Generalization Bounded Implicit Learning of Nearly Discontinuous Functions

Bibit Bianchini BIBIT@SEAS.UPENN.EDU
Mathew Halm MHALM@SEAS.UPENN.EDU
Nikolai Matni NMATNI@SEAS.UPENN.EDU
Michael Posa POSA@SEAS.UPENN.EDU
University of Pennsylvania, Philadelphia, PA 19104

Editors: R. Firoozi, N. Mehr, E. Yel, R. Antonova, J. Bohg, M. Schwager, M. Kochenderfer

Abstract

Inspired by recent strides in empirical efficacy of implicit learning in many robotics tasks, we seek to understand the theoretical benefits of implicit formulations in the face of nearly discontinuous functions, common characteristics for systems that make and break contact with the environment such as in legged locomotion and manipulation. We present and motivate three formulations for learning a function: one explicit and two implicit. We derive generalization bounds for each of these three approaches, exposing where explicit and implicit methods alike based on prediction error losses typically fail to produce tight bounds, in contrast to other implicit methods with violation-based loss definitions that can be fundamentally more robust to steep slopes. Furthermore, we demonstrate that this violation implicit loss can tightly bound graph distance, a quantity that often has physical roots and handles noise in inputs and outputs alike, instead of prediction losses which consider output noise only. Our insights into the generalizability and physical relevance of violation implicit formulations match evidence from prior works and are validated through a toy problem, inspired by rigid-contact models and referenced throughout our theoretical analysis.

Keywords: Implicit learning, contact dynamics, learning nearly discontinuous functions, generalization error bounds

1. Introduction

Extreme stiffness or even discontinuity is abundant in physical robotics tasks in which systems make or break contact with the environment (Yang and Posa, 2021; Ubellacker et al., 2021; Khader et al., 2020; Kolev and Todorov, 2015). Implicit learning is increasingly common to represent these complex functions, as are loss formulations with embedded optimization problems (Florence et al., 2022; de Avila Belbute-Peres et al., 2018; Fazeli et al., 2017a,b; Amos and Kolter, 2017). These approaches have demonstrated significant empirical performance benefits over explicit approaches in real robotics tasks, in agreement with the effectiveness of implicit models in other fields like trajectory optimization (Posa et al., 2014; Patel et al., 2019) and rigid body dynamics modeling (Stewart and Trinkle, 1996; Anitescu and Potra, 1997; Chatterjee and Ruina, 1998). In particular, ContactNets (Pfrommer* et al., 2020) learns frictional contact behaviors implicitly, using a deep neural network (DNN) to represent inter-body distances and contact geometries. The empirical results from ContactNets show significantly improved sample complexity over explicit alternatives.

Other works motivate this implicit learning trend by exposing inadequacies of explicit approaches for nearly discontinuous functions (Parmar et al., 2021). Our goal is to expose why implicit

formulations are often better suited for these tasks, through the lenses of generalizability and relationship to graph distance, which we will motivate as a physically meaningful quantity. We focus on a specific class of functions f that can be stiff. This complicated, vector-valued, deterministic function $y = f(x)$ has structure such that it can be formulated as an implicit optimization problem defined by the following system of equations,

$$y = g(x, \lambda) \tag{1a}$$

$$\text{such that} \quad \lambda = \underset{\lambda \in \Lambda}{\arg\min}\ h\left(x, y, \lambda\right). \tag{1b}$$

This is a general formulation which can represent classes of smooth and discontinuous functions alike. We aim to understand the value of learning these implicit functions g and h over learning the function f directly and how loss designs can affect how beneficial implicit formulations can be.

1.1. Contributions and Outline

We contribute two analyses of implicit formulations, applied to two implicit approaches as well as their explicit counterpart, introduced in Section 2. First, we derive generalization error bounds for these three approaches in Section 3, drawing from extensive literature in uniform convergence (Shalev-Shwartz and Ben-David, 2014; Shalev-Shwartz et al., 2010), whose application to overparameterized DNNs has become better understood (Golowich et al., 2018; Neyshabur et al., 2018). With the results of these tools, we indicate a common failure mode of explicit approaches to generalize well, and more importantly we discuss why an implicit formulation can avoid this explosion of generalization error bounds even when learning a nearly discontinuous function. These theoretical results validate the sample efficiency observed by implicit methods in practice.

Second, in Section 4, we demonstrate that a violation-based implicit loss formulation can closely bound graph distance, a meaningful quantity grounded in the physical intuition for how to measure fit of a function. We motivate graph distance as potentially a more useful measure than standard prediction accuracy, specifically when functions are highly stiff. This analysis motivates selection of a hyperparameter ϵ introduced in the violation implicit loss. Application of the generalization error bound analysis from Section 3 indicates that at this choice of ϵ, the violation implicit approach can boast significant sample complexity benefits over prediction methods.

We ground our theoretical analysis with the presentation of a toy problem introduced in Section 2.2 and continually referenced throughout. This physically-motivated example provides demonstration of the utility of our general results: we generate a concrete value for ϵ and demonstrate data efficient generalization error bounds separated by over two orders of magnitude in dataset size in comparison to explicit approaches. We conclude in Section 5 with avenues for future work to build upon our results.

2. Problem Formulation

We seek to learn a function $f(x)$ given i.i.d. pairs drawn from a distribution $\{z_i = (x_i, y_i)\}_i^n \sim \mathcal{D}$ such that $y_i \approx f(x_i)$. We adopt an empirical risk minimization approach and consider the following three choices of model and training loss constructions.

Explicit approach (exp): Learn $f^\theta(x)$ directly, with explicit loss (on a data point of index i)

$$l^\theta_{\text{exp}}(x_i, y_i) = \left\| y_i - f^\theta(x_i) \right\|^2. \tag{2}$$

This is a common $l2$ or prediction loss many works in the literature employ including for regression (Fernández-Delgado et al., 2019), as it is a direct measure of the output error of a learned function acting on a provided input.

Naive implicit approach (nimp): Learn the implicit $g^\theta(x, \lambda)$ and $h^\theta(x, y, \lambda)$ where λ satisfies (1b) for the learned h. This loss is

$$l^\theta_{\text{nimp}}(x_i, y_i) = \left\| y_i - g^\theta(x_i, \lambda_i) \right\|^2 \tag{3a}$$

$$\text{such that} \quad \lambda_i = \underset{\lambda \in \Lambda}{\arg\min}\ h^\theta(x_i, g^\theta(x_i, \lambda), \lambda). \tag{3b}$$

Many prior works including Amos and Kolter (2017) take this approach, which requires differentiating through an $\arg\min$. Like the explicit approach, this also measures the output error of a function acting on a provided input, only now the output also depends on a learned implicit variable which minimizes (3b). The motivation for such a constraint is that it embeds complex relationships, often physically-motivated, between hidden states that explicit approaches do not supervise (Raissi and Karniadakis, 2018).

We note at this point that if the parameters represent the same physical quantities across both approaches, the explicit loss is equivalent to the naive implicit loss. **Prediction approach (pred)** refers jointly to both of these when their parameters are shared.

Violation implicit approach (vimp): Learn the implicit $g^\theta(x, \lambda)$ and $h^\theta(x, y, \lambda)$ with loss

$$l^\theta_{\text{vimp}}(x_i, y_i) = \min_{\lambda \in \Lambda} \left\| y_i - g^\theta(x_i, \lambda) \right\|^2 + \frac{1}{\epsilon} h^\theta(x_i, y_i, \lambda), \tag{4}$$

which introduces a hyperparameter, ϵ, that weights the relative importance between the two terms. Together, this violation loss allows and balances violation of the prediction matching term defined by g with the λ constraints defined by h. In contrast with the explicit and naive implicit approaches, Section 4 shows that this approach addresses errors in both the inputs and outputs of the functions. Notably with a realizable h, this approach also maintains the true global minimum: zero noise and a correct model correspond to zero loss. Works including ContactNets (Pfrommer* et al., 2020) employ this violation implicit structure.

2.1. Notation and Assumptions

Let $\mathcal{D}$ denote a distribution over input and output data points $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$, with $(x_1, y_1), \ldots (x_n, y_n)$ as n i.i.d. samples from $\mathcal{D}$. We assume bounds on θ and λ as B_θ and B_λ, respectively. All other bounds we impose to be $\forall z \in \mathcal{D}, \forall \|\lambda\| \leq B_\lambda, \forall \|\theta\| \leq B_\theta$, including bounds on f, g, h, l_{exp}, l_{nimp}, and l_{vimp} as B_f, B_g, B_h, B_{exp}, B_{nimp}, and B_{vimp}, respectively.

Let $\mathcal{F}$, $\mathcal{G}$, and $\mathcal{H}$ be parametric function spaces $f : \mathcal{X} \to \mathcal{Y}$, $g : \mathcal{X} \times \Lambda \to \mathcal{Y}$, and $h : \mathcal{X} \times \mathcal{Y} \times \Lambda \to \mathbb{R}^+$, respectively, for all parameters, θ. For example, the parametric function class is $\mathcal{F} = \{f^\theta : \theta \in \mathbb{R}^k, \|\theta\| \leq B_\theta, \frac{df}{d\theta} \text{ is bounded } \forall \|\theta\| \leq B_\theta, \forall (x, y) \in \mathcal{Z}\}$, and similarly for $\mathcal{G}$ and $\mathcal{H}$. We assume that $\min_{\lambda,x,y} h^\theta = 0$ and $\min_\lambda h^\theta(x, g^\theta(x, \lambda), \lambda) = 0$, $\forall x$. All loss function classes, $\mathcal{L}$ are $l : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$, for all parameters, θ, defining the f, g, and/or h therein.

We define specific Lipschitz constants as $\text{Lip}_* d(x, .)$ to refer to the Lipschitz constant of a function d with respect to $*$, allowing $*$ to change any inputs noted as '.' and holding all other

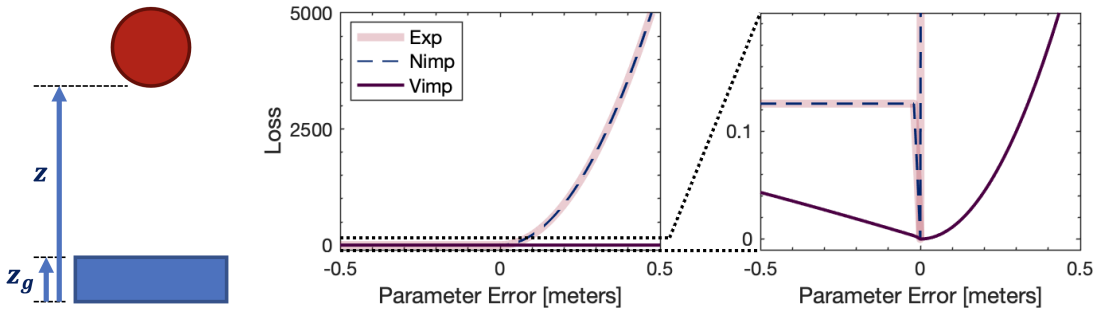

Figure 1: Toy problem of a 1-dimensional point mass falling under gravity until colliding inelastically with the ground. The zeroed loss landscapes of the three approaches are depicted with a zoomed in view at right. The explicit and naive implicit approach for this toy problem result in the same loss landscape. The violation implicit approach in contrast features a more balanced ascent from its minimum on both sides of the optimal parameter value, better suited for optimization via stochastic gradient descent.

inputs constant (in this example, x). Those required are

$$L_{f,\theta} = \text{Lip}_\theta\, f(x;.), \qquad L_{g,\theta} = \text{Lip}_\theta\, g(x,\lambda;.), \qquad L_{h,\theta} = \text{Lip}_\theta\, h(x,y,\lambda;.), \tag{5a}$$

$$L_{g,\lambda} = \text{Lip}_\lambda\, g(x,.;\theta), \qquad L_{h,\lambda} = \text{Lip}_\lambda\, h(x,y,.;\theta), \qquad L^{(\text{nimp})}_{\lambda,\theta} = \text{Lip}_\theta\, \lambda^*_{\text{nimp}}(x,y;.), \tag{5b}$$

where λ^*_{nimp} is the unique minimizing solution to (1b) (similarly for defining $L^{\text{vimp}}_{\lambda,\theta}$ through λ^*_{vimp} using $\|y-g\|^2 + \frac{h}{\epsilon}$ instead of h). If $\frac{\partial^2 h}{\partial \lambda^2}$ is invertible and if there are no constraints on λ, sensitivity analysis (Gould et al., 2016) shows that $L^{(\text{nimp})}_{\lambda,\theta}$ is given by

$$L^{(\text{nimp})}_{\lambda,\theta} = \sup_{x,y,\lambda;\theta} \left[\frac{d\lambda^*}{d\theta}\right] = \sup_{x,y,\lambda;\theta} \left[-\left(\frac{\partial^2 h}{\partial \lambda^2}\right)^{-1} \frac{\partial^2 h}{\partial \theta \partial \lambda}\bigg|_{x,y,\lambda;\theta}\right]. \tag{6}$$

Additionally, we will use the functions $\text{pos}(.) = \max(0,.)$ and $\text{neg}(.) = -\min(0,.)$ which represent the positive and negative part of their inputs. $\text{pos}(x)$ is equivalent to the rectified linear unit activation $\text{ReLU}(x) = \{0 \text{ if } x < 0,\ x \text{ if } x \geq 0\}$ (Nair and Hinton, 2010).

2.2. Toy Problem

To ground the general analysis with a physically-relevant manifestation, we present and refer to an example throughout this paper. Consider the following toy problem: a 1-dimensional point mass falling under gravity, with a flat ground whose interactions with the point mass are perfectly inelastic. See Figure 1 for a schematic of this scenario.

We parameterize this problem with a scalar θ that represents the approximated height of the ground, such that $\theta^* = z_g$. The state of the system $x = [z;\ v]$ is the point mass' position and velocity, and $y = v'$ is the velocity after a time step Δt. When the mass is under free fall, it undergoes projectile motion due to gravity (assume other continuous forces are negligible). Using rigid body time stepping simulation approaches via a linear complementarity program (LCP) (Stewart and Trinkle, 1996), $f^\theta(x)$ takes the form

$$f^\theta(x) = v - a_{\text{grav}}\Delta t + \text{pos}\left(-v + a_{\text{grav}}\Delta t + \frac{\theta - z}{\Delta t}\right). \tag{7}$$

2.2.1. IMPLICIT REPRESENTATION

In the LCP rigid body simulation technique (Stewart and Trinkle, 1996), an implicit scalar variable $\lambda \geq 0$ corresponds to the contact impulse between the ground and the point mass. The function g produces a dynamics prediction from the state x over the time step Δt as

$$g^\theta(x, \lambda) = v - a_{\text{grav}}\Delta t + \frac{1}{m}\lambda, \tag{8}$$

with m as the system mass and where λ satisfies (1b) with an unconstrained h,

$$h^\theta\left(\begin{bmatrix} z \\ v \end{bmatrix}, v', \lambda\right) = \frac{1}{2}\text{neg}(z + v'\Delta t - \theta)^2 + \frac{1}{2}\text{neg}(\lambda)^2 + \text{pos}(z + v'\Delta t - \theta)\text{pos}(\lambda), \tag{9}$$

which enforces no contact forces at a distance, no pulling contact forces, and no interpenetration of the mass with the rigid ground at the end of the time step.

3. Generalization Bounds

We wish to quantify a maximum bound for the generalization error of our general models. Generalization error is the difference between an empirical error observed on training data and the expected error on the true distribution of data. Define the generalization error $\Delta^{\mathcal{S}}_{gen} := l^\theta(\mathcal{D}) - l^\theta(\mathcal{S})$ for $l^\theta(\mathcal{D})$ the expected error on the true distribution of data (population risk), and $l^\theta(\mathcal{S})$ the measured error on a dataset $\mathcal{S} = \{z_i\}_{i=1}^n \sim \mathcal{D}^n$ (empirical risk). Bartlett and Mendelson (2002) produce a bound on the generalization error in terms of Rademacher complexity (Shalev-Shwartz and Ben-David, 2014) of the hypothesis function class, and application of Dudley's entropy integral (Dudley, 2014) can convert Rademacher complexity into quantifiable properties of the parameters, loss formulation, and function class (details and proof of the following theorem in **??**). Combining these steps, we can bound generalization error via the following theorem.

Theorem 1 *Fix a failure probability $\delta \in (0, 1)$, and assume that the loss function $l_{\text{approach}} \in \mathcal{L}_{\text{approach}} : \mathcal{Z} \to [0, B_{\text{approach}}]\}$, with $L_{\text{approach},\theta}$ as its Lipschitz constant with respect to its parameters $\theta \in \mathbb{R}^k$, is acting on a parametric hypothesis function class with data $\mathcal{S} = \{z_1, \ldots z_n\} \sim \mathcal{D}^n$. Then with probability at least $1 - \delta$ and $\forall l_{\text{approach}} \in \mathcal{L}_{\text{approach}}$ and for all hypothesis functions in the class,*

$$\Delta^{\mathcal{S}}_{\text{gen, approach}} \leq 44 L_{\text{approach},\theta} B_\theta \sqrt{\frac{k}{n}} + B_{\text{approach}} \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{10}$$

This approach is most suitable for underparameterized models ($n > k$). Similar analysis better suited for overparameterized DNNs exists (Golowich et al., 2018; Neyshabur et al., 2018) but is not used in this section. With Theorem 1, generalization bounds for each approach can be reduced to the loss Lipschitz constant with respect to the function class parameters, θ. Applying Theorem 1 and differentiating through the embedded optimization problems, we can compute these Lipschitz constants for the three losses as

$$L_{\text{exp},\theta} = 2B_{\text{exp}} L_{f,\theta}, \tag{11}$$

$$L_{\text{nimp},\theta} = 2B_{\text{nimp}} \left(L_{g,\lambda} L^{(\text{nimp})}_{\lambda,\theta} + L_{g,\theta}\right), \tag{12}$$

$$L_{\text{vimp},\theta} = 2B_{\text{nimp}} \left(L_{g,\lambda} L^{(\text{vimp})}_{\lambda,\theta} + L_{g,\theta}\right) + \frac{1}{\epsilon}\left(L_{h,\lambda} L^{(\text{vimp})}_{\lambda,\theta} + L_{h,\theta}\right). \tag{13}$$

Full derivations of these Lipschitz constants are provided in Appendix **??**.

For a steep function f, the difficulty in generalization can be clearly seen in the presence of $L_{f,\theta}$ in (11). For the naive implicit form, particularly in problems like our toy example where λ defines that stiffness, this difficulty has been shifted to $L^{(\text{nimp})}_{\lambda,\theta}$ in (12). This corresponds to a poorly conditioned $\frac{\partial^2 h}{\partial \lambda^2}$, whose inverse appears in (6). The violation implicit approach, on the other hand, can avoid this problem because g can act as a regularizing term on h: the second partial derivative of $\|y-g\|^2 + \frac{h}{\epsilon}$ with respect to λ can be much better conditioned than that of h on its own, allowing $L^{(\text{vimp})}_{\lambda,\theta}$ to be small even when $L^{(\text{nimp})}_{\lambda,\theta}$ is not. Here we remember the role of the hyperparameter ϵ: at large values, it relaxes the optimality criteria in (1b); at small values, the violation implicit approach approximates the naive implicit and thus explicit approaches.

It is important to note that the primary advantage of the naive implicit form over the explicit form can come from a difference in parameterizations. Parameters in the form of neural network weights and biases, for example, are not shared across different approaches, so the Lipschitz constants and thus generalization error bounds would differ for DNNs learning f versus g and h.

3.1. Generalization Bounds for Toy Problem

This difference, however, is not present in our toy problem, where the parameter θ represents ground height in all three approaches. This demonstrates the possible generalization error bound advantage of the violation implicit approach over the alternatives. The required loss Lipschitz constants with respect to the learned parameters are given by (11)-(13). Values for all of the Lipschitz constants are provided in Table 1 in algebraic and numerical form, given the following parameter choices: $\phi_{\max} = 8$ meters, $v_{\max} = 15$ m/s, $a_{\text{grav}} = 9.81$ m/s^2, $\Delta t = 0.005$ seconds, $m = 1$ kilogram, and $\lambda_{\max} = m(v_{\max} + g\Delta t) = 15.05$ Newton-seconds. Derivations of these quantities are provided in Appendix **??**. The $\max$ appears for some parameters due to specifying the h function as piece-wise continuous over different domains to remove the embedded neg and pos functions.

Substituting these values into Equations (11)-(13), the loss Lipschitz constants are

$$L_{\text{vimp},\theta} = \frac{1}{\epsilon}\left(mB_{\text{nimp}} + \lambda_{\max}\left(1 + \frac{m^2}{2\epsilon}\right)\right), \tag{14}$$

$$L_{\text{nimp},\theta} = 2B_{\text{nimp}}\frac{1}{\Delta t}, \tag{15}$$

$$L_{\text{exp},\theta} = 2B_{\text{exp}}\frac{1}{\Delta t}. \tag{16}$$

This reveals the scaling of the explicit (and naive implicit, which behave identically in terms of generalization due to their equivalent parameterization) generalization bounds with $\frac{1}{\Delta t}$, an artifact of discretizing the impact event. These Lipschitz constants can get arbitrarily large with small time steps; an unfortunate characteristic since small time steps are preferred for simulation accuracy. In contrast, the violation implicit approach avoids this poor scaling. Its generalization does depend on Δt since the included $\lambda_{\max}$ parameter in its expression is usually described as a function of time. This is typically calculated as $\lambda_{\max} = m(v_{\max} + a_{\text{grav}}\Delta t)$, which we see $\lambda_{\max} \approx m v_{\max}$ as $\Delta t \to 0$. Thus, as Δt gets arbitrarily small, the generalization error for the violation implicit approach does not grow unboundedly as it can for the explicit approach. Given these are upper bounds, we can only guarantee the violation implicit approach has provably easy generalization regardless of Δt. Such a guarantee cannot be made for the explicit and naive implicit approaches.

Table 1: Lipschitz Constants in Toy LCP Model

Parameter	Expression	Numerical Value
$L_{f,\theta}$	$\frac{1}{\Delta t}$	200
$L_{g,\lambda}$	$\frac{1}{m}$	1
$L_{g,\theta}$	0	0
$L_{h,\lambda}$	$\max\{\phi_{\max},\, \lambda_{\max}\}$	15.05
$L_{h,\theta}$	$\max\{\phi_{\max},\, \lambda_{\max}\}$	15.05
$L^{\text{nimp}}_{\lambda,\theta}$	$\max\left\{\frac{m\Delta t}{m^2+\Delta t^2},\, \frac{m}{\Delta t}\right\}$	200
$L^{\text{vimp}}_{\lambda,\theta}$	$\frac{m^2}{2\epsilon}$	1

We note that increasing the hyperparameter ϵ to tighten the violation implicit generalization bounds comes at the expense of relaxing the physical feasibility of the solution. This means at high ϵ values, the optimization could select unrealistic contact impulses corresponding to contact forces at a distance, penetration of the rigid mass with the ground, or contact forces that pull the objects together instead of pull apart. We will discover however that this approximation error can be bounded when an upper bound is placed on ϵ, derived in the next section.

4. Graph Distance

We present in this section the mathematical foundation to relate the violation implicit loss in (4) to a physically meaningful characteristic. Relating this loss to the concept of graph distance ensures that our aims of minimizing generalization error and training loss are in the pursuit of a quality model.

4.1. Merits of Graph Distance and Prediction Losses

Consider an arbitrary predictive model of the form $y = f^\theta(x)$, defined either explicitly or implicitly. The set of all input-output pairs predicted by this model is the *graph* of the function $f^\theta(x)$:

$$G_\theta = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} : y = f^\theta(x) \right\} . \tag{17}$$

If the noise component of a datapoint (x_i, y_i) is most likely to be small (e.g., when x_i and y_i have Gaussian white noise), then a necessary condition for the datapoint to match the model $y = f^\theta(x)$ well is that there exists an input-output pair $(\bar{x}_i, \bar{y}_i) \approx (x_i, y_i)$ in the graph of $f^\theta(x)$. *Graph distance* is therefore a natural metric to capture model accuracy:

$$d_{G_\theta}(x_i, y_i) = \text{dist}\left(\begin{bmatrix} x_i \\ y_i \end{bmatrix}, G_\theta \right) = \min_x \left\| \begin{bmatrix} x - x_i \\ f^\theta(x) - y_i \end{bmatrix} \right\| . \tag{18}$$

Particularly, when applied to predictive models where x is a system state and y is the next state at a following time step, it is often reasonable to believe there are similar noise distributions on the input and output, and thus weight the difference in x and y equally in (18). By contrast, *prediction losses* like the explicit loss and naive implicit loss only consider noise in the output:

$$l^\theta_{\text{exp}}(x_i, y_i) = \|y_i - f(x_i)\|^2 . \tag{19}$$

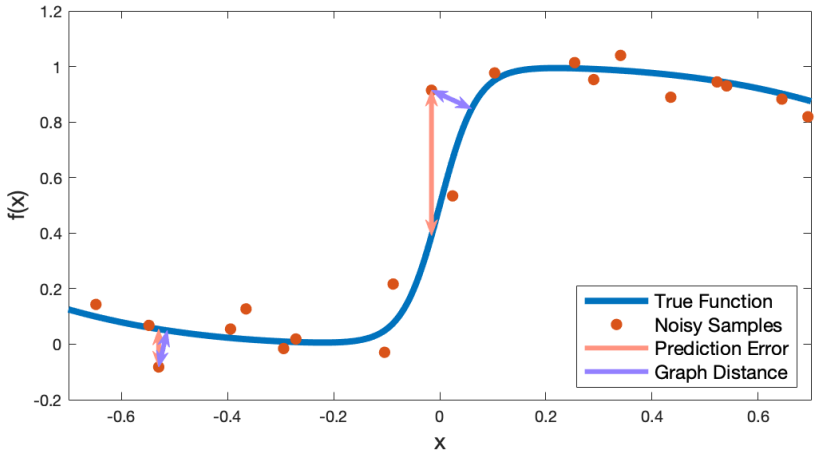

Figure 2: Illustration of graph distance versus prediction error on a function with a stiff region. For the annotated data point at left where the function is smooth, graph distance is similar to prediction error. In contrast, the annotated data point at the stiff region demonstrates a drastic disparity between these metrics.

While graph distance is a less ubiquitous performance metric than prediction loss, it has a rich history in statistical analysis through *errors-in-variables* modelling (Cifarelli, 1988), dating as far back as Adcock's 1878 method for linear regression (Adcock, 1878).

One potential reason graph distance-type losses are not more widely used is the complexity introduced in the form of an optimization problem in the loss (18). For non-stiff systems, it is not clear that this complexity is warranted; in fact, for $f^\theta(x)$ with a small Lipschitz constant, prediction and graph distance losses are equal up to a small constant (See Appendix **??** for a detailed proof):

Lemma 2 *If $f^\theta(x)$ is L_f-Lipschitz in x, $l^\theta_{\text{exp}}(x_i, y_i) \geq d_{G_\theta}(x_i, y_i)^2 \geq \frac{l^\theta_{exp}(x_i,y_i)}{1+L_f^2}$.*

However, we find that stiff, implicit models offer a unique combination of properties that make prediction loss a less suitable tool for analysis. First, as the implicit model already embeds an optimization problem in prediction loss, it no longer has a significant computational advantage over graph distance in many cases. Second, the stiffness of the model can induce a large discrepancy between prediction loss and graph distance, which scales with the Lipschitz constant L_f of $f^\theta(x)$ (Lemma 2). This can result in large prediction loss even when the model closely matches the data (see Figure 2 for an illustrative example). Finally, we have seen in Section 3 that when $L_f >> 1$, common analyses of prediction loss can fail to guarantee small generalization error, and thus cannot provide a guarantee of low test set graph distance. We now show that the better-behaved violation loss in fact can be used to tightly bound test set graph distance, even for such stiff cases.

4.2. Violation Loss as a Proxy for Graph Distance

Given that graph distance is a key metric of interest for a predictive model, and that the violation implicit loss can be shown to generalize well, we now establish conditions under which low graph distance can be guaranteed by low violation loss. For implicit models of the form in (1a)–(1b), the graph G_θ of the model is defined as

$$G_\theta = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} : \exists \lambda^* \in \Lambda,\, g^\theta(x, \lambda^*) = y \wedge h^\theta(x, y, \lambda^*) = 0 \right\}. \tag{20}$$

We will show that a relationship between the violation loss and graph distance can be established if l^{θ}_{vimp} has a *quadratic growth* behavior as defined in Karimi et al. (2016) and reproduced below.

Definition 3 *A function $f(x)$ has μ-Quadratic Growth (μ-QG) if for all x,*

$$\text{dist}\left(x, \arg\min_{\bar{x}} f(\bar{x})\right)^2 \leq \frac{2}{\mu}\left(f(x) - \min_{\bar{x}} f(\bar{x})\right). \tag{21}$$

For our purposes, we assume that $l^{\theta}_{\text{vimp}}(x_i, y_i)$ is $\mu(\epsilon)$-QG for some $\mu(\epsilon) > 0$, noting that this value depends on the ϵ in the violation loss. This condition holds for instance when g^{θ} is affine in (x, λ) and h is strictly convex, but is not limited to such strict assumptions; we will see in Section 4.3 that it holds on the toy example, despite the fact that h^{θ} in this case is non-convex and non-smooth.

To construct an inequality between violation loss and graph distance, we first prove (See Appendix **??**) the that the minimization of $l^{\theta}_{\text{vimp}}(x_i, y_i)$ is related to the model's graph G_{θ}:

Lemma 4 $\min_{x,y} l^{\theta}_{vimp}(x, y) = 0$, *and* $\arg\min_{x,y} l^{\theta}_{vimp}(x, y) = G_{\theta}$.

Lemma 4 follows directly from the definition of the model's graph G_{θ} and does not rely on any model assumptions beyond those in Section 3. However, under the assumption that l^{θ}_{vimp} has quadratic growth, this result can be extended to a comparison with graph distance:

Theorem 5 *Assume l^{θ}_{vimp} is $\mu(\epsilon)$-QG. Then for any datapoint (x_i, y_i), $d_{G_{\theta}}(x_i, y_i)^2 \leq \frac{2}{\mu(\epsilon)} l^{\theta}_{vimp}(x_i, y_i)$.*

Proof This claim follows directly from Lemma 4, as we can substitute $\arg\min_{x,y}\ l^{\theta}_{\text{vimp}}(x, y) = G_{\theta}$ and $\min_{x,y}\ l^{\theta}_{\text{vimp}}(x, y) = 0$ directly into the definition of quadratic growth. ■

Many of the properties that would allow for the violation loss optimization problem (4) to be solved, such as strong convexity, quadratic error bound, Polyak-Łojasiewicz, or Kurdyka-Łojasiewicz, are in fact equally or more strict that quadratic growth (Karimi et al., 2016). Thus while an additional assumption is required, Theorem 5 often holds when the violation loss is computationally tractable.

The inequality provided in Theorem 5 is particularly useful because it separates squared graph distance and violation loss by a constant factor *uniformly* over any possible data point. This inequality therefore is preserved under expectation, such that

$$\mathbb{E}_{[x;\, y]\sim\mathcal{D}}\left[d_{G_{\theta}}(x, y)^2\right] \leq \frac{2}{\mu(\epsilon)} \mathbb{E}_{[x;\, y]\sim\mathcal{D}}\left[l^{\theta}_{\text{vimp}}(x, y)\right].$$

Given that Section 3 guarantees l^{θ}_{vimp} generalizes well for large ϵ, guaranteeing low test-set graph distance therefore reduces to showing that l_{vimp} is $\mu(\epsilon)$-QG with both ϵ and $\mu(\epsilon)$ sufficiently large.

4.3. Graph Distance Property for Toy Problem

With derivations in Appendix **??**, l^{θ}_{vimp} is 1-QG if we select $\epsilon \leq \min\left(\frac{1}{4}, \frac{m^2}{2}\right)$. While this choice is not required, the feasible bounds for μ are $(0, 2)$, with 0 corresponding to infinite ϵ and 2 corresponding to $\epsilon = 0$. Our work provides the understanding that $\epsilon \to \infty$ means the violation implicit approach ignores optimality constraints and thus loses its relationship to graph distance, while $\epsilon = 0$ converts the formulation to the naive implicit approach, which loses tight generalization guarantees. Selecting $\mu = 1$ balances these two important characteristics, and thus can inform a selection of ϵ.

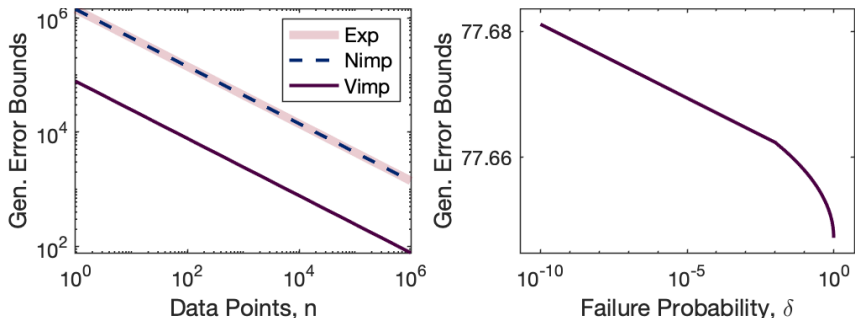

Figure 3: Generalization error bounds as a function of dataset size for the 3 approaches (left) and as a function of failure probability for the violation implicit approach (right) applied to the 1-D toy problem. The ϵ value is chosen via analysis relating the violation loss to graph distance: $\epsilon = \min(\frac{1}{4}, \frac{m^2}{2})$ which is $\frac{1}{4}$ for the values in the toy problem. The explicit and naive implicit approaches have identical bounds which require over two orders of magnitude more data to achieve the violation implicit approach's generalization error guarantees.

With $\epsilon = \min\left(\frac{1}{4}, \frac{m^2}{2}\right)$ and $m = 1$ kilogram for the specifics of the toy problem, we select $\epsilon = \frac{1}{4}$ to balance the violation implicit approach's ability to generalize and its physical meaning as a loss. The loss landscapes and bounds depicted in Figures 1 and 3, respectively, feature this value of ϵ; Figure 3 in particular still shows that the generalization of the violation implicit approach can be drastically better than the other approaches at this choice of ϵ.

5. Conclusion

We introduced and motivated three different approaches for learning a model: explicit, naive implicit, and violation implicit. We demonstrated benefits of the violation implicit approach in terms of its ability to generalize more reliably and its close relationship to graph distance, a metric that better suits functions with uncertainty in both outputs and inputs. Together, these two benefits motivate a theoretically-grounded value for the violation implicit approach's hyperparameter, ϵ, balancing better generalization with tighter relationship to graph distance. Our inelastic contact toy problem demonstrates this choice of ϵ results in significant generalization error bound improvements over either alternative approach as well as twice its loss upper bounds graph distance squared.

This paper focused on generalization error bounds, but Figure 1 also illustrates that the optimization landscape of a violation implicit loss itself is improved from that of a prediction loss. We have observed this benefit empirically from our work on contact learning (Parmar et al., 2021; Pfrommer* et al., 2020). Given typical properties associated with ease of learnability like convexity (Boyd et al., 2004) or smoothness (Karimi et al., 2016) are not commonly met by formulations of interest (including our toy problem), we seek a more general explanation in our future work.

Acknowledgments

This work was supported by the National Defense Science and Engineering Graduate Fellowship and an NSF Graduate Research Fellowship under Grant No. DGE-1845298. Toyota Research Institute also provided funds to support this work. Nikolai Matni is funded by NSF awards CPS-2038873, CAREER award ECCS-2045834, and a Google Research Scholar award.

Appendices

Appendices are included on arXiv at: `https://arxiv.org/abs/2112.06881`

References

R. J. Adcock. A problem in least squares. *Analyst*, 5:53, 1878.

Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning*, pages 136–145. PMLR, 2017.

Mihai Anitescu and Florian A Potra. Formulating dynamic multi-rigid-body contact problems with friction as solvable linear complementarity problems. *Nonlinear Dynamics*, 14(3):231–247, 1997.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Anindya Chatterjee and Andy Ruina. A new algebraic rigid-body collision law based on impulse space considerations. 1998.

Giulio Cifarelli. Measurement error models. *Journal of Applied Econometrics*, 3(4):315–317, 1988. ISSN 08837252, 10991255. URL http://www.jstor.org/stable/2096647.

Filipe de Avila Belbute-Peres, Kevin Smith, Kelsey Allen, Josh Tenenbaum, and J Zico Kolter. End-to-end differentiable physics for learning and control. *Advances in neural information processing systems*, 31:7178–7189, 2018.

Richard M Dudley. *Uniform central limit theorems*, volume 142. Cambridge university press, 2014.

Nima Fazeli, Roman Kolbert, Russ Tedrake, and Alberto Rodriguez. Parameter and contact force estimation of planar rigid-bodies undergoing frictional contact. *The International Journal of Robotics Research*, 36(13-14):1437–1454, 2017a.

Nima Fazeli, Samuel Zapolsky, Evan Drumwright, and Alberto Rodriguez. Learning data-efficient rigid-body contact models: Case study of planar impact. In *Conference on Robot Learning*, pages 388–397. PMLR, 2017b.

Manuel Fernández-Delgado, Manisha Sanjay Sirsat, Eva Cernadas, Sadi Alawadi, Senén Barro, and Manuel Febrero-Bande. An extensive experimental survey of regression methods. *Neural Networks*, 111:11–34, 2019.

Pete Florence, Corey Lynch, Andy Zeng, Oscar A Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *Conference on Robot Learning*, pages 158–168. PMLR, 2022.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

Stephen Gould, Basura Fernando, Anoop Cherian, Peter Anderson, Rodrigo Santa Cruz, and Edison Guo. On differentiating parameterized argmin and argmax problems with application to bi-level optimization. *arXiv preprint arXiv:1607.05447*, 2016.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

Shahbaz Abdul Khader, Hang Yin, Pietro Falco, and Danica Kragic. Data-efficient model learning and prediction for contact-rich manipulation tasks. *IEEE Robotics and Automation Letters*, 5(3): 4321–4328, 2020.

Svetoslav Kolev and Emanuel Todorov. Physically consistent state estimation and system identification for contacts. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pages 1036–1043. IEEE, 2015.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

Mihir Parmar, Mathew Halm, and Michael Posa. Fundamental challenges in deep learning for stiff contact dynamics. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5181–5188. IEEE, 2021.

Amir Patel, Stacey Leigh Shield, Saif Kazi, Aaron M Johnson, and Lorenz T Biegler. Contact-implicit trajectory optimization using orthogonal collocation. *IEEE Robotics and Automation Letters*, 4(2):2242–2249, 2019.

Samuel Pfrommer*, Mathew Halm*, and Michael Posa. ContactNets: Learning of Discontinuous Contact Dynamics with Smooth, Implicit Representations. In *The Conference on Robot Learning (CoRL)*, 2020. URL https://proceedings.mlr.press/v155/pfrommer21a.html.

Michael Posa, Cecilia Cantu, and Russ Tedrake. A direct method for trajectory optimization of rigid bodies through contact. *The International Journal of Robotics Research*, 33(1):69–81, 2014.

Maziar Raissi and George Em Karniadakis. Hidden physics models: Machine learning of nonlinear partial differential equations. *Journal of Computational Physics*, 357:125–141, 2018.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

David E Stewart and Jeffrey C Trinkle. An implicit time-stepping scheme for rigid body dynamics with inelastic collisions and coulomb friction. *International Journal for Numerical Methods in Engineering*, 39(15):2673–2691, 1996.

Wyatt Ubellacker, Noel Csomay-Shanklin, Tamas G Molnar, and Aaron D Ames. Verifying safe transitions between dynamic motion primitives on legged robots. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8477–8484. IEEE, 2021.

William Yang and Michael Posa. Impact invariant control with applications to bipedal locomotion. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5151–5158. IEEE, 2021.