# Reinforcement Learning of Optimal Delivery Destination Estimation

Presenters: AJ Bull, Chen Li

# Outline

1. **Background**
2. Methods
3. Results
4. Conclusion

# Package Theft

- Package theft is an emerging type of crime
  - Increasing volume of package delivered directly to a home[1,2]
  - Increasing concern among online retailers[3]
- Potential methods to reduce package theft:
  - Reducing award
  - Increasing risk
  - Increasing efforts

# Autonomous Delivery Robots

- Autonomous delivery robots for package delivery
  - Last mile delivery[4,5]
  - Fast and cheap[6]
  - Less impacted by traffic congestion
- Package theft prevention: delivery spot selection

# Package Theft Prevention as a Bandit Problem

- Unrealistic to solve analytically
  - Different house designs
  - Varied individual behaviors
  - Large number of packages
- Formulate package spot selection problem as an adversarial bandit problem
  - The choice of package placement has no inherent state
  - Each porch can be viewed as a different context
  - Actions are defined as choices of package placement

# Adversarial Bandit Problem Algorithms [8-10]

- Exp3: exponential weights
  - Exp3.P: Exp3 with high probability regret
  - Exp3.IX: Exp3 with high probability regret, using Implicit Exploration
  - Exp3.S: Exp3 for sequence of actions
- Exp4: exponential weights with experts
- We choose Exp4 Algorithm for the adversarial bandit problem
  - Homeowners can be modeled as experts
  - Thieves can also be modeled as experts in reverse

# Outline

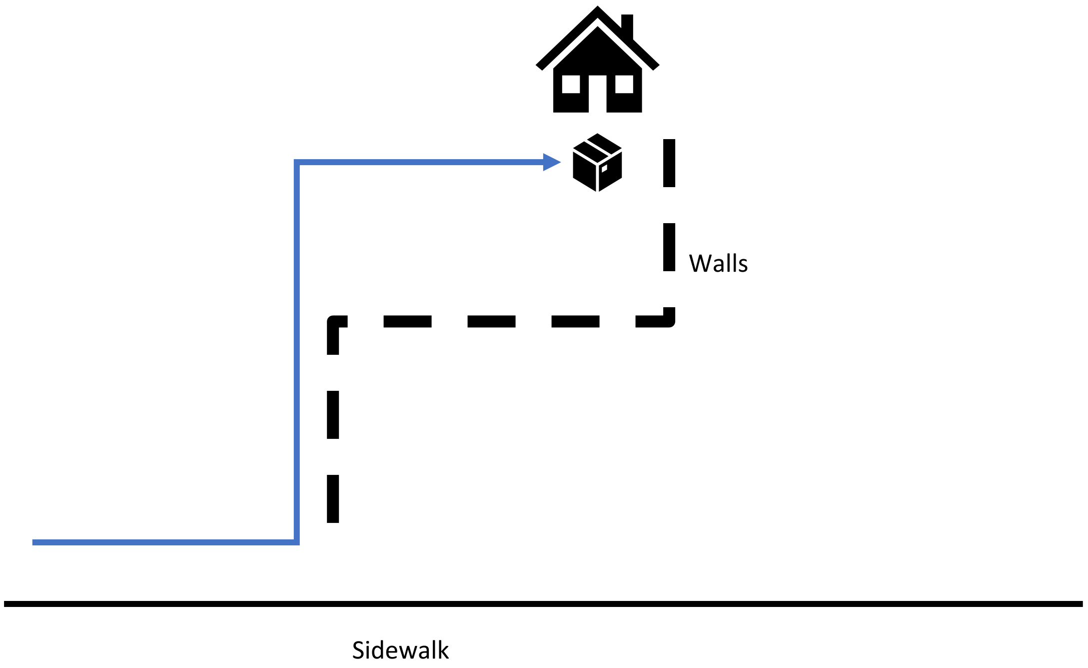
1. Background
2. **Methods**
3. Results
4. Conclusion

# Exp4 (Lattimore and Szepesvári)

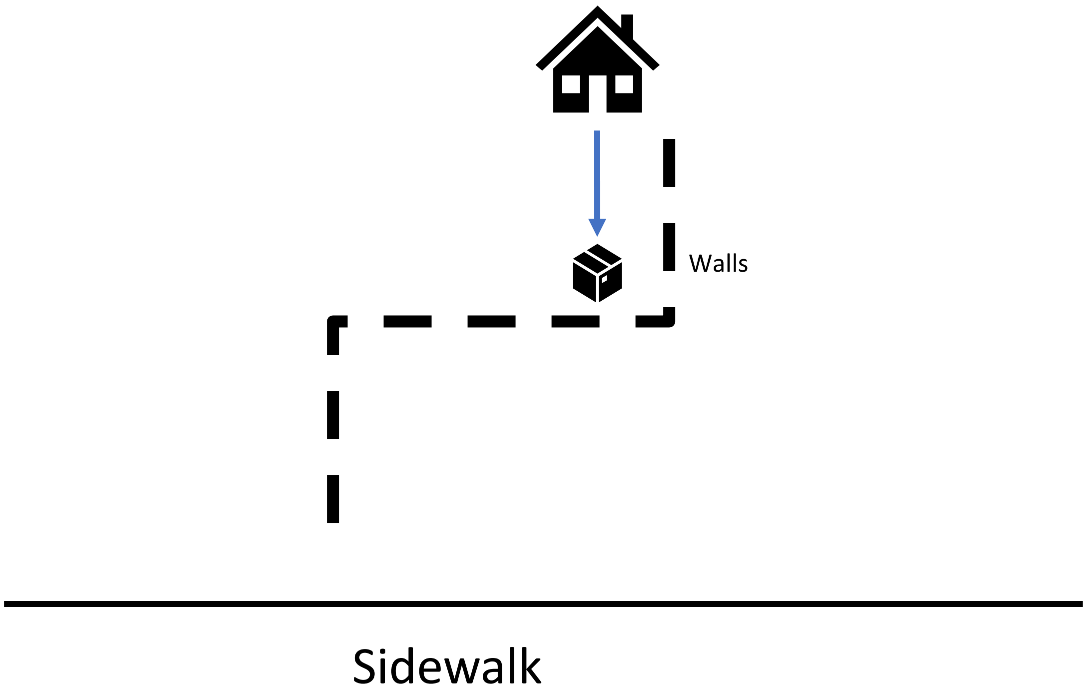
1. Take Real $\gamma \in (0,1]$
2. Initialize weights $W_0$
3. For each time t:
   A. Get expert advices $E(t)$
   B. Choose action $\mathrm{A}(t) \sim P(t) = W(t)E(t)$
   C. Receive reward $\mathrm{X}(t)$ for action $\mathrm{A}(t)$
   D. Estimate action rewards $\widehat{X}_i(t) = 1 - \frac{(1-X_i(t))}{P_i(t)}$, if $\mathrm{A}(t) = i$
   E. Update $\hat{Y}(t) = E(t)\hat{X}(t)$

$$W_i(t+1) = \frac{W_i(t)e^{\gamma \widehat{Y}_i(t)}}{\sum_j W_j(t)e^{\gamma \widehat{Y}_j(t)}}$$

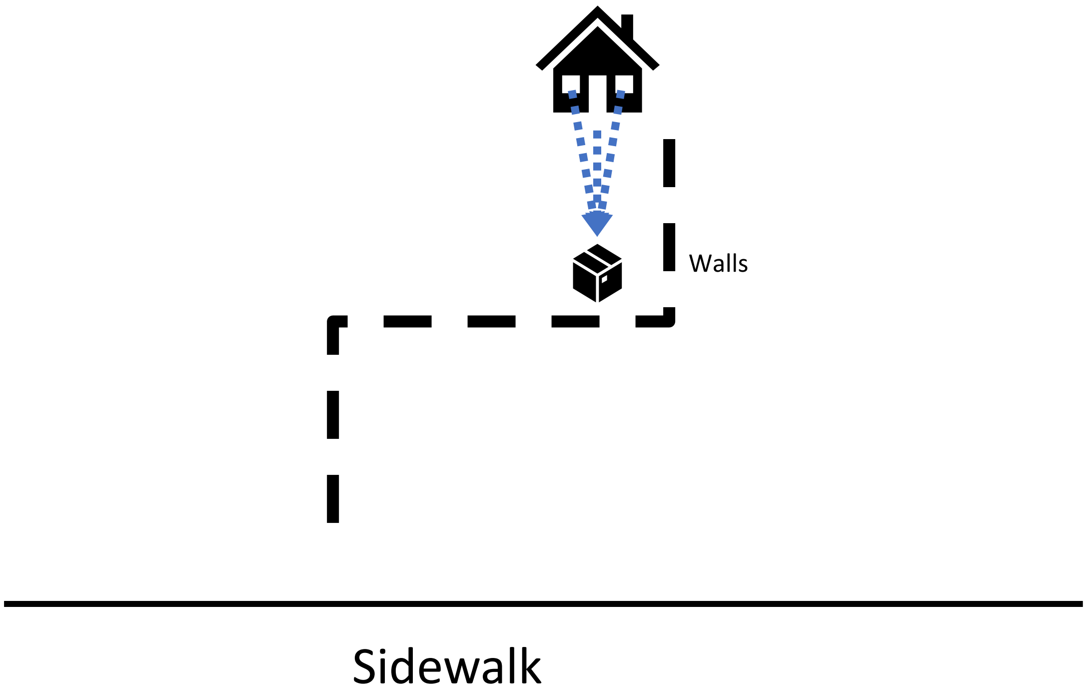
# Expert 1: Distance from sidewalk



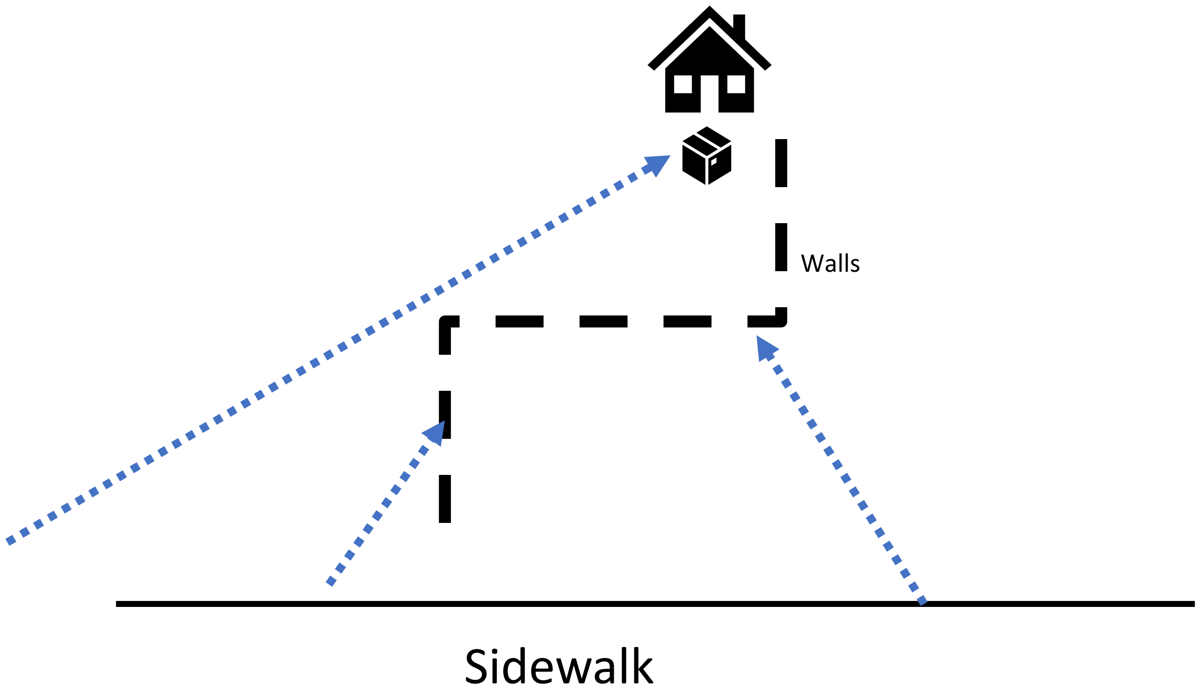
# Expert 2: Distance from doors



# Expert 3: Visibility from doors and windows



# Expert 4: Visibility from sidewalk



# Regret Analysis

Let $n$ be the number of iterations running Exp4, $M$ the set of experts, and $K$ the number of actions.

**Lemma.** For any expert $m^* \in M$,

$$\sum_{t=1}^{n} \hat{Y}_{m^*}(t) - \sum_{t=1}^{n} \sum_{m=1}^{|M|} W_m(t) \hat{Y}_m(t) \leq \frac{\log M}{\gamma} + \frac{\gamma}{2} \sum_{t=1}^{n} \sum_{m=1}^{|M|} W_m(t) (1 - \hat{Y}_m(t))^2$$

Since the experts are not oblivious, there exists a best-performing expert $m^*$ in hindsight.

# Regret Analysis

$$\mathbb{E}[\hat{Y}(t)] = \mathbb{E}[E(t)\hat{X}(t)] = E(t)\mathbb{E}[\hat{X}(t)] = E(t)X(t)$$

Define $L(Z) = 1_k - Z$. Then

$$\begin{aligned} E(t)L_k(\hat{X}(t)) &= E(t)1_k - E(t)\hat{X}(t) \\ &= 1_k - \hat{Y}(t) \\ &= L_k(\hat{Y}(t)) \end{aligned}$$

$$L(\hat{X}_i(t)) = \frac{L(X_i(t))}{P_i(t)} \text{ when } i = A(t), \text{ else } 0$$

# Regret Analysis

$$\mathbb{E}[L(\hat{Y}_m(t))^2] = \mathbb{E}\left[\left(E_{mi}(t)\frac{(1-X_i(t))}{P_i(t)}\right)^2\right]$$

$$\leq \sum_{i=1}^{K} P_i(t) E_{mi}(t)^2 \frac{L(X_i(t))^2}{P_i(t)^2}$$

$$\leq \sum \frac{E_{mi}(t)}{P_i(t)}$$

$$\mathbb{E}\left[\sum_{m=1}^{|M|} W_m(t)(1-\hat{Y}_m(t))^2\right] \leq \mathbb{E}\left[\sum_{m=1}^{|M|} W_m(t) \sum_{i=1}^{K} \frac{E_{mi}(t)}{P_i(t)}\right]$$

$$= \mathbb{E}\left[\sum_{i=1}^{K} \frac{\sum_{m=1}^{|M|} W_m(t) E_{mi}(t)}{P_i(t)}\right] = K$$

# Regret Analysis

$$\sum_{t=1}^{n}\hat{Y}_{m^*}(t)-\sum_{t=1}^{n}\sum_{m=1}^{|M|}W_m(t)\hat{Y}_m(t)\le\frac{\log M}{\gamma}+\frac{\gamma}{2}\sum_{t=1}^{n}\sum_{m=1}^{|M|}W_m(t)(1-\hat{Y}_m(t))^2$$

$$\Rightarrow R_n\le\frac{\log M}{\gamma}+\frac{\gamma}{2}\sum_{t=1}^{n}\mathbb{E}\left[\sum_{m=1}^{|M|}W_m(t)(1-\hat{Y}_m(t))^2\right]$$

$$\Rightarrow R_n\le\frac{\log M}{\gamma}+\frac{\gamma nK}{2}$$

$$\text{Let }\gamma=\sqrt{\frac{2\log M}{nK}},\text{ then }R_n\le\sqrt{2nK\log M}$$

# Comparison to Exp3

Exp3:

$$\gamma = \sqrt{\frac{2\log n}{nK}}$$

$$R_n \leq \sqrt{2nK\log K}$$

Exp4:

$$\gamma = \sqrt{\frac{2\log M}{nK}}$$

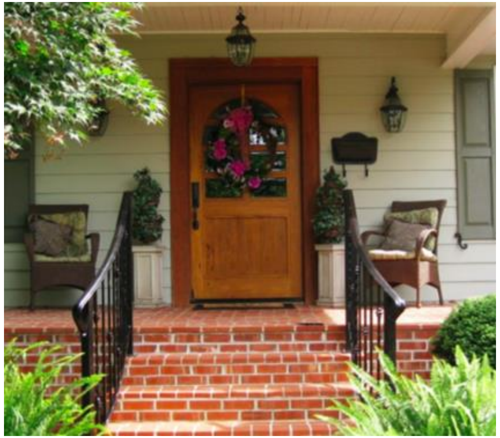
$$R_n \leq \sqrt{2nK\log M}$$

# Outline

1. Background
2. Methods
3. **Results**
4. Conclusion

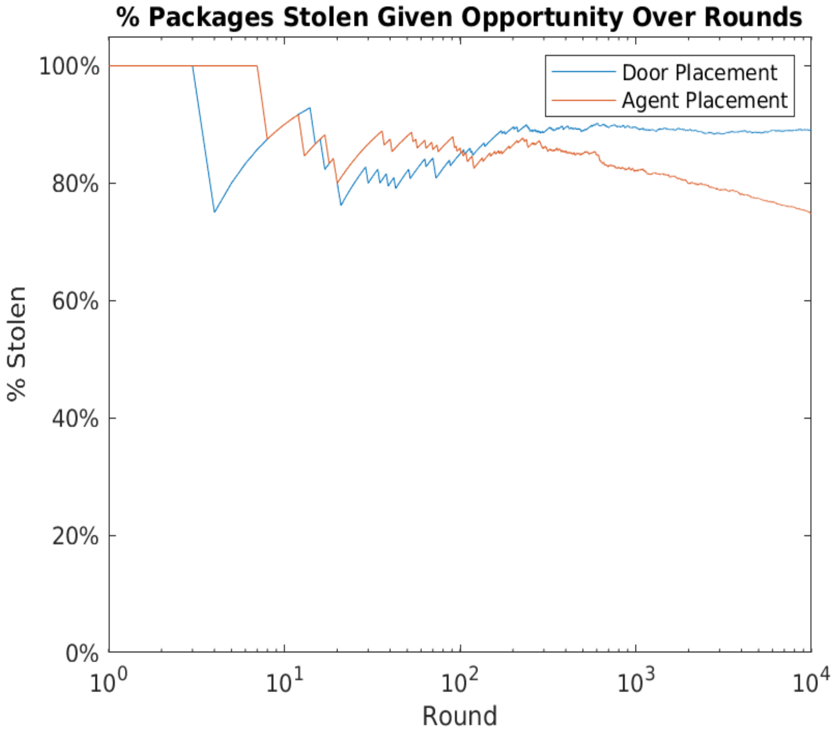
# Experimental Setup

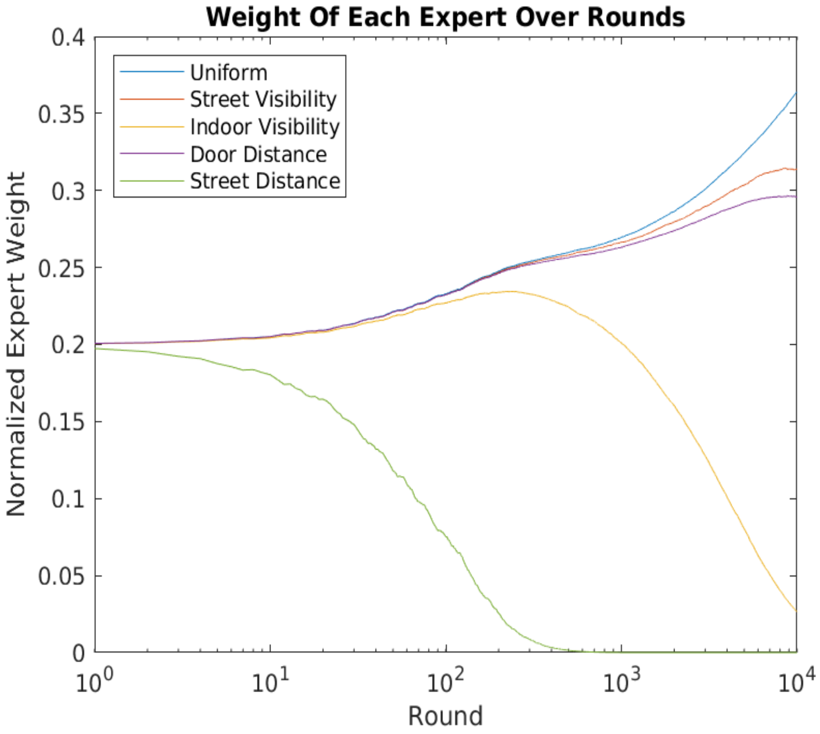
- Discretized Semantic 2D Maps
  - 40 x 40 grids, each cell is 20 in. x 20 in.
  - Each cell is a semantic class
  - Different maps represent significantly different architectures
- Simulated reward function
  - If package is visible from sidewalk and not too far out of reach (40 grid cells path length), it is stolen
  - Simulates probability of theft **given opportunity**
- Random Context
  - Each iteration, agent is given one of nine random maps
  - Agent placement vs. door placement
  - Mitigated by adding Uniform Expert



```
B W B W B B D D B B W B W B
1 0 0 0 0 0 0 0 0 0 0 0 0 1
1 0 0 0 0 0 0 0 0 0 0 0 0 1
1 0 0 0 0 0 0 0 0 0 0 0 0 1
1 0 0 0 0 0 0 0 0 0 0 0 0 1
C C C C H S S S S H C C C C
C C C C H S S S S H C C C C
C C C C H S S S S H C C C C
C C C C 0 0 0 0 0 0 C C C C
C C C C 0 0 0 0 0 0 C C C C
```

# Preliminary Results



# Outline

1. Background
2. Methods
3. Results
4. **Conclusion**

# Conclusion

- Current performance is marginally better than baseline
  - Difficulty simulating agent in action
  - Result has not converged after 10,000 iterations
    - May be due to choice of experts, or choice of maps
    - “Failed” experts do not recover
- Future
  - Experimental Regret
  - More experts (e.g. visibility kernel)
  - More maps

# Reference


[1] B. Stickle, M. Hicks, A. Stickle, and Z. Hutchinson, “Porch pirates:  Examining unattended package theft through crime script analysis, “Criminal Justice Studies, vol. 33, no. 2, pp. 79–95,2020.

[2] “11 million u.s. homeowners experienced package theft within the last year, august home study reveals.” https://www.businesswire.com/news/home/20161025005648/en/11-Million-U.S.-Homeowners-Experienced-Package-Theft. Accessed: 2021-02-28.

[3] M. Hicks, B. Stickle, and J. Harms, “Assessing the fear of package theft,” American Journal of Criminal Justice, pp. 1–20.

[4] M. Poeting, S. Schaudt, and U. Clausen, “Simulation of an optimized last-mile parcel delivery network involving delivery robots,” in Interdisciplinary Conference on Production, Logistics and Traffic, pp. 1–19, Springer, 2019.

[5] S. M. Shavarani, M. G. Nejad, F. Rismanchian, and G. Izbirak, “Application of hierarchical facility location problem for optimization of a drone delivery system: a case study of Amazon Prime Air in the city of San Francisco,” The International Journal of Advanced Manufacturing Technology, vol. 95, no. 9, pp. 3141–3153, 2018.

[6] D. Jennings and M. Figliozzi, “Study of road autonomous delivery robots and their potential effects on freight efficiency and travel,” Transportation Research Record,  vol. 2674,  no. 9,pp. 1019–1029, 2020.

[7] “10 tips to send porch pirates packing.” https://www.aarp.org/money/scams-fraud/info-2020/package-theft-holiday-season.html. Accessed: 2021-03-01.

[8] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, “The non-stochastic multi-armed bandit problem,” Society for Industrial and Applied Mathematics Journal On Computing, vol. 32, no. 1,pp. 48–77, 2002.

[9] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge: Cambridge University Press, 2020. pp. 228-230 https://tor-lattimore.com/downloads/book/book.pdf. Accessed: 2021-04-13.

[10] Neu, G., 2015. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. arXiv preprint arXiv:1506.03271,.