Guaranteeing Safety of Learned Perception Modules via Measurement-Robust Control Barrier Functions

[1]**Sarah Dean**   [2]**Andrew J. Taylor**   [2]**Ryan K. Cosner**

[1]**Benjamin Recht**   [2]**Aaron D. Ames**

[1]University of California, Berkeley   [2]California Institute of Technology
{dean_sarah,brecht}@berkeley.edu
{ajtaylor,rkcosner,ames}@caltech.edu

Abstract: Modern nonlinear control theory seeks to develop feedback controllers that endow systems with properties such as safety and stability. The guarantees ensured by these controllers often rely on accurate estimates of the system state for determining control actions. In practice, measurement model uncertainty can lead to error in state estimates that degrades these guarantees. In this paper, we seek to unify techniques from control theory and machine learning to synthesize controllers that achieve safety in the presence of measurement model uncertainty. We define the notion of a Measurement-Robust Control Barrier Function (MR-CBF) as a tool for determining safe control inputs when facing measurement model uncertainty. Furthermore, MR-CBFs are used to inform sampling methodologies for learning-based perception systems and quantify tolerable error in the resulting learned models. We demonstrate the efficacy of MR-CBFs in achieving safety with measurement model uncertainty on a simulated Segway system.

Keywords: safety, measurements, learning, perception

1   Introduction

Ensuring safety is of utmost importance in modern control systems, and as system complexity increases, it is necessary to rigorously encode safety during the controller design process. Examples of safety-critical control applications include autonomous vehicles, aerospace vehicles, and industrial robotics. In practice, these control systems rely on feedback involving imperfect or uncertain measurements models, which can lead to inaccurate state estimation and unsafe behavior if not properly accounted for in controller design. Furthermore, modern control challenges increasingly call for complex measurement systems incorporating data-driven learning methods, such as perception. Thus it is paramount that controllers and learning models be robustly designed to ensure safety in the presence of uncertain measurement models.

Control Barrier Functions (CBFs) [1, 2] have become increasingly popular [3, 4, 5] as a tool for achieving safety in the form of *set invariance* [6]. Furthermore, the integration of CBFs with machine-learning approaches for reducing model uncertainty has shown great promise theoretically [7, 8, 9] and in application [10, 11]. Synthesis of CBFs that encode certain safety properties have been explored from the perspective of backup sets [12] and data-driven methods [13]. In many of these settings it is assumed that controllers synthesized via CBFs have perfect access to measurements of the system state. In practice, these measurements are often corrupted due to uncertainty in measurement models and sensor noise. Safety guarantees in the presence of measurement noise has been addressed from a stochastic perspective [14, 15], and robustness to error in the estimate of the state was considered via a sub-tangentiality condition in [16]. The work in [17, 18] considers robust CBF formulations where worst-case disturbance bounds and perfect knowledge of the actuated dynamics enable robust safety. This existing work has not considered the case when the actuated dynamics are uncertain, nor a data-driven approach where deterministic error in the state estimate can be mitigated through learning.

In this work, we consider safety for a setting in which the system states are not directly observed. Inspired by examples of vision-based control [19, 20, 21], we suppose that state information is observed through a complex transformation, e.g. a camera image, and that an inverse mapping from measurement to state must be estimated. While many impressive demonstrations in the context of robotics rely on carefully calibrated systems in such settings [22], rigorous investigations relating data-driven calibration to learning in the context of control are just beginning to receive attention. This setting has been studied from the perspective of robustness and sample complexity for linear systems, where the main safety concern is stability [23, 24]. Other work has considered methods for ensuring obstacle avoidance under loss of observability for nonlinear systems [25], but do not consider learning directly. To the best of our knowledge, no such analysis exists for nonlinear systems through the perspective of CBFs.

The main contributions of this work are threefold. Firstly, we present the novel definition of Measurement-Robust Control Barrier Functions (MR-CBF) which modify the definition of CBFs to enable robustness to error in measurement models. Secondly, we show how MR-CBFs can be incorporated into convex optimization-based controllers that can be efficiently solved online. Lastly, we outline how MR-CBFs can be used to guide data sampling methodology in order to ensure that the resultant model yields tolerable error.

This remainder of this paper is organized as follows. In Section 2 we provide a review of Control Barrier Functions and their use in synthesizing safe controllers. Section 3 explores the impact of measurement model uncertainty on safety, and defines the MR-CBF as a tool for synthesizing controllers that guarantee safety in the absence of perfect state observation. In Section 4 we explain how data-driven learning methods can be used to reduce measurement model uncertainty and yield valid MR-CBFs. Finally, Section 5 presents simulation results which demonstrate the efficacy of MR-CBFs in enforcing safety in the presence of measurement model uncertainty for a Segway system. Additional details including proofs are provided in supplementary appendices.

2   Background

In this section we provide a review of safety and Control Barrier Functions (CBFs). These definitions will be used in quantifying how measurement uncertainty impacts safety guarantees.

Consider the nonlinear control affine system given by:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})\mathbf{u}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{u} \in \mathbb{R}^m$, and $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ and $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^{n \times m}$ are locally Lipschitz continuous on $\mathbb{R}^n$. Given a locally Lipschitz continuous state-feedback controller $\mathbf{k} : \mathbb{R}^n \to \mathbb{R}^m$, the closed-loop system dynamics are:

$$\dot{\mathbf{x}} = \mathbf{f}_{\mathrm{cl}}(\mathbf{x}) \triangleq \mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})\mathbf{k}(\mathbf{x}). \tag{2}$$

The assumption on local Lipschitz continuity of $\mathbf{f}$, $\mathbf{g}$, and $\mathbf{k}$ implies that $\mathbf{f}_{\mathrm{cl}}$ is locally Lipschitz continuous. Thus for any initial condition $\mathbf{x}_0 \triangleq \mathbf{x}(0) \in \mathbb{R}^n$ there exists a time interval $I(\mathbf{x}_0) = [0, t_{\max})$ such that $\mathbf{x}(t)$ is the unique solution to (2) on $I(\mathbf{x}_0)$ [26].

The notion of safety that we consider is formalized by specifying a *safe set* in the state space that the system must remain in to be considered safe. In particular, consider a set $\mathcal{C} \subset \mathbb{R}^n$ defined as the 0-superlevel set of a continuously differentiable function $h : \mathbb{R}^n \to \mathbb{R}$, yielding:

$$\mathcal{C} \triangleq \{\mathbf{x} \in \mathbb{R}^n : h(\mathbf{x}) \geq 0\}. \tag{3}$$

We refer to $\mathcal{C}$ as the *safe set*, and note $\partial \mathcal{C} \triangleq \{\mathbf{x} \in \mathbb{R}^n : h(\mathbf{x}) = 0\}$ and $\mathrm{Int}(\mathcal{C}) \triangleq \{\mathbf{x} \in \mathbb{R}^n : h(\mathbf{x}) > 0\}$. We assume that $\mathcal{C}$ is nonempty and has no isolated points, that is, $\mathrm{Int}(\mathcal{C}) \neq \emptyset$ and $\overline{\mathrm{Int}(\mathcal{C})} = \mathcal{C}$. This construction motivates the following definitions of forward invariance and safety:

Definition 1 (*Forward Invariant & Safety*). A set $\mathcal{C} \subset \mathbb{R}^n$ is *forward invariant* if for every $\mathbf{x}_0 \in \mathcal{C}$, the solution $\mathbf{x}(t)$ to (2) satisfies $\mathbf{x}(t) \in \mathcal{C}$ for all $t \in I(\mathbf{x}_0)$. The system (2) is *safe* with respect to the set $\mathcal{C}$ if the set $\mathcal{C}$ is forward invariant.

Before defining Control Barrier Functions as a tool for synthesizing safe controllers, we note that a continuous function $\alpha : (-b, a) \to \mathbb{R}$, with $a, b > 0$, is said to belong to *extended class* $\mathcal{K}$ ($\alpha \in \mathcal{K}_e$) if $\alpha(0) = 0$ and α is strictly monotonically increasing. If $a, b = \infty$, $\lim_{r \to \infty} \alpha(r) = \infty$, and

$\lim_{r \to -\infty} \alpha(r) = -\infty$, then α is said to belong to *extended class* $\mathcal{K}_\infty$ ($\alpha \in \mathcal{K}_{\infty,e}$). Furthermore, we note $c \in \mathbb{R}$ is referred to as a *regular value* of a continuously differentiable function $h : \mathbb{R}^n \to \mathbb{R}$ if $h(\mathbf{x}) = c \implies \frac{\partial h}{\partial \mathbf{x}}(\mathbf{x}) \neq \mathbf{0}$. These enable the definition of Control Barrier Functions as follows:

Definition 2 (*Control Barrier Function (CBF)*, [2]). Let $\mathcal{C} \subset \mathbb{R}^n$ be the 0-superlevel set of a continuously differentiable function $h : \mathbb{R}^n \to \mathbb{R}$ with 0 a regular value. The function h is a *Control Barrier Function* (CBF) for (1) on $\mathcal{C}$ if there exists $\alpha \in \mathcal{K}_{\infty,e}$ such that for all $\mathbf{x} \in \mathcal{C}$:

$$\sup_{\mathbf{u} \in \mathbb{R}^m} \dot{h}(\mathbf{x}, \mathbf{u}) \triangleq \underbrace{\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x})\mathbf{f}(\mathbf{x})}_{L_\mathbf{f} h(\mathbf{x})} + \underbrace{\frac{\partial h}{\partial \mathbf{x}}(\mathbf{x})\mathbf{g}(\mathbf{x})}_{L_\mathbf{g} h(\mathbf{x})} \mathbf{u} > -\alpha(h(\mathbf{x})). \tag{4}$$

This condition can be equivalently stated as:

$$\|L_\mathbf{g} h(\mathbf{x})\|_2 = 0 \implies L_\mathbf{f} h(\mathbf{x}) > -\alpha(h(\mathbf{x})). \tag{5}$$

The inequality in the definition of CBFs is strict to ensure that the controllers synthesized via CBFs are locally Lipschitz continuous [17]. Given a CBF h for (1) and a corresponding $\alpha \in \mathcal{K}_{\infty,e}$, we can consider the pointwise set of all control values that satisfy (4):

$$K_{\mathrm{cbf}}(\mathbf{x}) \triangleq \{\mathbf{u} \in \mathbb{R}^m \mid L_\mathbf{f} h(\mathbf{x}) + L_\mathbf{g} h(\mathbf{x})\mathbf{u} \geq -\alpha(h(\mathbf{x}))\}. \tag{6}$$

A main result in [27, 1] relates controllers taking values in $K_{\mathrm{cbf}}(\mathbf{x})$ to the safety of (2) on $\mathcal{C}$:

Theorem 1. *Given a set $\mathcal{C} \subset \mathbb{R}^n$ defined as the 0-superlevel set of a continuously differentiable function $h : \mathbb{R}^n \to \mathbb{R}$, if h is a CBF for (1) on $\mathcal{C}$, then any locally Lipschitz continuous controller $\mathbf{k} : \mathbb{R}^n \to \mathbb{R}^m$, such that $\mathbf{k}(\mathbf{x}) \in K_{cbf}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{C}$, renders the system (2) safe w.r.t. $\mathcal{C}$.*

This result motivates the construction of a pointwise optimal controller seeking to minimize a cost associated with the choice of input. To this end, we consider the *safety-critical control* formulation [28] that seeks to filter a hand-designed but potentially unsafe locally Lipschitz continuous controller, $\mathbf{k}_d : \mathbb{R}^n \to \mathbb{R}^m$, to find the nearest safe action:

$$\mathbf{k}(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{u} \in \mathbb{R}^m} \quad \frac{1}{2}\|\mathbf{u} - \mathbf{k}_d(\mathbf{x})\|_2^2 \tag{CBF-QP}$$
$$\text{s.t.} \quad L_\mathbf{f} h(\mathbf{x}) + L_\mathbf{g} h(\mathbf{x})\mathbf{u} \geq -\alpha(h(\mathbf{x})).$$

The validity of h as a CBF ensures the feasibility of this optimization problem, and the resulting controller is locally Lipschitz continuous [17].

3 Measurement-Robust Control Barrier Functions

In this section we explore the impact of measurement model uncertainty on safety guarantees, and propose the notion of a modified Control Barrier Function that is robust to such errors.

In many practical applications, the state $\mathbf{x}$ is not directly available to the controller, but rather a state-dependent sensor measurement:

$$\mathbf{y} = \mathbf{p}(\mathbf{x}), \tag{7}$$

where $\mathbf{p} : \mathbb{R}^n \to \mathbb{R}^k$ is assumed to be locally Lipschitz continuous. We assume the relationship between the measurement and the true state is deterministic, and note the application of CBFs in the context of stochastic differential equations has been considered in [14]. Future work will consider the unification of the results in this paper with the stochastic setting. We further assume there exists a locally Lipschitz continuous function $\mathbf{q} : \mathbb{R}^k \to \mathbb{R}^n$ such that for all $\mathbf{x} \in \mathbb{R}^n$, we have $\mathbf{q}(\mathbf{p}(\mathbf{x})) = \mathbf{x}$. This assumption implies that the state can be uniquely determined from any given measurement. This bijective relationship would allow the measurements to be redefined as the state of the system if the function $\mathbf{p}$ was known, but that is often not the case in many modern control applications (such as when using vision).

While the function $\mathbf{p}$ is often determined by the physical attributes of a system, a locally Lipschitz continuous estimate of the function $\mathbf{q}$, given by $\widehat{\mathbf{q}} : \mathbb{R}^k \to \mathbb{R}^n$, is constructed to determine an estimate of the state, $\widehat{\mathbf{x}}$. For notational simplicity we define the measurement-estimate function $\widehat{\mathbf{v}} : \mathbb{R}^n \to \mathbb{R}^k \times \mathbb{R}^n$ such that $\widehat{\mathbf{v}}(\mathbf{x}) = (\mathbf{p}(\mathbf{x}), \widehat{\mathbf{q}}(\mathbf{p}(\mathbf{x})))$. We also define the set $\mathbf{p}(\mathcal{C}) \subset \mathbb{R}^k$ as

the image of the safe set under the measurement function, the set $\widehat{\mathbf{q}}(\mathbf{p}(\mathcal{C})) \subset \mathbb{R}^n$ as the image of the safe set under the state estimate function, and $\widehat{\mathbf{v}}(\mathcal{C})$ as the image of the safe set under the measurement-estimate function.

The function $\widehat{\mathbf{q}}$ is constructed either via system and measurement models, or from data using learning methods, and thus its accuracy in estimating $\mathbf{q}$ degrades with imperfections in sensor fabrication and integration, or imperfections in learning models and training data. Thus we assume that our state estimate is related to the true state as follows:

$$\widehat{\mathbf{x}} \triangleq \widehat{\mathbf{q}}(\mathbf{y}) = \mathbf{x} + \mathbf{e}(\mathbf{x}), \tag{8}$$

for an unknown function $\mathbf{e} : \mathbb{R}^n \to \mathbb{R}^n$ that is defined implicitly via $\widehat{\mathbf{q}}$. In practice, the function $\mathbf{e}$ can often be characterized via upper bounds on model uncertainty or via data-driven arguments for learning models (discussed further in Section 4). In particular, we assume that while $\mathbf{e}(\mathbf{x})$ is not known for a particular value of $\mathbf{x}$, it is known that $\mathbf{e}(\mathbf{x}) \in \mathcal{E}(\mathbf{y})$ for a measurement dependent, compact pointwise set $\mathcal{E}(\mathbf{y})$. This leads to the definition of the following two pointwise sets:

$$\widehat{\mathcal{X}}(\mathbf{x}) \triangleq \{\widehat{\mathbf{x}} \in \mathbb{R}^n \mid \exists\, \mathbf{e} \in \mathcal{E}(\mathbf{p}(\mathbf{x})) \text{ s.t. } \widehat{\mathbf{x}} = \mathbf{x} + \mathbf{e}\}, \tag{9}$$

$$\mathcal{X}(\mathbf{y}) \triangleq \{\mathbf{x} \in \mathbb{R}^n \mid \exists\, \mathbf{e} \in \mathcal{E}(\mathbf{y}) \text{ s.t. } \widehat{\mathbf{x}} = \mathbf{x} + \mathbf{e}\}. \tag{10}$$

The first of these two pointwise sets can be interpreted as all possible state estimates corresponding to a particular state, restricted by the possible error dictated by $\mathcal{E}(\mathbf{p}(\mathbf{x}))$. While it is not directly computable without knowledge of $\mathbf{p}$, this set will play an important conceptual role in Section 4 to argue about how data can be used to determine error bounds. The second pointwise set consists of all potential states that may yield a measurement-state estimate pair.

Since a controller enforcing the CBF condition (4) requires exact knowledge of the state $\mathbf{x}$, we propose an alternative condition which depends on only the set $\mathcal{X}(\mathbf{y})$ and the state estimate $\widehat{\mathbf{x}}$. To ensure safety with a CBF, it is sufficient for the following condition to hold for all $\mathbf{y} \in \mathbf{p}(\mathcal{C})$:

$$\sup_{\mathbf{u} \in \mathbb{R}^m} \inf_{\mathbf{x} \in \mathcal{X}(\mathbf{y})} \frac{\partial h}{\partial \mathbf{x}}(\mathbf{x})(\mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})\mathbf{u}) + \alpha(h(\mathbf{x})) \geq 0 . \tag{11}$$

This condition implies that there exists a control input that renders the system safe for all possible states corresponding to a given state estimate. Verifying that this condition holds can be difficult for an arbitrary CBF, and it is not easily (or possibly) enforced in a convex-optimization based controller. To resolve these problems, we introduce the following definition:

Definition 3 (*Measurement-Robust Control Barrier Function (MR-CBF)*)**.** Let $\mathcal{C} \subset \mathbb{R}^n$ be the 0-superlevel set of a continuously differentiable function $h : \mathbb{R}^n \to \mathbb{R}$ with 0 a regular value. The function h is a *Measurement-Robust Control Barrier Function* (MR-CBF) for (1) on $\mathcal{C}$ with *parameter function* $(a, b) : \mathbb{R}^k \to \mathbb{R}_+^2$ if there exists $\alpha \in \mathcal{K}_{\infty,e}$ such that for all $(\mathbf{y}, \widehat{\mathbf{x}}) \in \widehat{\mathbf{v}}(\mathcal{C})$:

$$\sup_{\mathbf{u} \in \mathbb{R}^m} L_{\mathbf{f}} h(\widehat{\mathbf{x}}) + L_{\mathbf{g}} h(\widehat{\mathbf{x}})\mathbf{u} - (a(\mathbf{y}) + b(\mathbf{y})\|\mathbf{u}\|_2) > -\alpha(h(\widehat{\mathbf{x}})). \tag{12}$$

Verifying this condition over $\widehat{\mathbf{v}}(\mathcal{C})$ may not be possible, but as will be seen in Section 4, the set $\widehat{\mathcal{X}}(\mathbf{x})$ can be used to provide sufficient conditions under which (12) is met. The definition of a MR-CBF introduces the non-positive term $-(a(\mathbf{y}) + b(\mathbf{y})\|\mathbf{u}\|_2)$ to the control barrier condition, requiring that a stronger degree of safety be enforced compared to the typical control barrier condition. Furthermore, the norm of the input appears in this term, indicating that for large values of b large inputs can lead to unsafe behavior. This condition is equivalently stated as:

$$\|L_{\mathbf{g}} h(\widehat{\mathbf{x}})\|_2 \leq b(\mathbf{y}) \implies L_{\mathbf{f}} h(\widehat{\mathbf{x}}) > -\alpha(h(\widehat{\mathbf{x}})) + a(\mathbf{y}). \tag{13}$$

In contrast to the implication in (5), the size of set for which the antecedent in (13) is met may be larger, requiring the natural dynamics to be safe ($L_{\mathbf{f}} h(\widehat{\mathbf{x}}) > -\alpha(h(\widehat{\mathbf{x}})) + a(\mathbf{y})$) in a larger region. Given a MR-CBF h for (1) on $\mathcal{C}$ with parameter function (a, b) and a corresponding $\alpha \in \mathcal{K}_{\infty,e}$, we can consider the pointwise set of all control values that satisfy (12):

$$K_{\text{mr-cbf}}(\mathbf{y}, \widehat{\mathbf{x}}) \triangleq \{\mathbf{u} \in \mathbb{R}^m \mid L_{\mathbf{f}} h(\widehat{\mathbf{x}}) + L_{\mathbf{g}} h(\widehat{\mathbf{x}})\mathbf{u} - (a(\mathbf{y}) + b(\mathbf{y})\|\mathbf{u}\|_2) \geq -\alpha(h(\widehat{\mathbf{x}}))\}, \tag{14}$$

for $(\mathbf{y}, \widehat{\mathbf{x}}) \in \widehat{\mathbf{v}}(\mathcal{C})$. Given this construction, we have the following result relating the existence of a MR-CBF to safety under the presence of measurement model uncertainty:

4

Theorem 2. *Let a set $\mathcal{C} \subset \mathbb{R}^n$ be defined as the 0-superlevel set of a continuously differentiable function $h : \mathbb{R}^n \to \mathbb{R}$. Assume the functions $L_{\mathbf{f}}h : \mathbb{R}^n \to \mathbb{R}$, $L_{\mathbf{g}}h : \mathbb{R}^n \to \mathbb{R}^m$, and $\alpha \circ h : \mathbb{R}^n \to \mathbb{R}$ are Lipschitz continuous on $\mathcal{C}$ with Lipschitz coefficients $\mathfrak{L}_{L_{\mathbf{f}}h}$, $\mathfrak{L}_{L_{\mathbf{g}}h}$, and $\mathfrak{L}_{\alpha \circ h}$, respectively. Further assume there exists a locally Lipschitz function $\epsilon : \mathbb{R}^k \to \mathbb{R}_+$, such that $\max_{\mathbf{e} \in \mathcal{E}(\mathbf{y})} \|\mathbf{e}\|_2 \leq \epsilon(\mathbf{y})$ for all $\mathbf{y} \in \mathbf{p}(\mathcal{C})$. If h is a MR-CBF for (1) on $\mathcal{C}$ with parameter function $(\epsilon(\mathbf{y})(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h}), \epsilon(\mathbf{y})\mathfrak{L}_{L_{\mathbf{g}}h})$, then any locally Lipschitz continuous controller $\mathbf{k} : \mathbb{R}^k \times \mathbb{R}^n \to \mathbb{R}^m$, such that $\mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}}) \in K_{mr\text{-}cbf}(\mathbf{y}, \widehat{\mathbf{x}})$ for all $(\mathbf{y}, \widehat{\mathbf{x}}) \in \widehat{\mathbf{v}}(\mathcal{C})$, renders the system (2) safe w.r.t. $\mathcal{C}$.*

A proof of this theorem can be found in Appendix A. To more clearly see how the upper bound on the estimate error, $\epsilon(\mathbf{y})$, manifests in the MR-CBF condition, we note the particular condition that must be satisfied for this theorem is given by:

$$\sup_{\mathbf{u} \in \mathbb{R}^m} L_{\mathbf{f}}h(\widehat{\mathbf{x}}) + L_{\mathbf{g}}h(\widehat{\mathbf{x}})\mathbf{u} - \epsilon(\mathbf{y})(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h} + \mathfrak{L}_{L_{\mathbf{g}}h}\|\mathbf{u}\|_2) > -\alpha(h(\widehat{\mathbf{x}})). \tag{15}$$

Thus as $\epsilon(\mathbf{y})$ becomes smaller, the level of robustness required by an MR-CBF approaches that of a regular CBF for the same set $\mathcal{C}$, and recovers the original control barrier condition with no estimate error. Furthermore, smaller values of $\epsilon(\mathbf{y})$ can be interpreted as leading to an enlarging of the region over which the condition (13) holds.

One advantage of this approach for resolving the impact of measurement model uncertainty on safety is that the constraint in (14) remains convex. This constraint can then be directly integrated into an optimization based controller as follows:

$$\mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}}) = \underset{\mathbf{u} \in \mathbb{R}^m}{\mathrm{argmin}} \quad \frac{1}{2}\|\mathbf{u} - \mathbf{k}_d(\widehat{\mathbf{x}})\|_2^2 \tag{MR-OP}$$
$$\text{s.t.} \quad L_{\mathbf{f}}h(\widehat{\mathbf{x}}) - (\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})\epsilon(\mathbf{y}) + L_{\mathbf{g}}h(\widehat{\mathbf{x}})\mathbf{u} - \mathfrak{L}_{L_{\mathbf{g}}h}\epsilon(\mathbf{y})\|\mathbf{u}\|_2 \geq -\alpha(h(\widehat{\mathbf{x}})).$$

This problem is in fact a second-order cone program (SOCP), with an explicit conversion to standard form provided in Appendix B. As this constraint is non-smooth, existing methods for computing closed-form solutions via Lagrangian duality and assessing Lipschitz continuity [17] are not applicable. Future work will consider methods from variational analysis to study the Lipschitz continuity of solutions to this problem [29]. In practice, a slack variable, δ, is often added to ensure constraint feasibility. This relaxation is penalized in the cost with a large coefficient $p \in \mathbb{R}_{++}$:

$$\mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}}) = \underset{(\mathbf{u}, \delta) \in \mathbb{R}^m \times \mathbb{R}}{\mathrm{argmin}} \quad \frac{1}{2}\|\mathbf{u} - \mathbf{k}_d(\widehat{\mathbf{x}})\|_2^2 + p\delta^2 \tag{R-MR-OP}$$
$$\text{s.t.} \quad L_{\mathbf{f}}h(\widehat{\mathbf{x}}) - (\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})\epsilon(\mathbf{y}) + L_{\mathbf{g}}h(\widehat{\mathbf{x}})\mathbf{u} - \mathfrak{L}_{L_{\mathbf{g}}h}\epsilon(\mathbf{y})\|\mathbf{u}\|_2 \geq -\alpha(h(\widehat{\mathbf{x}})) - \delta.$$

While this relaxed controller does not necessarily enforce the desired safety constraint, if δ remains small the impact on safety can be understood through the notion of projection-to-state safety [7]. Furthermore, this relaxation ensures that the resulting controller is locally Lipschitz continuous as made explicit in Appendix C via the methods in [30].

4 Learning for Measurement Model Uncertainty Reduction

In this section we explore how data-driven learning methods can be used to reduce measurement model uncertainty and yield valid MR-CBFs.

The previous section provides a method for guaranteeing safety in the absence of perfect state observation which relies on specifying a valid MR-CBF h. The condition in (12) restricts admissible h and $\mathcal{C}$ with a stronger condition than the typical CBF condition. A CBF h which is valid when states are perfectly observed may no longer be valid in the presence of measurement model uncertainty. In this section, we reconsider the requirement that h be a valid MR-CBF as a specification on measurement or calibration errors. Clearly, as measurement model uncertainty becomes arbitrarily close to 0, any valid CBF will be a valid MR-CBF. We now make this intuition precise.

By applying the logic of the implication (13), we see that (15) is true as long as for all $(\mathbf{y}, \widehat{\mathbf{x}}) \in \widehat{\mathbf{v}}(\mathcal{C})$,

$$\epsilon(\mathbf{y}) < \max\left\{ \frac{\|L_{\mathbf{g}}h(\widehat{\mathbf{x}})\|_2}{\mathfrak{L}_{L_{\mathbf{g}}h}}, \frac{L_{\mathbf{f}}h(\widehat{\mathbf{x}}) + \alpha(h(\widehat{\mathbf{x}}))}{\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h}} \right\}. \tag{16}$$

This expression gives the maximum admissible error under which the MR-CBF condition will hold for a given set of dynamics and function h. It is not easy to reason about this quantity directly, since the measurement-estimate set $\widehat{\mathbf{v}}(\mathcal{C})$ is difficult to characterize without knowledge of $\mathbf{p}$. Instead, we reformulate this expression to depend on the associated underlying state $\mathbf{x}$. We do so by interpreting ϵ as a function of the state via $\epsilon(\mathbf{y}) = \epsilon(\mathbf{p}(\mathbf{x})) \triangleq \varepsilon(\mathbf{x})$ and appealing conceptually to the set $\widehat{\mathcal{X}}(\mathbf{x})$ defined in (9) to consider all possible observations at a given state.

Theorem 3. *Let $\mathcal{C}$, h, $\mathfrak{L}_{L_{\mathbf{f}}h}$, $\mathfrak{L}_{L_{\mathbf{g}}h}$, $\mathfrak{L}_{\alpha \circ h}$, and ϵ be defined as in Theorem 2. Then h is a MR-CBF for (1) on $\mathcal{C}$ with parameter function $(\epsilon(\mathbf{y})(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h}), \epsilon(\mathbf{y})\mathfrak{L}_{L_{\mathbf{g}}h})$ if for all $\mathbf{x} \in \mathcal{C}$:*

$$\varepsilon(\mathbf{x}) < \max \left\{ \frac{\|L_{\mathbf{g}}h(\mathbf{x})\|_2}{2\mathfrak{L}_{L_{\mathbf{g}}h}}, \frac{L_{\mathbf{f}}h(\mathbf{x}) + \alpha(h(\mathbf{x}))}{2(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})} \right\} \triangleq \bar{\varepsilon}(\mathbf{x}) . \tag{17}$$

Even though the true states will not be observed during operation, this result gives a sufficient condition on error with respect to states in the safe set $\mathcal{C}$. Comparing this expression with the bound in (16), there is an additional factor of 2. This corresponds to doubling the radius of the uncertainty set, which occurs because it is necessary to consider the entire set $\widehat{\mathcal{X}}(\mathbf{x})$ for each $\mathbf{x}$.

We now draw an explicit connection between measurement model uncertainty and learning from data. Recall that the error $\mathbf{e}(\mathbf{x})$ is implicitly defined by the estimated $\widehat{\mathbf{q}}$, so it can be viewed as arising from imperfectly approximating the inverse map $\mathbf{q}$:

$$\mathbf{e}(\mathbf{x}) = \mathbf{x} - \widehat{\mathbf{q}}(\mathbf{p}(\mathbf{x})) = \mathbf{q}(\mathbf{y}) - \widehat{\mathbf{q}}(\mathbf{y}). \tag{18}$$

Therefore, we treat $\varepsilon(\mathbf{x})$ as a pointwise error bound on the learned map $\widehat{\mathbf{q}}$ over the space of measurements. In a supervised learning setting, this map arises from training data, which we denote as $\mathcal{S} = \{(\mathbf{y}_i, \bar{\mathbf{x}}_i)\}_{i=1}^N$. The recorded system outputs $\mathbf{y}_i$ play the role of the independent variables, while the corresponding recorded states $\bar{\mathbf{x}}_i$ play the role of dependent variables. We denote the recorded states as $\bar{\mathbf{x}}_i$ to allow for the possibility that they are imperfect measurements of the state. For example, the recorded values could be noisy according to some distribution $\mathcal{D}$:

$$\bar{\mathbf{x}}_i = \mathbf{x}_i + \mathbf{w}_i, \quad \mathbf{w}_i \overset{\text{iid}}{\sim} \mathcal{D}, \tag{19}$$

for $i = 1, \ldots, N$. This scenario corresponds to using a simple noisy sensor to calibrate a more complex sensor by learning the inverse map $\widehat{\mathbf{q}}$. Denote the true states as $\mathcal{X}_{\mathcal{S}} = \{\mathbf{x}_i\}_{i=1}^N$.

Our work is motivated by instances of control-from-pixels, where the system outputs $\mathbf{y}$ are camera images [19, 20, 21]. Therefore, we consider maps $\widehat{\mathbf{q}}$ which are non-parametric and thus highly data dependent, like neural networks.[1] Though necessary for guaranteeing safety, providing pointwise bounds on errors as in (18) is often not the focus of machine learning analyses, which favor a mean-error perspective. We point to this issue to highlight an area of future work, and rely here on a simplified model. The following definition models training data-dependent errors:

Definition 4. A *non-parametric error bound* with parameters $(L, \sigma) \in \mathbb{R}_+^2$ and hyperparameter $\gamma \in \mathbb{R}_+$ has the form:

$$\|\mathbf{e}(\mathbf{x})\|_2 \leq L\gamma + \frac{\sigma}{\sqrt{\#\{\mathbf{x}_i \in \mathcal{X}_{\mathcal{S}} \mid \|\mathbf{x} - \mathbf{x}_i\| \leq \gamma\}}} . \tag{20}$$

In Appendix D, we relate this definition to a high probability bound for a particular class of non-parametric regression models analysed in [24] for the noisy sensor setting described by (19). Combining this definition with Theorem 3 yields a condition on the way that training data is collected.

Corollary 1. *Consider the setting of Theorem 3 and suppose that $\widehat{\mathbf{q}}$ is learned with a method that satisfies the (L, σ) non-parametric error bound with hyperparameter γ. Further suppose that the training data is collected from states $\mathcal{X}_{\mathcal{S}}$ such that for all $\mathbf{x} \in \mathcal{C}$,*

$$\#\{\mathbf{x}_i \in \mathcal{X}_{\mathcal{S}} \mid \|\mathbf{x} - \mathbf{x}_i\|_2 \leq \gamma\} > \frac{4\sigma^2}{\left(\max\left\{ \frac{1}{\mathfrak{L}_{L_{\mathbf{g}}h}} \|L_{\mathbf{g}}h(\mathbf{x})\|_2, \frac{L_{\mathbf{f}}h(\mathbf{x}) + \alpha(h(\mathbf{x}))}{(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})} \right\} - 2L\gamma \right)^2} . \tag{21}$$

---

[1]While we suppose that $\widehat{\mathbf{q}}$ is learned entirely from data, our results are equally applicable to learning the *residual errors* of an existing perception component (see, e.g. surveyed methods in [31]).
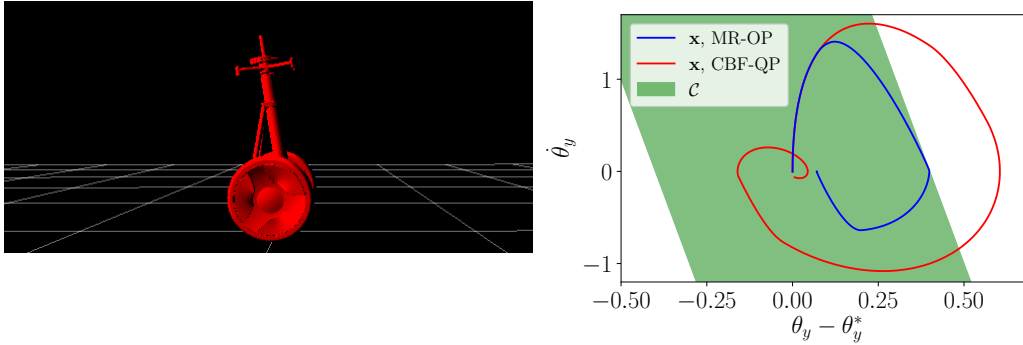
Figure 1. (Left) The Segway model used in simulation from the perspective of the fixed virtual camera used to estimate its state. **(Right)** Simulation results for worst-case measurement model uncertainty of $\epsilon = 0.2$ subtracted from the true pitch angle θ_y when measured. A state trajectory generated using the the Standard CBF Filter (red) and the MR-OP Filter (blue) are shown as projections onto their pitch angle and pitch rate components. The safe set is plotted in green. Given the same initial condition, the MR-OP filter ensured safety of the trajectory whereas the Standard CBF Filter did not.

Then h is a valid MR-CBF satisfying the conditions of Theorem 2.

The statement above characterizes a density condition for data used to calibrate complex sensors. To ensure the feasibility of the MR-CBF procedure, training data should be sampled non-uniformly according to the right hand side of (21). Future work will seek to connect this sampling scheme with dynamically feasible trajectory planning targeted towards data density.

5 Simulation Results

In this section we present simulation results using MR-CBFs and data-driven learning models on a simulated robotic Segway platform.

The Segway can be seen in Figure 1, and is modeled with system dynamics derived using the unconstrained Euler-Lagrange Equations. The system is constrained to planar motion by providing identical input torques about both wheels. The resulting degrees of freedom are the Segway's horizontal position r, horizontal velocity $\dot{r}$, pitch angle θ_y, and pitch rate $\dot{\theta}_y$. The nominal controller $\mathbf{k}_d$ is a simple proportional-derivative (PD) controller as in [7]. The simulation was written in a Robot Operating System (ROS) based environment using a mixture of C++ and Python to mimic the code structure of the existing hardware platform [28].[2]

The safe set for the simulation was defined as $\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^7 : h_1(\mathbf{x}) \geq 0, \ h_2(\mathbf{x}) \geq 0\}$ with:

$$h_1 = -\dot{\theta}_y + \alpha_e(c - \theta_y + \theta_y^\star) \qquad h_2 = \dot{\theta}_y + \alpha_e(c + \theta_y - \theta_y^\star) \qquad (22)$$

where $c \in \mathbb{R}_{++}$, $\alpha_e \in \mathbb{R}_{++}$, and $\theta_y^\star$ is the pitch angle at equilibrium. The MR-OP Filter constraint in (MR-OP) was applied simultaneously to both safety functions (22) and was then implemented using the ECOS Second Order Cone Problem solver [32]. The Lipschitz constants in this constraint were estimated by sampling $\mathfrak{L}_{L_{\mathbf{f}}h}$, $\mathfrak{L}_{L_{\mathbf{g}}h}$ and $\alpha \circ h$ on a set of gridded values around the system's equilibrium point by taking the maximum of the slopes between any two adjacent grid points. As a baseline comparison, a CBF-QP Filter (CBF-QP) was implemented and applied using both safety functions (22). We considered the two following testing scenarios:

Worst-Case Synthetic Measurement Model Uncertainty: In this testing scenario we assumed that direct measurements of the pitch angle θ_y were offset by a constant factor of $\epsilon > 0$, such that $\widehat{\theta}_y = \theta_y - \epsilon$. Implementing the MR-OP Filter for $\epsilon > 0$ ensures safety for this worst-case error of up to ϵ. The result of this type of worst-case measurement model uncertainty in the Segway system with a standard CBF-QP Filter and an MR-OP Filter can be seen in Figure 1.

---

[2]The full simulation code can be found at `https://github.com/rkcosner/cyberpod_sim_ros.git`
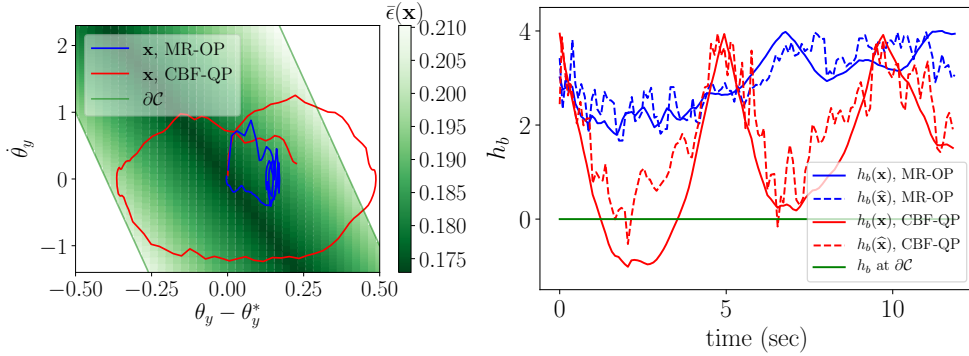
Figure 2. Simulation results demonstrating the ability of the MR-OP Filter to mitigate the impact of imperfect learned perception models on safety. **(Left)** The state trajectory generated using the Standard CBF Filter (red) and MR-OP Filter (blue) are shown as projections onto their pitch angle and pitch rate components. Given the same initial condition, the MR-OP Filter generated a safe trajectory whereas the Standard CBF-QP Filter did not. The state estimates for each trajectory had a maximum error of 0.183 and 0.201 respectively. **(Right)** The Boolean composition, $h_b = \min\{h_{e1}, h_{e2}\} = h_{e1} \wedge h_{e2}$ as defined in [34], is plotted for the CBF-QP Filter and MR-OP Filter trajectories for the true and estimated states, $\mathbf{x}$ (solid line) and $\widehat{\mathbf{x}}$ (dotted line). The safety violation of the Standard CBF-QP Filter can be seen where $h_b(\mathbf{x})$ crosses 0.

Data-Driven Sensor Calibration: In this scenario a more realistic form of measurement model uncertainty is introduced through the use of a learned model to estimate the position r and pitch angle θ_y from camera images. In simulation, a virtual camera and lighting source were implemented to provide a 15 Hz video feed with a fixed perspective, an example of which can be seen in Figure 1. The labels for this supervised-learning problem were noisy measurements of the position and pitch angle generated by the system's inertial measurement unit, corrupted by Gaussian noise with standard deviation 0.1. We use `sklearn`'s Kernel Ridge Regression with radial basis functions [33] trained using a set of 800 labelled images associated with a gridded range of position and pitch angle values to ensure dense coverage. The hyperparameter values $\alpha = 0, \gamma = 5.4 \times 10^{-8}$ were selected to minimize the average error on an 80% - 20% random train-test split.

The result of the learning-induced errors in the Segway system with a standard CBF Filter and an MR-OP Filter with $\epsilon = 0.2$ can be seen in Figure 2. In the left panel, the safe set $\mathcal{C}$ is shaded according to the upper bound on error $\bar{\varepsilon}(\mathbf{x})$ under which feasibility is guaranteed. As the errors in the learned map do not exceed this value empirically, the set $\mathcal{C}$ is rendered invariant. The expression for $\bar{\varepsilon}(\mathbf{x})$ in this experimental scenario is presented in a result analogous to Theorem 3 in Appendix E, along with an empirical validation of learned model errors.

6   Conclusion

In conclusion, we presented the notion of a Measurement-Robust Control Barrier Function as a tool for ensuring safety in the presence of error in the model relating measurements to state estimates. The resulting safety condition required by a MR-CBF can be directly incorporated into an optimization based controller as a second order cone constraint, preserving the convexity of typical CBF based controllers. We explore how worst case error in learned measurement models can be quantified in terms of data density, and demonstrate how learning-based perception models can be trained on noisy state data and incorporated with MR-CBFs to achieve safe, perception based control.

Future work will seek to explore the concepts in this paper from both a practical and a theoretical standpoint. From a practical perspective, we will demonstrate the feasibility of MR-CBFs and perception-based control on unstable systems such as the Segway in real world experiments. From a theoretical perspective, we will explore the impact of noise, targeted data acquisition through trajectory generation, Lipschitz properties of the resulting optimization based controllers, and strategies for non-invertible observation models such as dynamic estimators (Kalman filters).

References

[1] A. Ames, J. Grizzle, and P. Tabuada. Control barrier function based quadratic programs with application to adaptive cruise control. In *Conference on Decision & Control (CDC)*, pages 6271–6278. IEEE, 2014.

[2] A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada. Control barrier function based quadratic programs for safety critical systems. *Transactions on Automatic Control*, 62(8):3861–3876, 2017.

[3] A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada. Control barrier functions: Theory and applications. In *European Control Conference (ECC)*, pages 3420–3431. IEEE, 2019.

[4] A. Singletary, Y. Chen, and A. D. Ames. Control barrier functions for sampled-data systems with input delays. *arXiv preprint arXiv:2005.06418*, 2020.

[5] Q. Nguyen and K. Sreenath. Exponential control barrier functions for enforcing high relative-degree safety-critical constraints. In *American Control Conference (ACC)*, pages 322–328. IEEE, 2016.

[6] F. Blanchini. Set invariance in control. *Automatica*, 35(11):1747–1767, 1999.

[7] A. J. Taylor, A. Singletary, Y. Yue, and A. D. Ames. A control barrier perspective on episodic learning via projection-to-state safety. *arXiv preprint arXiv:2003.08028*, 2020.

[8] C. Cheng, X. Yan, E. Theodorou, and B. Boots. Accelerating imitation learning with predictive models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

[9] R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Conference on Artificial Intelligence*, volume 33, pages 3387–3395. AAAI, 2019.

[10] A. J. Taylor, A. Singletary, Y. Yue, and A. D. Ames. Learning for safety-critical control with control barrier functions. *Proceedings of Machine Learning Research (PMLR)*, 120:708–717, 2020.

[11] J. Choi, F. Castañeda, C. J. Tomlin, and K. Sreenath. Reinforcement learning for safety-critical control under model uncertainty, using control lyapunov functions and control barrier functions. *arXiv preprint arXiv:2004.07584*, 2020.

[12] T. Gurriet, M. Mote, A. D. Ames, and E. Feron. An online approach to active set invariance. In *Conference on Decision & Control (CDC)*, pages 3592–3599. IEEE, 2018.

[13] A. Robey, H. Hu, L. Lindemann, H. Zhang, D. V. Dimarogonas, S. Tu, and N. Matni. Learning control barrier functions from expert demonstrations. *arXiv preprint arXiv:2004.03315*, 2020.

[14] A. Clark. Control barrier functions for complete and incomplete information stochastic systems. In *American Control Conference (ACC)*, pages 2928–2935. IEEE, 2019.

[15] P. Nilsson and A. D. Ames. Lyapunov-like conditions for tight exit probability bounds through comparison theorems for sdes. In *American Control Conference (ACC)*, pages 5175–5181. IEEE, 2020.

[16] T. Gurriet, P. Nilsson, A. Singletary, and A. D. Ames. Realizable set invariance conditions for cyber-physical systems. In *American Control Conference (ACC)*, pages 3642–3649. IEEE, 2019.

[17] M. Jankovic. Robust control barrier functions for constrained stabilization of nonlinear systems. *Automatica*, 96:359–367, 2018.

[18] R. Takano and M. Yamakita. Robust constrained stabilization control using control lyapunov and control barrier function in the presence of measurement noises. In *Conference on Control Technology and Applications (CCTA)*, pages 300–305. IEEE, 2018.

[19] F. Codevilla, M. Miiller, A. López, V. Koltun, and A. Dosovitskiy. End-to-end driving via conditional imitation learning. In *International Conference on Robotics and Automation (ICRA)*, pages 1–9. IEEE, 2018.

[20] A. Lambert, A. Shaban, A. Raj, Z. Liu, and B. Boots. Deep forward and inverse perceptual models for tracking and prediction. In *International Conference on Robotics and Automation (ICRA)*, pages 675–682. IEEE, 2018.

[21] S. Tang, V. Wüest, and V. Kumar. Aggressive flight with suspended payloads using vision-based control. *Robotics and Automation Letters*, 3(2):1152–1159, 2018.

[22] Z. Roth, B. Mooring, and B. Ravani. An overview of robot calibration. *Journal on Robotics and Automation*, 3(5):377–385, 1987.

[23] S. Dean, N. Matni, B. Recht, and V. Ye. Robust guarantees for perception-based control. *arXiv preprint arXiv:1907.03680*, 2019.

[24] S. Dean and B. Recht. Certainty-equivalent perception-based control. *Manuscript submitted for publication*, 2020.

[25] F. Laine, C.-Y. Chiu, and C. Tomlin. Eyes-closed safety kernels: Safety for autonomous systems under loss of observability. *arXiv preprint arXiv:2005.07144*, 2020.

[26] L. Perko. *Differential equations and dynamical systems*, volume 7. Springer Science & Business Media, 2013.

[27] X. Xu, P. Tabuada, J. W. Grizzle, and A. D. Ames. Robustness of control barrier functions for safety critical control. *IFAC-PapersOnLine*, 48(27):54–61, 2015.

[28] T. Gurriet, A. Singletary, J. Reher, L. Ciarletta, E. Feron, and A. Ames. Towards a framework for realizable safety critical control through active set invariance. In *International Conference on Cyber-Physical Systems (ICCPS)*, pages 98–106. IEEE Press, 2018.

[29] B. S. Mordukhovich, J. V. Outrata, and M. E. Sarabi. Full stability of locally optimal solutions in second-order cone programs. *Journal on Optimization*, 24(4):1581–1613, 2014.

[30] A. L. Dontchev and W. W. Hager. Lipschitzian stability in nonlinear control and optimization. *Journal on Control and Optimization*, 31(3):569–603, 1993.

[31] A. De Maio and S. Lacroix. On learning visual odometry errors. 2019.

[32] A. Domahidi, E. Chu, and S. Boyd. Ecos: An socp solver for embedded systems. In *European Control Conference (ECC)*, pages 3071–3076. IEEE, 2013.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[34] P. Glotfelter, J. Cortés, and M. Egerstedt. Boolean composability of constraints and control synthesis for multi-robot systems via nonsmooth control barrier functions. In *Conference on Control Technology and Applications (CCTA)*, pages 897–902. IEEE, 2018.

[35] G. C. Calafiore and L. El Ghaoui. *Optimization models*. Cambridge university press, 2014.

A Measurement-Robust Control Barrier Function Safety Proof

In this appendix we provide a proof of Theorem 2 relating MR-CBFs to safety.

Recall that Theorem 2 is given by:

Theorem (2). *Let a set $\mathcal{C} \subset \mathbb{R}^n$ be defined as the 0-superlevel set of a continuously differentiable function $h : \mathbb{R}^n \to \mathbb{R}$. Assume the functions $L_{\mathbf{f}}h : \mathbb{R}^n \to \mathbb{R}$, $L_{\mathbf{g}}h : \mathbb{R}^n \to \mathbb{R}^m$, and $\alpha \circ h : \mathbb{R}^n \to \mathbb{R}$ are Lipschitz continuous on $\mathcal{C}$ with Lipschitz coefficients $\mathfrak{L}_{L_{\mathbf{f}}h}$, $\mathfrak{L}_{L_{\mathbf{g}}h}$, and $\mathfrak{L}_{\alpha \circ h}$, respectively. Further assume there exists a locally Lipschitz function $\epsilon : \mathbb{R}^k \to \mathbb{R}_+$, such that $\max_{\mathbf{e} \in \mathcal{E}(\mathbf{y})} \|\mathbf{e}\|_2 \le \epsilon(\mathbf{y})$ for all $\mathbf{y} \in \mathbf{p}(\mathcal{C})$. If h is an MR-CBF for (1) on $\mathcal{C}$ with parameter function $(\epsilon(\mathbf{y})(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h}), \epsilon(\mathbf{y})\mathfrak{L}_{L_{\mathbf{g}}h})$, then any locally Lipschitz continuous controller $\mathbf{k} : \mathbb{R}^k \times \mathbb{R}^n \to \mathbb{R}^m$, such that $\mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}}) \in K_{mr\text{-}cbf}(\mathbf{y}, \widehat{\mathbf{x}})$ for all $(\mathbf{y}, \widehat{\mathbf{x}}) \in \widehat{\mathbf{v}}(\mathcal{C})$, renders the system (2) safe with respect to the set $\mathcal{C}$.*

Proof of Theorem 2. Define the function $c : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ as

$$c(\mathbf{x}, \mathbf{u}) = \frac{\partial h}{\partial \mathbf{x}}(\mathbf{x})\Big(\mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})\mathbf{u}\Big) + \alpha(h(\mathbf{x})),$$
$$= L_{\mathbf{f}}h(\mathbf{x}) + L_{\mathbf{g}}h(\mathbf{x})\mathbf{u} + \alpha(h(\mathbf{x})).$$

This proof will follow from Theorem 1, in that for any $\mathbf{x} \in \mathcal{C}$, with $(\mathbf{y}, \widehat{\mathbf{x}}) = \widehat{\mathbf{v}}(\mathbf{x})$, we will show:

$$c(\mathbf{x}, \mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}})) \ge 0. \tag{23}$$

To show that (23) is true, consider a measurement-state estimate pair $(\mathbf{y}, \widehat{\mathbf{x}}) \in \widehat{\mathbf{v}}(\mathcal{C})$. A sufficient condition for (23) to hold is given by:

$$\inf_{\mathbf{x} \in \mathcal{X}(\mathbf{y})} c(\mathbf{x}, \mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}})) \ge 0 .$$

Recalling that we define $\widehat{\mathbf{x}} = \mathbf{x} + \mathbf{e}(\mathbf{x})$, we have:

$$\inf_{\mathbf{x} \in \mathcal{X}(\mathbf{y})} c(\mathbf{x}, \mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}})) = \inf_{\mathbf{e} \in \mathcal{E}(\mathbf{y})} c(\widehat{\mathbf{x}} - \mathbf{e}, \mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}})),$$
$$= c(\widehat{\mathbf{x}}, \mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}})) + \inf_{\mathbf{e} \in \mathcal{E}(\mathbf{y})} c(\widehat{\mathbf{x}} - \mathbf{e}, \mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}})) - c(\widehat{\mathbf{x}}, \mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}})),$$
$$\ge c(\widehat{\mathbf{x}}, \mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}})) - \sup_{\mathbf{e} \in \mathcal{E}(\mathbf{y})} |c(\widehat{\mathbf{x}} - \mathbf{e}, \mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}})) - c(\widehat{\mathbf{x}}, \mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}}))|.$$

The assumption on Lipschitz continuity of $L_{\mathbf{f}}h$, $L_{\mathbf{g}}h$, and $\alpha \circ h$ enables the following bound:

$$|c(\mathbf{x}', \mathbf{u}) - c(\mathbf{x}, \mathbf{u})| = |L_{\mathbf{f}}h(\mathbf{x}') - L_{\mathbf{f}}h(\mathbf{x}) + L_{\mathbf{g}}h(\mathbf{x}')\mathbf{u} - L_{\mathbf{g}}h(\mathbf{x})\mathbf{u} + \alpha(h(\mathbf{x}')) - \alpha(h(\mathbf{x}))|,$$
$$\le |L_{\mathbf{f}}h(\mathbf{x}') - L_{\mathbf{f}}h(\mathbf{x})| + |L_{\mathbf{g}}h(\mathbf{x}')\mathbf{u} - L_{\mathbf{g}}h(\mathbf{x})\mathbf{u}| + |\alpha(h(\mathbf{x}')) - \alpha(h(\mathbf{x}))|,$$
$$\le \mathfrak{L}_{L_{\mathbf{f}}h}\|\mathbf{x}' - \mathbf{x}\|_2 + \|L_{\mathbf{g}}h(\mathbf{x}') - L_{\mathbf{g}}h(\mathbf{x})\|_2\|\mathbf{u}\|_2 + \mathfrak{L}_{\alpha \circ h}\|\mathbf{x}' - \mathbf{x}\|_2,$$
$$\le (\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{L_{\mathbf{g}}h}\|\mathbf{u}\|_2 + \mathfrak{L}_{\alpha \circ h})\|\mathbf{x}' - \mathbf{x}\|_2.$$

Therefore, using the definition of $\epsilon(\mathbf{y})$ we have:

$$\sup_{\mathbf{e} \in \mathcal{E}(\mathbf{y})} |c(\widehat{\mathbf{x}} - \mathbf{e}, \mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}})) - c(\widehat{\mathbf{x}}, \mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}}))| \le \sup_{\mathbf{e} \in \mathcal{E}(\mathbf{y})} (\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{L_{\mathbf{g}}h}\|\mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}})\|_2 + \mathfrak{L}_{\alpha \circ h})\|\mathbf{e}\|_2,$$
$$\le (\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{L_{\mathbf{g}}h}\|\mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}})\|_2 + \mathfrak{L}_{\alpha \circ h})\epsilon(\mathbf{y}).$$

Thus:

$$\inf_{\mathbf{x} \in \mathcal{X}(\mathbf{y})} c(\mathbf{x}, \mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}})) \ge c(\widehat{\mathbf{x}}, \mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}})) - (\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{L_{\mathbf{g}}h}\|\mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}})\|_2 + \mathfrak{L}_{\alpha \circ h})\epsilon(\mathbf{y}).$$

By the MR-CBF condition and the design of $\mathbf{k}$ we have that:

$$c(\widehat{\mathbf{x}}, \mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}})) - \epsilon(\mathbf{y})(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h}) - \epsilon(\mathbf{y})\mathfrak{L}_{L_{\mathbf{g}}h}\|\mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}})\|_2 \ge 0,$$

implying the condition (23).

$\square$

B MR-OP Second Order Cone Conversion

In this appendix we show how the optimization problem specified by (MR-OP) can be written in the standard form for a Second-Order Cone Program. The MR-OP Filter is given by:

$$\mathbf{k}(\mathbf{y},\widehat{\mathbf{x}}) = \underset{\mathbf{u}\in\mathbb{R}^m}{\mathrm{argmin}} \quad \frac{1}{2}\|\mathbf{u} - \mathbf{k}_d(\widehat{\mathbf{x}})\|_2^2 \tag{MR-OP}$$

$$\text{s.t.} \quad L_{\mathbf{f}}h(\widehat{\mathbf{x}}) - (\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha\circ h})\epsilon(\mathbf{y}) + L_{\mathbf{g}}h(\widehat{\mathbf{x}})\mathbf{u} - \mathfrak{L}_{L_{\mathbf{g}}h}\epsilon(\mathbf{y})\|\mathbf{u}\|_2 \geq -\alpha(h(\widehat{\mathbf{x}})).$$

First, the constant term can be removed from the cost such that it becomes: $\frac{1}{2}\|\mathbf{u}\|_2^2 - \mathbf{k}_d(\widehat{\mathbf{x}})^\top\mathbf{u}$. Additionally, the constraint can be written in terms of a second order cone:

$$\mathbf{q} = \begin{bmatrix} L_{\mathbf{g}}h(\widehat{\mathbf{x}}) \\ \mathfrak{L}_{L_{\mathbf{g}}h}\epsilon(\mathbf{y})I_m \end{bmatrix}\mathbf{u} + \begin{bmatrix} \alpha(h(\widehat{\mathbf{x}})) + L_{\mathbf{f}}h(\widehat{\mathbf{x}}) - (\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha\circ h})\epsilon(\mathbf{y}) \\ 0_m \end{bmatrix}, \quad \mathbf{q} \in Q^{m+1},$$

where $Q^{m+1} \triangleq \{(q_0, \mathbf{q}_1) \in \mathbb{R}\times\mathbb{R}^m \mid q_0 \geq \|\mathbf{q}_1\|_2\}$ is the second-order cone in $\mathbb{R}^{m+1}$. Furthermore, by adding the decision variable $t \in \mathbb{R}$, the quadratic cost function can be converted to an equivalent linear cost, $t - \mathbf{k}_d(\widehat{\mathbf{x}})^T\mathbf{u}$, with a rotated second order cone constraint, $\|\mathbf{u}\|_2^2 \leq 2t$, that can be converted to a standard second order cone constraint via a rotation matrix $\mathbf{R} \in \mathbb{R}^{(m+2)\times(m+2)}$ [35]. Thus the additional constraint can be written as:

$$\mathbf{r} = \mathbf{R}\begin{bmatrix} t & 1 & \mathbf{u}^\top \end{bmatrix}^\top, \ \mathbf{r} \in Q^{m+2}.$$

Combining the two second order cone constraints, the problem can be written as:

$$\mathbf{k}(\mathbf{y},\widehat{\mathbf{x}}) = \underset{(\mathbf{u},t,\mathbf{s})\in\mathbb{R}^{3m+4}}{\mathrm{argmin}} \quad \begin{bmatrix} 1 & 0 & -\mathbf{k}_d(\widehat{\mathbf{x}})^\top \end{bmatrix}\begin{bmatrix} t & 1 & \mathbf{u}^\top \end{bmatrix}^\top$$

$$\text{s.t.} \quad \mathbf{G}(\mathbf{y},\widehat{\mathbf{x}})\begin{bmatrix} t \\ 1 \\ \mathbf{u} \end{bmatrix} + \mathbf{s} = \mathbf{h}(\mathbf{y},\widehat{\mathbf{x}}),$$

$$\mathbf{s} \in Q^{m+2} \times Q^{m+1},$$

where:

$$\mathbf{G}(\mathbf{y},\widehat{\mathbf{x}}) = -\begin{bmatrix} \mathbf{R}_1 & 0 & \mathbf{R}_{3:m+2} \\ 0 & 0 & L_{\mathbf{g}}h(\widehat{\mathbf{x}}) \\ \mathbf{0}_m & \mathbf{0}_m & \mathfrak{L}_{L_{\mathbf{g}}h}\epsilon(\mathbf{y})\mathbf{I}_m \end{bmatrix},$$

$$\mathbf{h}(\mathbf{y},\widehat{\mathbf{x}}) = \begin{bmatrix} \mathbf{R}_2 \\ \alpha(h(\widehat{\mathbf{x}})) + L_{\mathbf{f}}h(\widehat{\mathbf{x}}) - (\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha\circ h})\epsilon(\mathbf{y}) \\ \mathbf{0}_m \end{bmatrix}.$$

Here $\mathbf{R}_i$ represents the ith column of $\mathbf{R}$, and $\mathbf{R}_{3:m+2}$ represents the columns 3 through $m+2$. This equivalent formulation of (MR-OP) is in standard Second-Order Cone Program form as in [32].

C R-MR-OP Lipschitz Continuity Proof

In this appendix we show that the controller (R-MR-OP) is locally Lipschitz continuous in terms of the true state $\mathbf{x}$. Recall this controller is given by:

$$\mathbf{k}(\mathbf{y},\widehat{\mathbf{x}}) = \underset{(\mathbf{u},\delta)\in\mathbb{R}^m\times\mathbb{R}}{\mathrm{argmin}} \quad \frac{1}{2}\|\mathbf{u} - \mathbf{k}_d(\widehat{\mathbf{x}})\|_2^2 + p\delta^2 \tag{R-MR-OP}$$

$$\text{s.t.} \quad L_{\mathbf{f}}h(\widehat{\mathbf{x}}) - (\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha\circ h})\epsilon(\mathbf{y}) + L_{\mathbf{g}}h(\widehat{\mathbf{x}})\mathbf{u} - \mathfrak{L}_{L_{\mathbf{g}}h}\epsilon(\mathbf{y})\|\mathbf{u}\|_2 \geq -\alpha(h(\widehat{\mathbf{x}})) - \delta.$$

where $\mathbf{y} = \mathbf{p}(\mathbf{x})$ and $\widehat{\mathbf{q}}(\mathbf{p}(\mathbf{x}))$, with $\mathbf{p}$ and $\widehat{\mathbf{q}}$ locally Lipschitz continuous. Thus if we prove that $\mathbf{k}$ is locally Lipschitz with respects to its arguments, the composition of $\mathbf{k}$ with the measurement functions will be locally Lipschitz. We also note that this optimization is always feasible, with a unique minimizer as it is convex.

The constraint in this optimization problem can be restated as a conic constraint:

$$\mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}}) = \underset{(\mathbf{u}, \delta) \in \mathbb{R}^m \times \mathbb{R}}{\operatorname{argmin}} \quad \frac{1}{2}\|\mathbf{u} - \mathbf{k}_d(\widehat{\mathbf{x}})\|_2^2 + p\delta^2 \qquad \text{(R-MR-OP)}$$

$$\text{s.t.} \quad \begin{bmatrix} L_\mathbf{g} h(\widehat{\mathbf{x}})\mathbf{u} + \delta + \alpha(h(\widehat{\mathbf{x}})) + L_\mathbf{f} h(\widehat{\mathbf{x}}) - (\mathfrak{L}_{L_\mathbf{f} h} + \mathfrak{L}_{\alpha \circ h})\epsilon(\mathbf{y}) \\ \mathfrak{L}_{L_\mathbf{g} h}\epsilon(\mathbf{y})\mathbf{u} \end{bmatrix} \in Q^{m+1}.$$

To prove local Lipschitz continuity of this controller, we will make use of Theorem 4 in [30], stated below. To draw parallels with the notation of this work, we define the following:

$$\mathbf{w} \triangleq (\mathbf{u}, \delta) \in \mathbb{R}^{m+1}, \tag{24}$$

$$C_{(\mathbf{y},\widehat{\mathbf{x}})}(\mathbf{w}) \triangleq \frac{1}{2}\|\mathbf{u} - \mathbf{k}_d(\widehat{\mathbf{x}})\|_2^2 + p\delta^2, \tag{25}$$

$$G_{(\mathbf{y},\widehat{\mathbf{x}})}(\mathbf{w}) \triangleq \begin{bmatrix} L_\mathbf{g} h(\widehat{\mathbf{x}})\mathbf{u} + \delta + \alpha(h(\widehat{\mathbf{x}})) + L_\mathbf{f} h(\widehat{\mathbf{x}}) - (\mathfrak{L}_{L_\mathbf{f} h} + \mathfrak{L}_{\alpha \circ h})\epsilon(\mathbf{y}), \\ \mathfrak{L}_{L_\mathbf{g} h}\epsilon(\mathbf{y})\mathbf{u} \end{bmatrix} \tag{26}$$

The optimization problem can then be written as:

$$\mathbf{k}(\mathbf{y}, \widehat{\mathbf{x}}) = \underset{\mathbf{w} \in \mathbb{R}^{m+1}}{\operatorname{argmin}} \quad C_{(\mathbf{y},\widehat{\mathbf{x}})}(\mathbf{w}) \tag{27}$$

$$\text{s.t.} \quad G_{(\mathbf{y},\widehat{\mathbf{x}})}(\mathbf{w}) \in Q^{m+1}.$$

We note that $G_{(\mathbf{y},\widehat{\mathbf{x}})} : \mathbb{R}^{m+1} \to \mathbb{R}^{m+1}$, with $\mathbb{R}^{m+1}$ a Banach space, and that $Q^{m+1} \subset \mathbb{R}^{m+1}$ is a closed, convex cone with its vertex at the origin. We further have that $C_{(\mathbf{y},\widehat{\mathbf{x}})}$ and $G_{(\mathbf{y},\widehat{\mathbf{x}})}$ are twice differentiable with respect to $\mathbf{w}$, and these derivatives are locally Lipschitz continuous with respect to $(\mathbf{y}, \widehat{\mathbf{x}})$ by local Lipschitz continuity of $\mathbf{k}_d$, $L_\mathbf{f} h$, $L_\mathbf{g} h$, $\alpha \circ h$, and ϵ.

The $H_{(\mathbf{y},\widehat{\mathbf{x}})}$ denote the Lagrangian:

$$H_{(\mathbf{y},\widehat{\mathbf{x}})}(\mathbf{w}, \boldsymbol{\lambda}) = C_{(\mathbf{y},\widehat{\mathbf{x}})}(\mathbf{w}) - \langle \boldsymbol{\lambda}, G_{(\mathbf{y},\widehat{\mathbf{x}})}(\mathbf{w}) \rangle, \tag{28}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^m$. The first-order necessary conditions associated with a solution $(\mathbf{w}^\star_{(\mathbf{y},\widehat{\mathbf{x}})}, \boldsymbol{\lambda}^\star_{(\mathbf{y},\widehat{\mathbf{x}})})$ to (27) can be expressed as:

$$\frac{\partial H_{(\mathbf{y},\widehat{\mathbf{x}})}}{\partial \mathbf{w}}(\mathbf{w}^\star_{(\mathbf{y},\widehat{\mathbf{x}})}, \boldsymbol{\lambda}^\star_{(\mathbf{y},\widehat{\mathbf{x}})}) = \mathbf{0}, \tag{29}$$

$$\langle \boldsymbol{\lambda}^\star_{(\mathbf{y},\widehat{\mathbf{x}})}, G_{(\mathbf{y},\widehat{\mathbf{x}})}(\mathbf{w}^\star_{(\mathbf{y},\widehat{\mathbf{x}})}) \rangle = 0, \tag{30}$$

$$G_{(\mathbf{y},\widehat{\mathbf{x}})}(\mathbf{w}^\star_{(\mathbf{y},\widehat{\mathbf{x}})}) \in Q^{m+1^+}, \tag{31}$$

where $\mathbf{w}^\star_{(\mathbf{y},\widehat{\mathbf{x}})} \in \mathbb{R}^m + 1$ and $\boldsymbol{\lambda}^\star_{(\mathbf{y},\widehat{\mathbf{x}})} \in Q^{m+1}$. We also have that the following coercivity condition is met:

$$\left\langle \frac{\partial H_{(\mathbf{y},\widehat{\mathbf{x}})}}{\partial \mathbf{w}}(\mathbf{w}^\star_{(\mathbf{y},\widehat{\mathbf{x}})}, \boldsymbol{\lambda}^\star_{(\mathbf{y},\widehat{\mathbf{x}})})(\mathbf{v}_2 - \mathbf{v}_1), \mathbf{v}_2 - \mathbf{v}_1 \right\rangle \geq \|\mathbf{v}_2 - \mathbf{v}_1\|_2^2, \tag{32}$$

for any $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^{m+1}$ as:

$$\frac{\partial H_{(\mathbf{y},\widehat{\mathbf{x}})}}{\partial \mathbf{w}}(\mathbf{w}^\star_{(\mathbf{y},\widehat{\mathbf{x}})}, \boldsymbol{\lambda}^\star_{(\mathbf{y},\widehat{\mathbf{x}})}) = \mathbf{I}_{m+1}, \tag{33}$$

for all $(\mathbf{y}, \widehat{\mathbf{x}})$. Lastly, we note that:

$$\frac{\partial G_{(\mathbf{y},\widehat{\mathbf{x}})}}{\partial \mathbf{w}}(\mathbf{w}^\star_{(\mathbf{y},\widehat{\mathbf{x}})}) = \begin{bmatrix} L_\mathbf{g} h(\widehat{\mathbf{x}}) & 1 \\ \mathfrak{L}_{L_\mathbf{g} h}\epsilon(\mathbf{y})\mathbf{I}_m & \mathbf{0}_{m \times 1} \end{bmatrix}. \tag{34}$$

Under the assumption that $\epsilon(\mathbf{y}) \neq 0$ for all $\mathbf{y} \in \mathbf{p}(\mathcal{C})$ (or that every measurement has some amount of worst case error), we have the map in $\frac{\partial G_{(\mathbf{y},\widehat{\mathbf{x}})}}{\partial \mathbf{w}}(\mathbf{w}^\star_{(\mathbf{y},\widehat{\mathbf{x}})})$ is surjective from $\mathbb{R}^{m+1}$ to $\mathbb{R}^{m+1}$. Thus we have that all of the conditions of the following theorem are met:

Theorem 4 (Lipschitz Continuity of SOCP [30]). *If $\frac{\partial G_{(\mathbf{y},\widehat{\mathbf{x}})}}{\partial \mathbf{w}}(\mathbf{w}^\star_{(\mathbf{y},\widehat{\mathbf{x}})})$ is surjective and the coercivity condition (32) holds, then there exists $s > 0$ such that (27) has a strict local minimizer $\mathbf{w}^\star_{(\mathbf{y},\widehat{\mathbf{x}})}$ for each $(\mathbf{y}', \widehat{\mathbf{x}}') \in B_s((\mathbf{y}, \widehat{\mathbf{x}}))$, and both $\mathbf{w}^\star_{(\mathbf{y},\widehat{\mathbf{x}})}$, and the associated (unique) multiplier $\boldsymbol{\lambda}^\star_{(\mathbf{y},\widehat{\mathbf{x}})}$*

satisfying the first-order necessary condition (29), are Lipschitz continuous functions of $(\mathbf{y}', \widehat{\mathbf{x}}') \in$ $B_s((\mathbf{y}, \widehat{\mathbf{x}}))$.

As $(\mathbf{y}, \widehat{\mathbf{x}})$ were arbitrary, and are locally Lipschitz continuous functions of the true state $\mathbf{x}$, we have that $\mathbf{k}$ is locally Lipschitz with respect to the state.

D Learning and Uncertainty Reduction Results

In this section, we provide a proof of Theorem 3, give a motivation for the definition of the nonparametric error bound, and provide a proof of Corollary 1.

Recall that Theorem 3 is given by:

Theorem (3). *Let* $\mathcal{C}$, h, $\mathfrak{L}_{L_{\mathbf{f}}h}$, $\mathfrak{L}_{L_{\mathbf{g}}h}$, $\mathfrak{L}_{\alpha \circ h}$, *and* ϵ *be defined as in Theorem 2. Then* h *is a MR-CBF for (1) on* $\mathcal{C}$ *with parameter function* $(\epsilon(\mathbf{y})(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h}), \epsilon(\mathbf{y})\mathfrak{L}_{L_{\mathbf{g}}h})$ *if for all* $\mathbf{x} \in \mathcal{C}$:

$$\varepsilon(\mathbf{x}) < \max \left\{ \frac{\|L_{\mathbf{g}}h(\mathbf{x})\|_2}{2\mathfrak{L}_{L_{\mathbf{g}}h}}, \frac{L_{\mathbf{f}}h(\mathbf{x}) + \alpha(h(\mathbf{x}))}{2(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})} \right\} \triangleq \bar{\varepsilon}(\mathbf{x}) \ .$$

Proof. For completeness, we begin by proving the condition (16). The function h is a MR-CBF if

$$\sup_{\mathbf{u} \in \mathbb{R}^m} L_{\mathbf{g}}h(\widehat{\mathbf{x}})\mathbf{u} - b(\mathbf{y})\|\mathbf{u}\|_2 > -L_{\mathbf{f}}h(\widehat{\mathbf{x}}) - \alpha(h(\widehat{\mathbf{x}})) + a(\mathbf{y}). \tag{35}$$

Considering the left hand side, and doing a change of variables with $s = \|\mathbf{u}\|_2$ and $\mathbf{v} = \mathbf{u}/\|\mathbf{u}\|_2$,

$$\sup_{\mathbf{u} \in \mathbb{R}^m} L_{\mathbf{g}}h(\widehat{\mathbf{x}})\mathbf{u} - b(\mathbf{y})\|\mathbf{u}\|_2 = \sup_{s>0} s \cdot \left(\max_{\|\mathbf{v}\|_2=1} L_{\mathbf{g}}h(\widehat{\mathbf{x}})\mathbf{v} - b(\mathbf{y}) \right),$$
$$= \sup_{s>0} s \cdot (\|L_{\mathbf{g}}h(\widehat{\mathbf{x}})\|_2 - b(\mathbf{y})) \ .$$

If $\|L_{\mathbf{g}}h(\widehat{\mathbf{x}})\|_2 - b(\mathbf{y}) > 0$, then s can be chosen arbitrarily large, implying (35) holds. If $\|L_{\mathbf{g}}h(\widehat{\mathbf{x}})\|_2 - b(\mathbf{y}) \le 0$, then (35) only holds if

$$0 > -L_{\mathbf{f}}h(\widehat{\mathbf{x}}) - \alpha(h(\widehat{\mathbf{x}})) + a(\mathbf{y}).$$

Substituting $a(\mathbf{y}) = \epsilon(\mathbf{y})\mathfrak{L}_{L_{\mathbf{g}}h}$ and $b(\mathbf{y}) = \epsilon(\mathbf{y})(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})$ and combining these conditions, we see that (35) is satisfied if and only if

$$\epsilon(\mathbf{y}) = \varepsilon(\mathbf{x}) < \max \left\{ \frac{\|L_{\mathbf{g}}h(\widehat{\mathbf{x}})\|_2}{\mathfrak{L}_{L_{\mathbf{g}}h}}, \frac{L_{\mathbf{f}}h(\widehat{\mathbf{x}}) + \alpha(h(\widehat{\mathbf{x}}))}{(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})} \right\} \ .$$

It remains to find a sufficient condition in which the right hand side depends only on $\mathbf{x}$. Notice that

$$\|L_{\mathbf{g}}h(\mathbf{x})\|_2 = \|L_{\mathbf{g}}h(\widehat{\mathbf{x}} - \mathbf{e}(\mathbf{x}))\|_2 \le \|L_{\mathbf{g}}h(\widehat{\mathbf{x}})\|_2 + \mathfrak{L}_{L_{\mathbf{g}}h}\epsilon,$$

therefore,

$$\|L_{\mathbf{g}}h(\mathbf{x})\|_2 > 2\mathfrak{L}_{L_{\mathbf{g}}h}\varepsilon(\mathbf{x}) \implies \|L_{\mathbf{g}}h(\widehat{\mathbf{x}})\|_2 > \mathfrak{L}_{L_{\mathbf{g}}h}\varepsilon(\mathbf{x}).$$

By a similar argument, we also have that

$$L_{\mathbf{f}}h(\mathbf{x}) + \alpha(h(\mathbf{x})) > 2(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})\varepsilon(\mathbf{x}) \implies L_{\mathbf{f}}h(\widehat{\mathbf{x}}) + \alpha(h(\widehat{\mathbf{x}})) > (\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})\varepsilon(\mathbf{x}).$$

Combining these two expressions gives a sufficient condition depending only on $\mathbf{x}$ and yields the desired result. $\square$

D.1 Motivation for Non-Parametric Error Bound

Consider the simple Nadarya-Watson non-parametric regressor defined from a set of training data $\mathcal{S} = \{(\mathbf{y}_i, \bar{\mathbf{x}}_i)\}_{i=1}^N$ as

$$\widehat{\mathbf{q}}(\mathbf{y}) = \sum_{i=1}^N \frac{\mathbf{1}\{\rho(\mathbf{y}_i, \mathbf{y}) \le \alpha\}}{s_N(\mathbf{y})} \cdot \bar{\mathbf{x}}_i, \quad s_N(\mathbf{y}) = \sum_{i=1}^N \mathbf{1}\{\rho(\mathbf{y}_i, \mathbf{y}) \le \alpha\} \ , \tag{36}$$

where $\rho : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}_+$ is a metric with respect to which the functions $\mathbf{p}$ and $\mathbf{q}$ are smooth:

$$\rho(\mathbf{p}(\mathbf{x}), \mathbf{p}(\mathbf{x}')) \leq \mathfrak{L}_{\mathbf{p}} \|\mathbf{x} - \mathbf{x}'\|_2, \quad \|\mathbf{q}(\mathbf{y}) - \mathbf{q}(\mathbf{y}')\|_2 \leq \mathfrak{L}_{\mathbf{q}} \rho(\mathbf{y}, \mathbf{y}') . \tag{37}$$

This regressor performs local averaging over training data to predict the underlying state as a function of the system output $\mathbf{y}$.

Lemma 1 (Adapted from [24]). *For a system satisfying* (37) *and a learned perception map of the form* (36) *with noisy labels as in* (19) *with* $\mathbb{E}[\mathbf{w}_i] = 0$ *and* $\mathbb{E}[\mathbf{w}_i \mathbf{w}_i^\top] = \sigma_w^2$, *we have with probability at least* $1 - \delta$ *that for any fixed* $\mathbf{y}$ *with* $s_N(\mathbf{y}) \neq 0$,

$$\|\widehat{\mathbf{q}}(\mathbf{y}) - \mathbf{q}(\mathbf{y})\|_\infty \leq \alpha \mathfrak{L}_{\mathbf{q}} + \frac{n\sigma_w}{\sqrt{s_N(\mathbf{y})}} \sqrt{n \log\left(n\sqrt{s_N(\mathbf{y})}/\delta \right)} . \tag{38}$$

To see how this applies to the error bound presented in Definition 4, we first note that

$$
\begin{aligned}
s_N(\mathbf{y}) &= \sum_{i=1}^{N} \mathbf{1}\{\rho(\mathbf{p}(\mathbf{x}_i), \mathbf{p}(\mathbf{x})) \leq \alpha\}, \\
&\geq \sum_{i=1}^{N} \mathbf{1}\{\|\mathbf{x}_i - \mathbf{x}\|_2 \leq \tfrac{\alpha}{\mathfrak{L}_{\mathbf{p}}}\}, \\
&= \#\{\mathbf{x}_i \in \mathcal{X}_{\mathcal{S}} \mid \|\mathbf{x} - \mathbf{x}_i\| \leq \tfrac{\alpha}{\mathfrak{L}_{\mathbf{p}}}\}.
\end{aligned}
$$

Therefore, Lemma 1 implies a pointwise high probability bound on error of the form in Definition 4 with

$$L = \mathfrak{L}_{\mathbf{p}} \mathfrak{L}_{\mathbf{q}}, \quad \sigma = n\sigma_w \sqrt{n \log(n\sqrt{N}/\delta)}, \quad \gamma = \tfrac{\alpha}{\mathfrak{L}_{\mathbf{p}}} .$$

Extending this probabilistic result from pointwise to hold over a compact set is nontrivial. The subtleties involved are handled more carefully in [24], and we focus only on this illustrative connection here.

D.2 Proof of Corollary 1

This result follows by combining the expression (17) with the nonparametric error bound introduced in Definition 4. Using this definition, we have that

$$\varepsilon(\mathbf{x}) = \sup_{\mathbf{e} \in \mathcal{E}(\mathbf{y})} \|\mathbf{e}\|_2 \leq L\gamma + \frac{\sigma}{\sqrt{\#\{\mathbf{x}_i \in \mathcal{X}_{\mathcal{S}} \mid \|\mathbf{x} - \mathbf{x}_i\| \leq \gamma\}}} .$$

Combining this with the sufficient condition,

$$\frac{\sigma}{\sqrt{\#\{\mathbf{x}_i \in \mathcal{X}_{\mathcal{S}} \mid \|\mathbf{x} - \mathbf{x}_i\| \leq \gamma\}}} < \max\left\{ \frac{\|L_{\mathbf{g}}h(\mathbf{x})\|_2}{2\mathfrak{L}_{L_{\mathbf{g}}h}}, \frac{L_{\mathbf{f}}h(\mathbf{x}) + \alpha(h(\mathbf{x}))}{2(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})} \right\} - L\gamma .$$

The result follows by rearranging terms.

E Further Experimental Details

In this section, we provide additional plots related to the learned model and present a result analogous to Theorem 3 specialized to the experimental setting.

An empirical validation of the learned model is presented in Figure 3, which displays the training data, the model predictions of θ_y, and the errors over the validation set.

We now present a result analogous to Theorem 3 to derive an expression for $\bar{\varepsilon}(\mathbf{x})$ in the experimental scenario. For control of the robotic Segway under planar motion, the input torques about both wheels were constrained to be identical. Furthermore, the MR-OP Filter constraint in (MR-OP) was applied simultaneously to both safety functions (22). Therefore, the MR-OP Filter requires the feasibility of
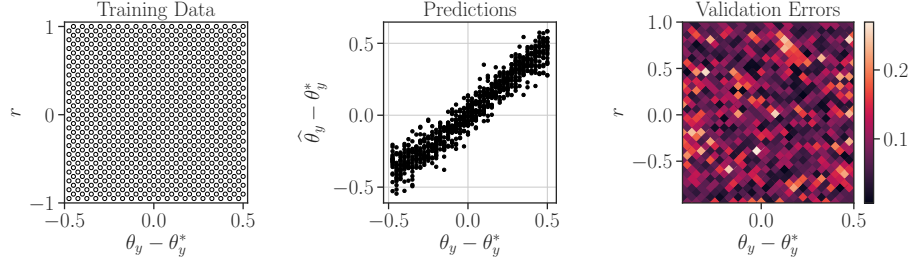
Figure 3. Learned model results. (**Left**) Illustration of the training data, which was collected from a gridded range of position r and pitch angle θ_y. (**Middle**) The estimated vs. actual pitch angles on a validation set. (**Right**) The empirical model errors $\varepsilon(\mathbf{x})$ measured on a validation set.

the following set of constraints, for $u \in \mathbb{R}$:

$$\mathbf{u} = u\mathbf{1},$$
$$L_{\mathbf{f}}h_1(\widehat{\mathbf{x}}) + L_{\mathbf{g}}h_1(\widehat{\mathbf{x}})\mathbf{u} - (\mathfrak{L}_{L_{\mathbf{f}}h_1} + \mathfrak{L}_{\alpha \circ h_1})\epsilon(\mathbf{y}) - \mathfrak{L}_{L_{\mathbf{g}}h_1}\epsilon(\mathbf{y})\|\mathbf{u}\|_2 \geq -\alpha(h_1(\widehat{\mathbf{x}})),$$
$$L_{\mathbf{f}}h_2(\widehat{\mathbf{x}}) + L_{\mathbf{g}}h_2(\widehat{\mathbf{x}})\mathbf{u} - (\mathfrak{L}_{L_{\mathbf{f}}h_2} + \mathfrak{L}_{\alpha \circ h_2})\epsilon(\mathbf{y}) - \mathfrak{L}_{L_{\mathbf{g}}h_2}\epsilon(\mathbf{y})\|\mathbf{u}\|_2 \geq -\alpha(h_2(\widehat{\mathbf{x}})).$$

Making use of the fact that $\frac{\partial h_1}{\partial \mathbf{x}} = -\frac{\partial h_2}{\partial \mathbf{x}}$, we can simplify and rewrite these constraints as

$$L_{\mathbf{g}}h_1(\widehat{\mathbf{x}})\mathbf{1}u - \mathfrak{L}_{L_{\mathbf{g}}h}\epsilon(\mathbf{y})\sqrt{2}|u| \geq -\alpha(h_1(\widehat{\mathbf{x}})) - L_{\mathbf{f}}h_1(\widehat{\mathbf{x}}) + (\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})\epsilon(\mathbf{y}),$$
$$-L_{\mathbf{g}}h_1(\widehat{\mathbf{x}})\mathbf{1}u - \mathfrak{L}_{L_{\mathbf{g}}h}\epsilon(\mathbf{y})\sqrt{2}|u| \geq -\alpha(h_2(\widehat{\mathbf{x}})) + L_{\mathbf{f}}h_1(\widehat{\mathbf{x}}) + (\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})\epsilon(\mathbf{y}).$$

(39)

Proposition 1. The set of constraints (39) will be feasible for all $\mathbf{x} \in \mathcal{C}$ if for all $\mathbf{x} \in \mathcal{C}$:

$$\varepsilon(\mathbf{x}) < \frac{1}{2}\max\left\{\min\left\{\frac{\alpha(h_1(\mathbf{x})) + L_{\mathbf{f}}h_1(\mathbf{x})}{(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})}, \frac{\alpha(h_2(\mathbf{x})) - L_{\mathbf{f}}h_1(\mathbf{x})}{(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})}\right\},\right.$$
$$\min\left\{\frac{\alpha(h_1(\mathbf{x})) + L_{\mathbf{f}}h_1(\mathbf{x})}{(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})},\right.$$
$$\left.\frac{|L_{\mathbf{g}}h_1(\mathbf{x})\mathbf{1}|(\alpha(h_1(\mathbf{x})) + \alpha(h_2(\mathbf{x})))}{\mathfrak{L}_{L_{\mathbf{g}}h}\sqrt{2}(\alpha(h_2(\mathbf{x})) - \alpha(h_1(\mathbf{x})) - 2L_{\mathbf{f}}h_1(\mathbf{x})) + 2(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})|L_{\mathbf{g}}h_1(\mathbf{x})\mathbf{1}|}\right\},$$
$$\min\left\{\frac{\alpha(h_2(\mathbf{x})) - L_{\mathbf{f}}h_1(\mathbf{x})}{(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})},\right.$$
$$\left.\left.\frac{|L_{\mathbf{g}}h_1(\mathbf{x})\mathbf{1}|(\alpha(h_1(\mathbf{x})) + \alpha(h_2(\mathbf{x})))}{\mathfrak{L}_{L_{\mathbf{g}}h}\sqrt{2}(\alpha(h_1(\mathbf{x})) - \alpha(h_2(\mathbf{x})) + 2L_{\mathbf{f}}h_1(\mathbf{x})) + 2(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})|L_{\mathbf{g}}h_1(\mathbf{x})\mathbf{1}|}\right\}\right\}.$$

Proof. First, we adapt constraints in (39) to be in terms of the true state $\mathbf{x}$ rather than the estimated $\widehat{\mathbf{x}}$. By a smoothness argument,

$$L_{\mathbf{g}}h_1(\mathbf{x} - \mathbf{e})\mathbf{1}u - \mathfrak{L}_{L_{\mathbf{g}}h}\varepsilon(\mathbf{x})\sqrt{2}|u| \geq -\alpha(h_1(\mathbf{x} - \mathbf{e})) - L_{\mathbf{f}}h_1(\mathbf{x} - \mathbf{e}) + (\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})\varepsilon(\mathbf{x})$$
$$\impliedby L_{\mathbf{g}}h_1(\mathbf{x})\mathbf{1}u - 2\mathfrak{L}_{L_{\mathbf{g}}h}\varepsilon(\mathbf{x})\sqrt{2}|u| \geq -\alpha(h_1(\mathbf{x})) - L_{\mathbf{f}}h_1(\mathbf{x}) + 2(\mathfrak{L}_{L_{\mathbf{f}}h} + \mathfrak{L}_{\alpha \circ h})\varepsilon(\mathbf{x}).$$

A similar argument holds for the second constraint. Therefore, to find a guarantee depending only on $\mathbf{x}$, we effectively double the radius of error to $2\varepsilon(\mathbf{x})$.

Then the result follows from Lemma 2, which applies to this problem with

$$\epsilon = 2\varepsilon(\mathbf{x}),$$
$$a = L_{\mathbf{g}}h_1(\mathbf{x})\mathbf{1},$$
$$b = \mathfrak{L}_{L_{\mathbf{g}}h}\sqrt{2},$$
$$d_1 = \alpha(h_1(\mathbf{x})) + L_{\mathbf{f}}h_1(\mathbf{x}),$$

16

$$d_2 = \alpha(h_2(\mathbf{x})) - L_{\mathbf{f}} h_1(\mathbf{x}),$$
$$L = (\mathfrak{L}_{L_{\mathbf{f}} h} + \mathfrak{L}_{\alpha \circ h}).$$

$\square$

Lemma 2. *The following two constraints*

$$au - b\epsilon|u| \geq -d_1 + L\epsilon,$$
$$-au - b\epsilon|u| \geq -d_2 + L\epsilon,$$

are simultaneously feasible for some $u \in \mathbb{R}$ if and only if

$$\epsilon \leq \max\left\{ \min\left\{ \frac{d_1}{L}, \frac{d_2}{L} \right\}, \min\left\{ \frac{d_1}{L}, \frac{|a|(d_1 + d_2)}{b(d_2 - d_1) + 2L|a|} \right\}, \min\left\{ \frac{d_2}{L}, \frac{|a|(d_1 + d_2)}{b(d_1 - d_2) + 2L|a|} \right\} \right\}.$$

Proof. It is equivalent to consider

$$-d_1 + L\epsilon \leq \max_u \quad au - b\epsilon|u|$$
$$\text{s.t.} \quad -au - b\epsilon|u| \geq -d_2 + L\epsilon.$$

First, notice that the first condition implies that $u = 0$ satisfies the expression.

Otherwise, consider the case that $-d_2 + L\epsilon \leq 0$, and see that we will set u to have the same sign as a. Then the optimization problem is solved by

$$(|a| - b\epsilon)\frac{-d_2 + L\epsilon}{-(|a| + b\epsilon)} = \max_{|u|} \quad |u|(|a| - b\epsilon)$$
$$\text{s.t.} \quad -(|a| + b\epsilon)|u| \geq -d_2 + L\epsilon.$$

This yields the third condition after some algebra:

$$(|a| - b\epsilon)\frac{-d_2 + L\epsilon}{-(|a| + b\epsilon)} \geq -d_1 + L\epsilon,$$
$$(|a| - b\epsilon)(d_2 - L\epsilon) \geq (-d_1 + L\epsilon)(|a| + b\epsilon),$$
$$|a|d_2 - b\epsilon d_2 - L\epsilon|a| + bL\epsilon^2 \geq -d_1|a| + L\epsilon|a| + -d_1 b\epsilon + Lb\epsilon^2,$$
$$|a|(d_2 + d_1) \geq (2L|a| - d_1 b + d_2 b)\epsilon.$$

Reversing the roles of d_1 and d_2 and setting u to have the opposite sign of a yields the second condition. $\square$