Learning Risk Level Set Parameters from Data Sets for Safer Driving

Alyssa Pierson[1], Wilko Schwarting[1], Sertac Karaman[2], and Daniela Rus[1]

***Abstract*— This paper examines how vehicles can quickly quantify the level of congestion in their environment for planning. We use risk level sets to define a metric of congestion for the vehicles. Using this metric, we can quickly identify distributions of environment and driver features, such as velocities and number of neighbors, based on risk within human driving data sets. We use the NGSIM and highD data sets to study how risk influences behaviors in city and highway driving. From these data sets, we learn common risk thresholds for classifying low, medium, and high-risk situations. Using these thresholds, we develop simulations of an autonomous vehicle driving along a highway, and demonstrate how the chosen risk threshold influences the autonomous vehicle behavior.**

I. Introduction

Driver-assist systems can increase vehicle safety and mitigate the risks to a driver. Risk assessment needs to be done rapidly for these systems to handle critical traffic situations. Location, time of day, and other environmental factors can change the risk. Traffic data sets contain a wealth of information, but one challenge is understanding how to extract relevant information and relate this data to risk. This paper proposes the use of risk level sets to analyze human driving data, to extract relevant parameters for computing risk online, and using the results for the purpose of planning safer driving.

In our prior work, we propose risk level sets to quantify the level of clutter within the environment and safely navigate among dynamic obstacles [1]. We use a clutter metric to define a cost function that maps occupancy and movement to risk for a vehicle. We let vehicles to choose their allowable risk threshold, which manifests as tuning behavior to be more aggressive or conservative. Here, we present a modified formulation of the risk level sets that better encapsulates dynamic obstacles and allows for greater variations in velocity. We apply this cost to data sets in order to learn parameters and classify the data sets by their risk. In this paper, we use the highD [2] and the NGSIM [3] data sets, although our methods easily extend to any driving data. This approach allows us to examine how the driving conditions change based on low, medium, and high-risk situations. Figure 1 illustrates the contours of the risk thresholds for an ego vehicle (red) from the highD data set. We identify the risk thresholds from the data, and then demonstrate how we can tune the conservativeness of an autonomous vehicle navigating a highway by adjusting its allowable risk to enable safer driving. This method can be used to enhance driver assist systems or autonomous driving.

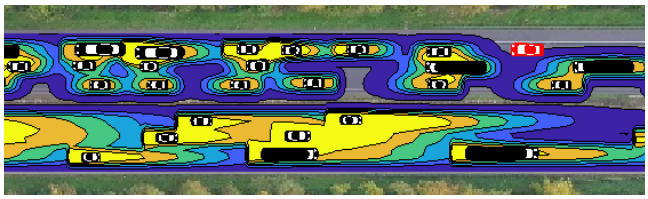


Fig. 1. Applying risk level sets allows us to quickly measure the congestion cost and corresponding risk to a vehicle in the environment. Here, we use the highD data set to study highway driving.

The main contributions of this paper are: (1) Risk level sets learned from human driving data; (2) An analysis of NGSIM and highD data distributions based on low, medium, and high-risk; and (3) Autonomous highway driving algorithms and simulations with varying conservativeness utilizing learned risk thresholds.

A. Related Work

Human response to congested traffic depends on their assessment and tolerance to risk. When the the driver's risk tolerance is known, it is possible to predict maneuvers and differentiate driving styles [4], [5], [6], [7]. The NGSIM data set is useful for understanding cooperative traffic flow [8], and behavior modeling [9], [10], [11]. Despite flaws in the raw data, NGSIM yields useful information with proper processing and filtering [12], [13], [14]. The highD data set provides higher-fidelity tracking data [2], and can be used for modeling and generating driver trajectories [15]. Here, instead of extracting behaviors or features from the entire data set, we score points in the data set with our risk metric, then divide the data by observed low, medium, and high-risk thresholds. Different cost thresholds yield different distributions of environment features, which can be used to better inform driver-assist systems or autonomous driving.

Using insights from the data sets, we simulate highway driving for an autonomous vehicle, extending the work of [1]. For autonomous highway driving, prior work focuses on individual lane changes [16], [17], [18], [19], [20], whereas we study the sequence of lane changes to see how vehicles may move over time using their risk threshold. Each vehicle uses risk to construct a cost map, then navigates through the environment along the lowest-cost path. Overall, we show that vehicles with a higher risk tolerance make more lane changes and navigate through traffic faster than those with a lower risk tolerance.

The remainder of the paper is organized as follows: Section II gives the problem formulation and defines risk level sets along with collision, safety, and planning thresholds.

[1]Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA [apierson, wilkos, rus]@mit.edu

[2]Laboratory of Information and Decision Systems (LIDS), MIT, Cambridge, MA, USA sertac@mit.edu

This work supported in part NSF Grant 1723943, the Office of Naval Research (ONR) Grant N00014-18-1-2830, Amazon Research, NVIDIA Corporation, and Toyota Research Institute (TRI). This article solely reflects the opinions and conclusions of its authors and not TRI, Toyota, or any other entity. Their support is gratefully acknowledged.

Section III uses the NGSIM and highD data sets to instantiate the parameters of the risk level sets and analyze how it generates distinct distributions within the data. Section IV details how to determine the planning thresholds for driver-assist systems or autonomous driving. The simulation results are presented in Section V.

II. Problem Formulation

Our approach treats all objects and other agents in the environment as dynamic obstacles with respect to the ego agent. Consider an environment $Q \subset \mathbb{R}^N$, with points in Q denoted q. We denote the positions of all agents and obstacles in the environment as p_i, for $i \in \{1, ..., n\}$, with the position of the ego agent denoted p_e. We refer to position of all agents as $p = [p_1^T \mid .... \mid p_n^T]^T$. We model the agents with double-integrator dynamics, such that

$$\ddot{p}_i = u_i,$$

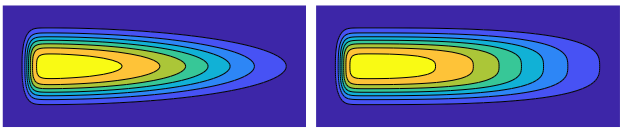
where u_i is the control input for the agent. We also assume that all agents have some maximum speed and acceleration, such that $\|\ddot{p}_i\| < a_i$ and $\|\dot{p}_i\| < v_i$, where a_i and v_i are the maximum acceleration and velocity, respectively. We assume all agents have some braking distance R_b, and some safety radius r_c. Let $d_{ij} = \|p_i - p_j\|$ be the distance between two agents. A collision between agents is defined by $d_{ij} < r_c$.

In [1], we construct the occupancy cost due to all agents as a Gaussian peak centered around the agent's position skewed by a logistic function in the direction of the agent's velocity. This works well for agents with a small footprint, but for agents with larger footprints, the cost along the edge of the agent's body had a lower cost than the center. In this paper, we modify the cost function using a higher-order Gaussian, which flattens the top of the peak so the cost is equal across all points on the agent's body. Without loss of generality, we compute this in the reference frame of the ego vehicle, letting the origin of the system be the ego agent's position, and all dynamics relative to the ego agent's motion. Consider

$$\mathcal{H}\left(q, p_i, \dot{p}_i\right) = \sum_{i=1}^{n} \frac{\exp\left(-\left(\left(q-p_i\right)^T \Omega_i \left(q-p_i\right)\right)^{\beta}\right)}{1+\exp\left(\alpha \dot{p}_i^T \left(q-p_i\right)\right)}, \tag{1}$$

where p_i is the position of agent i, Ω_i is the diagonal matrix of the inverse square of the standard deviation, $\alpha > 0$ is a scaling parameter, and $\beta > 1$ is the exponential parameter. Higher values of β will increase the flatness of the peak. Intuitively, the cost function skews the values in the direction of motion, such that points in the environment carry a higher cost in the immediate path of the agent, and lower cost where the agent is not moving.

The cost function in (1) serves as a proxy for the immediate actions of other agents in the system. Instead of requiring the ego agent to predict the intentions of all other agents in the system before choosing its control policy, we propose that the ego agent can use (1) to quickly assess safe and unsafe regions to restrict its planning space. To compute $\mathcal{H}$, we assume the ego agent is able to observe the state (position) of other agents and obstacles, and can estimate its velocity and acceleration from prior observations. Since



(a) Elliptical Peak (b) Rectangular Peak

Fig. 2. Comparison of the contours generate from the (a) elliptical peak in (1) versus (b) rectangular peak in (3).

$\mathcal{H}$ is defined in the reference frame of the ego agent, these measurements may be relative. We also assume all agents are self-preserving, such that given sufficient braking distance, an agent will stop to avoid collisions rather than collide with another agent. While this assumption does not account for malicious agents in the system, it is reasonable in the context of autonomous driving. Throughout this paper, we refer to several thresholds of the cost function, defined as follows:

Definition 1 (Congestion Cost, $\mathcal{H}$)**.** *A metric to quantify the positions and movement of agents in the system, relative to the ego agent. Higher values of $\mathcal{H}$ relate to more risky areas of the environment.*

Definition 2 (Collision Threshold, $\mathcal{H}_C$)**.** *The cost an ego agent experiences at the point of collision with another agent.*

Definition 3 (Safety Threshold, $\mathcal{H}_T$)**.** *The maximum allowable cost an ego agent experiences while maintaining collision avoidance.*

Definition 4 (Planning Threshold, $\mathcal{H}_P$)**.** *The chosen allowable cost by an ego agent for planning based on aggressiveness preferences, with $\mathcal{H}_P \leq \mathcal{H}_T$.*

Definition 5 (Risk Level Set, $L_{\bar{P}}$)**.** *The set of all points in the environment with a congestion cost less than the chosen planning threshold $\mathcal{H}_P$. The ego agent uses this set of points to restrict its planning.*

$$L_{\bar{P}} = \{q \mid \mathcal{H}(q, p, \dot{p}) \leq \mathcal{H}_P\}, \tag{2}$$

A. Rectangular vs Elliptical Higher-Order Gaussians

For the congestion cost in (1), the peaks of the cost function are elliptical in shape. In certain situations, it may be more desirable to use a rectangular-shaped peak, which we denote as $\hat{\mathcal{H}}$. It is possible to create the rectangular variant by modifying the cost function as

$$\hat{\mathcal{H}}\left(q, p_i, \dot{p}_i\right) = \sum_{i=1}^{n} \frac{\exp\left(-\left(\frac{(q_x-p_{x,i})^2}{\sigma_{x,i}^2}\right)^{\beta} - \left(\frac{(q_x-p_{y,i})^2}{\sigma_{y,i}^2}\right)^{\beta}\right)}{1+\exp\left(\alpha \dot{p}_i^T \left(q-p_i\right)\right)}, \tag{3}$$

where p_x and p_y are the x, y components of the position p for $\mathbb{R}^2$. Figure 2 compares the shape of the cost function for both the elliptical and rectangular form of the cost function.

In Section IV, we present the derivation of the safety thresholds using (1). When using the rectangular variant, we know that $\hat{\mathcal{H}}$ is upper- and lower-bounded by elliptical cost functions, defind by Proposition 1.

Proposition 1. *For rectangular variant of the congestion cost $\hat{\mathcal{H}}$ in* (3)*, the cost is lower-bounded by $\mathcal{H}$ in* (1) *and upper-bounded by*

$$\mathcal{W}(q,p,\dot{p}) = \sum_{i=1}^{n} \frac{\exp\left(-2^{1-\beta}\left((q-p_i)^T \Omega_i (q-p_i)\right)^{\beta}\right)}{1+\exp\left(\alpha \dot{p}_i^T (q-p_i)\right)},$$

such that

$$\mathcal{H}(q,p,\dot{p}) \leq \hat{\mathcal{H}}(q,p,\dot{p}) \leq \mathcal{W}(q,p,\dot{p}).$$

Proof. To prove Proposition 1, consider variables $a \geq 0,\ b \geq 0$ and $\beta > 1$. We know that

$$(a+b)^{\beta} \geq a^{\beta} + b^{\beta}.$$

Furthermore, using the Minkowski Inequality [21], we find

$$2^{1-\beta}(a+b)^{\beta} \leq a^{\beta} + b^{\beta}.$$

By the properties of the exponential function, we know

$$\exp\left(-2^{1-\beta}(a+b)^{\beta}\right) \geq \exp\left(-\left(a^{\beta}+b^{\beta}\right)\right).$$

Let $a_i = \frac{(q_x - p_{x,i})^2}{\sigma_{x,i}^2}$ and $b_i = \frac{(q_y - p_{y,i})^2}{\sigma_{x,i}^2}$. The cost function in (1) can also be written as

$$\mathcal{H}(q,p,\dot{p}) = \sum_{i=1}^{n} \frac{\exp\left(-\left(a_i+b_i\right)^{\beta}\right)}{1+\exp\left(\alpha \dot{p}_i^T (q-p_i)\right)},$$

and it is easily seen that

$$\mathcal{H}(p,\dot{p}) \leq \hat{\mathcal{H}}(p,\dot{p}) \leq \mathcal{W},$$

thus completing the proof. □

Proposition 1 implies that we can use the bounds of $\mathcal{H}$ and $\mathcal{W}$ to quickly generate conservative collision, safety, and planning thresholds for an ego agent using the rectangular variant $\hat{\mathcal{H}}$.

III. Learning from Human Driver Data

The previous section introduced our risk level sets that allow a vehicle to quickly assess the congestion in the environment. A rapid computation of risk is valuable not only for autonomous vehicle planning, but could also be utilized by driver-assistance systems to determine when to intervene. In either case, the vehicle must choose an appropriate response to perceived risk. Here, we examine how features of the environment change with different risk thresholds. For example, learning the underlying distributions of the vehicle velocities at certain risk thresholds will better inform an autonomous or assistive system in speed regulation. Learning how risk maps to lanes would also inform a risk-averse vehicle to avoid lanes that have a higher-than-average risk.

We use the highD [2] and NGSIM [3] data sets, but our approach generalizes to any set of human data. We compare across these two since they both include highway driving from an "eagle-eye" perspective. However, our congestion cost function is relative to the ego agent and does not require a global reference frame, so these techniques can also be applied to data sets from the vehicle perspective. The highD data provides highly accurate trajectory information of over 100,000 vehicles at six different highway segments in Germany. The NGSIM data also provides examples of city and highway driving which we can use for case studies. Although the NGSIM data is known to contain numerous tracking errors and has noisier data ([12], [13], [14]), we performed extensive filtering, smoothing, and reconstruction on the data before using in analysis, and cull the data analyzed to valid trajectories.

Using both data sets, we tune our cost function parameters to fit the human data and find thresholds corresponding to low, medium, and high-risk driving. Next, we use these risk thresholds to parse the data set among environmental features, such as velocity, number of neighbors, and lane preference. These underlying distributions by risk threshold provide a better understanding of the environment conditions. Finally, we examine two high-risk case studies from the data sets: one where a car must swerve into another lane to avoid collision, and another with two lane-splitting motorcycles passing between lanes of cars.

A. Tuning Cost Function Parameters

In both NGSIM and highD data sets, the cars are represented by rectangular bounding boxes. For this reason, we use the rectangular variant of the cost function in (3). Furthermore, we allow the parameters for σ_x, σ_y to vary with the relative velocity of the vehicles. Let $w_{x,i}$ be the width of the vehicle i, with velocity $\dot{p}_i$. We let $\sigma_{x,i} = \frac{w_{x,i}}{2} + |\dot{p}_{x,i}|$ when computing the contribution to the congestion cost from the i-th agent. We find that $\alpha = 0.8$ and $\beta = 1.5$ provided the best fit of the cost function to the expected behavior from cars in the data set. We multiply the cost function by a scaling factor of $A = 15$ to create rounded integer values when describing our risk thresholds. The scaling factor does not affect the shape of the cost function.

During our analysis of the driving data, we found several natural risk thresholds of our cost function, corresponding to low, medium, and high-risk situations for the ego vehicle. We define the low-risk threshold as $\mathcal{H}_1 = \{\mathcal{H} < 1\}$, which includes free-space driving where the cost is expected to be zero. The medium-risk threshold is $\mathcal{H}_2 = \{1 < \mathcal{H} < 5\}$, chosen as it incorporates most vehicles throughout the data set performing average maneuvers. When the value of the cost function exceeded $\mathcal{H}_2$, this typically corresponded to a vehicle quickly swerving into the other lane, tailgating, and other risky behaviors, thus we define the high-risk threshold as $\mathcal{H}_3 = \{\mathcal{H} > 5\}$. Our analysis shows that using these thresholds creates distinct distributions among features of the data, detailed in the following section.

B. Underlying Risk-Based Variations in Data

For each vehicle at each frame, we compute the cost in (3) that the vehicle experiences due to all other vehicles on the road, thus assessing the risk to each vehicle at each point in the data set. The features we examine include: lateral road position (lanes), forward and lateral velocities, nearest neighbor to the ego vehicle, and the total neighbors within a $50m$ radius. These features are typically used in behavior and maneuver classification of vehicles, so understanding the underlying relationship between risk and these features may

TABLE I

MEAN VALUES ACROSS COST THRESHOLDS

		$\mathcal{H}_1$	$\mathcal{H}_2$	$\mathcal{H}_3$
highD	**Fwd-Velo (m/s)**	28.08	24.62	19.98
	Lat-Velo (>0.5 m/s)	0.78	0.88	1.11
	Nearest Neighbor (m)	17.67	8.16	7.38
	Neighbors 50m	3.50	4.61	6.68
NGSIM	**Fwd-Velo (m/s) (ATL)**	5.44	2.86	4.41
	Fwd-Velo (m/s) (LA)	6.11	3.86	6.67
	Fwd-Velo (m/s) (Hwy101)	9.74	7.34	7.01
	Fwd-Velo (m/s) (I-80)	6.91	5.37	5.01
	Nearest Nbr (m) (ATL)	9.73	4.79	4.70
	Nearest Nbr (m) (LA)	7.82	4.61	4.89
	Nearest Nbr (m) (Hwy101)	5.90	4.42	4.24
	Nearest Nbr (m) (I-80)	5.83	4.66	4.14
	Nbrs 50m (ATL)	8.65	11.95	11.46
	Nbrs 50m (LA)	19.17	23.05	20.94
	Nbrs 50m (Hwy101)	22.36	26.63	28.57
	Nbs 50m (I-80)	29.73	33.11	33.12

TABLE II

TWO-SAMPLE KOLMOGOROV-SMIRNOV TEST STATISTICS

		$\mathcal{H}_1, \mathcal{H}_2$	$\mathcal{H}_1, \mathcal{H}_3$	$\mathcal{H}_2, \mathcal{H}_3$
highD	**Fwd-Velo**	0.12	0.41	0.30
	Lat-Velo	0.17	0.51	0.36
	Nearest Neighbor	0.36	0.44	0.09
	Neighbors (50m)	0.11	0.31	0.23
NGSIM	**Lat-Pos (ATL)**	0.06	0.03	0.06
	Lat-Pos (LA)	0.11	0.10	0.05
	Lat-Pos (Hwy101)	0.10	0.14	0.13
	Lat-Pos (I-80)	0.08	0.14	0.07
	Fwd-Velo (ATL)	0.25	0.16	0.23
	Fwd-Velo (LA)	0.19	0.14	0.28
	Fwd-Velo (Hwy101)	0.22	0.30	0.11
	Fwd-Velo (I-80)	0.19	0.24	0.06
	Nearest Nbr (ATL)	0.48	0.44	0.13
	Nearest Nbr (LA)	0.39	0.33	0.09
	Nearest Nbr (Hwy101)	0.30	0.28	0.14
	Nearest Nbr (I-80)	0.24	0.32	0.17
	Nbrs 50m (ATL)	0.27	0.26	0.04
	Nbrs 50m (LA)	0.18	0.11	0.09
	Nbrs 50m (Hwy101)	0.23	0.33	0.13
	Nbs 50m (I-80)	0.16	0.15	0.02

improve further behavioral classifications. It also allows us to quickly filter the data set along the desired risk thresholds. A key finding from our analysis here is that dividing the data set by the low, medium, and high-risk thresholds of $\mathcal{H}$ yields distinct, different distributions among the features. We use the two-sample Kolmogorov-Smirnov (KS) test statistics [22] to verify that the distributions taken along the risk thresholds are distinct from one another. The KS test measures the maximum distance between the CDFs of the distributions. While we do not include the critical threshold for the KS scores, all values reported in the table are above their respective critical threshold, which were typically around $D_{\text{crit}} < 0.03$ for these distributions.

Table I summarizes the mean values for various features of the data set across the different cost threshold. In Table I, the columns $\mathcal{H}_1$, $\mathcal{H}_2$, and $\mathcal{H}_3$ refer to the low, medium, and high-risk thresholds presented in the previous section. To compute the mean value, all instances of that feature were sorted first by the associated cost at that point, then the statistics calculated. Table II reports the KS test statistic between the different thresholds. For example, within the highD data set, we compare how the distribution of nearest neighbor to the ego vehicle varies by risk threshold. For the $\mathcal{H}_1$ threshold, the mean distance to the nearest neighbor was $d = 17.67m$, while for $\mathcal{H}_2$ it is $d = 8.16m$ and for $\mathcal{H}_3$ it is $d = 7.38m$. From Table II, we see that the KS statistic between the $\mathcal{H}_1, \mathcal{H}_2$ distributions is $D = 0.36$, which implies the two distributions are very distinct. However, the KS statistic between $\mathcal{H}_2, \mathcal{H}_3$ is only $D = 0.09$, confirming the two distributions are closer.

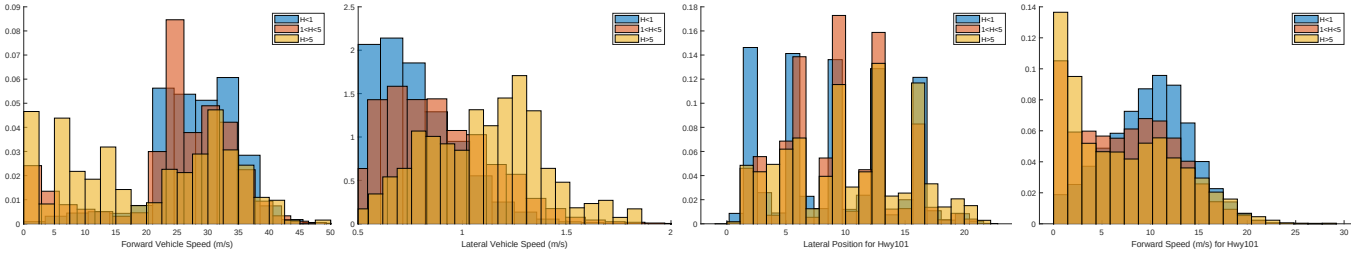
C. Key Results from highD

In the highD data set, we have high-fidelity information on the locations and forward and lateral velocities of all vehicles. Although the data set includes multiple highway segments, we analyze the data across all locations as a single data set. The four features we highlight are: forward velocity (fwd-velo), lateral velocity (lat-velo), nearest neighbor distance, and neighbors within a 50m radius.

When comparing the forward velocity by risk threshold, one might assume that higher velocities carry a higher risk. Our results show a different trend, that the average forward velocity decreases from the low-risk threshold to the high-risk threshold. Figure 3a displays the histograms of the distributions of forward velocity of all vehicles, separated by cost threshold. From the histogram, we notice that the high-risk threshold $\mathcal{H}_3$ appears to have a bimodal distribution, with one peak at higher speeds, and one peak at low speeds. The lower speed peak corresponds to stop-and-go traffic along the road, an inherently riskier traffic mode than driving along an open road. Figure 3b plots the distributions of lateral vehicle speeds as related to the cost threshold. Here, we look specifically at values of $v_{\text{lat}} > 0.5$, which approximates when vehicles are changing lanes. At higher lateral speeds, we see higher values of $\mathcal{H}$, corresponding to an increase in risk. We note this as a correlation, but further analysis is needed to determine if the vehicles change lanes at higher speeds to avoid a high-risk situation, or are inducing a higher risk from their faster lane change. From Table II, the difference between $\mathcal{H}_1$ and $\mathcal{H}_3$ is one of the largest values for any of the distributions pairs tested. Finally, we examine the overall congestion the vehicles experience. We compute both the nearest neighbor, which is the distance to the nearest car traveling in the same direction, as well as the number of neighbors within a 50m radius around the ego vehicle. From Table I, we see that the distance to the nearest neighbor decreases as the threshold increases, while the total number of neighbors increases as risk increases. However, from Table II, we note the difference is much less significant between the thresholds $\mathcal{H}_2$ and $\mathcal{H}_3$, indicating only a small difference in the overall distribution.

D. Key Results from NGSIM

The NGSIM data set comprises four distinct locations across the US: Peachtree St in Atlanta (ATL), Lankershim Boulevard in Los Angeles (LA), Highway 101 in Los Angeles (Hwy101), and Interstate 80 in Emeryville (I-80). The



(a) highD v_{fwd} (b) highD v_{lat} (c) NGSIM x_{lat} (d) NGSIM v_{fwd}

Fig. 3. For all histograms, $\mathcal{H}_1$ is shown in blue, $\mathcal{H}_2$ in red, and $\mathcal{H}_3$ in yellow. (a) Histogram of the forward velocity of all vehicles in the highD data set, categorized by the various thresholds of the cost function. (b) Histogram of the lateral velocity (for velocities $v_{\text{lat}} > 0.5\ m/s$) of all vehicles in the highD data set, categorized by the various thresholds of the cost function. (c) Distributions of lateral position by various cost thresholds in the Highway 101 data. (d) Forward velocity distributions by various cost thresholds along Highway 101.

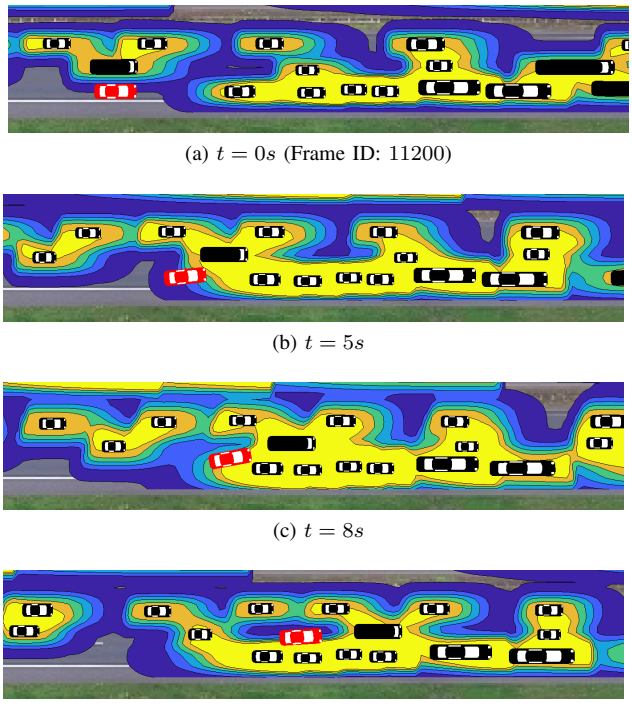
first two locations are city traffic through signalized intersections, while the latter two are highways. In general, we find the highway driving to include more congestion than the city driving, as indicated by the average number of neighbors in Table I. We analyze the NGSIM data by location, since the locations are more distinct in their topologies.

Similar to the highD data, we examine trends that emerge across features in the data set. Here, we look at the lateral position instead of lateral velocity. Unlike highD, the NGSIM data set only includes forward velocity and does not include reliable lane markings for the vehicles. However, by examining the lateral position, we can infer lanes along the road, as well as whether certain lane positions correlate with higher risk. Figure 3c plots the histograms of the lateral position distributions for Highway 101. From the histograms, we notice that for the high-risk distribution, the lateral positions are more dispersed between lanes. This is in part to lane-splitting motorcycles between the left-most and second lanes along the Highway 101 section, discussed further in the case studies. Figure 3d shows the distributions of the forward velocity for the Highway 101 data. Similar to the highD data in Figure 3a, we notice that the higher risk distribution includes two peaks, one at lower speeds, and one at the higher speeds. Overall, the NGSIM data set has much slower highway speeds, due to the higher levels of traffic and congestion present in this data set. In particular, the mean number of highway neighbors across all NGSIM highway data is $n = 28.92$, versus the average of $n = 4.93$ neighbors within 50 m for the highD data set.

E. Case Studies

In addition to analyzing broad trends and distributions in the data sets, we can also examine case studies to understand high-risk situations. Our risk thresholds provides a rapid assessment of the data, and high values of $\mathcal{H}$ indicate points in the data set where interactions occur, allowing us to quickly find examples of high-risk situations which may be difficult to identify if we were only sorting on speed or neighbor distance alone. Here, we examine an evasive lane change in the highD data set, and lane-splitting motorcycles from the NGSIM.

1) Evasive Lane Change: From the highD data set, we noticed spikes in a vehicle's risk over time that correlated with lane changes. Here, we present one example where a high risk value occurred when a vehicle needed to make an evasive lane change. This particular instance may be located as car 1034 within Track 25 of the highD data files. Figure 4 illustrates a vehicle maneuvering into the other lane after rapidly braking to avoid the other vehicles.



(a) $t = 0s$ (Frame ID: 11200)

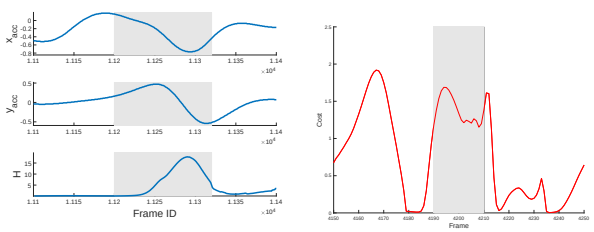
(b) $t = 5s$

(c) $t = 8s$

(d) $t = 10s$

Fig. 4. Here, we see the ego vehicle approach a backed-up line of cars. It rapidly brakes to avoid collision (a-b), then changes lanes to decrease its cost (c-d).

In Figure 4, the ego vehicle (red) approaches a line of stopped cars, and needs to break quickly to avoid collision. The preceding line of cars is stopped, and so the ego vehicle waits for the truck in the next lane to pass before sliding into the other lane. By changing lanes, it moves from a high-risk area in the environment to a lower-risk lane. Figure 5a plots the forward (x) and lateral (y) accelerations over this time span, as well as the computed values of $\mathcal{H}$. The accelerations demonstrate the car's braking and lane swerve, which results in the cost decreasing after the rapid increase as it approaches



(a) (b)

Fig. 5. (a) Forward (x) and lateral (y) accelerations and overall cost $\mathcal{H}$ for Car 1034. The shaded region corresponds to the frames shown in Figure 4. Note the car slows down and changes lane to decrease its cost. (b)

the vehicles. By applying our risk level sets to the highD data set, we were quickly able to identify this situation for further analysis.

2) Lane-Splitting Motorcycle: In the state of California, it is legal and commonplace for motorcycles to ride between two lanes of traffic, also known as lane-splitting. In fact, from the instances in the NGSIM data set, motorcycles often prefer to use this lane-splitting technique in congestion, using primarily the region adjacent to the outside lane.

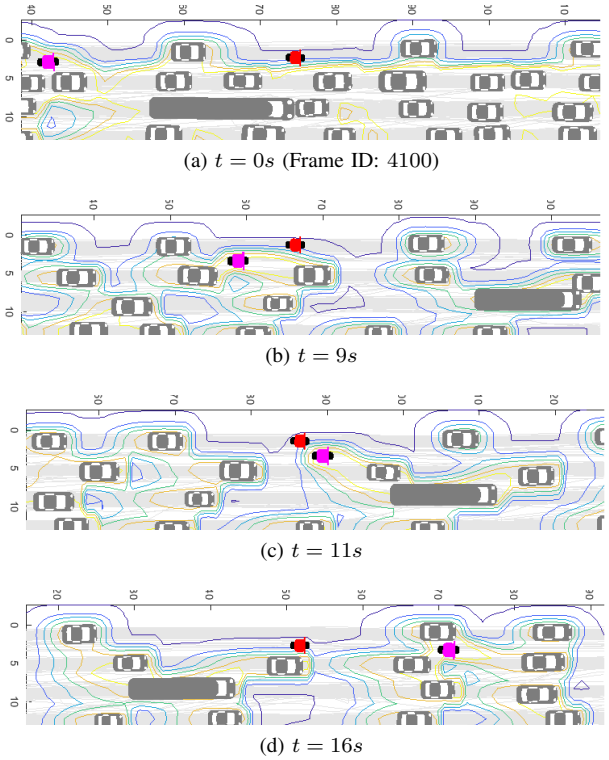
To illustrate this preference for lane-splitting, Figure 5 shows one example of a slow motorcycle (red) lane-splitting, until a faster motorcycle (pink) approaches from behind. While it appears that the leftmost lane is free for the red motorcycle, the motorcycle only enters the lane to let the pink motorcycle pass, then returns to lane-splitting. Figure 5b plots the cost of the red motorcycle through the interaction. At the beginning of the interaction, the red motorcycle slides over, which decreases it risk, until the pink motorcycle passes. The grey region in Figure 5b corresponds to Figure 5(b-c). The passing motorcycle increases the the risk of the red motorcycle, but once it masses, the red motorcycle returns to lane-splitting.

IV. Planning with Risk Level Sets

From the NGSIM and highD data sets, we found thresholds of our cost function that corresponded to low, medium, and high-risk situations. For driver-assist and autonomous systems, we also need to know the collision, $\mathcal{H}_C$ and safety, $\mathcal{H}_T$, thresholds of the cost function to determine a safe planning threshold, $\mathcal{H}_P$. This section presents how to compute these thresholds, as well as an overview of how obeying the planning threshold can guarantee collision avoidance under the assumption of self-preserving agents. By computing the collision threshold $\mathcal{H}_C$, we show that for agents starting within their risk level set $L_{\bar{P}}$ and only choosing actions within this set, the agent can avoid collisions [1].

A. Two-Agent System

We start with a two-agent system to demonstrate our collision avoidance guarantees, then generalize to a multi-agent system. In the reference frame of the ego vehicle, the two-agent system comprises the ego agent, located at $p_e = 0$, and another agent, located at p. To compute the threshold values, we follow the proof techniques similar to [1] for (1), but omit some of the proofs for brevity. For an ego agent to avoid collisions, we know the distance between itself and the



(a) $t = 0s$ (Frame ID: 4100)

(b) $t = 9s$

(c) $t = 11s$

(d) $t = 16s$

Fig. 6. Example of lane-splitting motorcycles from the NGSIM data set along Interstate 80. (a-d) Here, we see the red motorcycle slide over into a lane to let the pink motorcycle pass between lanes.

other agent must satisfy $d > r_c$. By the properties of $\mathcal{H}$, we show that a collision is defined by the maximum cost $\mathcal{H}_C$, and that so long as the agent has a lower value of $\mathcal{H}$, it cannot be in collision. Furthermore, by restricting its planning set by (2), the agent will never experience a cost greater than the collision cost, thereby avoiding collisions.

Lemma 1. *([1]) For a two-agent system with cost $\mathcal{H}$ defined in* (1)*, safety radius r_c, $\alpha > 0$, $\beta > 1$, and $\sigma_m = \max\{\sigma_x, \sigma_y\}$, the maximum cost before collision is*

$$\mathcal{H}_C = \frac{\exp\left(-\left(\frac{r_c^2}{\sigma_m^2}\right)^\beta\right)}{1 + \exp\left(-\alpha v_{\max} r_c\right)}. \tag{4}$$

From Lemma 1, the maximum cost occurs when an agent is traveling at its maximum velocity and is located $d = r_c$ away from the ego agent. If our agents have instantaneous braking, we could define our planning sets based on the minimum cost before collision. However, for agents that cannot brake instantaneously, experiencing a cost of $\mathcal{H}_C$ would result in a collision. Instead, we define the safety threshold $\mathcal{H}_T$ as the cost where agents may still avoid collisions given braking distance R_b, defined in Lemma 2.

Lemma 2. *([1]) For the two-agent system with $\mathcal{H}$ defined in* (1)*, braking distance R_b, $\alpha > 0$, $\beta > 1$ and $\sigma_\ell = \min\{\sigma_x^2, \sigma_y^2\}$ the safety threshold for avoiding collisions is*

$$\mathcal{H}_T = \frac{\exp\left(-\left(\frac{R_b^2}{\sigma_\ell^2}\right)^\beta\right)}{2}. \tag{5}$$

From the safety threshold $\mathcal{H}_T$, the ego agent chooses its allowable planning threshold $\mathcal{H}_P \leq \mathcal{H}_T$. Choosing a lower value for $\mathcal{H}_P$ will increase the buffer around each agent, which in turn results in more conservative actions for the ego agent. In Theorem 1, we show that when an ego agent stays within its chosen planning threshold $\mathcal{H}_P$, its cost will not exceed $\mathcal{H}_C$, which implies it never collides with any other agent.

Theorem 1. *([1]) For an ego agent starting with some initial cost $\mathcal{H} < \mathcal{H}_P$ computed in* (1)*, and choosing control actions within the set of points defined by $L_{\bar{P}}$ in* (2)*, the total cost experienced by the ego agent will never exceed $\mathcal{H}_C$.*

Proof. The proof is a straightforward extension of the proof of Theorem 1 in [1] with our cost function in (1), and omitted here for space constraints. □

B. Multi-Agent System

When considering multiple agents and obstacles, a simple approach would be to construct the risk level set as the intersection of all pairwise level sets. Consider a system of $i = \{1, ..., n\}$ agents interacting with an ego agent. Let $L_{\bar{P},i}$ denote the pairwise level set computed between the ego agent and agent i using (2). The overall risk level set is

$$L_{\bar{P}} = \bigcap_{i=1}^{n} L_{\bar{P},i}. \tag{6}$$

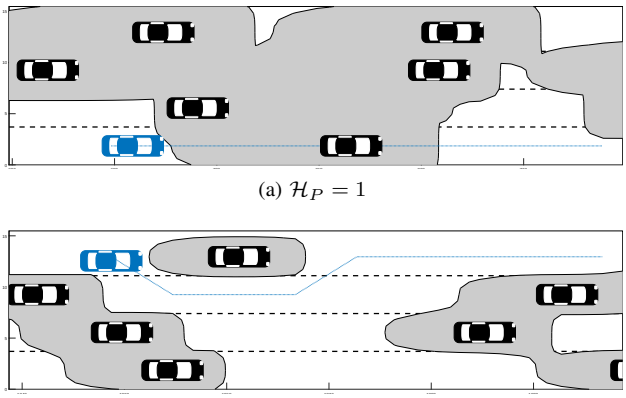
It is a straightforward extension of Theorem 1 to show this also guarantees collision avoidance. However, it maybe become computationally intractable to compute all pairwise level sets for large numbers of agents. In [1], we use the circle-packing problem to determine a planning threshold approximation when computing the cost from (1). Using

$$H_P \leq \frac{3\exp\left(-\left[\frac{R_b^2}{\sigma_\ell}\right]^\beta\right)}{2}$$

creates a conservative estimate of for the ego agent for $n \geq 6$ other agents in the environment.

V. Autonomous Vehicle Simulations

For our autonomous vehicle simulations, we generated a multi-lane highway environment in Matlab. Here, we wish to test the impact of the choice of planning threshold on a car's navigation and lane changes over a highway segment. We implement a lane-change planner similar to [1], where the vehicle computes a weighted graph of the discretized environment, using $\mathcal{H}$ to determine the edge weights. The vehicle then plans its sequence of lane changes by choosing the lowest-cost path through the graph. We generate the environment with a mix of "active cars" planning their sequence of lane changes, and "obstacle cars" which stay in their lane. For the obstacle cars, we use the intelligent driver model (IDM) of traffic flow [23] to simulate fluctuations in traffic. Figure 7 illustrates an example of a vehicle on a four-lane highway. As shown in the figure, the choice of planning threshold determines the available area the ego agent (blue) may use in planning its path.



(a) $\mathcal{H}_P = 1$

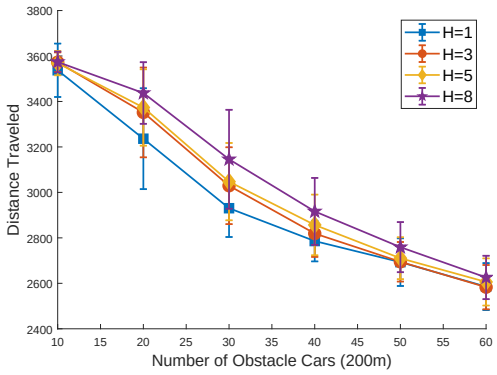
(b) $\mathcal{H}_P = 8$

Fig. 7. Comparison of (a) low-risk and (b) high-risk planning thresholds for the ego agent (blue). Here, the grey region are points deemed too risky. Low-risk agents have a more restricted planning space.

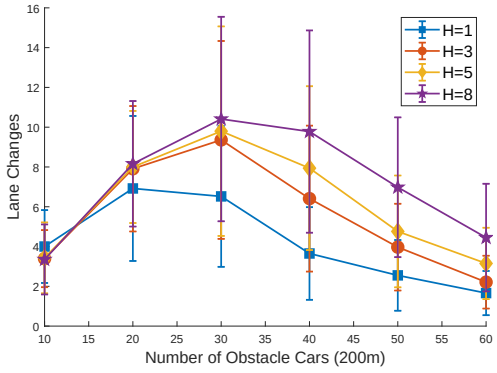
For our simulations, we set the ego vehicle's desired speed at $v_{\text{des}} = 30\ m/s$, yet restrict the maximum speed of all the lanes to $v_o \leq 25\ m/s$ for obstacle vehicles. This encourages the ego vehicle to weave between lanes in order to reach its goal. While the obstacle vehicles use IDM for velocity management within the lane, the active vehicles use their computed planning set $L_{\bar{P}}$, decelerating to avoid entering into regions above their allowed risk threshold, and accelerating in free space. We ran the simulations for $t = 120s$, with $\Delta t = 0.1s$ increments, and measure the ego vehicle's progress along the road segment at the end of the time limit.

We choose the four planning thresholds $\mathcal{H}_P = \{1, 3, 5, 8\}$ based on the low, medium, and high-risk thresholds observed in the human driving data. All cost function parameters used in the simulation are identical to those used in analyzing the data set. We ran the simulation with varying numbers of obstacle cars along the roadway, with randomized initial positions for each trial (all initial positions were checked to ensure valid, collision-free vehicle arrangements). For each of the 24 scenarios (6 possible obstacles, 4 possible risk thresholds), we ran 100 randomized trials and recorded the total distance traveled by the ego vehicle, as well as the total number of lane changes. Figure 8a and 8b plot the mean and standard deviations of the distance traveled and lane changes, respectively, across the different scenarios. Each trend line corresponds to a specific risk threshold.

Across the simulations, we observe that vehicles with a higher risk threshold make more lane changes and travel further along the road segment than vehicles with the lower risk threshold. The performance difference between high-risk and low-risk vehicles was dependent on the number of obstacle cars in the environment. At $n = 10$ obstacle cars, the performance is identical across all thresholds, as the cars are in effectively free-space driving and need minimal lane changes to avoid the other cars. At $n = 60$ obstacle cars, we see the high-risk vehicles still attempt more lane changes, but the traffic is too dense for them to make significantly more progress along the highway segment, as they are trapped behind slower traffic.



(a)



(b)

Fig. 8. (a) Mean total distance traveled in $t = 120s$ with varying number of obstacles. Overall, higher-risk vehicles travel further than the lower-risk vehicles, with the total distance traveled dependent on congestion. (b) Total lane changes for varying number of obstacles. At low congestion, the ego vehicle does not need to make lane changes to travel at its desired speed, and at high congestion, the traffic is too dense for safe lane changes.

VI. Conclusions

This paper applies our risk level sets to human driving data, using both the highD and NGSIM data sets. By studying human data, we generate a new formulation of the congestion cost function applicable to highway and city driving environments. Furthermore, the risk level sets present a unique way to study the distribution of data set features corresponding by risk. We can quickly identify high-risk situations within the data, both to further examine these scenarios and better understand the environment. For driver-assist or autonomous vehicle systems, knowing these thresholds will better inform the choice of behavior. We outline how the risk thresholds may be used in planning, and using the thresholds learned from the data set, present autonomous highway driving simulations. While we focus on two data sets in this paper, our results generalize to any data set on human driving, and future work may explore risk level sets using data from the perspective of the vehicle.

References

[1] A. Pierson, W. Schwarting, S. Karaman, and D. Rus, "Navigating congested environments with risk level sets," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018.

[2] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems," in *2018 IEEE 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018.

[3] FWHWA and US Department of Transportation. (2017) Ngsim-next generation simulation. [Online]. Available: https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm

[4] F. Guo and Y. Fang, "Individual driver risk assessment using naturalistic driving data," *Accident Analysis & Prevention*, vol. 61, pp. 3 – 9, 2013, emerging Research Methods and Their Application to Road Safety Emerging Issues in Safe and Sustainable Mobility for Older Persons The Candrive/Ozcandrive Prospective Older Driver Study: Methodology and Early Study Findings.

[5] J. Wang, Y. Zheng, X. Li, C. Yu, K. Kodaka, and K. Li, "Driving risk assessment using near-crash database through data mining of tree-based model," *Accident Analysis & Prevention*, vol. 84, pp. 54 – 64, 2015.

[6] T. A. Dingus, F. Guo, S. Lee, J. F. Antin, M. Perez, M. Buchanan-King, and J. Hankey, "Driver crash risk factors and prevalence evaluation using naturalistic driving data," *Proceedings of the National Academy of Sciences*, vol. 113, no. 10, pp. 2636–2641, 2016. [Online]. Available: https://www.pnas.org/content/113/10/2636

[7] G. Li, S. E. Li, B. Cheng, and P. Green, "Estimation of driving style in naturalistic highway traffic using maneuver transition probabilities," *Transportation Research Part C: Emerging Technologies*, vol. 74, pp. 113 – 125, 2017.

[8] J. Monteil, A. Nantes, R. Billot, J. Sau, and N. E. Faouzi, "Microscopic cooperative traffic flow: calibration and simulation based on a next generation simulation dataset," *IET Intelligent Transport Systems*, vol. 8, no. 6, pp. 519–525, Sep. 2014.

[9] D. J. Phillips, T. A. Wheeler, and M. J. Kochenderfer, "Generalizable intention prediction of human drivers at intersections," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, June 2017, pp. 1665–1670.

[10] R. P. Bhattacharyya, D. J. Phillips, B. Wulfe, J. Morton, A. Kuefler, and M. J. Kochenderfer, "Multi-agent imitation learning for driving simulation," *arXiv preprint arXiv:1803.01044*, 2018.

[11] S. Su, K. Muelling, J. Dolan, P. Palanisamy, and P. Mudalige, "Learning vehicle surrounding-aware lane-changing behavior from observed trajectories," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, June 2018, pp. 1412–1417.

[12] C. Thiemann, M. Treiber, and A. Kesting, "Estimating acceleration and lane-changing dynamics from next generation simulation trajectory data," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2088, pp. 90–101, 2008.

[13] M. Montanino and V. Punzo, "Making ngsim data usable for studies on traffic flow theory: Multistep method for vehicle trajectory reconstruction," *Transportation Research Record*, vol. 2390, no. 1, pp. 99–111, 2013. [Online]. Available: https://doi.org/10.3141/2390-11

[14] B. Coifman and L. Li, "A critical evaluation of the next generation simulation (ngsim) vehicle trajectory dataset," *Transportation Research Part B: Methodological*, vol. 105, pp. 362 – 377, 2017.

[15] R. Krajewski, T. Moers, D. Nerger, and L. Eckstein, "Data-driven maneuver modeling using generative adversarial networks and variational autoencoders for safety validation of highly automated vehicles," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, Nov 2018, pp. 2383–2390.

[16] K. Ahmed, M. Ben-Akiva, H. Koutsopoulos, and R. Mishalani, "Models of freeway lane changing and gap acceptance behavior," *Transportation and traffic theory*, vol. 13, pp. 501–515, 1996.

[17] W. Schwarting and P. Pascheka, "Recursive conflict resolution for cooperative motion planning in dynamic highway traffic," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, Oct 2014, pp. 1039–1044.

[18] G. Schildbach and F. Borrelli, "Scenario model predictive control for lane change assistance on highways," in *2015 IEEE Intelligent Vehicles Symposium (IV)*, June 2015, pp. 611–616.

[19] M. Wang, Z. Wang, S. Paudel, and M. Schwager, "Safe distributed lane change maneuvers for multiple autonomous vehicles using buffered input cells," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 1–7.

[20] W. Schwarting, J. Alonso-Mora, and D. Rus, "Planning and decision-making for autonomous vehicles," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, pp. 187–210, 2018.

[21] G. Hardy, J. Littlewood, and G. Pólya, *Inequalities*, ser. Cambridge Mathematical Library. Cambridge University Press, 1988.

[22] I. M. Chakravarti, R. Laha, and J. Roy, *Handbook of Methods of Applied Statistics, Volume I*. John Wiley and Sons, 1967.

[23] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Phys. Rev. E*, vol. 62, pp. 1805–1824, Aug 2000.