Predicting Hard Disk Drive Access Times with Fourier Analysis and Neural Nets

Adam Crume†, Carlos Maltzahn† Lee Ward‡, Thomas Kroeger‡, Matthew Curry‡, Ron Oldfield‡, Patrick Widener‡
†University of California, Santa Cruz
‡Sandia National Laboratories, Livermore, CA

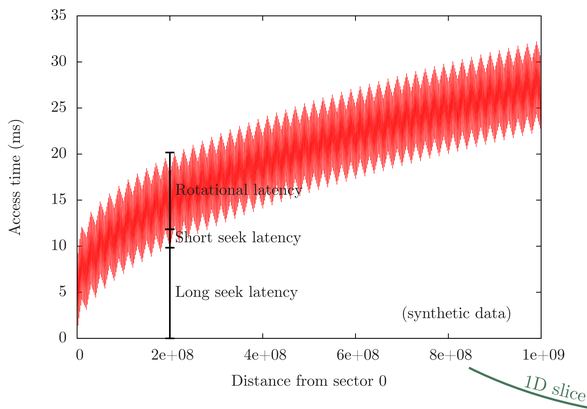
Baskin Engineering UC SANTA CRUZ

Problem

Hard disk drive performance models are often used as part of a much larger system simulation. The most accurate models are white-box models such as DiskSim. Modern hard disk drives are complex, meaning that these models are also complex, and parameterizing the models is a long and difficult process. Tools such as DIG can extract this information. Unfortunately, they require assumptions about the internal structure of the hard disk drive. This structure is likely to change in the future due to the introduction of shingled hard disk drives or other optimizations, as has happened in the past with Zoned Bit Recording, serpentine layouts, etc. Manufacturers do not release this information, so researchers must reverse-engineer a device before modifying DIG and DiskSim to support the new layout.

A more desirable approach is to use machine learning to reproduce the behavior with as few assumptions as possible. Some progress has been made in behavioral modeling of hard disk drive performance, but none can accurately model individual requests.

Cause

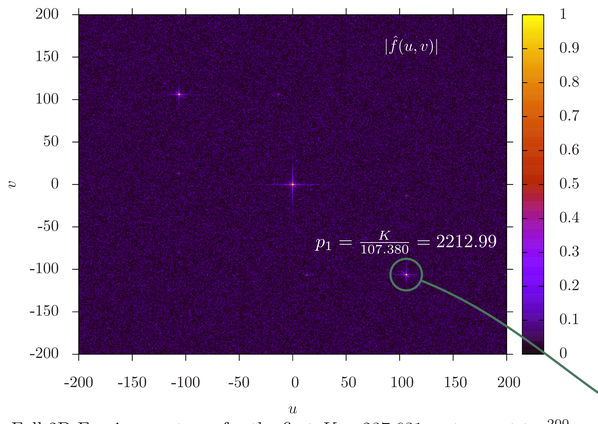


A significant component of the access time is rotation time, which is very high frequency (about a million oscillations, even worse than what is depicted above). Most machine learning approaches cannot learn periodic patterns, except by memorization of every period.

Solution

To find the frequencies, we first perform a Fourier analysis of access times in a random read workload. We then augment the input vector (which originally contained only the start and end sectors) with sines and cosines of the start and end sectors using the strong frequencies.

Fourier transform



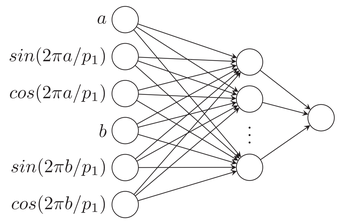
Full 2D Fourier spectrum for the first $K = 237{,}631$ sectors out to $\frac{200}{K} \times \frac{200}{K} = 8.42 \cdot 10^{-4} \times 8.42 \cdot 10^{-4}$, which corresponds to periods of at least $1/(8.42 \cdot 10^{-4}) = 1188.155$ sectors. Plot is clipped to magnitude 1 to show detail, but central spike goes up to 8.6, and other diagonal spikes go up to 3.9. Because our input data is sparse, we cannot use the FFT and must compute the Fourier transform with brute force. However, note that strong frequencies lie on the diagonal $v = -u$. By searching only this line, we reduce computation time considerably.

$$\hat{f}(u,v) = \iint f(a,b) e^{-2\pi i(au+bv)} \, da \, db$$

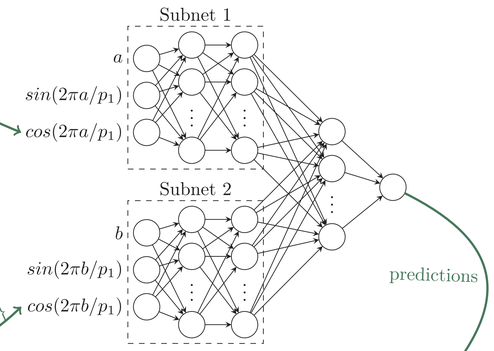
spatial data $f(a,b)$

p_1

Neural net with shared weights

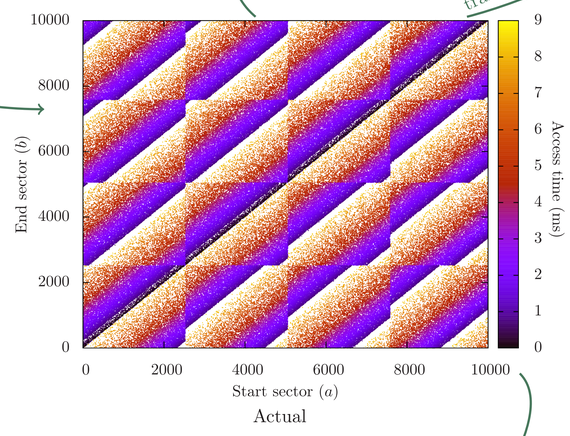


We know that we have two groups of inputs that are structurally identical. By encoding this into the neural net, we reduce the error. Weight sharing means that all the connection weights in subnet 1 are equal to the corresponding connection weights in subnet 2.

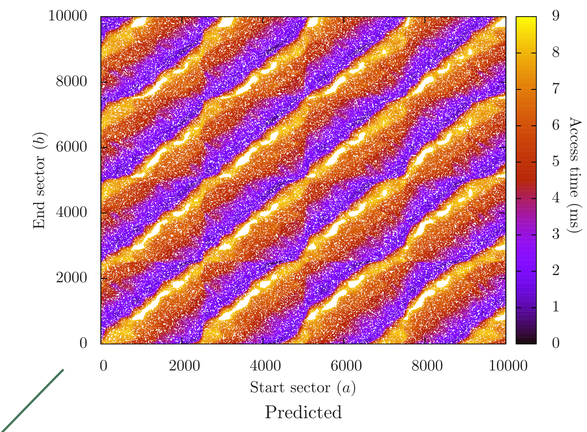


training data $\{(a, b, f(a,b))\}$

Access times over the first 4 tracks



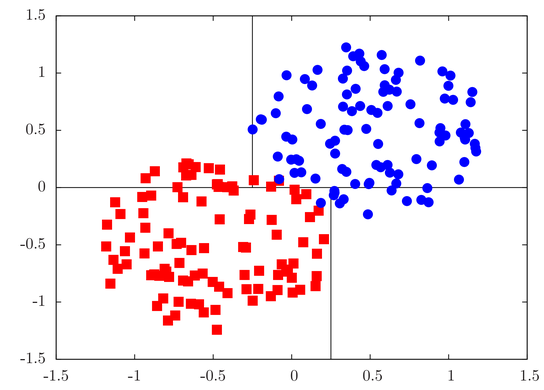
Actual



Predicted

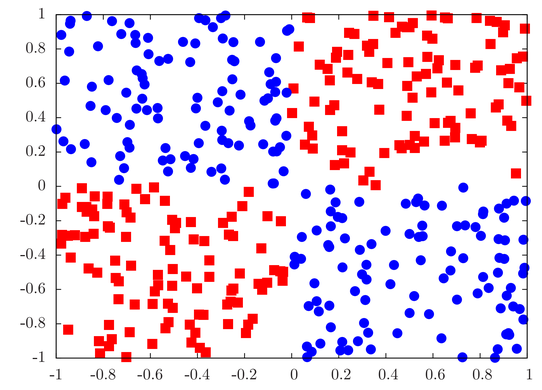
1D slice of

Decision trees



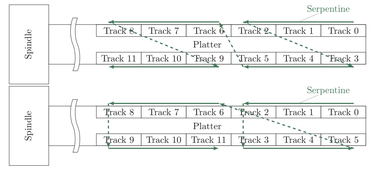
Decision trees work by recursively partitioning the parameter space. For efficiency reasons, they usually use axis-orthogonal splits, and they are greedy (meaning they always choose the split that results in the best immediate partitioning).

Interdependence



Our input parameters are highly interdependent. This means that knowledge of a single value gives you no knowledge of the output value. Knowledge of multiple input values are necessary to gain any knowledge of the output value. This is a hard case for decision trees.

Serpentines

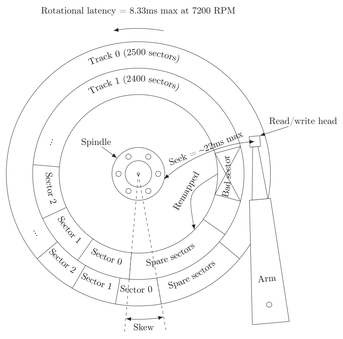


Results

	Configuration	Error (ms)
	constant value	2.013 ± 0.000
Decision trees	no periods, without bagging	2.075 ± 0.014
	no periods, with bagging	2.067 ± 0.001
	6 random periods, without bagging	2.019 ± 0.013
	6 random periods, with bagging	2.015 ± 0.013
	6 periods, without bagging	1.649 ± 0.154
	6 periods, with bagging	1.123 ± 0.009
Neural nets	no periods, without subnets	2.014 ± 0.034
	no periods, with subnets	2.012 ± 0.019
	6 random periods, without subnets	1.924 ± 0.176
	6 random periods, with subnets	1.992 ± 0.059
	6 periods, without subnets	0.954 ± 0.052
	6 periods, with subnets	0.830 ± 0.031

RMS errors for predictions over the first 237,631 sectors (94 tracks) with a random read workload.

Hard drive internals



Funding

This work was supported by Sandia National Laboratories. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.