_d}, \hat{C}_D = 1 + \frac{\mathbb{E}[X_D^2] - (\mathbb{E}[X_D])^2 + \sigma_d^2}{(\mathbb{E}[X_D] + \mu_d)^2} = 1 + \frac{\frac{2}{9}r^2 + \sigma_d^2}{\left(\frac{4}{3}r + \mu_d\right)^2}, \tag{G.1}$$

when the courier serves dedicated delivery, and

$$\hat{\mu}_B = \frac{1}{\mathbb{E}[X_B] + \mu_d} = \frac{1}{\frac{4(32+15\pi)}{45}r + \mu_d}, \hat{C}_B = \frac{1}{2} + \frac{\mathbb{E}[X_B^2] - (\mathbb{E}[X_B])^2 + \sigma_d^2}{(\mathbb{E}[X_B] + \mu_d)^2} = \frac{1}{2} + \frac{5.428r^2 - \left(\frac{4(32+15\pi)}{45}r\right)^2 + \sigma_d^2}{\left(\frac{4(32+15\pi)}{45}r + \mu_d\right)^2}, \tag{G.2}$$

when serving batch. Then the expected wait time of the two delivery modes can be written as $W_D(\lambda, \hat{\mu}_D, \hat{C}_D)$ and $W_B(\lambda, \hat{\mu}_B, \hat{C}_B)$, respectively.

We can recreate the analysis in our base model using this alternative setup. Take Section 4 with exogenous demand for example. We can still show that there are thresholds on demand rate λ and service radius r, respectively, such that below which serving dedicated is optimal and above which the courier should serve batch. We only provide a sketch of proof here, using a similar method as the proofs of Propositions 1 and 2.

Consider function

$$\begin{aligned} f(\lambda, r) &:= W_D(\lambda, \hat{\mu}_D, \hat{C}_D) - W_B(\lambda, \hat{\mu}_B, \hat{C}_B) \\ &= \frac{1}{180\pi\lambda(-3 + 4r\lambda + 3\lambda\mu_d)(64r\lambda + 15\pi(-3 + 2r\lambda + 3\lambda\mu_d))} \\ &\quad \Big[8192r^2\lambda^2(-3 + 4r\lambda + 3\lambda\mu_d) - 1920\pi r\lambda(-9 + \lambda(2r^2\lambda - 9\mu_d(-3 + \lambda\mu_d) - 12r(-2 + \lambda\mu_d) + 9\lambda\sigma_d^2)) \\ &\quad + 225\pi^2(4(-10 + 9a)r^3\lambda^3 - 54(-1 + \lambda\mu_d)^2 + 3r^2\lambda^2(12 - 9a - 28\lambda\mu_d + 9a\lambda\mu_d) \\ &\quad - 36r\lambda(-3 + \lambda(\mu_d + \lambda\mu_d^2 - \lambda\sigma_d^2))) \Big] - \frac{r}{3}, \end{aligned} \tag{G.3}$$

where $a = \mathbb{E}[X_B^2] - (\mathbb{E}[X_B])^2 \approx 0.416$. One can easily verify that $f(\lambda, r) = 0$ is a cubic equation with respect to either λ or r. Furthermore, it has a unique real solution for λ or r, just as in the proof of Propositions 1 and 2. Lastly, we only need to verify that $W_B(\lambda, \hat{\mu}_B, \hat{C}_B) < W_D(\lambda, \hat{\mu}_D, \hat{C}_D)$ when λ (or r) approaches zero, but $W_B(\lambda, \hat{\mu}_B, \hat{C}_B) > W_D(\lambda, \hat{\mu}_D, \hat{C}_D)$ when λ (or r) is sufficiently large.

As we can see, with the addition of extra delays independent of the delivery trip travel time, the analysis becomes a lot messier due to complicated algebra operations. Thus, we decide not to include these features into our base model to preserve a clean yet informative analysis.

H. Non-linear Wait Cost

In our main model, we assume that the customers' value function has a linear penalty with respect to the wait time. In this section, we relax that assumption and consider another wait cost function with practical motivations. In particular, a customer's utility can be

$$v - p - c[(w-\theta)^+]^2, \tag{H.1}$$

where v is still the realization of a customer's valuation; p is the price; w is the expected wait time; and θ is a threshold on the wait time. Based on this new utility function, a customer does not incur wait cost when the expected wait time is below θ, but incurs quadratic wait cost for every unit of time beyond θ.

First, if the demand is exogenous, just like in the base model, we only need to compare the expected wait time under the two delivery modes, and the shorter one leads to higher revenue. To see this, note that under the new utility function of customers, we need to modify the vendor's revenue functions under dedicated and batch delivery as the following

$$\hat{V}_D(\lambda, W_D(\lambda,\mu_D,C_D)) = \lambda\left[F^{-1}\left(1-\frac{\lambda}{\Lambda r^2}\right) - c\left[(W_D(\lambda,\mu_D,C_D)-\theta)^+\right]^2\right], \tag{H.2}$$

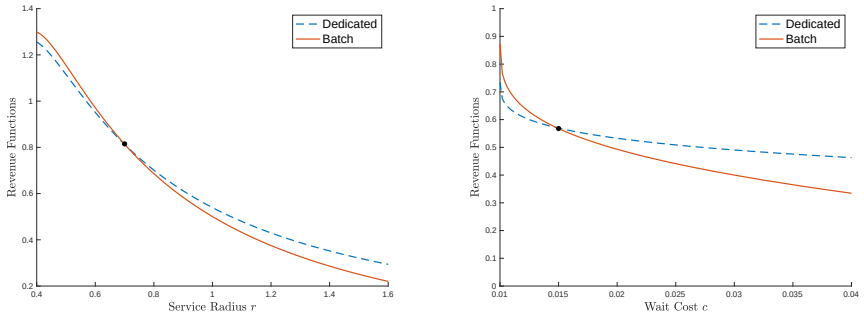
and

$$\hat{V}_B(\lambda, W_B(\lambda,\mu_B,C_B)) = \lambda\left[F^{-1}\left(1-\frac{\lambda}{\Lambda r^2}\right) - c\left[(W_B(\lambda,\mu_B,C_B)-\theta)^+\right]^2\right], \text{ respectively,} \tag{H.3}$$

where the two wait times W_D and W_B are derived in (5) and (12), respectively. Thus, a shorter wait time still leads to higher revenue under this setting. More specifically, when λ is exogenous, compared to the base model, the only difference is that we may encounter a trivial case in which an exogenous demand leads to both W_D and W_B no greater than θ, making the two delivery mode indifferent. Otherwise, all of the insights in Section 4 still hold. As long as customers' utility function is non-increasing in the wait time, our insights in Section 4 still hold.

When the demand is endogenously determined as part of the revenue maximization, we verify that our major insights in Section 5 still hold numerically. Establishing the analytical propositions in Section 5 is very challenging due to drastically increasing difficulties introduced by the "kink" resulted from the threshold θ and the higher order terms created by the quadratic wait penalties.

In Figure H.1, we plot revenue functions with respect to the service radius and the wait cost, respectively. As we can see, we still observe a threshold below which, serving batch is optimal and above which, dedicated service is optimal. Thus, the major insights of our base model still hold even under the non-linear wait cost setup.



(a) (b)

Figure H.1 Revenue functions when serving dedicated versus batch.
(a) $c = 0.01$, $\Lambda = 50$ **(b)** $r = 1$, $\Lambda = 50$

References

Dal Maso, G. (1993). *An Introduction to* $\Gamma-$*Convergence*. Birkhäuser, Basel.