

An Introduction to an Application of the Implicit Function Theorem

Justin Chiu
Cornell Tech
jtc257@cornell.edu

August 3, 2021

Abstract

Gradient-based learning forms the foundation of modern machine learning, and automatic differentiation allows ML practitioners to easily compute gradients. While automatic differentiation only costs a constant multiple of the time and space required to evaluate a function, it has its limitations. In particular, when evaluating a function itself is expensive, the direct application of automatic differentiation is infeasible. In this report, we review the implicit function theorem (IFT) and its use in reducing the cost of computing gradients in scenarios where function evaluation is expensive, focusing on the application of implicit differentiation to variational inference.

1 Introduction

Gradient-based learning underpins many of the recent successes in machine learning, particularly advances involving neural networks. The key to the success of gradient-based methods is automatic differentiation (AD), which greatly increases the development speed of machine learning research by allowing practitioners to circumvent the error-prone and time-consuming process of computing gradients manually. AD operates by reducing functions into compositions of atomic operations, for which we have a library of derivatives for, and composing those derivatives via the chain rule. The underlying concept behind AD is that a program's execution trace is a valid and powerful representation of a function [Griewank and Walther, 2008].

Storing the execution trace of a program allows AD systems to easily compute derivatives. However, longer execution traces can quickly consume a large amount of memory. Consider an iterative method, such as gradient descent, whose execution trace takes the form of an unrolled loop: Given an initial point $\theta = x_0$, iterates x_1, x_2, \dots, x_K are produced by running the gradient descent update for K iterations. In order to differentiate through this procedure with AD (i.e. compute $\frac{dx_K}{d\theta}$), we have to store all the x_k iterates as well the computation used to produce them. Thus, the memory complexity of storing this execution trace scales linearly in the number of iterations K as well as the dimensionality of the iterates x_k . For large K , this can be infeasible. One method for overcoming the space complexity's dependence on the number of iterations K in the above example

32 is to use the implicit function theorem (IFT), letting you throw away x_1 through x_{K-1} while still
33 being able to compute the derivative $\frac{dx_K}{d\theta}$.

34 In this report, we will cover the use of the IFT as a method for dealing with the space com-
35 plexity of AD in certain cases. We will focus on applying the IFT to differentiating the solutions to
36 optimization problems [Amos and Kolter, 2017, Agrawal et al., 2019].

37 2 Related Work

38 There are a variety of methods for overcoming the space limitations of AD, of which we only men-
39 tion three: checkpointing, reversible computation, and implicit differentiation. The first method,
40 checkpointing, improves space complexity at the cost of time [Griewank and Walther, 2008]. Rather
41 than storing all intermediate computation, checkpointing instead recomputes values when needed.
42 This can result in a large slowdown, and also requires careful choosing of which computational sub-
43 graphs to checkpoint. A second method is an improvement upon checkpointing, called reversible
44 computation [Maclaurin et al., 2015, Gomez et al., 2017], which improves space complexity at the
45 cost of expressivity, but not speed. Reversible computation ensures that the gradient with respect
46 to (wrt) the input depends only on the output, allowing the input to be discarded during function
47 evaluation. This is typically accomplished by ensuring that the input is easily reconstructed from
48 the output, restricting the expressivity of layers. A third method is implicit differentiation, which
49 potentially improves space complexity at the cost of stronger assumptions. Implicit differentiation
50 relies on the implicit function theorem (IFT), which gives conditions under which derivatives can be
51 computed independent of intermediate computation. The primary condition is the characterization
52 of the output as the solution to a system of equations.

53 One of the main applications where the space complexity of AD limits applicability is bilevel
54 optimization. Bilevel optimization problems are, as implied by the name, optimization problems
55 with another nested inner optimization problem embedded within. Methods for solving bilevel
56 optimization typically proceed iteratively. For every iteration when solving the outer optimization
57 problem, we must additionally solve an inner optimization problem.

58 The application we focus on in this report is expressing individual layers of a neural network
59 declaratively as the solution of an optimization problem [Amos and Kolter, 2017, Agrawal et al.,
60 2019, Gould et al., 2019]. This allows models to learn, without heavy manual specification, the
61 constraints of the problem in addition to the parameters of the objective. An example of this is
62 learning to play Sudoku from only input-output pairs.

63 Other applications that can be formulated as bilevel optimization problems are hyperparam-
64 eter optimization, metalearning, and variational inference. Hyperparameter optimization formulates
65 hyperparameter tuning as a bilevel optimization problem, as for each hyperparameter configura-
66 tion a new model must be trained as the inner loop [Maclaurin et al., 2015, Lorraine et al., 2019b,a,
67 Bertrand et al., 2020]. Gradients must then be propagated through the inner training loop to the outer
68 hyperparameter loop. Similarly, metalearning learns the parameters of a model such that the model
69 is able to quickly be adapted to a new task via gradient descent [Finn et al., 2017, Rajeswaran et al.,
70 2019]. This is accomplished by differentiating through the learning procedure of each new task.
71 Finally, a variant of variational inference follows a very similar format: semi-amortized variational



Figure 1: A circle, defined by the relation $a^2 + b^2 - 1 = 0$. While we cannot differentiate the relation directly, we can compute derivatives at the point a_1 using the local parameterization $f_1(a) = \sqrt{1 - b^2}$ that holds in a neighbourhood around a_1 , visualized as an arc. We cannot use the same parameterization at a_2 as the derivative is undefined. In general, the IFT is most useful in cases more complicated than the unit circle, where local parameterizations are too complex to write down.

inference (SAVI) aims to learn a model that is able to provide a good initialization for variational parameters that are subsequently updated iteratively to maximize a lower bound objective [Kim et al., 2018]. This is also accomplished by differentiating through the iterative optimization procedure applied to the variational parameters during inference.

In all the above applications, the inner-loop optimization problem is solved with an iterative method, except in rare, simple cases. The IFT reduces the memory footprint of automatic differentiation, which would otherwise be difficult to scale.

3 The Implicit Function Theorem

The implicit function theorem (IFT) has a long history, as well as many applications in a wide variety of fields such as economics and differential geometry. For an overview of the history of the IFT and some of its classical applications in mathematics and economics, see the book by Krantz and Parks [2003].

3.1 What is the IFT?

Consider the unit circle, governed by the relation $F(a, b) = a^2 + b^2 - 1 = 0$, which can be interpreted as a system of equations. As F fails the vertical line test, we cannot write b as a function of a globally. This prevents us from taking derivatives, for example $\frac{db}{da}$. However, we can use local parameterizations: $f_1(a) = \sqrt{1 - b^2}$ if $b > 0$ or $f_2(a) = -\sqrt{1 - b^2}$ if $b < 0$. Note that the local parameterizations hold only within a neighbourhood of a particular solution (a, b) . These local parameterizations then allow us to take derivatives $\frac{db}{da}$ at particular points (a, b) using the corresponding parameterization. See Fig. 1 for an illustration. The IFT generalizes this example,

and formalizes the conditions under which there exist continuous local parameterizations for a given relation or system of equations.

While the unit circle in this example has very simple local parameterizations, in general local parameterizations can be more complicated: For example, a series of iterative updates is a local parameterization. Additionally, the IFT does not give the form of the local parameterizations; it only guarantees the existence of one around a point and a way to compute its derivative. The local parameterization is left implicit, hence the ‘implicit’ in IFT.

Formally, given a system of equations $F(\theta, x) = \mathbf{0}_m$,¹ and a solution point $(\theta, x) \in \mathbb{R}^n \times \mathbb{R}^m$, the IFT gives sufficient conditions under which x can locally be written as a function of just the parameters θ within a neighbourhood of the solution point (θ, x) .² We refer to this function $x^*(\theta) = x$ as a solution mapping, and informally refer to x as a solution, as θ are parameters. These conditions are as follows:

1. We have a solution point (θ, x) that satisfies the system of equations $F(\theta, x) = 0$.
2. F has at least continuous first derivatives: $F \in \mathcal{C}^1$.
3. The Jacobian matrix of F wrt x evaluated at the solution point (θ, x) is nonsingular: $\det \frac{dF(\theta, x)}{dx} \neq 0$.

Given these conditions hold for F, θ , and x , the IFT asserts the existence of the implicit solution mapping $x^*(\theta) : \mathbb{R}^n \rightarrow \mathbb{R}^m$, and that the derivative of the solution mapping is given by $\frac{dx^*(\theta)}{d\theta} = -[\frac{dF(\theta, x)}{dx}]^{-1} \frac{dF(\theta, x)}{d\theta}$.

still searching for a good intuitive explanation.

3.2 How is the IFT useful?

As mentioned in Sec. 1, the IFT is useful for saving memory when performing automatic differentiation. The main benefit of the IFT is that once we have a solution to a system of equations, we can forget about how we obtained that solution but still compute derivatives.

In particular, we will show how to compute derivatives of the solution of an optimization problem wrt parameters of the problem without storing intermediate computations, as illustrated in Fig. 2. We will use the optimality criteria of the optimization problem to define a system of equations, then apply the IFT to compute the Jacobian of the solution wrt the parameters. This methodology allows us to use the solution to an optimization problem as the output of a layer within a neural network, as done in OptNet [Amos and Kolter, 2017].

Afterwards, we will cover an application of the IFT to speeding up variational inference.

¹ We denote vectors and matrices of all 1s and 0s by $\mathbf{1}_S$ and $\mathbf{0}_S$, where S denotes the shape, i.e. $\mathbf{0}_m \in \mathbb{R}^m$.

² Going back to the unit circle example, the local parameterizations were only valid around a particular (a, b) point. In that example, a was the parameter and b the solution.



Figure 2: An example relationship between the parameters θ , solution x_K after K iterations of an iterative method, and implicit function $x^*(\theta)$ of the IFT. The rectangle depicts a space which, in this example, contains both θ and x_K . This is not necessary for the IFT, but simplifies illustration. The parameters θ provide an initial point, which is then iteratively refined into solution x_K , shown by the squiggly line. If x_K satisfies the conditions of the IFT, then the IFT both guarantees the existence of the implicit solution mapping $x_K = x^*(\theta)$ (dashed line) and tells us how to compute $\frac{dx^*(\theta)}{d\theta}$. This is useful if the iterative procedure (squiggly line), is too expensive to store in memory for use in automatic differentiation.

4 Embedding Optimization inside a Neural Network

As an introductory example, we will replace the softmax layer of a neural network with an equivalent function defined as the output of an optimization problem, then derive derivatives using the IFT. We will start by reviewing softmax and its expression as an optimization problem. After checking the conditions of the IFT hold, we can then compute derivatives. Since the Jacobian of softmax is known, we can directly verify that the IFT gives the correct answer.

4.1 Learning via the chain rule

Before we get into the details of softmax, we first motivate the learning problem. Neural networks consist of stacks of layers, all of which take the output of the previous layer as input. In order to train neural networks we rely on reverse mode AD, or backpropagation. Reverse mode AD propagates error information backward through the network via the chain rule. Key to this process is the Jacobian matrix of each layer, the matrix of partial derivatives of the layer's outputs wrt its inputs.

Given a function $F : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^m$, we denote the Jacobian evaluated at the point $(\theta, x) \in \mathbb{R}^n \times \mathbb{R}^m$ as

$$\frac{dF(\theta, x)}{d(\theta, x)} = \begin{bmatrix} \frac{dF(\theta, x)}{d\theta} & \frac{dF(\theta, x)}{dx} \end{bmatrix},$$

where we have the matrix of partial derivatives

$$\frac{dF(\theta, x)}{d\theta} = \begin{bmatrix} \frac{dF_1(\theta, x)}{d\theta_1} & \dots & \frac{dF_1(\theta, x)}{d\theta_n} \\ \vdots & \ddots & \vdots \\ \frac{dF_m(\theta, x)}{d\theta_1} & \dots & \frac{dF_m(\theta, x)}{d\theta_n} \end{bmatrix},$$

and similarly for $\frac{dF(\theta, x)}{dx} \in \mathbb{R}^{m \times m}$.

In order to embed an optimization layer inside a neural network, we must provide the Jacobian of our layer for use in an AD system.³ With this in mind, we proceed to review softmax and its equivalent formulation as an optimization problem.

4.2 Softmax

Softmax is often used to parameterize categorical distributions within neural networks, such as in attention layers. It has its origins in statistical mechanics and decision theory, and functions as a differentiable surrogate for argmax.

Softmax assumes that we have n items with independent utilities, $\theta \in \mathbb{R}^n$, which indicate preferences. Softmax then gives the following distribution over items: $z_i = \frac{\exp(\theta_i)}{\sum_j \exp(\theta_j)}$, with $z \in \mathbb{R}^n$. While there is a closed-form equation for both softmax and its Jacobian, we use it as an introduction to the mechanism behind OptNet (and other differentiable optimization layers) [Amos and Kolter, 2017, Agrawal et al., 2019].

The output of softmax is the solution of the following optimization problem [Gao and Pavel, 2018]:

$$\begin{aligned} & \text{maximize} && z^\top \theta + H(z) \\ & \text{subject to} && z^\top \mathbf{1} = 1 \\ & && z_i \geq 0, \forall i, \end{aligned} \tag{1}$$

where $H(z) = -\sum_i z_i \log z_i$ is the entropy.⁴ We will refer to this as the softmax problem.⁵

Our goal is to compute the Jacobian of softmax $\frac{dz}{d\theta} = \frac{d\text{softmax}(\theta)}{d\theta}$ using the IFT and the optimization problem above, as an exercise. Applying the IFT to optimization problems consists of four steps:

1. Find a solution to the optimization problem.
2. Write down the system of equations.
3. Check that the conditions of the IFT hold.
4. Compute the derivative of the implicit solution mapping wrt the parameters.

³ This is a slight simplification. We often need to provide at least the vector-Jacobian product.

⁴ We have taken care to use the variable z for the primal variable of the softmax problem, rather than x . This is because, as we will soon see, the solution point of the resulting system of equations must also include dual variables.

⁵ Removing the entropy regularization term results in an optimization problem whose solution is argmax: maximize $z^\top \theta$, subject to $z^\top \mathbf{1}_n = 1$ and $z \succeq 0$.

157 We assume the first step has been done for us, and we have some solution z to the softmax
 158 problem. We will then use the IFT to compute gradients of z wrt the parameters θ by following the
 159 rest of the steps.

160 4.3 The KKT conditions determine F

161 Given an optimization problem, the Karush-Kuhn-Tucker (KKT) conditions determine a system of
 162 equations that the solution must satisfy, i.e the optimality criteria [Karush, 1939, Kuhn and Tucker,
 163 1951]. They are, roughly, stationarity (the gradient should be 0 at a local optima) and feasibility
 164 (the constraints of the problem should not be violated). For a thorough introduction to the KKT
 165 conditions, see chapter 5 of Boyd and Vandenberghe [2004] or the Wikipedia article.⁶

We will use the KKT conditions of the softmax problem in Eqn. 1 to determine the function F in the IFT. First, we introduce dual variables $u \in \mathbb{R}, v \in \mathbb{R}^n$ and write out the Lagrangian:

$$\mathcal{L}(\theta, z, u, v) = z^\top \theta + H(z) + u(z^\top \mathbf{1} - 1) + v^\top z.$$

We therefore have the solution point (θ, z, u, v) , with parameters θ and solution $x = (z, u, v)$. We then have the following necessary conditions for a solution (z, u, v) , i.e. the KKT conditions:

$$\begin{aligned} \frac{d}{dx} \mathcal{L}(\theta, z, u, v) &= \mathbf{0}_n \quad (\text{stationarity}) \\ u(z^\top \mathbf{1} - 1) &= 0 \quad (\text{primal feasibility, equality}) \\ \text{diag}(v)z &= \mathbf{0}_n \quad (\text{complementary slackness}) \\ z &\succeq \mathbf{0}_n \quad (\text{primal feasibility, inequality}) \\ v &\succeq \mathbf{0}_n \quad (\text{dual feasibility}) \end{aligned} \tag{2}$$

166 As we only need a system of equations with $2n + 1$ equations to determine the $2n + 1$ solution
 167 variables $x = (z, u, v)$, we use the first three conditions: stationary, primal feasibility (equality),
 168 and complementary slackness.

In full, the system of equations $F(\theta, z, u, v) = 0$ is

$$\begin{aligned} \theta + -\log(z) - 1 + u\mathbf{1}_n + v &= \mathbf{0}_n \\ u(z^\top \mathbf{1}_n - 1) &= 0 \\ \text{diag}(v)z &= \mathbf{0}_n. \end{aligned} \tag{3}$$

169 Note that the first and third equations are vector-valued.

170 Now we can check the conditions of the IFT. We assumed we were given an optimal and there-
 171 fore feasible solution $x = (z, u, v)$ which must satisfy $F(\theta, z, u, v) = 0$,⁷ and F has continuous first
 172 derivatives, i.e. $F \in \mathcal{C}^1$. All that remains is to check that the Jacobian matrix $\frac{dF(\theta, z, u, v)}{d(z, u, v)}$ (evaluated
 173 at the solution point) is non-singular.

⁶ If this is your first time seeing the KKT conditions, please check out one of the suggested references and make your way back here to flex your new knowledge.

⁷ Recall that softmax also has a closed form expression.

The Jacobian $\frac{dF(\theta, z, u, v)}{d(z, u, v)} \in \mathbb{R}^{2n+1 \times 2n+1}$ is given by

$$\frac{dF}{d(z, u, v)} = \begin{bmatrix} \text{diag}(z)^{-1} & -\mathbf{1}_n & -I_{n \times n} \\ u\mathbf{1}_n^\top & z^\top \mathbf{1}_n - 1 & 0 \\ \text{diag}(v) & 0 & \text{diag}(z) \end{bmatrix}. \quad (4)$$

Since a solution must be feasible, we know that $z^\top \mathbf{1} - 1 = 0$ and $u > 0$. However, the upper left block, $\text{diag}(z)^{-1}$, contains a divide-by-zero term if any component $z_i = 0$.⁸ To avoid this, we consider only strictly positive $z \succ 0$ for the IFT.⁹ Given the strict positivity of z , we can deduce that the Jacobian of F is full rank and therefore has nonzero determinant. This shows that the conditions of the IFT hold.

4.4 The Jacobian of Softmax

Now that we have shown that the conditions of the IFT hold, we can proceed to apply the second part of the IFT in order to compute $\frac{dz}{d\theta}$. Recall that we have the solution $x = (z, u, v)$; we will switch to x for brevity. The second part of the IFT tells us that we can compute the Jacobian of the solution mapping $\frac{dx}{d\theta} = \frac{dx^*(\theta)}{d\theta} = \left[\frac{dF(\theta, x)}{dx} \right]^{-1} \frac{dF(\theta, x)}{d\theta}$, then pick out the relevant components.

The second term, $\frac{dF(\theta, x)}{d\theta}$, is simple. Since θ only appears in the first vector-valued function of F (see Eqn. 3), we have

$$\frac{dF(\theta, x)}{d\theta} = \begin{bmatrix} I_{n \times n} \\ \mathbf{0}_{(n+1) \times (n+1)} \end{bmatrix}. \quad (5)$$

The large amount of sparsity allows us to skip some computation further down.¹⁰

Next, we have to invert the Jacobian from Eqn. 4:

$$\left[\frac{dF(\theta, x)}{dx} \right]^{-1} = \begin{bmatrix} \text{diag}(z)^{-1} & -\mathbf{1}_n & -I_{n \times n} \\ u\mathbf{1}_n^\top & x^\top \mathbf{1}_n - 1 & 0 \\ \text{diag}(v) & 0 & \text{diag}(z) \end{bmatrix}^{-1}. \quad (6)$$

The remainder of this section is compute-intensive; feel free to skip ahead to Sec. 4.5 for a discussion on the limitations of applying the IFT. We use the block-wise inversion formula

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & 0 \\ 0 & (D - CA^{-1}B)^{-1} \end{bmatrix} \begin{bmatrix} I & -BD^{-1} \\ -CA^{-1} & I \end{bmatrix},$$

where

$$\begin{aligned} A &= \begin{bmatrix} \text{diag}(z)^{-1} & -\mathbf{1}_n \\ u\mathbf{1}_n^\top & 0 \end{bmatrix} & B &= \begin{bmatrix} -I_{n \times n} \\ 0 \end{bmatrix} \\ C &= [\text{diag}(v) \quad 0] & D &= \text{diag}(z). \end{aligned}$$

⁸ This term was obtained by differentiating the entropy term of the Lagrangian, $H(z) = \sum_i z_i \log z_i$. While we could use the convention $0 \log 0 = 0$, this does not fix the divide-by-zero issue with the second derivative, which we see here.

⁹ We saw a similar issue in the unit circle example, where the derivative $\frac{db}{da}$ was undefined when $a = 0$ (see Fig. 1).

¹⁰ This sparsity is due to the simple constraints in the softmax problem, which is no longer available in more general optimization problems.

However, by complementary slackness, we have $v = 0$,¹¹ reducing the above to

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix} \begin{bmatrix} I_{(n+1) \times (n+1)} & -BD^{-1} \\ 0 & I_{n \times n} \end{bmatrix}.$$

As we are interested in computing $\frac{dz}{d\theta}$, rather than the full derivative $\frac{dx}{d\theta}$ (recall $x = (z, u, v)$), in addition to the sparsity of $\frac{dF}{d\theta}$, we only have to solve for the upper-left $n \times n$ block of $A^{-1} \in \mathbb{R}^{(n+1) \times (n+1)}$. To do so, we will repeat the same block-wise inverse computation. Let us denote

$$A = \begin{bmatrix} \text{diag}(z)^{-1} & -\mathbf{1}_n \\ u\mathbf{1}_n^\top & 0 \end{bmatrix} = \begin{bmatrix} E & F \\ G & H \end{bmatrix}.$$

First, we compute the Schur complement of A ,

$$A/E = H - GE^{-1}F = 0 + u\mathbf{1}_n^\top \text{diag}(z)\mathbf{1}_n = uz^\top \mathbf{1}_n. \quad (7)$$

Since z is feasible, we have $A/E = u$ due to the equality constraints (z must sum to 1 as a probability mass function). Then, we have

$$A^{-1} = \begin{bmatrix} \text{diag}(z)^{-1} & -\mathbf{1}_n \\ u\mathbf{1}_n^\top & 0 \end{bmatrix}^{-1} = \begin{bmatrix} E & F \\ G & H \end{bmatrix}^{-1} = \begin{bmatrix} E^{-1} + E^{-1}F(A/E)^{-1}GE^{-1} & -E^{-1}F(A/E)^{-1} \\ -(A/E)^{-1}GE^{-1} & (A/E)^{-1} \end{bmatrix}. \quad (8)$$

Plugging in,

$$\begin{aligned} A^{-1} &= \begin{bmatrix} \text{diag}(z) - \text{diag}(z)\mathbf{1}_n u^{-1} u\mathbf{1}_n^\top \text{diag}(z) & \text{diag}(z)\mathbf{1}_n u^{-1} \\ -u^{-1} u\mathbf{1}_n^\top \text{diag}(z) & u^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \text{diag}(z) - zz^\top & u^{-1}z \\ -z^\top & u^{-1} \end{bmatrix}. \end{aligned} \quad (9)$$

185 Pulling out the top-left $n \times n$ block yields the Jacobian $\frac{dz}{d\theta} = \text{diag}(z) - zz^\top$, which agrees with
 186 directly differentiating softmax [Martins and Astudillo, 2016].

187 Given the formulation of softmax as an optimization problem, we have successfully shown how
 188 to differentiate the output of softmax wrt the parameters in a solver-agnostic manner using the IFT.
 189 This concludes the exercise of using the IFT to differentiate through the softmax problem.

190 4.5 Limitations

191 In order to compute the derivative $\frac{dx}{d\theta}$, we had to invert the Jacobian of F . However, The first part
 192 of F (recall from Eqn. 3), the stationarity condition $\frac{d}{dx}\mathcal{L} = 0$, already involved the Jacobian of the
 193 Lagrangian \mathcal{L} . In general, this means that in order to apply the IFT to solutions of optimization prob-
 194 lems, we must compute the inverse Hessian of the Lagrangian (or at least a Hessian-vector-product).
 195 The Hessian is a matrix of size $O(n^2)$, and inverting this would take $O(n^3)$ computation.¹² Thank-
 196 fully, there are relatively cheap ways of approximating this computation, such as with approximate
 197 (inverse) Hessian-vector-product techniques [Rajeswaran et al., 2019, Lorraine et al., 2019a].

¹¹ We apply the IFT to solutions where $z \succ 0$ due to a divide-by-zero issue in $\frac{d^2 H(z)}{dz_i^2} = \frac{1}{z_i}$.

¹² The number of solution variables scales with the number of primal variables, but also the number of constraints. This potentially makes applying the IFT to optimization problems with exponentially many constraints difficult.

4.6 Extensions

The methods we covered can be extended to variations of argmax problems other than the softmax problem.¹³ The softmax problem altered the argmax problem by introducing entropy regularization. Rather than regularizing with entropy, one could instead alter the objective to find the Euclidean projection of the parameters onto the probability simplex, resulting in SparseMax [Martins and Astudillo, 2016]. While the output of softmax variants often have a closed form expression, the IFT provides another way of deriving their Jacobians and could potentially pave the way for differentiating through argmax problems that do not have closed-form expressions.

More generally, the IFT can be applied to cases where, unlike softmax, we do not have an explicit functional form (i.e., the unit circle), and outputs are governed only by a system of equations. This includes more general optimization problems, such as quadratic programs [Amos and Kolter, 2017] or other convex optimization problems [Agrawal et al., 2019].

5 OptNet

OptNet generalizes the methodology applied above to the softmax problem by extending the optimization problems considered, in particular including parameterized constraints. This allows us to learn not only the objective, but also the constraints. Explicitly incorporating families of constraints in models with optimization layers allows them to perform well on tasks with rigid constraints, such as learning to play Sudoku from only inputs and outputs [Amos and Kolter, 2017].

5.1 Quadratic Programs

OptNet applies the IFT to quadratic programs (QPs) in particular. As the simplest nonlinear optimization problem, QPs strike a balance between expressivity and computational tractability [Frank and Wolfe, 1956]. The methodology remains the same as the softmax problem: Given a QP and a solution, use the KKT conditions to produce a system of equations then apply the IFT / implicit differentiation to compute the derivative of the solution wrt the parameters of the objective and constraints.

Quadratic programs take the following form:

$$\begin{aligned} & \text{maximize} && \frac{1}{2}x^\top Qx + q^\top x \\ & \text{subject to} && Ax = b \\ & && Gx \leq h, \end{aligned} \tag{10}$$

where we optimize over x and the parameters are $\theta = \{Q, q, A, b, G, h\}$. Compared to the softmax problem in Eqn. 1, we see that the main difference is the learnable parameters in the constraints. As the application of the IFT is almost identical, we will not cover it in detail.

¹³ The argmax problem is given by maximize $z^\top \theta$, subject to $z^\top \mathbf{1}_n = 1$ and $z \succeq 0$.

The second contribution of OptNet was the extension of a state-of-the-art interior point solver [Amos and Kolter, 2017], and its adaptation to parallel machines (GPUs) and batch processing. While outside the scope of this report, see the paper by Amos and Kolter [2017] for the details.

6 Semi-Amortized Variational Inference (POSTPONED / dont read)

We now apply the IFT to variational inference.

Variational inference has found success in recent applications to generative models, in particular by allowing practitioners to depart from conjugate models and extend emission models with expressive neural network components. The main insight that led to this development is that inference can be amortized through the use of an inference network. One approach to variational inference, stochastic variational inference (SVI), introduces local, independent variational parameters for every instance of hidden variable. While flexible, the storage of all variational parameters is expensive, and the optimization of each parameter independently slow []. Amortized variational inference (AVI) solves that by instead using a hierarchical process. Variational parameters are produced hierarchically via an inference network, which in turn generates the local variational parameters []. The resulting local parameters may or may not be subsequently optimized.

Failure to further optimize local variational parameters may result in an amortization gap []. Prior work has shown that this gap can be ameliorated by performing a few steps of optimization on the generated local parameters obtained from the inference network, and even by propagating gradients through the optimization process. Optimizing through the inner optimization problem results in semi-amortized variational inference (SAVI) [].

As our main motivating example, we will examine whether we can apply the IFT to SAVI. We will start by formalizing the problem of variational inference for a simple model.

We will start with a model defined by the following generative process, used by Dai et al. [2019] to analyze posterior collapse:

1. Choose a latent code from the prior distribution $z \sim p(z) = N(0, I)$.

2. Given the code, choose an observation from the emission distribution $x | z \sim p_\theta(x | z) = N(\mu_x(z, \theta), \gamma I)$,

where $\mu_x(z, \theta) \equiv \text{MLP}(z, \theta)$ and $\gamma > 0$ is a hyperparameter. This yields the joint distribution $p(x, z) = p(x | z)p(z)$.

Since the latent code z is unobserved, training this model would require optimizing the evidence $p(x) = \int p(x, z)$. However, due to the MLP parameterized μ_x , the integral is intractable. Variational inference performs approximate inference by introducing variational distribution $q_\phi(z | x)$ and maximizing the following lower bound on $\log p(x)$:

$$\log p(x) - D_{\text{KL}}[q(z | x) || p(z | x)] = \mathbb{E}_{q_\phi(z|x)} \left[\log \frac{p_\theta(x, z)}{q_\phi(z | x)} \right] = L(\theta, \phi). \quad (11)$$

(Write out objective in full.)

While SVI introduces local parameters for each instance of z , and AVI uses a single $q(z | x)$ for all instances, we will follow the approach of SAVI. We will perform inference as follows: For each instance x , produce local variational parameter $z^{(0)} = g(x; \phi)$. Obtain z^* by solving $\mathcal{L}(\theta, z^{(0)}) = 2$, with (local) optima ℓ^* . Take gradients through the whole procedure, i.e. compute $\frac{\partial \ell^*}{\partial \phi} = \frac{\partial \ell^*}{\partial z^*} \frac{\partial z^*}{\partial z^{(0)}} \frac{\partial z^{(0)}}{\partial \lambda}$. The main difficulty lies in computing $\frac{\partial z^*}{\partial z^{(0)}}$. (Highlight challenge)

In order to avoid the memory costs of storing all intermediate computation performed in a solver, we will instead apply the IFT. In order to apply the IFT, we must satisfy the three conditions. First, we must have a solution point to a system of equations, $F(x_0, z_0) = 0$. In this setting, we will use the KKT conditions of the optimization problem to define F .

7 Limitations

References

- A. Agrawal, B. Amos, S. T. Barratt, S. P. Boyd, S. Diamond, and J. Z. Kolter. Differentiable convex optimization layers. *CoRR*, abs/1910.12430, 2019. URL <http://arxiv.org/abs/1910.12430>.
- B. Amos and J. Z. Kolter. Optnet: Differentiable optimization as a layer in neural networks. *CoRR*, abs/1703.00443, 2017. URL <http://arxiv.org/abs/1703.00443>.
- Q. Bertrand, Q. Klopfenstein, M. Blondel, S. Vaiter, A. Gramfort, and J. Salmon. Implicit differentiation of lasso-type models for hyperparameter optimization, 2020.
- S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004. ISBN 0521833787. URL <http://www.amazon.com/exec/obidos/redirect?tag=citeulike-20&path=ASIN/0521833787>.
- B. Dai, Z. Wang, and D. P. Wipf. The usual suspects? reassessing blame for VAE posterior collapse. *CoRR*, abs/1912.10702, 2019. URL <http://arxiv.org/abs/1912.10702>.
- C. Finn, P. Abbeel, and S. Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *CoRR*, abs/1703.03400, 2017. URL <http://arxiv.org/abs/1703.03400>.
- M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956. URL <https://EconPapers.repec.org/RePEc:wly:navlog:v:3:y:1956:i:1-2:p:95-110>.
- B. Gao and L. Pavel. On the properties of the softmax function with application in game theory and reinforcement learning, 2018.
- A. N. Gomez, M. Ren, R. Urtasun, and R. B. Grosse. The reversible residual network: Backpropagation without storing activations. *CoRR*, abs/1707.04585, 2017. URL <http://arxiv.org/abs/1707.04585>.

- 289 S. Gould, R. Hartley, and D. Campbell. Deep declarative networks: A new hope. *CoRR*,
290 abs/1909.04866, 2019. URL <http://arxiv.org/abs/1909.04866>.
- 291 A. Griewank and A. Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic*
292 *Differentiation*. SIAM, Philadelphia, 2nd edition, 2008.
- 293 W. Karush. *Minima of functions of several variables with inequalities as side conditions*. PhD thesis,
294 Thesis (S.M.)—University of Chicago, Department of Mathematics, December 1939., 1939.
- 295 Y. Kim, S. Wiseman, A. C. Miller, D. Sontag, and A. M. Rush. Semi-amortized variational autoen-
296 coders, 2018.
- 297 S. Krantz and H. Parks. The implicit function theorem : History, theory, and applications / s.g.
298 krantz, h.r. parks. 01 2003. doi: 10.1007/978-1-4612-0059-8.
- 299 H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley*
300 *Symposium on Mathematical Statistics and Probability, 1950*, pages 481–492, Berkeley and Los
301 Angeles, 1951. University of California Press.
- 302 J. Lorraine, P. Vicol, and D. Duvenaud. Optimizing millions of hyperparameters by implicit differ-
303 entiation. *CoRR*, abs/1911.02590, 2019a. URL <http://arxiv.org/abs/1911.02590>.
- 304 J. Lorraine, P. Vicol, and D. Duvenaud. Optimizing millions of hyperparameters by implicit differ-
305 entiation. *CoRR*, abs/1911.02590, 2019b. URL <http://arxiv.org/abs/1911.02590>.
- 306 D. Maclaurin, D. Duvenaud, and R. P. Adams. Gradient-based hyperparameter optimization through
307 reversible learning, 2015.
- 308 A. F. T. Martins and R. F. Astudillo. From softmax to sparsemax: A sparse model of attention and
309 multi-label classification. *CoRR*, abs/1602.02068, 2016. URL <http://arxiv.org/abs/1602.02068>.
- 311 A. Rajeswaran, C. Finn, S. M. Kakade, and S. Levine. Meta-learning with implicit gradients. *CoRR*,
312 abs/1909.04630, 2019. URL <http://arxiv.org/abs/1909.04630>.

313 **A Example Appendix**

314 Neural ODEs use reversibility.