Swamps of Numerical Integration
Monte Carlo Heights
Normalizing Flow
Unrolling Hills of Time
Backprop Bay
Forward-Backward Beach


2020
Swamps of Numerical Integration
Monte Carlo Heights
Normalizing Flow
Unrolling Hills of Time
Backprop Bay
Forward-backward Beach
Differentiation
Vale of Implicit

Implicit differentiation

Cheng Soon Ong
Marc Peter Deisenroth

December 2020

Motivation

- In machine learning, we use gradients to train predictors
- For functions $f(\boldsymbol{x})$ we can directly obtain its gradient $\nabla f(\boldsymbol{x})$
- **How to represent a constraint?**

$$G(x, y) = 0$$

E.g. conservation of mass

High school gradients

This is an equation

$$y = x^3 + 2x^2 + x + 4$$

What is the gradient $\frac{dy}{dx}$?

High school gradients

This is an equation

$$y = x^3 + 2x^2 + x + 4$$

What is the gradient $\frac{dy}{dx}$?

$$\frac{dy}{dx} = 3x^2 + 4x + 1$$

High school gradients

This is an equation

$$y = x^3 + 2x^2 + x + 4$$

What is the gradient $\frac{dy}{dx}$?

$$\frac{dy}{dx} = 3x^2 + 4x + 1$$

Observe that we can write the equation as

$$x^3 + 2x^2 + x + 4 - y = 0$$

which is of the form $G(x, y) = 0$.

Optimization with constraints

Given a constrained continuous optimization problem

$$\text{min}_{x,y} F(x,y)$$
$$\text{subject to } G(x,y) = 0$$

We can solve the equality constraint to get

$$y = g(x)$$

and substitute into the objective $F(x, g(x))$, and calculate the gradient

$$\frac{d}{dx} F(x, g(x))$$

Advanced high school gradients

This is an equation

$$y = x^3 + 2x^2 + xy + 4$$

What is the gradient $\frac{dy}{dx}$?

Advanced high school gradients

This is an equation

$$y = x^3 + 2x^2 + xy + 4$$

What is the gradient $\frac{dy}{dx}$?

Implicit function theorem

- Solve for $y = g(x)$ and use quotient rule
- Directly differentiate, and use product rule

Function of two variables

Consider a function

$$G(x, y) = 0$$

where $x, y \in \mathbb{R}$.
Assume that near a particular point x_0, we can write a closed form expression for y in terms of x, that is

$$y = g(x)\,.$$

Solve for one variable

Substituting $y = g(x)$ into $G(x, y)$, near x_0, we get

$$G(x, g(x)) = 0\,.$$

We calculate the derivative of G with respect to x using the **chain rule**,

$$G_x(x, g(x)) + G_y(x, g(x)) \cdot g'(x) = 0\,.$$

If $G_y \neq 0$, then

$$g'(x) = -\frac{G_x}{G_y}.$$

Implicit function theorem

The implicit function theorem **provides conditions** under which we can write $G(x, y) = 0$ as $y = g(x)$, and also conditions when $g'(x) = -\frac{G_x}{G_y}$.

Hand wavy intuition

Given three variables (could be any topology) x, y, z and

$$G(x, y) = z \,,$$

What are the conditions for the following to be well behaved?

$$\begin{aligned} g(x) &= y \\ G(x, g(x)) &= z \,. \end{aligned}$$

Implicit function theorem as *ansatz*

Ansatz

1. Explicitly solve one variable in terms of another
2. Chain the gradients

- Ansatz = a way to look at problems
- The implicit function theorem is a way to solve equations

Generalizations depending on where equations live (Krantz and Parks, 2013):

- Inverse function theorem
- Constant rank theorem
- Banach fixed point theorem
- Nash-Moser theorem

Pointers to literature

- From linear algebra to calculus (Hubbard and Hubbard, 2015; Spivak, 2008)
- Implicit function theorem from variational analysis (Dontchev and Rockafellar, 2014)
- Many versions of implicit function theorem (Krantz and Parks, 2013)
- Deep Declarative Networks
  `https://anucvml.github.io/ddn-cvprw2020/`

Augustin-Louis Cauchy, 1916

LA MÉCANIQUE CÉLESTE

ET SUR

UN NOUVEAU CALCUL APPELÉ CALCUL DES LIMITES (').

Gradient of loss w.r.t. optimal value

- Consider the problem of structured prediction (Nowozin et al., 2014)
- Let a sample be $(\boldsymbol{x}_n, \boldsymbol{y}_n)$
- Energy given the parameter of the learner $\boldsymbol{w}$ is $E(\boldsymbol{y}_n, \boldsymbol{x}_n, \boldsymbol{w})$
- We assume that the best predictor is found by an optimization algorithm over $\boldsymbol{y}$,

$$\boldsymbol{y}^*(\boldsymbol{x}_n, \boldsymbol{w}) = \mathrm{opt}_{\boldsymbol{y}} E(\boldsymbol{y}, \boldsymbol{x}_n, \boldsymbol{w}) \,.$$

- Measure the error with $\ell(\boldsymbol{y}, \boldsymbol{y}_n)$
- Loss per sample $L(\boldsymbol{x}_n, \boldsymbol{y}_n, \boldsymbol{w})$ (assume predict with $\boldsymbol{y}^*$)
- **Want to take the gradient of loss L w.r.t. paramters w**
- but have an optimization problem inside (to find $\boldsymbol{y}^*$)

Gradient of L w.r.t. w

By implicit differentation (Do et al., 2007; Samuel and Tappen, 2009), the gradient of the loss $L(\boldsymbol{x}_n, \boldsymbol{y}_n, \boldsymbol{w})$ with respect to the parameters $\frac{dL}{d\boldsymbol{w}}$ has a closed form.

Theorem

Let $\boldsymbol{y}^*(\boldsymbol{w}) =$ *argmin*$_{\boldsymbol{y}}\, E(\boldsymbol{y}, \boldsymbol{w})$, *and* $L(\boldsymbol{w}) = \ell(\boldsymbol{y}^*(\boldsymbol{w}))$. *Then*

$$\frac{dL}{d\boldsymbol{w}} = -\frac{\partial^2 E}{\partial \boldsymbol{w} \partial \boldsymbol{y}^\top} \left(\frac{\partial^2 E}{\partial \boldsymbol{y} \partial \boldsymbol{y}^\top} \right)^{-1} \frac{d\ell}{d\boldsymbol{y}} .$$

Sketch: chain rule

$$L(\boldsymbol{w}) = \ell\left(\mathsf{opt}_{\boldsymbol{y}} E(\boldsymbol{y}, \boldsymbol{w})\right)$$

By the chain rule,

$$\frac{\partial L}{\partial \boldsymbol{w}} = \frac{\partial \ell}{\partial \boldsymbol{y}} \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{w}}\,.$$

Sketch: gradient with respect to energy $\boldsymbol{E}$

Denote by

$$g(\boldsymbol{y}, \boldsymbol{w}) = \frac{\partial E(\boldsymbol{y}, \boldsymbol{w})}{\partial \boldsymbol{y}}$$

The gradient of g with respect to $\boldsymbol{w}$ is given by the chain rule again. Note that $\boldsymbol{y}$ is actually a function of $\boldsymbol{w}$, i.e. $g(\boldsymbol{y}(\boldsymbol{w}), \boldsymbol{w})$.

$$\frac{\partial}{\partial \boldsymbol{w}} g(\boldsymbol{y}(\boldsymbol{w}), \boldsymbol{w}) = \frac{\partial g}{\partial \boldsymbol{y}} \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{w}} + \frac{\partial g}{\partial \boldsymbol{w}} \, .$$

Sketch: Stationarity conditions

At optimality of $E(\boldsymbol{y}, \boldsymbol{w})$, its gradient is zero, i.e. $g(\boldsymbol{y}, \boldsymbol{w}) = 0$. Solving

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{w}} = -\frac{\partial g}{\partial \boldsymbol{w}} \left(\frac{\partial g}{\partial \boldsymbol{y}} \right)^{-1}$$

Substituting $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{w}}$ into $\frac{\partial L}{\partial \boldsymbol{w}}$,

$$\frac{\partial L}{\partial \boldsymbol{w}} = -\frac{\partial g}{\partial \boldsymbol{w}} \left(\frac{\partial g}{\partial \boldsymbol{y}} \right)^{-1} \frac{\partial \ell}{\partial \boldsymbol{y}} .$$

Sketch: Substitute back to obtain $\frac{\partial L}{\partial w}$

Recall the definition of g as gradient of energy E,

$$g(\boldsymbol{y}, \boldsymbol{w}) = \frac{\partial E(\boldsymbol{y}, \boldsymbol{w})}{\partial \boldsymbol{y}}$$

And hence we get the Hessian with respect to the energy

$$\frac{\partial L}{\partial \boldsymbol{w}} = -\frac{\partial^2 E}{\partial \boldsymbol{w} \partial \boldsymbol{y}^\top} \left(\frac{\partial^2 E}{\partial \boldsymbol{y} \partial \boldsymbol{y}^\top} \right)^{-1} \frac{d\ell}{d\boldsymbol{y}} .$$

Result: Gradient of L w.r.t. w

The gradient of $L(\boldsymbol{w}) = \ell(\text{argmin}_{\boldsymbol{y}} E(\boldsymbol{y}(\boldsymbol{w}), \boldsymbol{w}))$ with respect to w has a closed form.

Theorem

Let $\boldsymbol{y}^*(\boldsymbol{w}) =$ *argmin*$_{\boldsymbol{y}}\, E(\boldsymbol{y}, \boldsymbol{w})$, *and* $L(\boldsymbol{w}) = \ell(\boldsymbol{y}^*(\boldsymbol{w}))$. *Then*

$$\frac{dL}{d\boldsymbol{w}} = -\frac{\partial^2 E}{\partial \boldsymbol{w} \partial \boldsymbol{y}^\top} \left(\frac{\partial^2 E}{\partial \boldsymbol{y} \partial \boldsymbol{y}^\top} \right)^{-1} \frac{d\ell}{d\boldsymbol{y}}.$$

Domke (2012)

Summary

- We want to take the gradient with respect to an equality constraint
- Implicit function theorem gives conditions where we can "invert" a derivative
- Implicit function theorem as a way to solve equations
- Useful to take a gradient over an optimum

Backpropagation is just...

the implicit function theorem

References

Do, C. B., Foo, C.-S., and Ng, A. (2007). Efficient multiple hyperparameter learning for log-linear models. In *Neural Information Processing Systems*.

Domke, J. (2012). Generic methods for optimization-based modeling. In *AISTATS*.

Dontchev, A. L. and Rockafellar, R. T. (2014). *Implicit Functions and Solution Mappings: A View from Variational Analysis*. Springer, 2nd edition.

Hubbard, J. H. and Hubbard, B. B. (2015). *Vector Calculus, Linear Algebra, and Differential Forms: A Unified Approach*. Matrix Editions, 5th edition.

Krantz, S. G. and Parks, H. R. (2013). *The Implict Function Theorem: History, Theory, and Applications*. Springer.

Nowozin, S., Gehler, P. V., Jancsary, J., and Lampert, C. H. (2014). *Advanced Structured Prediction*. MIT Press.

References (cont.)

Samuel, K. and Tappen, M. (2009). Learning optimized map estimates in continuously-valued MRF models. In *CVPR*.

Spivak, M. (2008). *Calculus*. Publish or Perish, Inc., 4th edition.