Dynamic Programming with Google JAX

John Stachurski

June 2024

Topics

- 4 minute history of scientific computing
- What's JAX?
- Intro to JAX – hands on
- JAX for DP – hands on

Target audience: people who are new to / curious about JAX

Please feel free to question / debate / share your experiences

Slides, code:

https://github.com/QuantEcon/cef_2024_singapore

Quick poll:

- Python programmers?
- Julia?
- MATLAB?
- C?
- Fortran?

- JAX users?
- Regular GPU users?

Old school: static types & AOT compilers

Example. Consider

$$k_{t+1} = sk_t^{\alpha} + (1 - \delta)k_t \qquad \text{with } k_0 \text{ given}$$

Fortran code:

```fortran
program main
    implicit none
    integer, parameter :: dp=kind(0.d0)
    integer  :: n=1000
    real(dp) :: s=0.3_dp
    real(dp) :: a=1.0_dp
    real(dp) :: delta=0.1_dp
    real(dp) :: alpha=0.4_dp
    real(dp) :: k=0.2_dp
    integer  :: i
    do i = 1, n - 1
        k = a * s * k**alpha + (1 - delta) * k
    end do
    print *,'k = ', k
end program main
```

Relative merits?

Pros

- fast loops / arithmetic

Cons

- low interactivity
- time consuming to write / read / debug
- hard to parallelize

For comparison, the same operation in Python:

```python
α = 0.4
s = 0.3
δ = 0.1
n = 1_000
k = 0.2

for i in range(n):
    k = s * k**α + (1 - δ) * k

print(k)
```

Pros

- high interactivity
- easy to write / read / debug

Cons

- slow loops / arithmetic (in pure Python)

Why is pure Python slow?

Pros

- high interactivity
- easy to write / read / debug

Cons

- slow loops / arithmetic (in pure Python)

Why is pure Python slow?

Problem 1: Type checking

```
x, y = 1, 2
z = x + y

x, y = 1.0, 2.0
z = x + y

x, y = 'foo', 'bar'
z = x + y
```

How does Python know which operation to perform?

Answer: Python checks the type of the objects first

```
>> x = 1
>> type(x)
int
```

```
>> x = 'foo'
>> type(x)
str
```

In a large loop, this type checking generates massive overhead

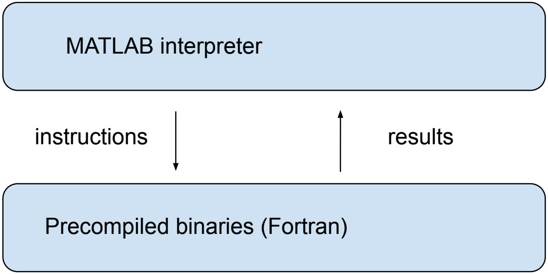
Problem 2: Memory management

```
>>> import sys
>>> x = [1.0, 2.0]
>>> sys.getsizeof(x) * 8          # number of bits
576                               # whaaaat???
```

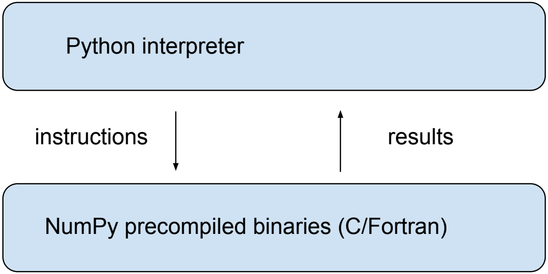
So how can we get

good execution speeds **and** high productivity / interactivity?

MATLAB's vectorization trick



Python + NumPy – stealing MATLAB's idea



Vectorization: pros and cons

Pros

- fast array operations
- high interactivity

Cons

- some tasks cannot be efficiently vectorized
- cannot adapt flexibly to function arguments / hardware

Julia — rise of the JIT compilers

```
function solow(k0, α=0.4, δ=0.1, n=1_000)
    k = k0
    for i in 1:(n-1)
        k = s * k^α + (1 - δ) * k
    end
    return k
end

solow(0.2)
```

Function `solow` is efficiently JIT compiled after the first call

Python + Numba copy Julia

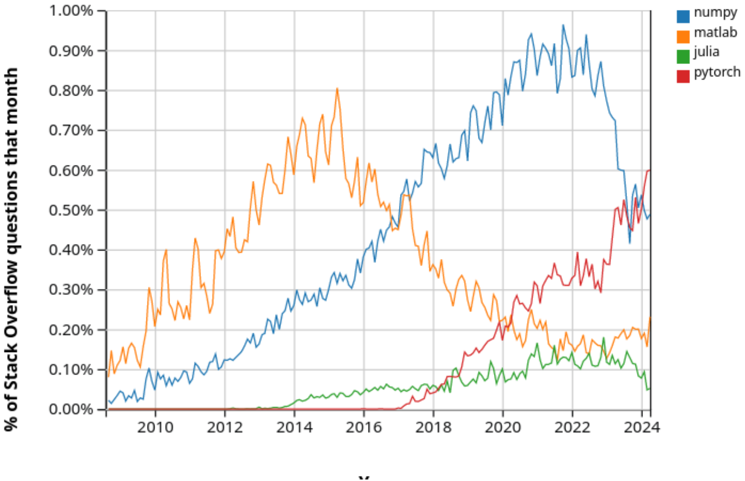
```
from numba import jit

@jit
def solow(k0, α=0.4, δ=0.1, n=1_000):
    k = k0
    for i in range(n-1):
        k = s * k**α + (1 - δ) * k
    return k

solow(0.2)
```

Runs at same speed as Julia / C / Fortran

Notice that most discussion so far is about execution speed on a **single** thread...

Later we'll discuss parallelization

Some trends:



numpy
matlab
julia
pytorch
% of Stack Overflow questions that month
1.00%
0.90%
0.80%
0.70%
0.60%
0.50%
0.40%
0.30%
0.20%
0.10%
0.00%
2010
2012
2014
2016
2018
2020
2022
2024

So where does JAX fit in?

Let's start with some motivation and background

AI-driven scientific computing

AI is changing the world

- image processing / computer vision
- speech recognition, translation
- forecasting and prediction
- generative AI

Plus killer drones, skynet, etc....

AI-driven scientific computing

AI is changing the world

- image processing / computer vision
- speech recognition, translation
- forecasting and prediction
- generative AI

Plus killer drones, skynet, etc....

Projected spending on AI in 2024:

- Google: \$48 billion
- Microsoft: \$60 billion
- Meta: \$40 billion
- etc.

Deep learning in two slides

We postulate a relationship between inputs and outputs

$$y = f(x) \qquad (x \in \mathbb{R}^d,\ y \in \mathbb{R})$$

Examples.

- $x =$ sequence of words, $y =$ next word
- $x =$ weather sensor data, $y =$ max temp tomorrow

Problem:

- observe $(x_i, y_i)_{i=1}^n$ and seek f such that $y_{n+1} \approx f(x_{n+1})$

Training: given parametric class $\{f_\theta\}_{\theta\in\Theta}$, minimize the loss

$$\ell(\theta) := \sum_{i=1}^{n}(y_i - f_\theta(x_i))^2 \quad \text{s.t.} \quad \theta \in \Theta$$

In the case of ANNs,

$$f_\theta = \sigma \circ A_1 \circ \cdots \circ \sigma \circ A_{k-1} \circ \sigma \circ A_k$$

where

- $A_i x = W_i x + b_i$ is an affine map
- σ is a nonlinear "activation" function

Training: given parametric class $\{f_\theta\}_{\theta\in\Theta}$, minimize the loss

$$\ell(\theta) := \sum_{i=1}^{n}(y_i - f_\theta(x_i))^2 \quad \text{s.t.} \quad \theta \in \Theta$$

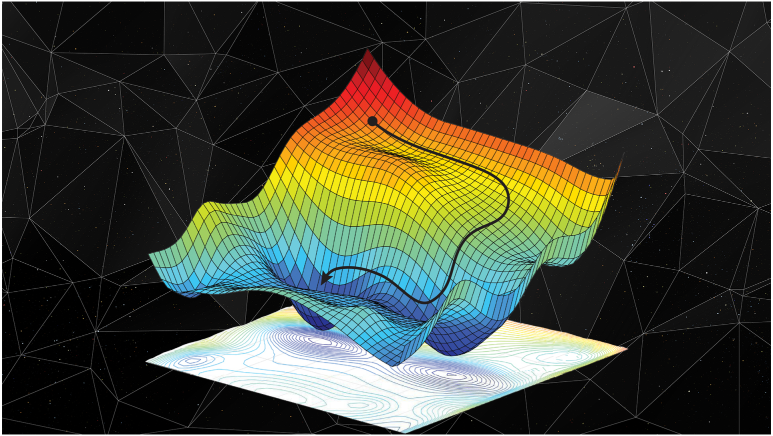
In the case of ANNs,

$$f_\theta = \sigma \circ A_1 \circ \cdots \circ \sigma \circ A_{k-1} \circ \sigma \circ A_k$$

where

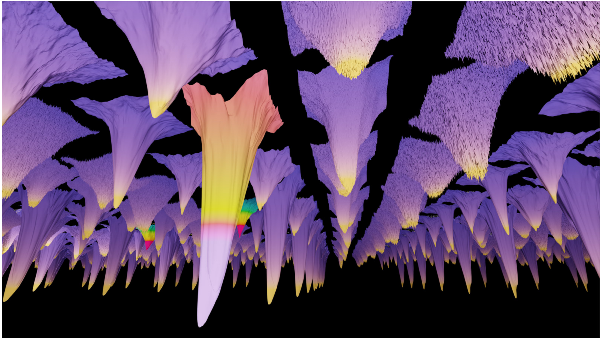
- $A_i x = W_i x + b_i$ is an affine map
- σ is a nonlinear "activation" function

Minimizing a smooth loss functions – what algorithm?



Source: https://danielkhv.com/

Deep learning: $\theta \in \mathbb{R}^d$ where $d = ?$



Source: https://losslandscape.com/gallery/

Hardware

"NVIDIA supercomputers are the factories of the AI industrial revolution." – Jensen Huang

- How many GPUs did OpenAI use to train ChatGPT 4?

Software

Core elements

- automatic differentiation (for gradient descent)
- parallelization — exploit parallel hardware!
- Compilers / JIT-compilers

Crucially, these components must be well integrated

One library with these features is...

Software

Core elements

- automatic differentiation (for gradient descent)
- parallelization — exploit parallel hardware!
- Compilers / JIT-compilers

Crucially, these components must be well integrated

One library with these features is...

```python
import jax.numpy as jnp
from jax import grad, jit

def f(θ, x):
  for W, b in θ:
    w = W @ x + b
    x = jnp.tanh(w)
  return x

def loss(θ, x, y):
  return jnp.sum((y - f(θ, x))**2)

grad_loss = jit(grad(loss))  # Now use gradient descent
```

Source: JAX readthedocs

JAX for economists

JAX is obviously useful if you do deep learning

But what about <u>me</u>?

I do mathematical modeling / optimization / simulation

My wishlist:

- exposes low level operations
- <u>automated</u> parallelization
- JIT compiler
- integrated autodiff
- <u>automatically</u> / transparently supports CPUs / GPUs / TPUs

JAX ticks these boxes