What About the Middleman?

Shrinking the gap between theory and practice in machine learning

Santiago Rodriguez

under mentorship from Steve Zdancewic and Stephen Mell

REPL, August 2023

Penn Engineering

The Philosophical Problem

Can an inanimate object learn? Does this question even make sense?

The Philosophical Problem

Can an inanimate object learn? Does this question even make sense?

The Actual Problem

Given some computational problem, how can we get a machine to automatically solve it?

The Philosophical Problem

Can an inanimate object learn? Does this question even make sense?

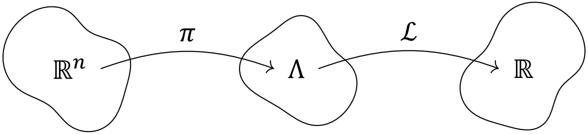
The Actual Problem

Given some computational problem, how can we get a machine to automatically solve it?

The Solution*

Carve out a manageable subset of programs and search for the one that's "good enough".

Let $\pi : \mathbb{R}^n \to \Lambda$ be a function from a parameter space (e.g. $\mathbb{R}^n$) to a set of programs Λ and let $\mathcal{L} : \Lambda \to \mathbb{R}$ be a measure of how poorly a program satisfies some problem specification $\mathbf{P}$, i.e. $\mathcal{L}(\lambda_1) < \mathcal{L}(\lambda_2)$ means λ_1 is better than λ_2 at satisfying $\mathbf{P}$.



We call $f := \mathcal{L} \circ \pi$ the *objective function*. The problem reduces to calculating

$$\overline{\theta} := \arg\min_{\theta} f(\theta)$$

where $\pi(\overline{\theta})$ is the best program that satisfies $\mathbf{P}$ (wrt metric $\mathcal{L}$).

Actually calculating $\arg\min_\theta f(\theta)$ is infeasible but we can get decent approximations using methods from optimization theory.

- Grid Search
- Gradient Descent
- Stochastic Variations

We even have strong theoretical bounds on the efficiency and optimality of these methods for various classes of objective functions.

- Smooth Functions
- Strongly Convex
- Lipschitz Functions

Poorly understood heuristic algorithms have consistently outperformed theoretically optimal algorithms in complex, real-world tasks. Even good ol' fashion gradient descent would sometimes outperform!

Poorly understood heuristic algorithms have consistently outperformed theoretically optimal algorithms in complex, real-world tasks. Even good ol' fashion gradient descent would sometimes outperform!

Is the math wrong?

Poorly understood heuristic algorithms have consistently outperformed theoretically optimal algorithms in complex, real-world tasks. Even good ol' fashion gradient descent would sometimes outperform!

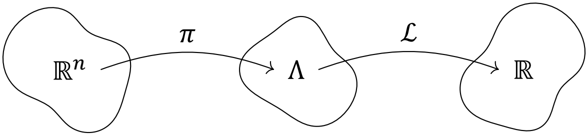
Is the math wrong? I hope not.

Poorly understood heuristic algorithms have consistently outperformed theoretically optimal algorithms in complex, real-world tasks. Even good ol’ fashion gradient descent would sometimes outperform!

Is the math wrong? I hope not.

The assumptions on our objective functions, however, are clearly flawed.

Instead of guessing what kind of properties our objective function $f := \mathcal{L} \circ \pi$ has, let's derive them from the parts!



But in order to do any analysis, we need a suitable structure on Λ that will allow us to talk about continuous and/or "differentiable" functions between the set of programs and the real line.

Coherence Spaces for Banach Spaces

Santiago Rodriguez 2.0
under mentorship from Steve Zdancewic and Stephen Mell

REPL, August 2023

DEF A *coherence space* $\mathbf{X}$ consists of a set of *tokens* $|\mathbf{X}|$ and a reflexive symmetric relation $\frown$ on $|\mathbf{X}|$ called the *coherence relation*. This forms a graph called the *web* of $\mathbf{X}$.

Abusing notation, an element of $\mathbf{X}$ is a clique in the web of $\mathbf{X}$, i.e.

$$\mathbf{X} \coloneqq \{\alpha \subseteq |\mathbf{X}| \mid \forall x, x' \in \alpha.\ x \frown x'\}.$$

Note coherence spaces provide a denotational semantics for typed λ-calculi like STLC. We interpret a type A as a coherence space $\mathbf{A}$ and a term $a : A$ as a clique $\alpha \in \mathbf{A}$.

Example

Bool Type

$$
\begin{tikzcd}[column sep=small]
& \emptyset \arrow[dl] \arrow[dr] & \\
\{\mathbf{T}\} & & \{\mathbf{F}\}
\end{tikzcd}
$$

Bool × **Bool** Type

$$
\begin{tikzcd}[column sep=small]
& \langle \emptyset, \emptyset \rangle \arrow[dl] \arrow[d] \arrow[dr] \arrow[drr] & & \\
\langle \{\mathbf{T}\}, \emptyset \rangle \arrow[d] \arrow[dr] & \langle \emptyset, \{\mathbf{T}\} \rangle \arrow[dl] \arrow[dr] & \langle \emptyset, \{\mathbf{F}\} \rangle \arrow[dl] \arrow[dr] & \langle \{\mathbf{F}\}, \emptyset \rangle \arrow[dl] \arrow[d] \\
\langle \{\mathbf{T}\}, \{\mathbf{T}\} \rangle & \langle \{\mathbf{T}\}, \{\mathbf{F}\} \rangle & \langle \{\mathbf{F}\}, \{\mathbf{T}\} \rangle & \langle \{\mathbf{F}\}, \{\mathbf{F}\} \rangle
\end{tikzcd}
$$

A coherence space paired with the inclusion relation $\subseteq$ forms a partial order (specifically a *Scott domain*). This gives us a rather natural topology.

DEF Let $\mathbf{X}$ be a coherence space. The *Scott topology* on $\mathbf{X}$ is the topology generated by the basis $\{\uparrow\alpha \mid \alpha \in \mathbf{X}_{\text{fin}}\}$ where $\mathbf{X}_{\text{fin}}$ is the set of finite cliques and $\uparrow\alpha := \{\beta \in \mathbf{X} \mid \alpha \subseteq \beta\}$.

Now we have continuous functions!

[Digress to ML]

$$\mathbb{R}^n \xrightarrow{\pi} \Lambda \xrightarrow{\mathcal{L}} \mathbb{R}$$

DEF Let $\mathbf{X}$ and $\mathbf{Y}$ be coherence spaces. A function $F : \mathbf{X} \to \mathbf{Y}$ is *stable* iff it is continuous and preserves pullback, i.e. $\alpha \cup \beta \in \mathbf{X}$ implies $F(\alpha \cap \beta) = F(\alpha) \cap F(\beta)$.

We denote by **Coh** the category of coherence spaces with stable maps.

We interpret a function between types A and B as a stable function between their corresponding coherence spaces $\mathbf{A}$ and $\mathbf{B}$.

Stability in multiple arguments motivates the definition of a *direct product* $\mathbf{A} \times \mathbf{B}$ which is indeed a product object in the categorical sense.

The set of stable maps from $\mathbf{A}$ to $\mathbf{B}$ can be presented as a coherence space $\mathbf{A} \Rightarrow \mathbf{B}$ which is indeed an exponential object in the categorical sense.

The denotational semantics are pretty intuitive at this point. We interpret the product type $A \times B$ as $\mathbf{A} \times \mathbf{B}$ and the arrow type $A \to B$ as $\mathbf{A} \Rightarrow \mathbf{B}$.

[Digress to ML]

$$\mathbb{R}^n \xrightarrow{\pi} \Lambda \xrightarrow{\mathcal{L}} \mathbb{R}$$

Penn Engineering

DEF Let $\mathbf{X}$ be a coherence space and S an arbitrary set. A *representation* of S is a partial surjective function $\rho :\subseteq \mathbf{X} \twoheadrightarrow S$ denoted $\mathbf{X} \xrightarrow{\rho} S$ or simply ρ.

DEF Let $\mathbf{X} \xrightarrow{\rho} S$ and $\mathbf{Y} \xrightarrow{\varphi} T$ be representations. A function $f : S \to T$ is *stably realizable* via ρ to φ iff $\exists F : \mathbf{X} \to_{\mathrm{st}} \mathbf{Y}$ such that the following diagram commutes.

$$\begin{tikzcd} \mathbf{X} \arrow[r, "F"] \arrow[d, "\rho"'] & \mathbf{Y} \arrow[d, "\varphi"] \\ S \arrow[r, "f"] & T \end{tikzcd}$$

We denote by **Rep** the category of representations with stably realizable maps.

Penn Engineering

[Digress to ML]

$$\mathbb{R}^n \xrightarrow{\pi} \Lambda \xrightarrow{\mathcal{L}} \mathbb{R}$$

Constructions over coherence spaces naturally carries over to representations. For example given representations $\mathbf{X} \xrightarrow{\rho} S$ and $\mathbf{Y} \xrightarrow{\varphi} T$, we can construct

- the product representation $\mathbf{X} \times \mathbf{Y} \xrightarrow{\rho\times\varphi} S \times T$ and
- the exponential representation $\mathbf{X} \Rightarrow \mathbf{Y} \xrightarrow{\rho\to\varphi} \mathrm{Hom}(\rho, \varphi)$

both of which are product and exponential objects in the categorical sense, respectively.

DEF Let $\mathbf{X} \xrightarrow{\rho} S$ be a representation. A *spanning forest* for ρ is a set $\mathcal{F} \subseteq \mathbf{X}_{\text{fin}}$ such that

1. $(\mathcal{F}, \subseteq)$ is a forest, in particular $\forall \alpha, \alpha' \in \mathcal{F}$ if $\alpha \frown \alpha'$ then $\alpha \subseteq \alpha'$ or $\alpha' \subseteq \alpha$, and
2. $\mathcal{F}$ spans $\mathrm{dom}(\rho)$, i.e. $\forall \alpha \in \mathbf{X}.\ \alpha \in \mathrm{dom}(\rho)$ iff $\exists$ a maximal chain $\{\alpha_j\} \subseteq \mathcal{F}$ such that $\alpha = \bigcup_j \alpha_j$.

If such an $\mathcal{F}$ exists then ρ is a *spanned representation*.

Example

Prefix Tree $2^\star$ of Cantor set 2^ω

$$
\begin{tikzcd}
\varepsilon \arrow[r] \arrow[dddr] & 0 \arrow[r] \arrow[ddr] & 00 \arrow[r] \arrow[dr] & 000 \arrow[r] & \cdots \\
 & & & 001 \arrow[r] & \cdots \\
 & & 01 \arrow[r] & \cdots & \\
 & 1 \arrow[r] \arrow[dr] & 10 \arrow[r] & \cdots & \\
 & & 11 \arrow[r] & \cdots &
\end{tikzcd}
$$

The Cantor set has a very natural presentation as a coherence space $\mathbf{C}$ defined by

$$|\mathbf{C}| := 2^\star \quad \text{with} \quad u \frown v \text{ iff } u \sqsubseteq v \text{ or } v \sqsubseteq u$$

where $\sqsubseteq$ is the prefix relation. A maximal clique is a maximal chain in the tree and its limit is precisely an infinite binary sequence.

The *Cantor representation* is $\mathbf{C} \xrightarrow{\rho_{\mathbf{C}}} 2^\omega$ defined by

$$\mathrm{dom}(\rho_{\mathbf{C}}) := \mathbf{C}_{\text{max}} \quad \text{with} \quad \rho_{\mathbf{C}}(\{u_n\}) := \lim_{n\to\infty} u_n$$

where $\mathbf{C}_{\text{max}}$ is the set of maximal cliques.

DEF Let $\mathbb{Y}$ be a topological space. A continuous spanned representation $\mathbf{Y} \xrightarrow{\varphi} \mathbb{Y}$ is *admissible* iff $\forall$ continuous spanned representations $\mathbf{X} \xrightarrow{\rho} \mathbb{Y}_0$ (given subspace $\mathbb{Y}_0 \subseteq \mathbb{Y}$) the inclusion function $i : \mathbb{Y}_0 \hookrightarrow \mathbb{Y}$ is stably realizable via ρ to φ.

$$\begin{tikzcd} \mathbf{X} \arrow[r, "I"] \arrow[d, "\rho"'] & \mathbf{Y} \arrow[d, "\varphi"] \\ \mathbb{Y}_0 \arrow[r, hook, "i"] & \mathbb{Y} \end{tikzcd}$$

Penn Engineering

THM Let $\mathbf{X} \xrightarrow{\rho} \mathbb{X}$ and $\mathbf{Y} \xrightarrow{\varphi} \mathbb{Y}$ be admissible representations. Then a function $f : \mathbb{X} \to \mathbb{Y}$ is stably realizable via ρ to φ iff f is sequentially continuous, i.e.

$$\text{if } x_n \to x \ (n \to \infty) \text{ then } f(x_n) \to f(x) \ (n \to \infty).$$

COR Let $\mathbf{X} \xrightarrow{\rho} \mathbb{X}$ and $\mathbf{Y} \xrightarrow{\varphi} \mathbb{Y}$ be admissible representations. Suppose $\mathbb{X}$ is a sequential space. Then a function $f : \mathbb{X} \to \mathbb{Y}$ is stably realizable via ρ to φ iff f is continuous.

[Digress to ML]

$$\mathbb{R}^n \xrightarrow{\pi} \Lambda \xrightarrow{\mathcal{L}} \mathbb{R}$$

DEF A Banach space $\mathbb{X}$ is a vector space paired with a norm $||\cdot||$ such that the induced topology is complete, i.e.

$$\text{if } \langle x_n \rangle \in \mathbb{X}^\omega \text{ is a Cauchy sequence then } \exists x \in \mathbb{X}.\ x_n \to x\ (n \to \infty).$$

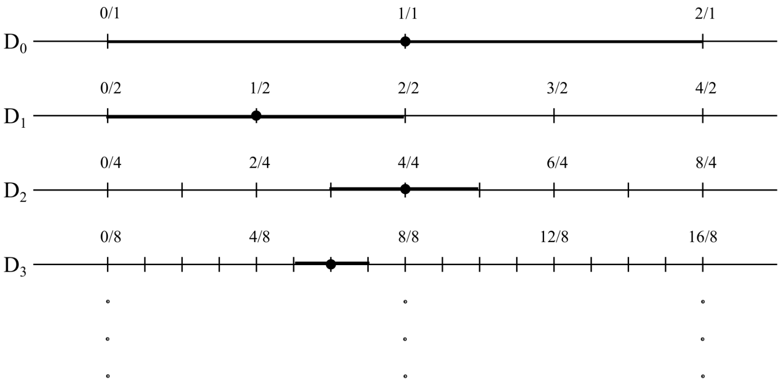
Banach spaces generalize topological spaces like $\mathbb{R}^n$ in that we can define derivatives between banach spaces (called *Fréchet derivatives*).

DEF Let $\mathbb{X}$ be a Banach space and $D \subseteq \mathbb{X}$ a dense subset. The *dyadic space* of $\mathbb{X}$ (wrt D) is the coherence space $\mathbf{X}$ defined by

$$|\mathbf{X}| := \mathbb{N} \times D \quad \text{with} \quad (m, x) \frown (n, x') \text{ iff } m \neq n \text{ and } \|x - x'\| \leq 2^{-m} + 2^{-n}.$$

Intuitively, cliques in $\mathbf{X}$ correspond to partially defined sequences, i.e. $(m, x) \in \alpha \in \mathbf{X}$ represents a partial sequence with the m^{th} term being x.

Example



A clique $\{(0,1),(1,{}^{1}\!/_{2}),(2,1),(3,{}^{3}\!/_{4}),\ldots\}$ in the dyadic space of $\mathbb{R}$ wrt dyadic rationals.

As such, a maximal clique will correspond to a fully defined sequence. But that's not all!

PROP Let $\mathbb{X}$ be a Banach space with dense subset $D \subseteq \mathbb{X}$ and dyadic space $\mathbf{X}$. Suppose $\{(n, x_n)\}_{n \in \mathbb{N}}$ is a maximal clique in $\mathbf{X}$. Then $\langle x_n \rangle$ is a (rapidly-converging) Cauchy sequence, i.e.

$$\forall \varepsilon > 0. \exists N \in \mathbb{N}. \forall m, n > N. \; \|x_m - x_n\| \leq 2^{-m} + 2^{-n} \leq 2^{-N} < \varepsilon.$$

Since Banach spaces are complete, $\langle x_n \rangle$ is a convergent sequence. Not only that, every $x \in \mathbb{X}$ has a Cauchy sequence consisting of elements in D that rapidly converge to x.

DEF Let $\mathbb{X}$ be a Banach space with dyadic space $\mathbf{X}$. The *Cauchy representation* of $\mathbb{X}$ over $\mathbf{X}$ is $\mathbf{X} \xrightarrow{\rho_{\mathbf{X}}} \mathbb{X}$ defined by

$$\mathrm{dom}(\rho_{\mathbf{X}}) \coloneqq \mathbf{X}_{\mathsf{max}} \quad \text{with} \quad \rho_{\mathbf{X}}(\{(n, x_n)\}) \coloneqq \lim_{n\to\infty} x_n$$

where $\mathbf{X}_{\mathsf{max}}$ is the set of maximal cliques.

THM Let $\mathbf{X} \xrightarrow{\rho_{\mathbf{X}}} \mathbb{X}$ be a Cauchy representation. Then $\rho_{\mathbf{X}}$ is admissible.

[Digress to ML]

$$\mathbb{R}^n \xrightarrow{\pi} \Lambda \xrightarrow{\mathcal{L}} \mathbb{R}$$