William Conner DiPaolo

301 Platt Boulevard
Claremont, CA 91711
United States of America
M +1 (949) 300 3774
E cdipaolo@hmc.edu
W github.com/cdipaolo
United States Citizen

Education

BSc Mathematics, *Harvey Mudd College*, Claremont, CA. **2015-2019**

3.81 overall GPA. Deans List.

Undergraduate Thesis

Title: *Randomized Numerical Linear Algebra and its Applications to Machine Learning*

Supervisors: Weiqing Gu

Description: Many machine learning problems reduce to linear algebra. We investigate the fundamentals of recent randomized algorithms for approaching these large matrix problems, for example proving lower bounds on randomized trace estimation, and apply these fundamentals to speeding up computations for large scale machine learning problems.

Awards

Giovanni Borrelli Mathematics Prize, *Harvey Mudd College*. **2018**

Courtney S. Coleman Prize, *Harvey Mudd College*. **2017**

Talks

A Briskly Paced Overview of Convex Optimization, *Yelp*, San Francisco, CA. **08/2016**

Teaching Experience

MATH173: Advanced Linear Algebra, **09-12/2018**
Teaching Assistant, Harvey Mudd College.

Prepared syllabus, wrote homeworks, and lectured course material.

MATH189R: Mathematics of Big Data, **09-12/2016**
Teaching Assistant, Harvey Mudd College.

Prepared syllabus, wrote homeworks and exam, and lectured material.

Research Experience

Randomized Numerical Linear Algebra, *Harvey Mudd College*, with Weiqing Gu. **09/2018-Present**

Numerical questions like computing the number of eigenvalues of a gigantic matrix in the interval $[0, 1]$ can be reduced approximately to computing traces of matrices for which matrix-vector products are much easier to

compute than the matrix itself. **Randomized algorithms** for estimating these traces in the matrix query model goes back to 1990 but have seen recent interest in the machine learning community. I proved answers to fundamental questions about optimality of these randomized algorithms in the form of **lower bounds**, as well as showing why randomization is necessary, by using some classic probability bounds and tools from optimization theory like resisting oracles.

Optical Communications, *Jet Propulsion Laboratory*, Deep Space Optical Communications Team. **06-08/2018**

Optical communications is slated to enter deep space for the first time with the Psyche mission in 2022. Fine-tune **control** of the telescope receiving this data is necessary to prohibit data loss at far range, but previous algorithms for this aren't designed for the high background noise and low resolution receivers. I closed this control loop by reducing our fine-tuning to signal intensity deconvolution via an **inverse problem**, proving these are equivalent using results from the 1980s concerning **mathematical program stability**. At this point, I was able to prove by concentration techniques **sample complexity bounds** on an existing deconvolution procedure, as well as making a small novel modification that speeds up computation by a factor of 2 at no loss of accuracy. Our paper is undergoing release, but regardless these algorithms will make our communications link robust to known jitter in the receiver for the 2022 launch.

Operator Theory, *Harvey Mudd College*, with Stephan Garcia. **05-12/2017**

Some physical systems can be modeled by self-adjoint operators on Hilbert space, but other times these systems are actually best modeled as less-studied complex-*symmetric* operators. This work studied the properties of the more general (m, C)-complex symmetric operators. I was able to characterize, for example, when scalar shifts of a complex anti-symmetric operator are (m, C)-symmetric.

Bayesian Statistics, *Harvey Mudd College*, with Weiqing Gu. **05-12/2016**

The densities of real world phenomena can often be written probabalistically as mixtures of simple distributions. In reality, however, both these simple distributions and their mixing weights vary over time or spacial information, and so accurate modeling requires incorporating variation in these parameters. I proposed using **latent Gaussian processes**, distributions over random functions, to incorporate these assumptions into **hierarchical models**. This, done intelligently, allowed us to specify accurate models for the movement of wolves under tagging measurements, or time-varying covariance between stock returns, for example.

Industry Work

Research Intern, *Jet Propulsion Laboratory*, Pasadena, CA, Deep Space Optical Communications Team. **06-08/2018**

Statistical algorithms and concentration/sample complexity bounds for estimation and control of an optical communications link. See 'Research Experience' section above for more details.

Intern, *Yelp*, San Francisco, CA, Ad Creative Team. **05-08/2017**

I was the first person at Yelp to seriously attack the problem of optimizing the photo we showed in these advertisements. I created both a **linear model** on photo metadata, as well as a **custom deep neural network** architecture based on a ResNet which linearly incorporated this same metadata, to predict click rates for every business' advertisement layout, also helping build **deployment infrastructure** for these notoriously hard-to-deploy models. These models were helpful to performance, with the deep neural network somewhat outperforming the linear model. Moreover, by maintaining **interpretibility within the deep model**, I was able to statistically interpret the affect of different metadata holding the *photo content* fixed, discovering a long standing issue in the cropping infrastructure without ever having to look at the internals. In addition to my modeling work, I was able to design, rally engineering effort, and help build a web-app for **automated Bayesian A/B testing** with the potential to be used across all of Ads.

Intern, *Yelp*, San Francisco, CA, Spam Team. **05-08/2016**

Multiple modelling teams at Yelp were building ad-hoc systems to copy large amounts of data into their own servers for easy access without impacting other teams. To unify these approaches, I was the primary contributor to a **database replication** service based on Yelp's then-new data pipeline which marshalled Kafka messages into each team's individual MySQL clusters. In addition to this infrastructure work, I was able to do modelling work retraining and performing performance analysis for a stale **language model** detecting **hate speech and spam** under a team wide model development/deployment framework I helped design. This model is (at least a year later) the best performing within the whole team.

Intern, *Veritone Media*, Newport Beach, CA. **07-08/2016**

I developed a sentiment analysis engine from scratch using my own machine learning library, with training data taken from IMDB reviews.

Lead Engineer, *Soulsoup*, Newport Beach, CA. **12/2015-08/2016**

I developed an API backend and mobile app for a 501(c)(3) nonprofit networking consumers who want to give with soup kitchens who need money. Much of my work involved moving data between an API written and Golang and a PostgreSQL database.

Open Source Work

GOML, *Creator*. `github.com/cdipaolo/goml`

Second most popular Golang machine learning library (>900 stars on Github), and the only machine learning library to use lightweight parallelism to easily train models in an online way from multiple data sources.

Languages

Native: English **Conversational**: Mandarin Chinese

Programming Languages and Technical Skills

Significant Experience: Python, LaTeX, Tikz, Golang

Other Experience: HTML, Javascript, R, Java, C

Tools: Numpy, Scipy, Matplotlib, Pandas, Scikit-Learn, Statsmodels, PyTorch, Flask

References

Furnished upon request.