# Safe Navigation of Mobile Robots in Crowded Uncertain Environments Utilizing Reinforcement Learning

Adrián Iribar[1] Dadi Hrannar Davidsson[1] Gayath Sanjula De Silva[1] Lasse Due Hornshøj[1] Oghuz Madinali[1]

Søren Haugaard[1]

***Abstract*—This study introduces an online self-learning control system for mobile robots to enhance safe navigation around humans in uncertain environments. The project faces this problem by integrating a safe reinforcement learning algorithm namely the LB-SGD algorithm into a mobile robot. The LB-SGD algorithm uses a Log barrier to ensure that the mobile robot is not able to hit an obstacle as the summation of the cost function expands logarithmically when approaching an obstacle. By making it theoretically impossible for the mobile robot to hit an obstacle the algorithm is then able to use a gradient decent function to calculate a path to the goal. This paper's contribution to this algorithm is applying it to a real robot which, to our knowledge, has not been done before.**

**Currently, the algorithm has not been implanted into a real robot but is being tested on the said robot in a simulated space, the theory has been crafted however the implementation is currently problematic.**

## I. Introduction

In the current age of mobile robotics the ability to move in an uncertain crowded environment has been widely researched but not solved. The research has shown that collision avoidance and safety are some of the key aspects. These aspects become even more important in highly dynamic and crowded environments, where the robot can be forced to deviate away from the pre-calculated global trajectory to safely avoid humans [1]. This paper is based on the aforementioned key aspects and thus puts importance on minimizing risks in autonomous decision-making. With the evolution of robotics in social environments, safe navigation has become crucial. In this context, the focus of this work is on the development of an advanced navigation system for the Boston Dynamics Spot robot with the goal of making it able to navigate autonomously in environments with static and dynamic obstacles, such as pedestrians. Successfully ensuring safe navigation involves integrating Safe Reinforcement Learning techniques that can be applied to Spot's dynamics. This project is focused on precisely adapting these techniques to enhance Spot's autonomous safe navigation. The implemented strategies and techniques will be detailed in the paper, also the results obtained exist both in virtual simulations and real-life environments. The main goal of this work is to contribute to the advancement of the investigation of secure navigation of robots in complex environments, providing valuable insights for the implementation of these in real-life scenarios. One of the key aspects to consider is people's lack of experience with robots. Taking this into account, in an environment where humans and robots share workspace, robots must consider the comfort and safety of people to be accepted and tolerated [2]. When people find themselves in unexpected situations tend to behave in a way that can be difficult to analyze and even more to predict. This is one of the main reasons why security is one of the most challenging issues to implement correctly.

Fig. 1. placeholder for image of full system, either spot the robot alone or with some movement visualisation as well

## II. RELATED WORK

In the field of robotics, various Reinforcement Learning approaches have been used for navigation within crowded environments. Vasquez et al. [3] learn a cost function, using Inverse Reinforcement Learning, to depict the dynamics of social navigation, utilizing a dataset obtained by teleoperating a robot across diverse real-world scenarios. They assess the effectiveness of their learned model in replicating trajectories characterized by social compliance. This compliance is modeled through a composite score derived from cost functions that represent aspects of human comfort. Two other approaches from Luber et al. [4] and Trautman et. al [5] acquire the capability to plan socially acceptable paths among humans by training on top-view pedestrian scenes, and learning human-

[1] Robotics, Department of Electronic Systems, Aalborg University, Aalborg, Denmark

like navigation. Truong and Ngo [6] combine elements from the social force model [7] and the reciprocal velocity model [8] to create a robot motion system that is socially aware in crowded scenes.

The before mentioned approaches all have advantages our approach could use. However, our approach is required to be safe while learning, thereby introducing the problem of online safe reinforcement learning. Usmanova et al. [9] seek to address the challenge of ensuring safety while learning. Their approach is based on logarithmic barriers, which they optimize using stochastic gradient descent with adaptive step sizes. This algorithm they have named LB-SGD. The fundamental insight lies in the fact that the descent direction of the logarithmic barrier serves to guide the iterates away from the boundary, at the same time converging towards an approximate Karush-Kuhn-Tucker (KKT) point in the context of non-convex scenarios and approximating a minimizer in the case of convex optimization. They analytically established the safety of the method throughout the learning process and conducted a comprehensive analysis of convergence rates for non-convex, convex, and strongly-convex problems. Through analytical demonstrations, they compared the performance of their method with existing approaches, revealing that, the sample and computational complexity scaled efficiently to high dimensions, and it maintained optimization iterates within the feasible set, with high probability. Furthermore, they showcased the effectiveness of the log barrier approach in addressing high-dimensional constrained reinforcement learning problems.

Our paper is based Usmanova et al. [9], and our contribution is a continuation of the LB-SGD approach. We have developed a more application-based approach, using the Boston Dynamic Spot robot. The main contribution is practical applications, specifically addressing the challenges associated with safe learning control for robots operating in uncertain environments with human interactivity. Our contribution lies in the conversion of the theoretical concepts introduced in the aforementioned research paper into a real-life application, focusing on the deployment of a robot. We adapt and implement the log barrier techniques to ensure the safety and reliability of blackbox optimization methods. Our approach aims to ensure the safety of both the robot and the human by employing the principles derived from LG-SGD to ensure safety for the robot's learning and decision-making processes. This application extension contributes to the practical viability and effectiveness of safe learning control strategies in crowded areas.

## III. Approach

As the main focus for our paper is the application of Usmanova et al. [9] approach, this section will present the necessary parts of their approach, as well as our contribution for the online safe reinforcement learning part. Section III-A briefly introduces the general constrained optimization problem, which provides the framework for addressing safe learning in scenarios where the constraints are unknown. Section III-B gives a short description of the logarithmic barrier used for repulsion. Section III-C introduces the structure of the LB-SGD algorithm developed by Usmanova et al. [9]. Lastly, section III-E will show our contribution and usage of the LB-SGD algorithm. Our paper follows the same assumptions made in Usmanova et al. [9], these assumptions can be seen below.

The **first assumption** is to consider $X$ with a bounded diameter, meaning there exists $R < 0$ such that for all $x, y \in X$, the inequality $||x - y|| \leq R$ holds.

The **second assumption** is that the objective and constraint functions $f_i(x)$ for $i \in \{0, \ldots, m\}$ are $M_i$ -smooth and $L_i$ -Lipschitz continuous on $X$ with constants $L_i, M_i > 0$. We denote $L := \max_{i \in \{0,\ldots,m\}} L_i$ and $M := \max_{i \in \{0,\ldots,m\}} M_i$.

The **third assumption** is that there exists a known starting point $x_0 \in \mathcal{X}$ at which $\max_{i \in [m]} f^i(x_0) \leq -\beta$, for $\beta > 0$.

The **fourth assumption** is letting $\mathcal{I}_\rho(x) := \left\{i \in [m] \mid -f^i(x) \leq \rho\right\}$ be the set of $\rho$-approximately active constraints at $x$ with $\rho > 0$. For some $\rho \in \left(0, \frac{\beta}{2}\right]$ and for any point $x \in \mathcal{X}$, there exists a direction $s_x \in \mathbb{R}^d : \|s_x\| = 1$, such that $\left\langle s_x, \nabla f^i(x)\right\rangle > l$ with $l > 0$, for all $i \in \mathcal{I}_\rho(x)$.

### A. Problem statement

A general constrained optimization problem can be seen below:

$$\begin{aligned} &\min_x f^0(x) \\ s.t.\ &f^i(x) \leq 0, i = 1, ..., m, \end{aligned} \tag{1}$$

where the objective function $f^0 : \mathbb{R}^d \rightarrow \mathbb{R}$ and the constraints $f^i : \mathbb{R}^d \rightarrow \mathbb{R}$ are unknown, possibly non-convex functions.

The objective is to address the safe learning problem by obtaining a solution to the constrained problem. This entails maintaining the feasibility of all iterates $(x_t)$ within the optimization procedure $(x_t \in X)$ with a high probability during the learning process.

*1) Measurements:* In the applications we normally examine, the information accessible to the learner is subject to noise. Specifically, the learner can observe perturbed gradients and values of $f^i(x)$, $\forall i = 0, ..., m$ at the designated points $(x_t)$. Therefore, we assume access to first-order stochastic measurements for each $f^i(x)$, which deliver a pair of value and gradient stochastic measurements:

$$O(f^i, x, \varepsilon) = (F^i(x, \varepsilon), G^i(x, \varepsilon)). \tag{2}$$

where $F^i(x, \varepsilon)$ is the unbiased stochastic value and $G^i(x, \varepsilon)$ is the biased stochastic gradient.

The mission for this paper, given the first-order stochastic information, is to approximate a solution to the constrained optimization problem, while only making value and gradient queries that satisfy the safety constraint i.e. lie inside the feasible set $X$ with high probability. For this to be possible, the logarithmic barrier optimization approach is introduced.

### *B. The logarithmic barrier*

The approach's key idea is to substitute the unconstrained log barrier surrogate $min_{x \in \mathbb{R}^d}$ $B_\eta(x)$ for the original constrained optimization problem, where $B_\eta(x)$ and its gradient $\nabla B_\eta(x)$ are defined as follows.

$$B_\eta(x) = f^0(x) + \eta \sum_{i=1}^{m} - \log(-f^i(x)), \tag{3}$$

$$\nabla B_\eta(x) = \nabla f^0(x) + \eta \sum_{i=1}^{m} \frac{\nabla f^i(x)}{-f^i(x)}. \tag{4}$$

In the content of this surrogate, $B_\eta(x)$ grows logarithmically as the argument converges to the boundary set of $X$. therefore, a benefit of this method is the automatic preservation of the feasibility of all iterates through careful selection of the optimization step-size.

### *C. Main approach, SGD*

It is suggested to use SGD with an adjustable step size in order to minimize the objective $B\eta$, which is the unconstrained log barrier. The method is called LB-SGD. That demonstrated for the target probability $1 - \hat{\delta}$, LB-SGD (with confidence parameter $\delta = \hat{\delta}/Tm$) produces the following convergence results:

*First Case:* For the non-convex case, in most $T = O\left(\frac{1}{\varepsilon^3}\right)$ iterations, and with $\sigma_i(n) = O\left(\varepsilon^2\right)$ and $\hat{\sigma}_i(n) = O(\varepsilon)$, LB-SGD outputs $x_{\hat{t}}$ which is $\varepsilon-$KKT point with probability $1-\hat{\delta}$. Finally, its required $N = \text{Tn} = O\left(\frac{1}{\varepsilon^7}\right)$ measurements queries $\mathcal{O}\left(f^i, x, \xi\right)$ for all $i \in \{0, \ldots, m\}$.

*Second Case:* For the convex case, in most $T = \tilde{O}\left(\frac{\|x_0 - x^*\|^2}{\varepsilon^2}\right)$ relations of LB-SGD, and with $\sigma_i(n) = \tilde{O}\left(\varepsilon^2\right)$ and $\hat{\sigma}_i(n) = \tilde{O}(\varepsilon)$, gets output $\bar{x}_T$ such that with probability $1 - \hat{\delta} : f^0\left(\bar{x}_T\right) - \min_{x \in \mathcal{X}} f^0(x) \leq \varepsilon$. In total, its required $N = Tn = \tilde{O}\left(\frac{1}{\varepsilon^6}\right)$ calls of the measurements $\mathcal{O}\left(f^i, x, \xi\right)$ for all $i \in \{0, \ldots, m\}$.

*Third Case:* For the $\mu$-strongly-convex case, in most $T = \tilde{O}\left(\frac{1}{\mu\varepsilon} \log \frac{1}{\varepsilon}\right)$ relations of LB-SGD with decreasing $\eta$, and with $\sigma_i(n) = \tilde{O}\left(\eta^2\right)$ and $\hat{\sigma}_i(n) = \tilde{O}(\eta)$, for the output $\hat{x}_K$ had with probability $1 - \hat{\delta} : f^0\left(\hat{x}_K\right) - \min_{x \in \mathcal{X}} f^0(x) \leq \varepsilon$. In total, its required $N = \tilde{O}\left(\frac{1}{\varepsilon^5}\right)$ calls of the measurements $\mathcal{O}\left(f^i, x, \xi\right)$ for all $i \in \{0, \ldots, m\}$.

*Forth Case:* Using finite difference to estimate the function gradients in the zeroth-order information scenario, it is derived the following limitations on the number of measurements:

- $N = O\left(\frac{d^2}{\varepsilon^7}\right)$ to obtain $\varepsilon$-approximate KKT point in the non-convex case;
- $N = \tilde{O}\left(\frac{d^2}{\varepsilon^6}\right)$ to obtain $\varepsilon$-approximate minimizer in the convex case;
- $N = \tilde{O}\left(\frac{d^2}{\varepsilon^5}\right)$ to obtain $\varepsilon$-approximate minimizer in the strongly-convex case;

*Fifth Case:* For any measurement in the aforementioned scenarios, there is a probability of $1 - \delta$ that ensures safety.

**Algorithm 1** Log Barrier Gradient estimator $O_\eta(x_t, n)$

```
Input: Oracles $F^i(\cdot, \xi), G^i(\cdot, \xi), \forall i \in \{0, \ldots, m\}, x_t \in \mathcal{X}, \eta > 0$, number of measurements $n$
$g_t \leftarrow G_n^0\left(x_t, \xi_t\right) + \eta \sum_{i=1}^{m} \frac{G_n^i(x_t, \xi_t)}{-F_n^i(x_t, \xi_t)}$;
Output: $g_t$
```

To lessen the variances, we permit taking a batch of measurements for each call and averaging them as specified. Whereas $\sigma_i^2(n) := \frac{\sigma_i^2}{n}, \hat{\sigma}_i^2(n) := \frac{\hat{\sigma}_i^2}{n}$.

### *D. The algorithms of LB-SGD*

The step size $\gamma_t$ is selected in a way that ensures $f^i\left(x_{t+1}\right) \leq f^i\left(x_t\right)/2$. $\underline{\alpha}_t^i$ according to the quadratic smoothness upper bound (black) on the requirement. Based on the mean estimator $\alpha_t^i$ is the lower bound on $\alpha_t^i = -f^i\left(x_t\right)$. We indicate the confidence interval for $\alpha_t^i$ by the orange interval. Based on the condition $f^i\left(x_{t+1}\right) \leq f^i\left(x_t\right)/2$. $\underline{\alpha}_t^i$, we indicate the adaptive region for $x_{t+1}$ by the green interval.

**Algorithm 2** LB-SGD $(x_0, \eta, T, n)$

```
Input: $M_i, \sigma_i, \hat{\sigma}_i, \hat{b}_i \in \mathbb{R}_+ \forall i \in \{0, \ldots, m\}, R \in \mathbb{R}_+, \eta \in \mathbb{R}_+, n \in \mathbb{N}, T \in \mathbb{N}, \delta \in [0, 1]$;
for $t = 1, \ldots, T$ do
    Set $g_t \leftarrow \mathcal{O}_\eta\left(x_t, n\right)$ by taking a batch of measurements of size $n$ at $x_t$;
    Calculate lower bounds $\underline{\alpha}_t^i := \bar{\alpha}_t^i - \sigma_i(n)\sqrt{\ln \frac{1}{\delta}}, \forall i \in [m]$;
    Calculate upper bounds $\hat{\theta}_t^i = \left|\left\langle G_n^i(x, \xi), \frac{g_t}{\|g_t\|}\right\rangle\right| + \hat{b}_i + \hat{\sigma}_i(n)\sqrt{\log \frac{1}{\delta}}, \forall i \in [m]$;
    Compute $\hat{M}_2\left(x_t\right)$;
    $\gamma_t \leftarrow \min\left\{\min_{i \in [m]}\left\{\frac{\underline{\alpha}_t^i}{2\left|\hat{\theta}_t^i\right| + \sqrt{\underline{\alpha}_t^i M_i}}\right\} \frac{1}{\|g_t\|}, \frac{1}{\hat{M}_2(x_t)}\right\}$
    $x_{t+1} \leftarrow x_t - \gamma_t g_t$
end for
Output: $\{x_1, \ldots, x_T\}$.
```

The smoothness constant $M_i$, of each function $f^i$ for $i \in \{0, \ldots, m\}$ the bounds on the variance of its value measurements $\sigma_i^2$ and gradient measurements $\hat{\sigma_i^2}$ and the upper bound on the bias of its gradient measurements $\hat{b}_i$ the bound on the diameter $R$ of the set $X$, the log barrier parameter $\eta$ the number of measurements per iteration $n$, the number of iterations $T$, and the confidence parameter $\delta$ are the input parameters used in the above.

### *E. Our Contribution*

What we have contributed is applying the algorithm to a simulated Spot robot as well as making it so that it is able to recognize obstacles by applying equations 5 and 6 to equation 1. (We have done more we didn't have time to write about :-))

$$f^0(x) = \sqrt{\vec{Goal}^2 - \vec{Robot}^2} \tag{5}$$

$$f^i(x) = 2 - \sqrt{\vec{Obstacle}^2 - \vec{Robot}^2} \tag{6}$$

## IV. Experiments

Experiments have been made in a simulation, as well as in real life on the physical robot Spot. The RL model is implemented in Python and the simulation has been made in a Webots environment. For testing, Spot has been equipped with two Spot GXP, a Spot CORE, and a TIM551 LiDAR. The github with the used code can be seen here: https://github.com/Lass6230/rob7_750

### A. Current Experiment Status

The RL model is implemented in Python and has currently been tested in a 2D environment with both static and dynamic obstacles. The objective for the experiments was for the simulated robot, to reach a set goal point from a known starting position without any collisions, after reaching the set goal point, a new, randomly generated, goal point will be made for the robot to reach, again without any collisions, this experiment was conducted with both static and dynamic obstacles. This experiment was conducted multiple times, but the results were not noted down, however, video recordings were taken, as we do not intend to use the 2D experiments for the final experiments section.

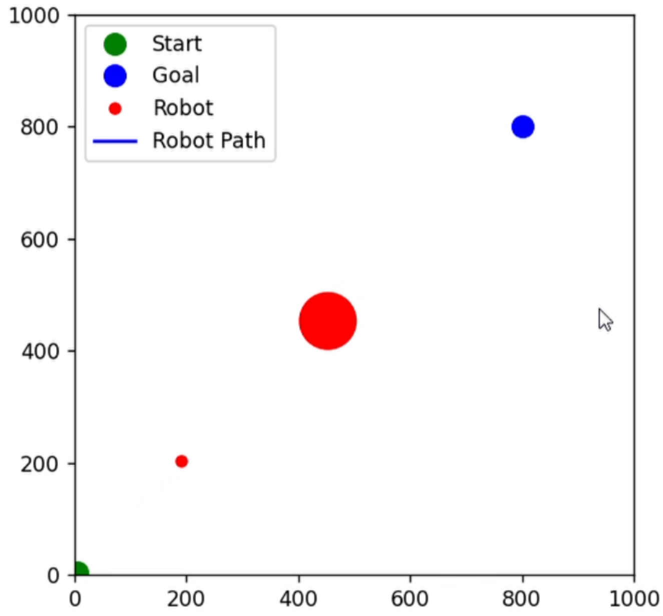

Fig. 2. Snippet from a recording of a 2D experiment with a single moving obstacle. The goal of this experiment was for the robot to continuously reach the newly generated goal points without colliding with the moving obstacle.)

The results can be seen here: https://drive.google.com/drive/folders/1cQoVSfRPREM-W6ug0LhN6RPsbMVIyZ8Z?usp=sharing

### B. Expected Experiments

We expect to make experiments for a simulated version of the model in a Webots environment with a simulated version of Spot with slightly different dynamics. After ensuring we are not going to break Spot, we will be making some real life experiments. Since the Webots simulation has access to robotic operating system 2 (ROS2), we expect similar results from the two experiments, but we are aware there is, at least for now, no randomness in Webots simulation, which will make the difference, between the simulated version and real life version. As for results, we expect our system, to be able to utilize safe learning during deployment, thereby avoiding collisions with both static and dynamic obstacles. We expect the system to minimize time lost, by firstly going the shortest distance to the goal, and secondly not having to stand still often. We have this expectation, because of the logarithmic barrier push. We expect the robot to be still if, we somehow enter the safety barrier of the logarithmic barrier, this can happen in a multitude of ways, varying from velocity to the amount of obstacles.

## V. CONCLUSION

This research has shown advancements in the development of a safe reinforcement learning network, as demonstrated through experiments in both virtual simulations and real-world scenarios with the Boston Dynamics Spot robot. The performance observed in adapting to dynamic environments highlights the successful integration of safe learning algorithms. These findings not only have implications for autonomous robotics but also offer promising applications in sectors requiring adaptable and secure robotic systems. However, there are challenges, like adapting to various environments and making safe learning more efficient, which are crucial for advancing and optimizing these achievements in advanced and secure robotic systems.

## References

[1] L. Liu, D. Dugas, G. Cesari, R. Siegwart and R. Dubé, ”Robot Navigation in Crowded Environments Using Deep Reinforcement Learning,” 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 2020, pp. 5671-5677, doi: 10.1109/IROS45743.2020.9341540.

[2] M. Kollmitz, T. Koller, J. Boedecker, W. Burgard, ”Learning Human-Aware Robot Navigation from Physical Interaction via Inverse Reinforcement Learning”, in press, October 2020

[3] D. Vasquez, B. Okal and K. O. Arras, ”Inverse Reinforcement Learning algorithms and features for robot navigation in crowds: An experimental comparison,” 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 2014, pp. 1341-1346, doi: 10.1109/IROS.2014.6942731.

[4] M. Luber, L. Spinello, J. Silva and K. O. Arras, ”Socially-aware robot navigation: A learning approach,” 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, Vilamoura-Algarve, Portugal, 2012, pp. 902-907, doi: 10.1109/IROS.2012.6385716.

[5] Trautman, P., Ma, J., Murray, R.M., & Krause, A. (2015). Robot navigation in dense human crowds: Statistical models and experimental studies of human–robot cooperation. The International Journal of Robotics Research, 34, 335 - 356.

[6] X. -T. Truong and T. D. Ngo, ”Toward Socially Aware Robot Navigation in Dynamic and Crowded Environments: A Proactive Social Motion Model,” in IEEE Transactions on Automation Science and Engineering, vol. 14, no. 4, pp. 1743-1760, Oct. 2017, doi: 10.1109/TASE.2017.2731371.

[7] D. Helbing and P. Molnár, Social force model for pedestrian dynamics, vol. 51, may 1995

[8] J. van den Berg, Ming Lin and D. Manocha, ”Reciprocal Velocity Obstacles for real-time multi-agent navigation,” 2008 IEEE International Conference on Robotics and Automation, Pasadena, CA, 2008, pp. 1928-1935, doi: 10.1109/ROBOT.2008.4543489.

[9] I. Usmanova, Y. As, M. Kamgarpour, A. Krause, ”Log barriers for safe black-box optimization with application to safe reinforcement learning”, in press, June 2023