Smooth Relaxation of Discrete Structure-Induced Discontinuities

Anton Kovalov

08.12.2025

Problematics

Let $f(x, \theta)$ be a map $f : \mathbb{R} \to \mathbb{R}$ parameterized by $\theta \in \mathbb{R}$.

The signature of its implementation as a numerical program, e.g., in Python could be:

```
def f(x: float, theta: float) -> float: ...
```

If we are interested in computing $\partial_\theta f(x, \theta)$ in many scenarios the structure of $f(x, \theta)$ is such that the tangent/adjoint-mode AD could be applied to obtain the derivative.

AD, however, has a **subtle**, but **essential requirement**, the result of $f(x, \theta)$ must have an **arithmetic-based data dependence** on θ, otherwise it does not enter the CG.

Problematics

If for instance θ appears exclusively as a part of the condition $c(x, \theta)$ that defines the branch, from AD perspective the computations that are based on θ are ‘as if invisible.’

Practical problems

The gradients are:

- ... either uninformative (though mathematically consistent)
- ... or actually incorrect

Problematics

Motif

In scenarios where it is not unreasonable to ask for $\partial_\theta f(x, \theta)$.

Problematics

Motif

Presence of parameterized predicates compromises the propagation of tangents or adjoints, preventing AD from the synthesizing needed derivatives.

Problematics

Motif

We seek to devise some form of treatment to deal with this problem, and enable the calculation of (useful?) AD-based gradients again.

Smoothing

Heaviside - a prototypical discontinuity

Let's begin with the simplest, but still instructive example - Heaviside:

$$z(x,\theta) = z(c(x,\theta)) = \begin{cases} 1, & c(x,\theta) > 0 \\ 0, & \text{else} \end{cases}$$

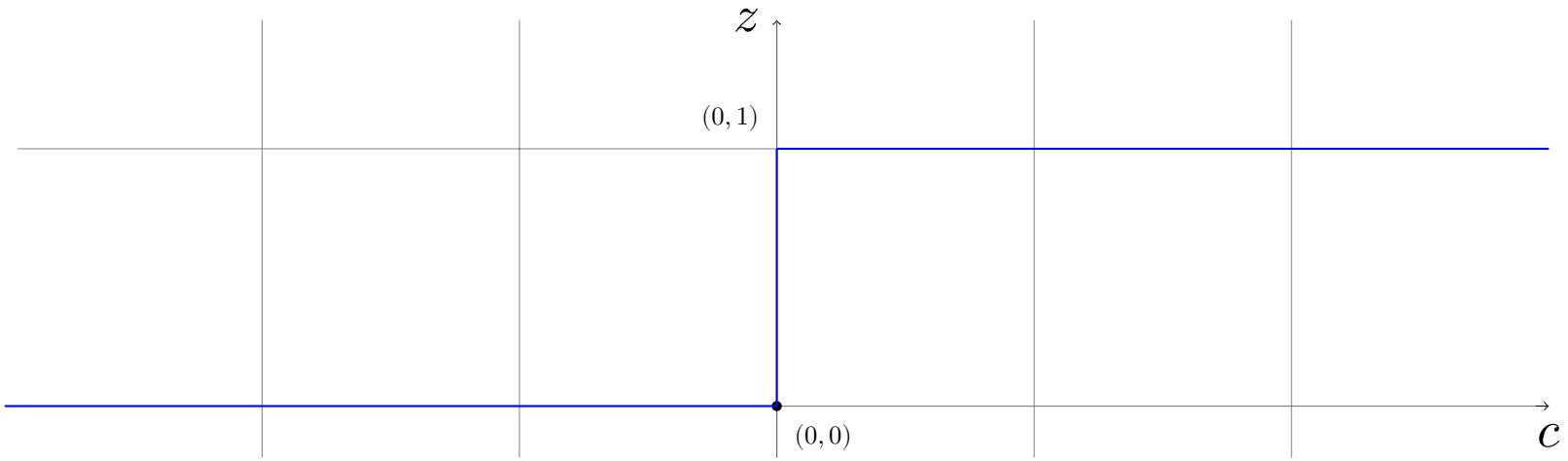
and

$$\partial_\theta z(c(x,\theta)) = \partial_c z \partial_\theta c = \delta(0)\partial_\theta c(x,\theta)$$

which any AD system would evaluate to **0 everywhere** (since the change occurs on the null set).

Smoothing

Heaviside - a prototypical discontinuity



Heaviside - a prototypical discontinuity

Thing is most (all?) conditionals can be recast into the form that reveal the heaviside function.

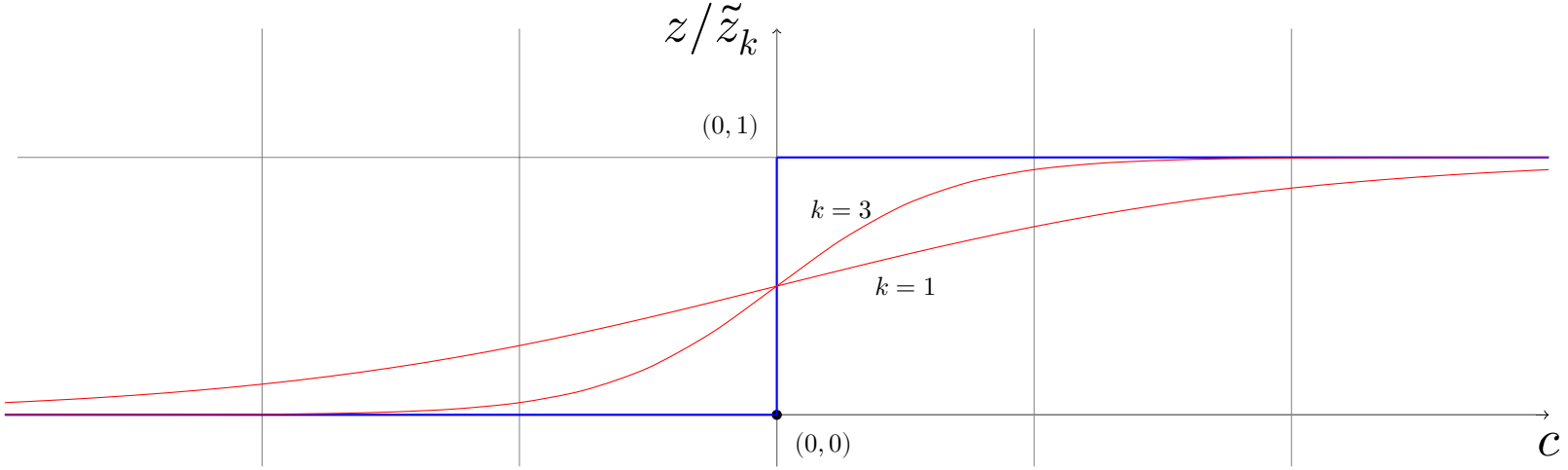
Treatment idea - smoothing

A simple treatment idea could be to approximate the discontinuous $z(c)$ with some smooth function that follows it's contour, e.g., a sigmoid

$$z(c(x,\theta)) \approx \tilde{z}_k(c(x,\theta)) = \frac{1}{1+e^{-k\cdot c(x,\theta)}}$$

Note: This trades the correctness of the model for some meaningful gradient, where 'meaningful' refers to adding some measure of 'anticipation' for the change to occur.

Treatment idea - smoothing



Ex. 1 - differentiation under integration

Problem

Let $f(x, \theta)$ be a discontinuous function

$$f(x, \theta) = \texttt{sign}(c(x, \theta)) := s(c(x, \theta))$$

with $c(x, \theta) := (x - \theta)(x + \theta)$ and

$$F(\theta) = \int_{-L}^{L} f(x, \theta) dx, \text{ for } \theta < L$$

solve

$$\min_{\theta} F(\theta)$$

Ex. 1 - differentiation under integration

1. $\partial_\theta F(\theta)$ is needed if one attacks this problem with a gradient-based method.

Ex. 1 - differentiation under integration

1. $\partial_\theta F(\theta)$ is needed if one attacks this problem with a gradient-based method.
2. AD through quadrature, i.e., numerical implementation of $F(\theta)$ gives 0, which is not just uninformative it is **wrong**.

Ex. 1 - differentiation under integration

The definition

$$F(\theta) = \int_{-L}^{L} f(x, \theta) dx, \text{ for } \theta < L$$

encodes an implicit split of the integral at all points $\{x_i^* = c_i^{-1}(0)\} = \{-\theta, \theta\}$,

$$F(\theta) = \int_{-L}^{-\theta} f(x, \theta) dx + \int_{-\theta}^{\theta} f(x, \theta) dx + \int_{\theta}^{L} f(x, \theta) dx$$

which prompts to invoke the Leibniz rule for differentiating the integral with 'moving' boundaries, i.e.,

$$\begin{aligned} d_\theta F(\theta) &= f^-(-\theta, \theta) \partial_\theta c(-\theta, \theta) - f^+(\theta, \theta) \partial_\theta c(\theta, \theta) \\ &= 1 \cdot (-2\theta) - (-1) \cdot (-2\theta) \\ &= -4\theta \end{aligned}$$

Ex. 1 - differentiation under integration

Treatment

Since $s(c(x,\theta))$ (`sign`) can be conveniently represented through $z(c(x,\theta))$ (Heaviside)

$$s(c(x,\theta)) = 2z(c(x,\theta)) - 1 := f(x,\theta)$$

we can immediately devise a smooth form of $s \to \tilde{s}$ by substituting $z \to \tilde{z}$

$$\tilde{s}_k(c(x,\theta)) = 2\tilde{z}_k(c(x,\theta)) - 1 := \tilde{f}_k(x,\theta)$$

Ex. 1 - differentiation under integration

Treatment

This allows a seamless smooth reformulation of the original problem as

$$\min_\theta F(\theta) \approx \min_\theta \tilde{F}_k(\theta) = \min_\theta \int_{-L}^{L} \tilde{f}_k(x, \theta) dx$$

which if attacked with a gradient-based method (via AD) would yield a meaningful gradient $d_\theta \tilde{F}$ which better $\approx d_\theta F$.

Ex. 1 - differentiation under integration

Meaningful as in consistent

$$\lim_{k\to\infty} d_\theta \tilde{F}_k(\theta) = d_\theta F(\theta)$$

Proof:

$$\begin{aligned}
\lim_{k\to\infty} d_\theta \tilde{F}(\theta) &= \lim_{k\to\infty} d_\theta \int_{-L}^{L} (2\tilde{z}(c(x,\theta)) - 1)dx \\
&= \lim_{k\to\infty} 2 \int_{-L}^{L} d_c\tilde{z} \cdot \partial_\theta c(x,\theta)dx \\
&= -4\theta \lim_{k\to\infty} \int_{-L}^{L} k\tilde{z}(1-\tilde{z})dx \\
&= -4\theta = d_\theta F(\theta)
\end{aligned}$$

Ex. 2 - differentiable bisection

Bisection 101

Given a function $g(x, \theta)$ bisection attempts to find an approximation to a certain root

$$x_i^* = g_i^{-1}(0, \theta)$$

contained within initial bracket $[a_0, b_0]$, by progressively halving this bracket via swapping one of the boundaries with the current bracket's mid point.

Ex. 2 - differentiable bisection

Bisection 101

The crux and the meat is of course the swapping decision

$$[a_{i+1}, b_{i+1}] = \begin{cases} [m_i, b_i], & c(a_i, m_i, \theta) := g(a_i, \theta) \cdot g(m_i, \theta) > 0 \\ [a_i, m_i], & \text{else} \end{cases}$$

which again can be conveniently represented via a $z(g(a_i, \theta) \cdot g(m_i, \theta)) := z_i$

$$[a_{i+1}, b_{i+1}] = z_i \cdot [m_i, b_i] + (1 - z_i) \cdot [a_i, m_i]$$

Ex. 2 - differentiable bisection

1. The issue – if we try to propagate tangents or adjoints the sensitivity obtained would be 0.

Ex. 2 - differentiable bisection

1. The issue – if we try to propagate tangents or adjoints the sensitivity obtained would be 0.
2. But what should we even expect?

Ex. 2 - differentiable bisection

Consistency criterion

Since we progressively shrink the bracket closing in on the root, i.e.,

$$[a_i, b_i] \to b_i - a_i := \frac{b_0 - a_0}{2^i}$$

with the error being bounded from above

$$|m_i - x^*| \leq \frac{b_i - a_i}{2}$$

then use IFT as a consistency criterion?

$$m_i \approx x^* := g^{-1}(0, \theta) = x^*(\theta)$$
$$d_\theta m_i \approx d_\theta x^* := d_\theta g^{-1}(0, \theta) = d_\theta x^*(\theta)$$

Ex. 2 - differentiable bisection

Consistency criterion

If we smooth:

$$[a_{i+1}, b_{i+1}] = z_i \cdot [m_i, b_i] + (1 - z_i) \cdot [a_i, m_i]$$

$$\Downarrow$$

$$[\widetilde{a_{i+1}, b_{i+1}}] = \tilde{z}_i \cdot [\widetilde{m_i, b_i}] + (1 - \tilde{z}_i) \cdot [\widetilde{a_i, m_i}]$$

should we expect?

$$d_\theta \tilde{m}_i \approx d_\theta x^* := d_\theta g^{-1}(0, \theta) = d_\theta x^*(\theta)$$

Ex. 2 - differentiable bisection

Concrete case #1

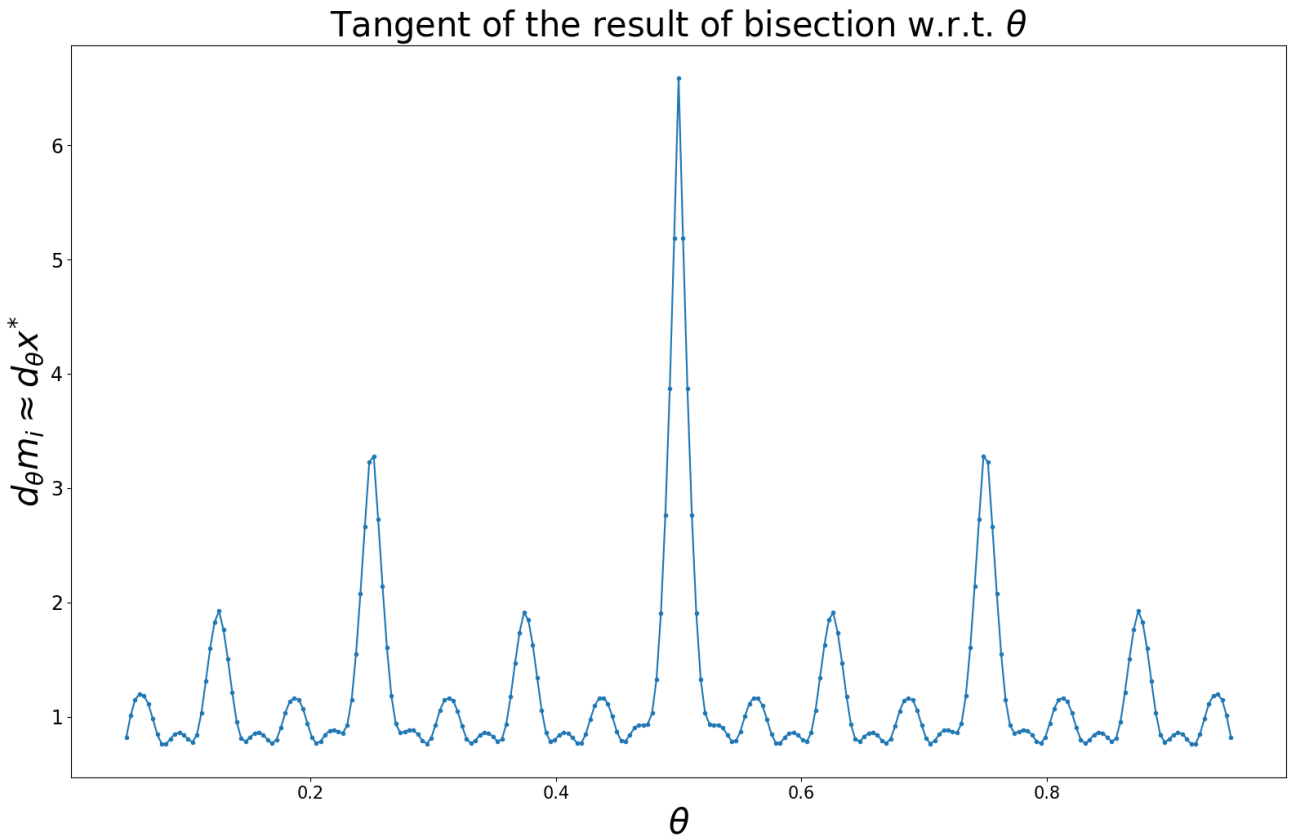
- Let $g(x, \theta) = x - \theta$, then $x^*(\theta) = \theta$, and $d_\theta x^*(\theta) = 1$ everywhere.

Ex. 2 - differentiable bisection

Concrete case #1

- Let $g(x, \theta) = x - \theta$, then $x^*(\theta) = \theta$, and $d_\theta x^*(\theta) = 1$ everywhere.
- Let's execute the smooth bisection algorithm on $[a_0, b_0] = [0, 1]$ to search for the root for a bunch of $\theta \in (0, 1)$ and plot tangents of the results.

Ex. 2 - differentiable bisection



Ex. 2 - differentiable bisection

Concrete case #2

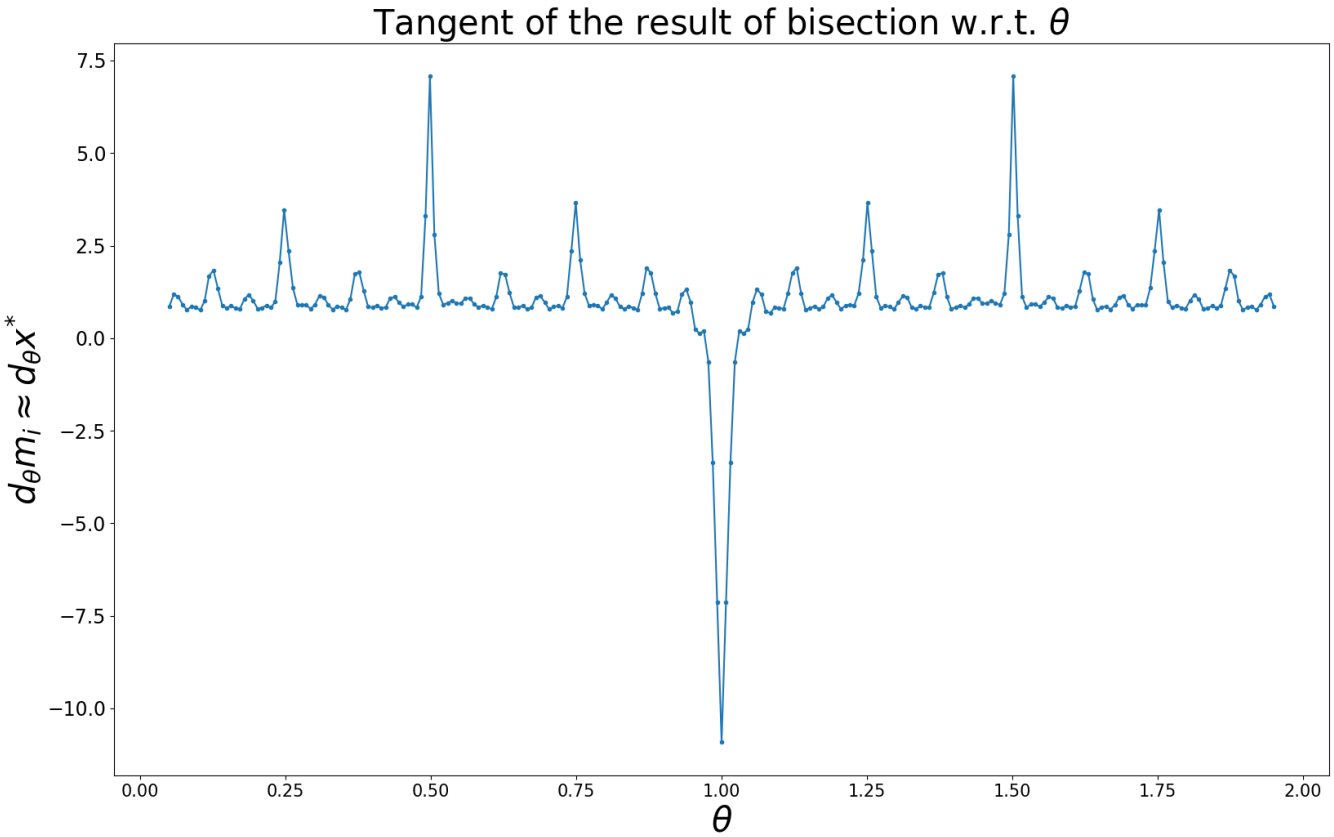
- Let $g(x, \theta) = (x - \theta)(x + 1 - \theta)$, then $d_\theta x^*(\theta) = 1$ for both roots.

Ex. 2 - differentiable bisection

Concrete case #2

- Let $g(x, \theta) = (x - \theta)(x + 1 - \theta)$, then $d_\theta x^*(\theta) = 1$ for both roots.
- Let's execute the smooth bisection algorithm on $[a_0, b_0] = [0, 1]$ to search for the root for a bunch of $\theta \in (0, 2)$ and plot tangents of the results.

Ex. 2 - differentiable bisection



The end

Outlook

- Variable dependent branch conditions jeopardize gradient quality with AD.
- Drop-in style smoothing can help deal with this issue.
- In certain scenarios sensitivities reflect **not mathematical** sensitivities, but **algorithmic** ones.
- Could these ideas be extended in useful sense to more involved problems, e.g., grid search, where the calculation paths are not just means to an end?

The end

Thank you!