POTs: Protective Optimization Technologies Supplementary Material

Bogdan Kulynych
EPFL
bogdan.kulynych@epfl.ch

Rebekah Overdorf
EPFL
rebekah.overdorf@epfl.ch

Carmela Troncoso
EPFL
carmela.troncoso@epfl.ch

Seda Gürses
TU Delft / KU Leuven
f.s.gurses@tudelft.nl

A SENSITIVITY TO MISSPECIFICATIONS

We theoretically estimate the impact of misspecifications on the severity of externalities. For that, we use influence functions from the toolkit of robust statistics [1].

We assume that the utility functions U, B, and $\hat{B}$ are strictly concave and twice-differentiable; and we strengthen the ideal property of the fair-by-design provider. Besides picking the Pareto-optimal solution that maximizes their model of social utility, we now assume the near-optimality of the social-utility objective for this solution: $\nabla\hat{B}(\theta^*) \approx 0$. That is, the system θ^* is close to the *ideal solution* for the social-utility objective $\hat{B}(\theta)$ [2].

Consider a *partially corrected* optimization objective:

$$\max_{\theta\in\Theta}\{U(\theta), \hat{B}(\theta) - \varepsilon\hat{B}(\theta) + \varepsilon B(\theta)\},$$

where we move a pointmass ε away from the provider's model of the social utility to its "god's view" value. Let us take its a Pareto-optimal solution θ^*_ε that has highest benefit. We define the *influence function* of ΔB as follows:

$$IF_\varepsilon \triangleq \frac{d}{d\varepsilon}[B(\theta^*_\varepsilon) - B(\theta^*)]\,.$$

This models how fast the magnitude of the externality grows as more weight is given to the corrected B in the optimization problem.

Let us restate a known property of Pareto-optimal solutions due to Kuhn and Tucker [3]:

Theorem A.1 ([3]). *Let θ^* be a Pareto-optimal solution to the optimization problem of the form:*

$$\max_{\theta\in\Theta}\{U(\theta), B(\theta)\}$$

Then there exists $\lambda \in [0, 1]$, such that the following holds:

$$\lambda\nabla U(\theta^*) + (1-\lambda)\nabla B(\theta^*) = 0$$

Let us denote by $\mathcal{H}f(x)$ the Hessian matrix of f at x, and for convenience set $\mathcal{H}_{\theta,\lambda} := \lambda\mathcal{H}U(\theta) + (1-\lambda)\mathcal{H}\hat{B}(\theta)$. Additionally, denote by $\Delta\theta := \theta^*_\varepsilon - \theta^*$ the difference in system parameters coming from the corrected and the original optimization problems. We can now present our estimate for the influence function.

Statement A.1. *Using linearization techniques, we can obtain the following approximation for the influence function for some $\lambda \in [0, 1]$:*

$$IF_\varepsilon \approx -\nabla B(\theta^*)^\intercal[\mathcal{H}_{\theta^*,\lambda}]^{-1}\nabla B(\theta^*)$$

Derivation. Using Theorem A.1, we can say there exists $\lambda \in [0, 1]$ such that:

$$\lambda\nabla U(\theta^*_\varepsilon) + (1-\lambda)\left(\nabla\hat{B}(\theta^*_\varepsilon) - \varepsilon\nabla\hat{B}(\theta^*_\varepsilon) + \varepsilon\nabla B(\theta^*_\varepsilon)\right) = 0 \qquad \text{(A.1)}$$

We now rewrite Eq. A.1 in terms of the original system parameters θ^* using a first-order Taylor approximation:

$$\begin{aligned}0 = {}& \lambda\nabla U(\theta^*) + (1-\lambda)\left(\nabla\hat{B}(\theta^*) - \varepsilon\nabla\hat{B}(\theta^*) + \varepsilon\nabla B(\theta^*)\right)\\ &+ \left[\lambda\mathcal{H}U(\theta^*) + (1-\lambda)\left(\mathcal{H}\hat{B}(\theta^*) - \varepsilon\mathcal{H}\hat{B}(\theta^*) + \varepsilon\mathcal{H}B(\theta^*)\right)\right]\cdot\Delta\theta\end{aligned}$$

We can rearrange and further approximate following Koh and Liang [4], keeping in mind that ε is small:

$$\begin{aligned}\Delta\theta &\approx -\varepsilon[\mathcal{H}_{\theta^*,\lambda}]^{-1}\nabla[B(\theta^*) - \hat{B}(\theta^*)]\\ &\approx -\varepsilon[\mathcal{H}_{\theta^*,\lambda}]^{-1}\nabla B(\theta^*)\end{aligned}$$

We now approximate the influence function using its first-order Taylor expansion and the obtained expression for $\Delta\theta$:

$$\begin{aligned}IF_\varepsilon = \frac{d}{d\varepsilon}[B(\theta^*_\varepsilon) - B(\theta^*)] &\approx \frac{d}{d\varepsilon}\left[B(\theta^*) + \nabla B(\theta^*)\cdot\Delta\theta\right]\\ &= \nabla B(\theta^*)\cdot\frac{d}{d\varepsilon}\Delta\theta\\ &= -\nabla B(\theta^*)^\intercal[\mathcal{H}_{\theta^*,\lambda}]^{-1}\nabla B(\theta^*)\end{aligned}$$

□

Statement A.2. *Given the assumptions on $U, B, \hat{B}$ and θ^*, our linear approximation for the influence function of ΔB is asymptotically lower bounded as follows:*

$$IF_\varepsilon = \Omega(\parallel \nabla B(\theta^*) \parallel^2)$$

Proof. As $\mathcal{H}_{\theta^*,\lambda}$ is negative-definite by the concavity assumption and the fact that convex combinations preserve concavity, so is its inverse. Hence, $-[\mathcal{H}_{\theta^*,\lambda}]^{-1}$ is positive-definite. By a lower bound of a symmetric positive-definite quadratic form we have:

$$\begin{aligned}\frac{d}{d\varepsilon}\Delta_B &\approx \nabla B(\theta^*)^\intercal[\mathcal{H}J(\theta^*)]^{-1}\nabla B(\theta^*)\\ &= \Omega(\parallel \nabla B(\theta^*) \parallel^2)\end{aligned}$$

□

For concave functions, $\parallel \nabla B(\theta^*) \parallel$ can serve as a measure of error of the solution θ^* [5], which confirms our intuition.

B DETAILS FOR THE TRAFFIC THWARTING CASE STUDY

B.1 MILP Formulation

With a simple reparameterization, it is possible to formulate the optimization problem as follows:

$$\begin{aligned} &\min_{\mu(\cdot)\in\{0,1\}} \sum_{(x,y)\in\mathbb{E}} c(x,y)\cdot\mu(x,y) \\ &\text{s.t.} \sum_{(x,y)\in\mathbf{e}} [t(x,y)+\mu(x,y)\cdot\Delta t(x,y)] \geq t^* \\ &\quad \text{for any path } \mathbf{e} \text{ from } a \text{ to } b, \end{aligned} \tag{B.1}$$

In this form, the optimization problem is the shortest-path interdiction problem [6, 7], and can be solved as an MILP [8, 9]:

$$\begin{aligned} &\min_{\mu(\cdot)\in\{0,1\}} \sum_{(x,y)\in\mathbb{E}} c(x,y)\cdot\mu(x,y) \\ &\text{s.t. } \pi(t)-\pi(s)\geq t^* \\ &\quad \pi(y)-\pi(x)\leq t(x,y)+\mu(x,y)\cdot\Delta t(x,y) \quad \forall (x,y)\in\mathbb{E} \\ &\quad \pi(v)\in\mathbb{R} \quad \forall v\in\mathbb{V}, \end{aligned} \tag{B.2}$$

where $\pi(v)$ are additional *vertex-potential variables* that represent the smallest time cost for getting from a to v in graph $\mathcal{G}'$.

Assume that each edge (x, y) is associated with a length defined by $s(x, y)$, and a speed limit $v(x, y)$. In the case of changing the speed limits through Δv, $\Delta t(x, y)$ can be obtained from s, v and Δv as follows:

$$\Delta t(x,y) = \frac{s(x,y)\cdot\Delta v(x,y)}{v(x,y)^2 - v(x,y)\cdot\Delta v(x,y)} \tag{B.3}$$

B.2 Evaluation Details for Fremont, California

The graph for Fremont, CA, USA, is much larger than for the other towns considered in our evaluation, with a total of 9,215 nodes and 19,313 edges. The normal time from a to b in the town is 8.5 minutes. Figure 1 shows the optimal set of roads to lower the speed limit on for $t^* = 15.25$, using a 75% decrease in time for the allowed road changes.

C DETAILS FOR THE LOAN APPROVAL CASE STUDY

We detail the heuristic algorithm we use to solve the optimization problem of the POT in Algorithm 1. To compute the scores, we retrain a classifier for each example $(x, y) \in X_{pool}$. In our case of the logistic regression as the bank's model, retraining is inexpensive. For more complex models, approximation techniques can be used [10].

Algorithm 1 Algorithm for selecting poisoning loan applications in order to reduce the false-negative rate on the target group.

(1) S = PriorityQueue()
(2) for $(x, y) \in X_{pool}$:
(3) continue if $f(x;\theta') \neq$ 'accept'
(4) simulate $\theta^*_{(x,y)} = \arg\min \sum_{X\cup\{(x,y)\}} L(x', y';\theta)$
(5) compute the score $J(\theta^*_{(x,y)})$
(6) add (x, y) to S along with the computed score
(7) $X_{pot} := \{\}$
(8) while $|X_{pot}| \leq n$:
(9) remove (x^*, y^*) with the lowest score from S
(10) add (x^*, y^*) to R.

REFERENCES

[1] Peter J Huber. *Robust statistics*. Springer, 2011.
[2] Kaisa Miettinen. *Nonlinear multiobjective optimization*. Springer Science & Business Media, 2012.
[3] Harold Kuhn and Albert W Tucker. Nonlinear programming. In *Traces and emergence of nonlinear programming*, pages 247–258. 2014.
[4] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning (ICML)*, 2017.
[5] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
[6] Bruce Golden. A problem in network interdiction. *Naval Research Logistics Quarterly*, 1978.
[7] Delbert Ray Fulkerson and Gary C Harding. Maximizing the minimum source-sink path subject to a budget constraint. *Mathematical Programming*, 1977.
[8] Eitan Israeli and R Kevin Wood. Shortest-path network interdiction. *Networks: An International Journal*, 2002.
[9] Xiangyu Wei, Cheng Zhu, Kaiming Xiao, Quanjun Yin, and Yabing Zha. Shortest path network interdiction with goal threshold. *IEEE Access*, 2018.
[10] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks. In *USENIX Security Symposium*, 2019.

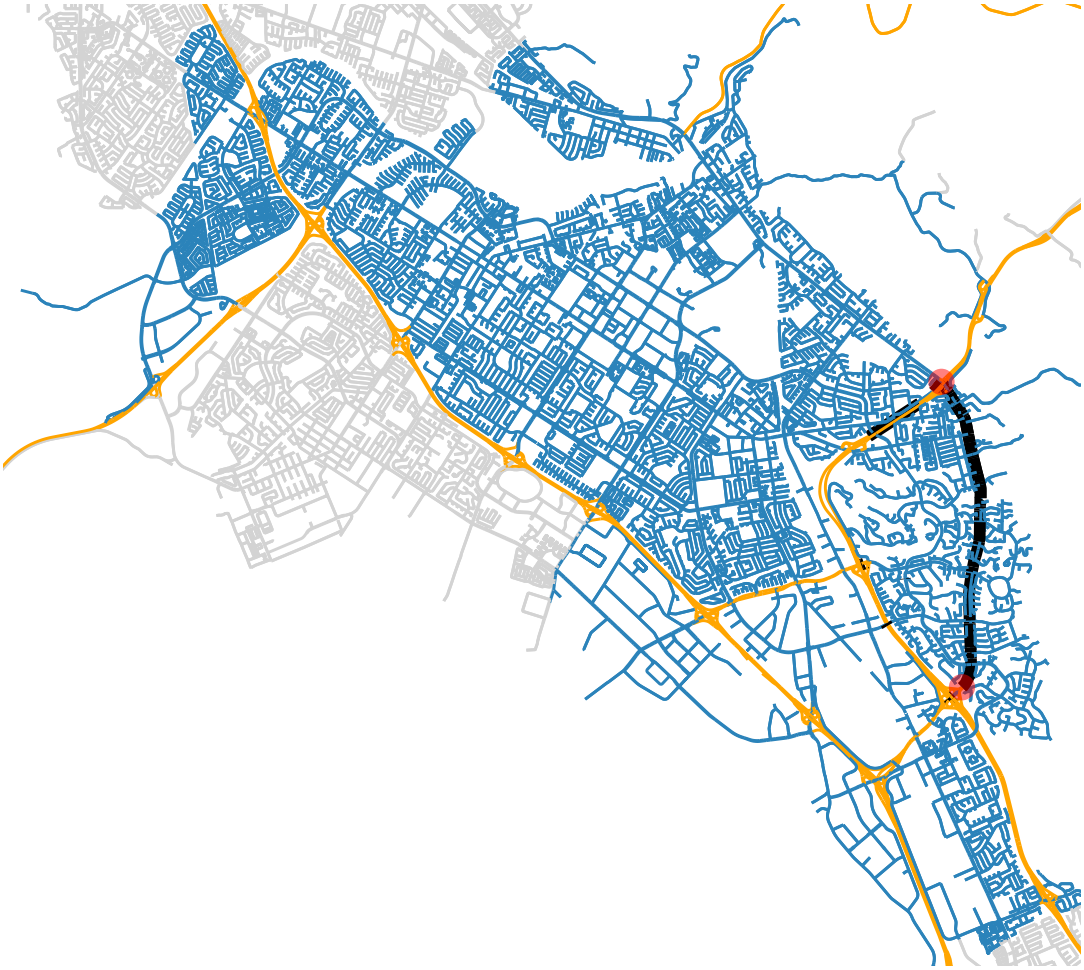

Figure 1: Solution for Fremont, California (in black)