

0.Intro_pdf

March 25, 2025

1 STAT 207: Advanced Statistical Computing

1.0.1 Zhe Fei (zhe.fei@ucr.edu)

1.0.2 Logistics

- Lecture: Tu/Th 11:00 - 12:20 PM
- Discussion: Tu 12:30 - 1:20 PM
- Location: Olmsted 1431
- Instructor: Zhe Fei (Olmsted 1344), zhef@ucr.edu
- Office Hours: W/F 11:00 - 12:00 PM, or by appointment. Zoom
(<https://ucr.zoom.us/j/99515225321>)

1.0.3 Description

4 Units, Lecture, 3 hours; discussion, 1 hour. Prerequisite(s): STAT 201A and STAT 206 or equivalents; or consent of instructor. Computational algorithms for research in statistics. Topics include Numerical linear algebra, Numerical optimization, Monte Carlo methods, Bootstrap and resampling methods.

1.0.4 Textbook

- Kenneth Lange, Numerical Analysis for Statisticians, 2nd Edition, Springer (2010), available online, [UCR library link](#)
- James Gentle, Computational Statistics, Springer (2009), available online, [UCR library link](#)

1.0.5 Discussion

- Coding and implementation;
- Journal club: review a paper.

1.0.6 Assignments

Expected 4 homework assignments (40%).

You are welcomed, and encouraged, to work with each other on the problems, but you must turn in your own work. If you copy someone else's work, both parties will receive a 0 for the homework grade as well as being reported to the [Student Conduct & Academic Integrity Programs \(SCAIP\)](#).

Submission instructions: You will turn in your homework on elearn.ucr.edu, with both **source codes** and **output files**.

1.0.7 Final project

A real data application of topics learned in class, with detailed coding, implementation, analysis, and interpretation.

1.0.8 Attendance

Attendance to both lectures and discussions is mandatory. If you are not able to attend a lecture or lab due to medical or technical reasons, please proactively communicate with the instructor and TA and notify us your circumstances at earliest chance.

1.0.9 Grading

- Homework, 40%
- Final project, 30%
- Paper presentation, 10%
- Attendance, 20%

Grades may be curved at the end. Cumulative numerical averages of 90-100 are guaranteed at least an A-, 80-89 at least a B-, and 70 - 79 at least a C-. However the exact ranges for letter grades will be determined after the final project.

1.0.10 Work Load and Teamwork

You are expected to put in about 3 - 4 hours of work outside of class for each hour of lecture. Some of you will do well with less time, and some might need more. You are encouraged to study with your classmates. But remember that anything that is not explicitly a team assignment must be your own work.

1.0.11 Policies

- You are responsible for checking announcements and accessing course materials on Canvas.
- Late work policy for the homework and labs reports:
 - next day: **lose 50%** of total possible points
 - later than next day: lose all points
- There will be no make-ups for homework, labs, quizzes, or exams. If the midterm exam must be missed, absence must be officially excused **in advance**, in which case the missing exam score will be imputed using the final exam score. This policy only applies to the midterm. All other missed assessments will receive a grade of 0. The final exam must be taken at the stated time. You must take the final exam to pass this course.
- Please be considerate of your classmates by arriving on time. If you arrive after at least one student has finished the exam and left the room, you will NOT be allowed to sit for the exam, and will receive a “0”. Turn off cell phones before entering the exam room. If your cell phone rings during the exam, you will lose points on the exam.
- Use of disallowed materials (textbook, class notes, web references, any form of communication with classmates or other persons, etc.) during exams will not be tolerated. This will result in

a 0 on the exam for all students involved, possible failure of the course, and will be reported to the [Student Conduct & Academic Integrity Programs \(SCAIP\)](#). If you have any questions about whether something is or is not allowed, ask me beforehand.

1.0.12 Tentative Topics

- Numerical linear algebra: vectors and matrices; eigenvalues; SVD;
- Numerical optimization: MM; EM; Newton’s method;
- Linear Programming;
- Resampling methods: Monte Carlo; bootstrap; cross validation;
- Advanced optimization topics

1.0.13 Quotes

“All models are wrong, but some are useful.” - George Box

“Statistics is partly empirical and partly mathematical. It is now almost entirely computational.” - Kenneth Lange

“A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.” - Josh Wills of Cloudera

1.1 Large Language Models

- A groundbreaking evolution in artificial intelligence.
- These models are deep neural networks—often with billions of parameters—that have been pre-trained on vast datasets, enabling them to understand and generate human-like text.
- Built on the transformer architecture, LLMs excel at a wide range of language tasks, from translation and summarization to creative writing and dialogue generation.

1.1.1 Reasoning

- An LLM-based reasoning model is an LLM designed to **solve multi-step problems** by generating intermediate steps or structured “thought” processes.
- Unlike simple question-answering LLMs that just share the final answer, reasoning models either explicitly display their thought process or handle it internally, which helps them to perform better at complex tasks such as puzzles, coding challenges, and mathematical problems.

Inference-Time Scaling: humans give better responses when given more time to think.

- Chain-of-Thought

Regular prompting	Chain-of-thought prompting
<p>Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?</p> <p>A: The answer (arabic numerals) is</p> <p>(Output) 8 X</p>	<p>Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?</p> <p>A: Let's think step by step.</p> <p>(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓</p>

Pure Reinforcement Learning (RL):

- A mini-R1 example, <https://www.philschmid.de/mini-deepseek-r1>

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>
 To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both ...
 $(\sqrt{a - \sqrt{a + x}})^2 = x^2 \implies a - \sqrt{a + x} = x^2$.
 Rearrange to isolate the inner square root term:
 $(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$
 ...
Wait, wait. Wait. That's an aha moment I can flag here.
 Let's reevaluate this step-by-step to identify if the correct sum can be ...
 We started with the equation:
 $\sqrt{a - \sqrt{a + x}} = x$
 First, let's square both sides:
 $a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$
 Next, I could square both sides again, treating the equation: ...
 ...

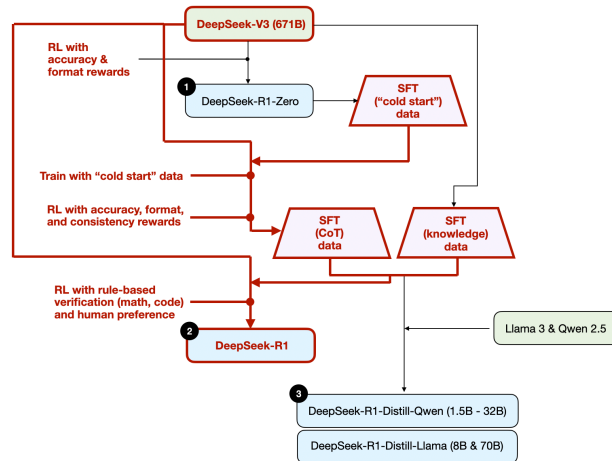
Table 3 | An interesting "aha moment" of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of reinforcement learning.

Supervised finetuning and reinforcement learning (SFT + RL)

- Typically, a model is first trained with SFT on high-quality instruction data and then further refined using RL to optimize specific behaviors.

Supervised fine-tuning and model distillation

- Fine-tuning smaller LLMs, such as Llama 8B and 70B and Qwen 2.5 models (0.5B to 32B), on a high-quality SFT dataset generated by larger LLMs.



1.2 Use of AI

AI tools can greatly improve our learning abilities, and work efficiency.

- Tell AI what to do, ask the right questions;
- AI as helpers;
- Never let AI do it for you;
- Always verify and double check AI answers.

1.3 How is it related to our class?

1.3.1 Transformers

- [Transformer architecture explained](#)
- [Transformer applications](#)
- [Transformer examples](#)

1.3.2 Foundations in Statistical Computing

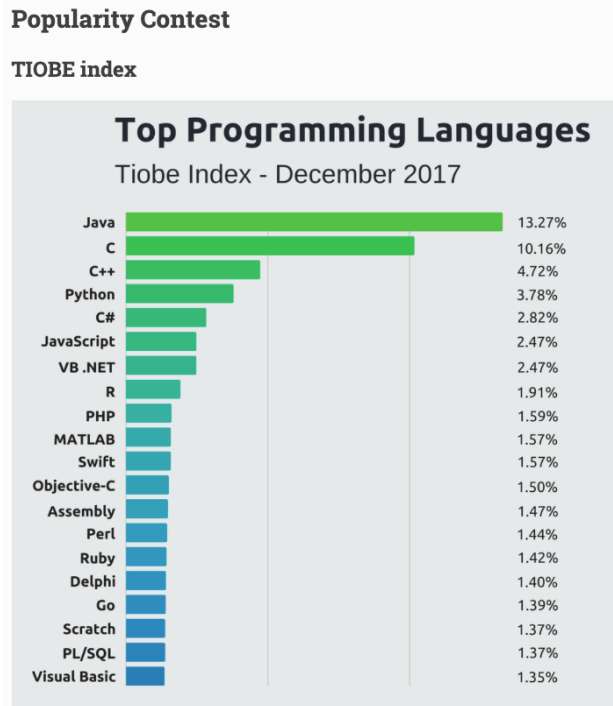
Large Language Models (LLMs) are built upon a wealth of statistical and computational techniques that are fundamental to advanced statistical computing.

- **Loss Functions:** Mean Squared Error (MSE), Cross-Entropy Loss, and Kullback-Leibler Divergence (KL Divergence).
- **Matrix Operations and Decompositions:** large-scale matrix operations with techniques like Singular Value Decomposition (SVD) and eigenvalue computations.
- **Optimization Algorithms:** Stochastic Gradient Descent (SGD) and Adam (Adaptive Moment Estimation); iteratively update millions (or billions) of parameters in LLMs or other deep learning models.
- **Regularization Techniques:** L1/L2 penalties or dropout. To prevent overfitting and enhances generalization.

1.4 Getting Started with Python

1.4.1 Popularity Contest

- 2017 [Top Programming Languages](#)



- March 2025 [TIOBE Index](#)

Mar 2025	Mar 2024	Change	Programming Language	Rating	Change
1	1		Python	23.85%	+8.22%
2	3	▲	C++	11.08%	+0.37%
3	4	▲	Java	10.36%	+1.41%
4	2	▼	C	9.53%	-1.64%
5	5		C#	4.87%	-2.67%
6	6		JavaScript	3.46%	+0.08%
7	8	▲	Go	2.78%	+1.22%
8	7	▼	SQL	2.57%	+0.65%
9	10	▲	Visual Basic	2.52%	+1.09%
10	15	▲	Delphi/Object Pascal	2.15%	+0.94%
11	14	▲	Fortran	1.70%	+0.48%
12	9	▼	Scratch	1.66%	+0.21%
13	12	▼	PHP	1.48%	+0.16%
14	17	▲	Rust	1.23%	+0.20%
15	13	▼	MATLAB	0.98%	-0.26%
16	21	▲	R	0.94%	+0.13%

1.5 Installation

1.5.1 Environment management

- What are environments: a Python environment is a self-contained directory that contains a specific Python version and various packages.
- Why it's crucial: different projects may require different versions of Python or libraries, and environments prevent conflicts between these requirements.

1.5.2 conda for environment management

- Download miniforge here, <https://github.com/conda-forge/miniforge>
- open your terminal/command prompt

```
conda create --name stat_207 python=3.10
```

```
conda activate stat_207
```

```
conda install numpy
```

```
conda deactivate
```

```
conda env remove --name myenv
```

1.5.3 Visual Studio Code (VS Code)

- a lightweight, cross-platform code editor with an extensive extension system
- get started with VS code, <https://code.visualstudio.com/docs/setup/setup-overview>

1.5.4 Jupyter lab/notebook

[website](#)

```
conda install matplotlib
```

```
conda install -c conda-forge jupyterlab
```

Once installed, launch JupyterLab with:

```
jupyter lab
```

Install the classic Jupyter Notebook with:

```
conda install notebook
```

To run the notebook:

```
jupyter notebook
```

1.5.5 (Optional)

- Google Colab: <https://colab.research.google.com/>

For deep learning and training deep neural networks,

- [TensorFlow](#)
- [PyTorch](#)