Re-evaluating scientific claims with a multiverse analysis

Guest Lecture, L48 Machine Learning for the Physical World

Samuel J. Bell

PhD Student, ML@CL, University of Cambridge

Outline

1. Motivation: efficient machine learning
2. Introducing the multiverse analysis
3. Modeling the machine learning multiverse
4. Case study 1: adaptive optimizers
5. Case study 2: large-batch generalization gap
6. Future stuff, discussion

Outline

1. Motivation: efficient machine learning
2. Introducing the multiverse analysis
3. Modeling the machine learning multiverse
4. Case study 1: adaptive optimizers
5. Case study 2: large-batch generalization gap
6. Future stuff, discussion

Are GANs Created Equal? A Large-Scale Study

Mario Lucic* Karol Kurach* Marcin Michalski Olivier Bousquet Sylvain Gelly
Google Brain

“Despite a very rich research activity leading to numerous interesting GAN algorithms…

Are GANs Created Equal? A Large-Scale Study

Mario Lucic* Karol Kurach* Marcin Michalski Olivier Bousquet Sylvain Gelly
Google Brain

“Despite a very rich research activity leading to numerous interesting GAN algorithms…

…we find that most models can reach similar scores with enough hyperparameter optimization and random restarts.”

On the State of the Art of Evaluation in Neural Language Models

Gábor Melis[†], Chris Dyer[†], Phil Blunsom[†‡]
{melisgl,cdyer,pblunsom}@google.com
[†]DeepMind
[‡]University of Oxford

"Ongoing innovations … state-of-the-art results on language modelling benchmarks…

ON THE STATE OF THE ART OF EVALUATION IN NEURAL LANGUAGE MODELS

Gábor Melis[†], **Chris Dyer**[†], **Phil Blunsom**[†‡]
{melisgl,cdyer,pblunsom}@google.com
[†]DeepMind
[‡]University of Oxford

"Ongoing innovations … state-of-the-art results on language modelling benchmarks…

…standard LSTM architectures, when properly regularised, outperform more recent models."

Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches

Maurizio Ferrari Dacrema
Politecnico di Milano, Italy
maurizio.ferrari@polimi.it

Paolo Cremonesi
Politecnico di Milano, Italy
paolo.cremonesi@polimi.it

Dietmar Jannach
University of Klagenfurt, Austria
dietmar.jannach@aau.at

"…difficult to keep track of what represents the state-of-the-art at the moment…

Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches

Maurizio Ferrari Dacrema
Politecnico di Milano, Italy
maurizio.ferrari@polimi.it

Paolo Cremonesi
Politecnico di Milano, Italy
paolo.cremonesi@polimi.it

Dietmar Jannach
University of Klagenfurt, Austria
dietmar.jannach@aau.at

"…difficult to keep track of what represents the state-of-the-art at the moment…

…recently proposed neural methods do not even outperform conceptually or computationally simpler, sometimes long-known, algorithms."

Do Transformer Modifications Transfer Across Implementations and Applications?

Sharan Narang[*] Hyung Won Chung Yi Tay William Fedus
Thibault Fevry[†] Michael Matena[†] Karishma Malkan[†] Noah Fiedel
Noam Shazeer Zhenzhong Lan[†] Yanqi Zhou Wei Li
Nan Ding Jake Marcus Adam Roberts Colin Raffel[†]

"The research community has proposed copious modifications to the Transformer architecture…

Do Transformer Modifications Transfer Across Implementations and Applications?

Sharan Narang* Hyung Won Chung Yi Tay William Fedus

Thibault Fevry† Michael Matena† Karishma Malkan† Noah Fiedel

Noam Shazeer Zhenzhong Lan† Yanqi Zhou Wei Li

Nan Ding Jake Marcus Adam Roberts Colin Raffel†

“The research community has proposed copious modifications to the Transformer architecture…

…we find that most modifications do not meaningfully improve performance…

…performance improvements may strongly depend on implementation details.”

Replication failures

- Each of these examples are replication failures

Replication failures

- Each of these examples are replication failures
- Every failure is wasted time, effort and resources

Replication failures

- Each of these examples are replication failures
- Every failure is wasted time, effort and resources
- This is bad for us as researchers, for scientific progress, and for society
  - e.g. the PhD student building on top of flawed foundations
  - e.g. wasted public funding poured into fruitless research
  - e.g. vast climate impact of pointless deep learning research

Robust scientific conclusions

- If we want our research to count, we need conclusions that are reproducible
  - i.e., other researchers can test the same claim and get the same result

Robust scientific conclusions

- If we want our research to count, we need conclusions that are reproducible
  - i.e., other researchers can test the same claim and get the same result

- But we also want conclusions that generalize
  - i.e., conclusions that hold in spite of irrelevant details changing

Robust scientific conclusions

- If we want our research to count, we need conclusions that are reproducible
  - i.e., other researchers can test the same claim and get the same result
- But we also want conclusions that generalize
  - i.e., conclusions that hold in spite of irrelevant details changing
- “Model X is the best” isn’t useful if only true under specific conditions
  - e.g., choice of benchmark, choice of hyperparameters, choice of architecture …

Outline

1. Motivation: efficient machine learning
2. Introducing the multiverse analysis
3. Modeling the machine learning multiverse
4. Case study 1: adaptive optimizers
5. Case study 2: large-batch generalization gap
6. Future stuff, discussion

An example from social psychology

An example from social psychology

- **Claim:** Fertility influences women’s religious & political preferences. [1]

[1] Durante et al. (2013). The Fluctuating Female Vote: Politics, Religion, and the Ovulatory Cycle. Psychological Science.

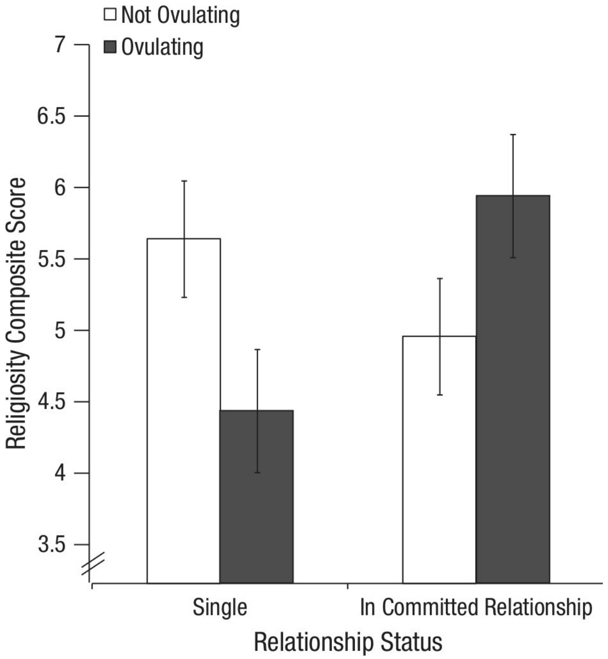
An example from social psychology

- **Claim:** Fertility influences women's religious & political preferences. [1]
- **Methods:** 502 women surveyed about religiosity, political attitudes, relationship status and start date of menstrual cycle.

[1] Durante et al. (2013). The Fluctuating Female Vote: Politics, Religion, and the Ovulatory Cycle. Psychological Science.

An example from social psychology

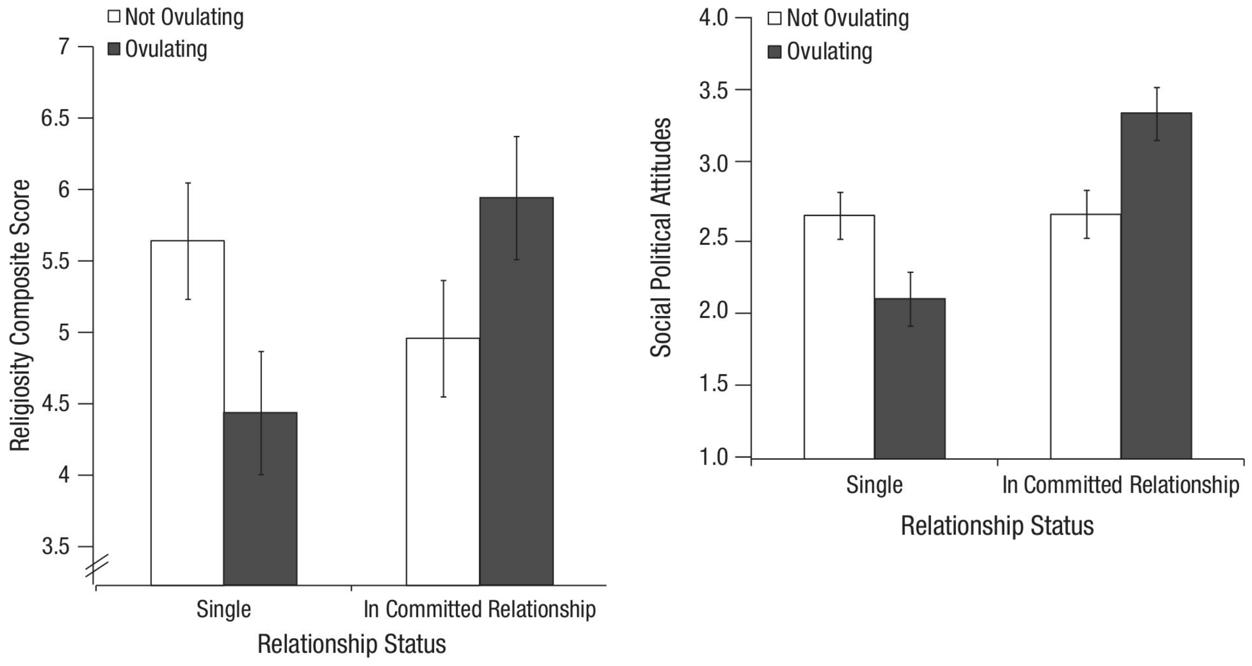
- **Claim:** Fertility influences women's religious & political preferences. [1]
- **Methods:** 502 women surveyed about religiosity, political attitudes, relationship status and start date of menstrual cycle.



[1] Durante et al. (2013). The Fluctuating Female Vote: Politics, Religion, and the Ovulatory Cycle. Psychological Science.

An example from social psychology

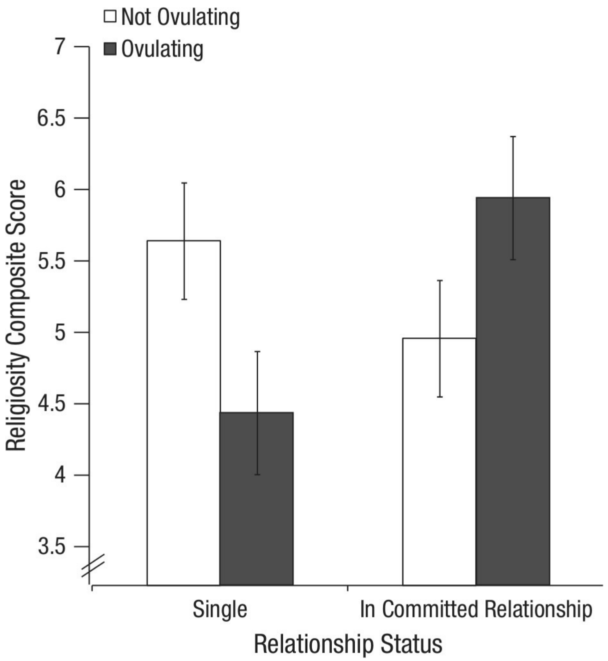
- **Claim:** Fertility influences women's religious & political preferences. [1]
- **Methods:** 502 women surveyed about religiosity, political attitudes, relationship status and start date of menstrual cycle.

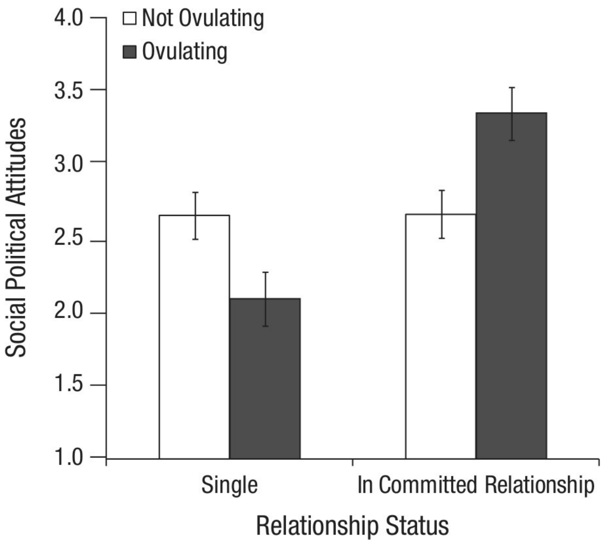


[1] Durante et al. (2013). The Fluctuating Female Vote: Politics, Religion, and the Ovulatory Cycle. Psychological Science.

An example from social psychology

- **Claim:** Fertility influences women's religious & political preferences. [1]
- **Methods:** 502 women surveyed about religiosity, political attitudes, relationship status and start date of menstrual cycle.



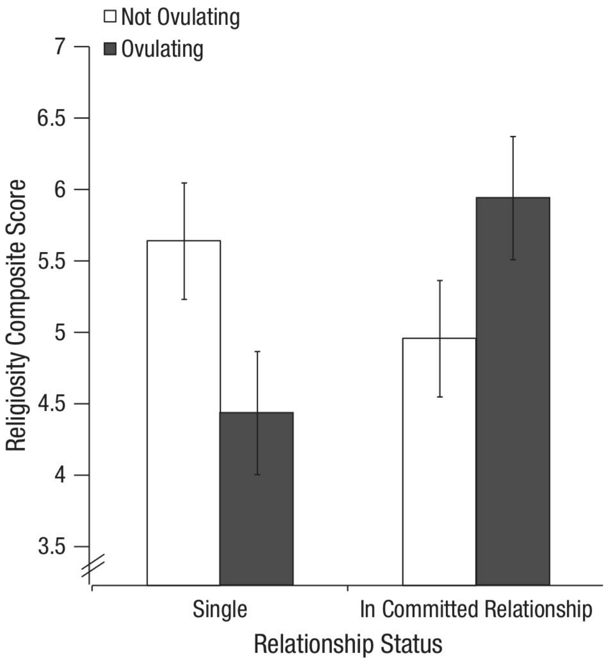


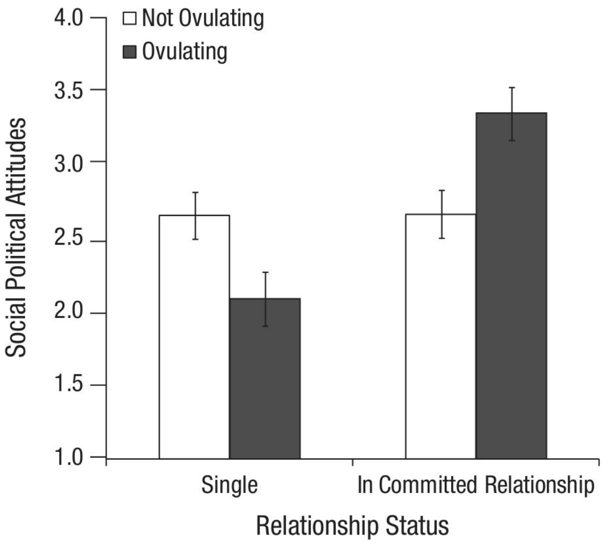
- Results: Fertility x rel. status interaction effect

[1] Durante et al. (2013). The Fluctuating Female Vote: Politics, Religion, and the Ovulatory Cycle. Psychological Science.

An example from social psychology

- **Claim:** Fertility influences women's religious & political preferences. [1]
- **Methods:** 502 women surveyed about religiosity, political attitudes, relationship status and start date of menstrual cycle.





- Results: Fertility x rel. status interaction effect
- Do we believe it?

[1] Durante et al. (2013). The Fluctuating Female Vote: Politics, Religion, and the Ovulatory Cycle. Psychological Science.

Durante’s degrees of freedom

- Durante et al. made a lot of choices about how to do their study [1]:

[1] Steegen et al. (2016). Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science.

Durante’s degrees of freedom

- Durante et al. made a lot of choices about how to do their study [1]:
- Which cycle days are considered “high fertility”?
  - Days 7-14, 6-14, 9-17 or 8-14?

[1] Steegen et al. (2016). Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science.

Durante's degrees of freedom

- Durante et al. made a lot of choices about how to do their study [1]:
- Which cycle days are considered "high fertility"?
  - Days 7-14, 6-14, 9-17 or 8-14?
- How to estimate next menstrual onset?
  - Reported or estimated cycle length?

[1] Steegen et al. (2016). Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science.

Durante’s degrees of freedom

- Durante et al. made a lot of choices about how to do their study [1]:
- Which cycle days are considered “high fertility”?
  - Days 7-14, 6-14, 9-17 or 8-14?
- How to estimate next menstrual onset?
  - Reported or estimated cycle length?
- What counts as “in a relationship”?
  - Does “dating” mean “single”?

[1] Steegen et al. (2016). Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science.

Durante’s degrees of freedom

- Durante et al. made a lot of choices about how to do their study [1]:
- Which cycle days are considered “high fertility”?
  - Days 7-14, 6-14, 9-17 or 8-14?
- How to estimate next menstrual onset?
  - Reported or estimated cycle length?
- What counts as “in a relationship”?
  - Does “dating” mean “single”?
- Outlier exclusion criteria

[1] Steegen et al. (2016). Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science.

Durante's degrees of freedom

- Durante et al. made a lot of choices about how to do their study [1]:
- Which cycle days are considered "high fertility"?
  - Days 7-14, 6-14, 9-17 or 8-14?
- How to estimate next menstrual onset?
  - Reported or estimated cycle length?
- What counts as "in a relationship"?
  - Does "dating" mean "single"?
- Outlier exclusion criteria

210 possible combinations

[1] Steegen et al. (2016). Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science.

Multiverse analysis

Religiosity (Study 2)

R1					R2					R3							
F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5			
0	0	0.03	0.02	0.08	0.02	0.02	0.03	0.05	0.06	0	0	0.02	0.01	0.04	EC1	ECL1	NMO1
0.01	0	0 .08	0.03	0.16	0.07	0.05	0.18	0.24	0.41	0.01	0	0.09	0.06	0.23	EC2		
0	0	0.06	0.04	0.37	0.02	0.03	0.07	0.08	0.21	0	0	0.05	0.03	0.23	EC1	ECL2	
0.01	0	0 .13	0.08	0.44	0.06	0.03	0.22	0.24	0.52	0.01	0	0.14	0.09	0.43	EC2		
0	0	0.03	0.01	0.08	0.15	0.07	0.17	0.07	0.14	0.02	0.01	0.06	0.02	0.07	EC1	ECL1	NMO2
0	0	0.02	0.01	0.06	0.2	0.05	0.42	0.23	0.44	0.03	0	0.14	0.05	0.19	EC2		
0.01	0.01	0.05	0.01	0.1	0.39	0.2	0.45	0.11	0.26	0.08	0.04	0.17	0.03	0.13	EC 1	ECL3	
0.01	0	0 .05	0.02	0.11	0.33	0.09	0.59	0.26	0.55	0.09	0.02	0.26	0.08	0.27	EC 2		
0.01	0.01	0.02	0.1	0.28	0.11	0.09	0.43	0.26	0.85	0.02	0.02	0.12	0.12	0.51	EC1	ECL1	NMO3
0.01	0.01	0	0 .07	0.06	0.07	0.1	0.11	0.14	0.23	0.01	0.02	0.02	0.06	0.08	EC2		
0.02	0.01	0.06	0.11	0.36	0.06	0.04	0.3	0.13	0.66	0.02	0.01	0.13	0.07	0.46	EC1	ECL2	
0.02	0.01	0.02	0.15	0.13	0.04	0.05	0.07	0.07	0.16	0.01	0.02	0.03	0.05	0.09	EC2		
0.07	0.04	0.12	0.09	0.16	0.16	0.11	0.54	0.32	0.77	0.07	0.04	0.25	0.13	0.39	EC1	ECL3	
0.02	0.02	0.01	0.06	0.02	0.06	0.1	0.07	0.16	0.17	0.02	0.03	0.02	0.07	0.05	EC2		

[1] Steegen et al. (2016). Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science.

Multiverse analysis

Social political attitudes

			R1					R2					R3				
			F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
NMO1	ECL1	EC 1	0	0	0	0	0	0.18	0.19	0.3	0.07	0.21	0.01	0.01	0.04	0	0.03
		EC 2	0.01	0.01	0.02	0	0.02	0.93	0.76	0.99	0.52	0.78	0.24	0.15	0.37	0.06	0.23
	ECL2	EC 1	0	0	0	0	0	0.17	0.26	0.3	0.13	0.27	0.02	0.02	0.06	0.01	0.05
		EC 2	0.02	0.01	0.04	0	0.01	0.89	0.86	0.98	0.47	0.66	0.24	0.2	0.42	0.05	0.19
NMO2	ECL1	EC 1	0.01	0.01	0.01	0	0	0.47	0.45	0.21	0.06	0.07	0.09	0.11	0.04	0.01	0.01
		EC 2	0.06	0.05	0.02	0.01	0	0.94	0.93	0.4	0.24	0.13	0.4	0.41	0.13	0.06	0.02
	ECL3	EC 1	0.01	0.01	0	0	0	0.49	0.45	0.22	0.08	0.1	0.1	0.11	0.04	0.01	0.02
		EC 2	0.1	0.11	0.02	0.02	0	0.87	0.91	0.32	0.27	0.18	0.42	0.48	0.12	0.09	0.04
NMO3	ECL1	EC 1	0	0	0.01	0	0.02	0.19	0.22	0.1	0.07	0.29	0.02	0.03	0.02	0.01	0.07
		EC 2	0.01	0	0.01	0	0.01	0.46	0.62	0.08	0.08	0.1	0.1	0.14	0.02	0.01	0.02
	ECL2	EC 1	0.01	0	0.01	0	0.01	0.37	0.27	0.18	0.09	0.23	0.07	0.05	0.03	0.01	0.04
		EC 2	0.03	0.02	0.01	0.02	0.01	0.76	0.84	0.13	0.16	0.1	0.24	0.29	0.04	0.04	0.02
	ECL3	EC 1	0.01	0.01	0.02	0	0	0.37	0.35	0.18	0.09	0.24	0.08	0.08	0.05	0.01	0.04
		EC 2	0.05	0.04	0.03	0.01	0	0.9	0.97	0.22	0.18	0.16	0.36	0.45	0.09	0.04	0.02

[1] Steegen et al. (2016). Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science.

Multiverse analysis

- So, Durante et al.’s claims aren’t robust

Multiverse analysis

- So, Durante et al.’s claims aren’t robust
  - They’re specific to *arbitrary implementation details*
  - Given a different set of choices, the conclusion could just as easily be false

Multiverse analysis

- So, Durante et al.’s claims aren’t robust
    - They’re specific to *arbitrary implementation details*
    - Given a different set of choices, the conclusion could just as easily be false

- **Multiverse analysis:** redoing the analysis at every point in the space of possible choices, and systematically reviewing the conclusions.

Multiverse analysis

- So, Durante et al.'s claims aren't robust
  - They're specific to *arbitrary implementation details*
  - Given a different set of choices, the conclusion could just as easily be false

- **Multiverse analysis:** redoing the analysis at every point in the space of possible choices, and systematically reviewing the conclusions.

- **What does this have to do with machine learning?**

The ML multiverse

- Just like Durante et al., we make decisions *all the time*

The ML multiverse

- Just like Durante et al., we make decisions *all the time*
- “Invention X improves model performance”
  - Model architectures
  - Baselines for comparison
  - Benchmark datasets
  - Training sets
  - Evaluation metrics
  - Termination criteria
  - Countless hyperparameters
  - Hyperparameter search spaces
  - Hyperparameter optimization approaches
  - Implementation libraries

  …

The ML multiverse

- Just like Durante et al., we make decisions *all the time*

- "Invention X improves model performance"
  - Model architectures
  - Baselines for comparison
  - Benchmark datasets
  - Training sets
  - Evaluation metrics
  - Termination criteria
  - Countless hyperparameters
  - Hyperparameter search spaces
  - Hyperparameter optimization approaches
  - Implementation libraries

  …

One
huge
multiverse

Outline

1. Motivation: efficient machine learning
2. Introducing the multiverse analysis
3. Modeling the machine learning multiverse
4. Case study 1: adaptive optimizers
5. Case study 2: large-batch generalization gap
6. Future stuff, discussion

Key challenge

Lots of choices

Key challenge

Lots of choices

+

Continuous dimensions (e.g., most hyperparameters)

Key challenge

Lots of choices

+

Continuous dimensions (e.g., most hyperparameters)

=

A large and **intractable** search space

Key challenge

Lots of choices

+

Continuous dimensions (e.g., most hyperparameters)

=

A large and **intractable** search space

Solution: *Model* the multiverse for efficient exploration

Efficient multiverse exploration

Definitions

- Evaluation function, ℓ

Efficient multiverse exploration

Definitions

- Evaluation function, ℓ
- Search space, $\mathcal{X}$

Efficient multiverse exploration

Definitions

- Evaluation function, ℓ
- Search space, $\mathcal{X}$

Approach

1. Sample an initial design, $X_0 \sim \mathcal{X}$
2. Evaluate ℓ at each point, $Y_0 = \ell(X_0)$
3. Fit a GP model f to X_0, Y_0

Efficient multiverse exploration

Definitions

- Evaluation function, ℓ
- Search space, $\mathcal{X}$

Approach

1. Sample an initial design, $X_0 \sim \mathcal{X}$
2. Evaluate ℓ at each point, $Y_0 = \ell(X_0)$
3. Fit a GP model f to X_0, Y_0
4. Use an acquisition function a on f to sample and evaluate a new batch X_i, Y_i

Efficient multiverse exploration

Definitions

- Evaluation function, ℓ
- Search space, $\mathcal{X}$

Approach

1. Sample an initial design, $X_0 \sim \mathcal{X}$
2. Evaluate ℓ at each point, $Y_0 = \ell(X_0)$
3. Fit a GP model f to X_0, Y_0
4. Use an acquisition function a on f to sample and evaluate a new batch X_i, Y_i
5. Repeat steps 2–4 until we have a high-confidence picture of the multiverse

Efficient multiverse exploration

Definitions

- Evaluation function, ℓ
- Search space, $\mathcal{X}$

Approach

1. Sample an initial design, $X_0 \sim \mathcal{X}$
2. Evaluate ℓ at each point, $Y_0 = \ell(X_0)$
3. Fit a GP model f to X_0, Y_0
4. Use an acquisition function a on f to sample and evaluate a new batch X_i, Y_i
5. Repeat steps 2–4 until we have a high-confidence picture of the multiverse

Bayesian experimental design

Bayesian experimental design

Bayesian experimental design

- **Initial design:** Sobol sequence is a *low-discrepancy* sequence

Bayesian experimental design

- **Initial design:** Sobol sequence is a *low-discrepancy* sequence
- **GP surrogate:**
  - $y_i = f(x_i) + \epsilon_i\,, \quad \epsilon_i \sim \mathcal{N}$
  - $f \sim \text{GP}(0, k)$

Bayesian experimental design

- **Initial design:** Sobol sequence is a *low-discrepancy* sequence
- **GP surrogate:**
  - $y_i = f(x_i) + \epsilon_i\,, \quad \epsilon_i \sim \mathcal{N}$
  - $f \sim \text{GP}(0, k)$
- **Acquisition function:** Integrated posterior variance reduction (IVR) [1]
  - Next point is the one which lowers the *overall* variance the most

[1] Sacks et al. (1989) Design and analysis of computer experiments. *Statistical Science.*

Bayesian experimental design

- **Initial design:** Sobol sequence is a *low-discrepancy* sequence
- **GP surrogate:**
  - $y_i = f(x_i) + \epsilon_i\,, \quad \epsilon_i \sim \mathcal{N}$
  - $f \sim \text{GP}(0, k)$
- **Acquisition function:** Integrated posterior variance reduction (IVR) [1]
  - Next point is the one which lowers the *overall* variance the most
  - $a\big(x_{i+1}; X_i, Y_i\big) = \int_{\mathcal{X}} \sigma^2\big(p; X_{i+1}, Y_{i+1}\big) - \sigma^2\big(p; X_i, Y_i\big)\, dp$
  - Monte Carlo approximate the integral over the whole search space

[1] Sacks et al. (1989) Design and analysis of computer experiments. *Statistical Science.*

To explore or optimize?

- In Bayesian optimization, we might use an optimization-focused acquisition function, like Upper Confidence Bound (UCB) [1]
  - Next point is either: expected high reward, or high information gain

[1] Srinivas et al. (2010) Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. ICML.

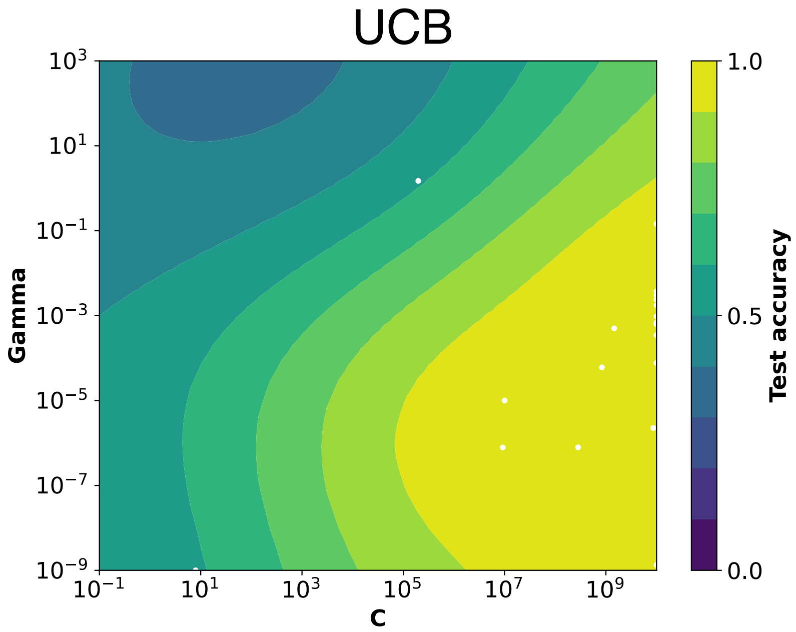
To explore or optimize?

- In Bayesian optimization, we might use an optimization-focused acquisition function, like Upper Confidence Bound (UCB) [1]
  - Next point is either: expected high reward, or high information gain
  - $a(x_{i+1}; X_i, Y_i) = \mu(x_{i+1}; X_i, Y_i) + \beta^{1/2}\, \sigma^2(x_{i+1}; X_i, Y_i)$

[1] Srinivas et al. (2010) Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. ICML.

To explore or optimize?

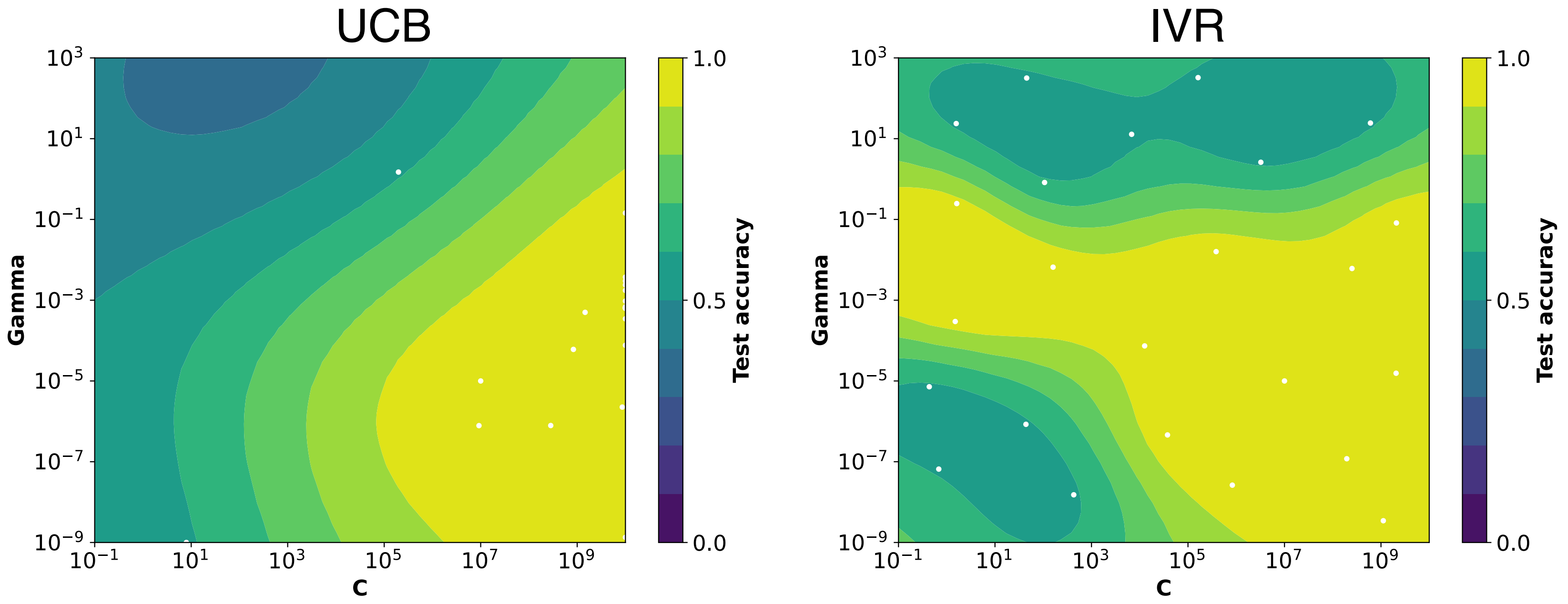
- In Bayesian optimization, we might use an optimization-focused acquisition function, like Upper Confidence Bound (UCB) [1]
  - Next point is either: expected high reward, or high information gain
  - $a(x_{i+1}; X_i, Y_i) = \mu(x_{i+1}; X_i, Y_i) + \beta^{1/2}\, \sigma^2(x_{i+1}; X_i, Y_i)$



[1] Srinivas et al. (2010) Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. ICML.

To explore or optimize?

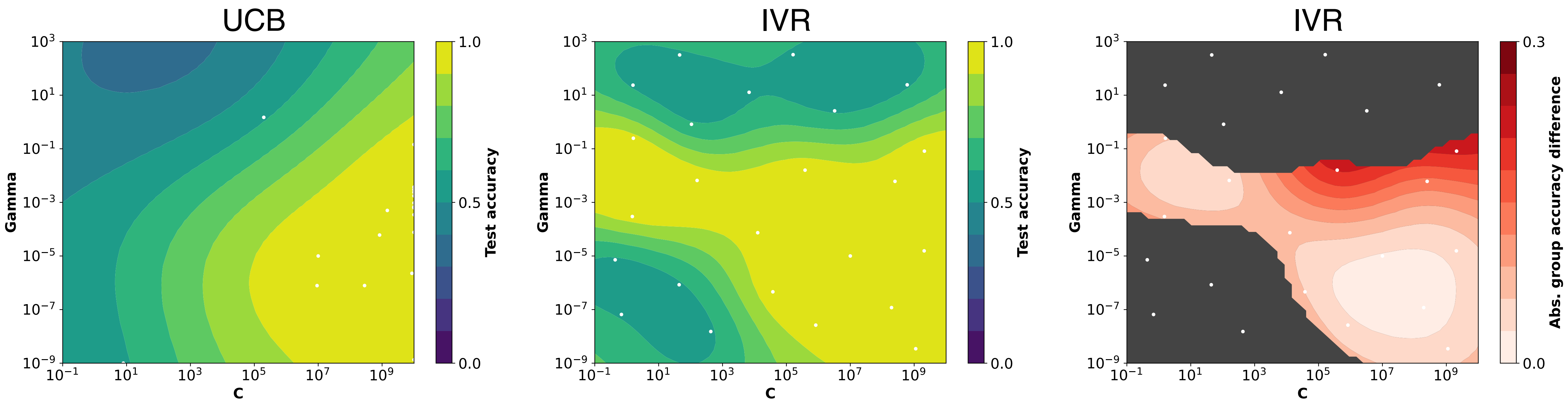
- In Bayesian optimization, we might use an optimization-focused acquisition function, like Upper Confidence Bound (UCB) [1]
  - Next point is either: expected high reward, or high information gain
  - $a(x_{i+1}; X_i, Y_i) = \mu(x_{i+1}; X_i, Y_i) + \beta^{1/2}\, \sigma^2(x_{i+1}; X_i, Y_i)$



[1] Srinivas et al. (2010) Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. ICML.

To explore or optimize?

- In Bayesian optimization, we might use an optimization-focused acquisition function, like Upper Confidence Bound (UCB) [1]
  - Next point is either: expected high reward, or high information gain
  - $a(x_{i+1}; X_i, Y_i) = \mu(x_{i+1}; X_i, Y_i) + \beta^{1/2}\,\sigma^2(x_{i+1}; X_i, Y_i)$



[1] Srinivas et al. (2010) Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. ICML.

Putting it together

- We want to understand the generality and robustness of conclusions
- So we explore the effect of researcher choices
- By modelling the multiverse using a GP surrogate
- Selecting the most informative points to evaluate using IVR

Outline

1. Motivation: efficient machine learning
2. Introducing the multiverse analysis
3. Modeling the machine learning multiverse
4. Case study 1: adaptive optimizers
5. Case study 2: large-batch generalization gap
6. Future stuff, discussion

Adam vs. SGD

- Two common optimizers for training deep neural networks:

Adam vs. SGD

- Two common optimizers for training deep neural networks:
  - SGD w. momentum
    - $\theta_t = \theta_{t-1} - \alpha d_t, \quad d_t = \mu d_{t-1} + g_t$

Adam vs. SGD

- Two common optimizers for training deep neural networks:
  - SGD w. momentum
    - $\theta_t = \theta_{t-1} - \alpha d_t , \quad d_t = \mu d_{t-1} + g_t$
  - Adam [1]
    - $d_t = \frac{\widehat{m_t}}{\sqrt{\widehat{v_t}}+\epsilon}$

[1] Kingma & Ba (2014). Adam: A method for stochastic optimization. *ICLR*.

Adam vs. SGD

- Two common optimizers for training deep neural networks:
  - SGD w. momentum
    - $\theta_t = \theta_{t-1} - \alpha d_t\,, \quad d_t = \mu d_{t-1} + g_t$
  - Adam [1]
    - $d_t = \frac{\widehat{m_t}}{\sqrt{\widehat{v_t}}+\epsilon}$
- Lots of back and forth about which is best e.g. [2, 3]

[1] Kingma & Ba (2014). Adam: A method for stochastic optimization. *ICLR*.
[2] Wilson et al. (2017). The marginal value of adaptive gradient methods in machine learning. *NeurIPS*.
[3] Choi et al. (2019). On empirical comparisons of optimizers for deep learning. *ICLR*.

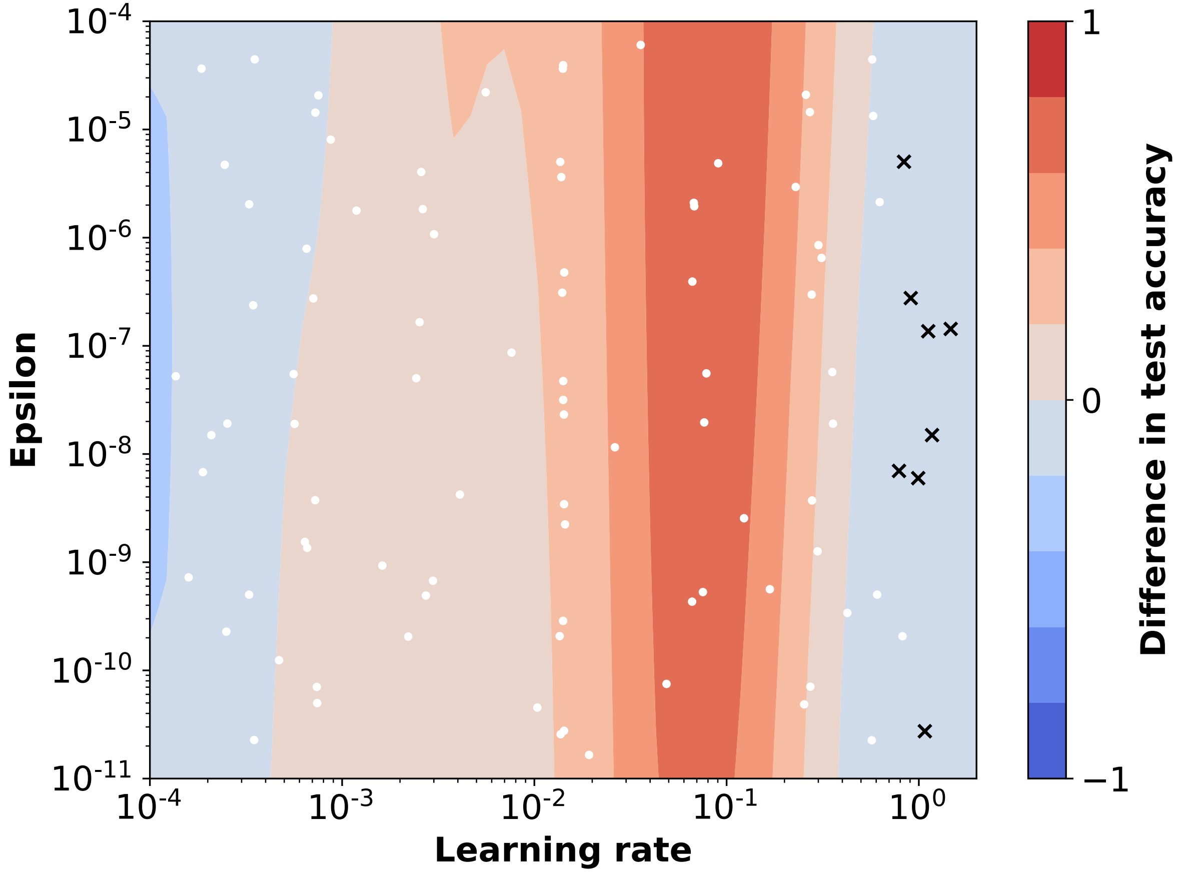
Adam vs. SGD

Multiverse 1: Are adaptive optimizers helpful?

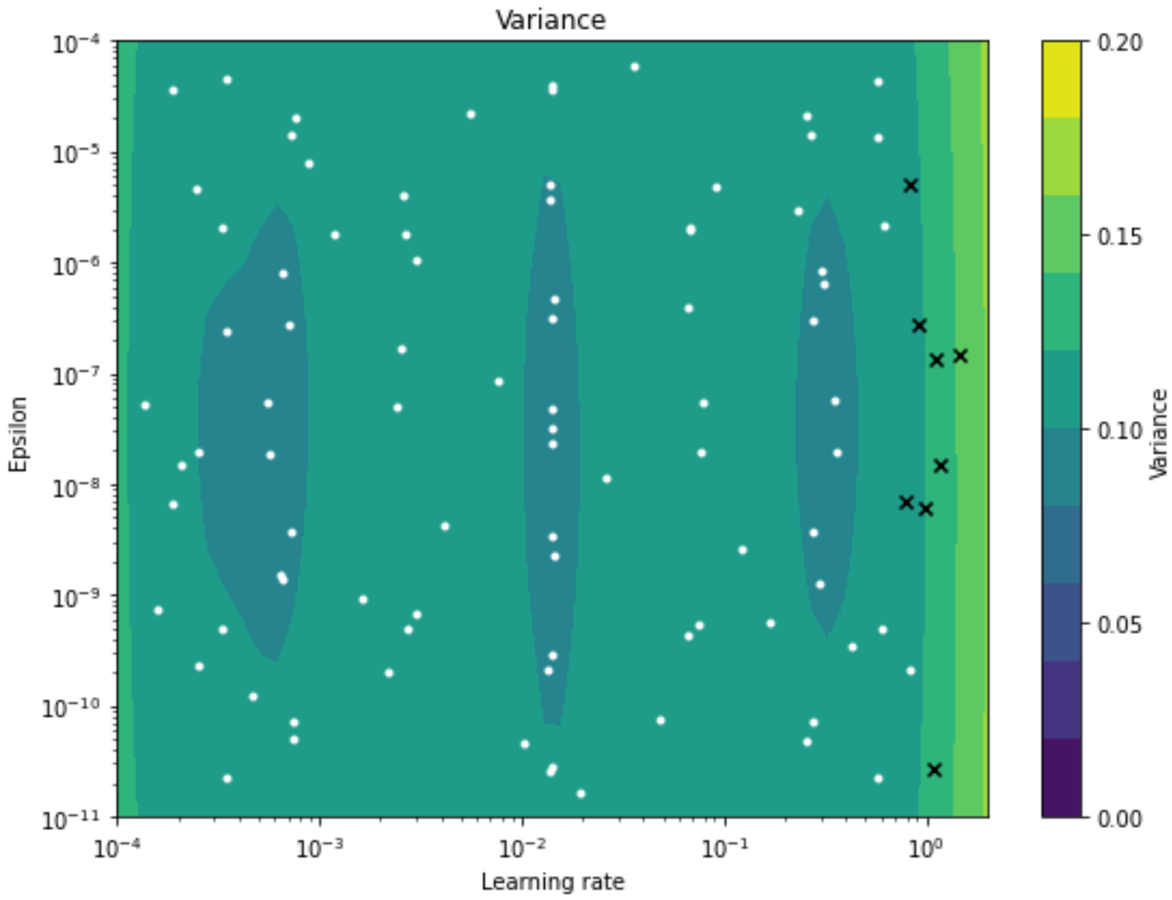
ℓ: $\mathrm{acc}_{SGD} - \mathrm{acc}_{Adam}$

X: $\mathrm{LR} \times \epsilon$

Adam vs. SGD

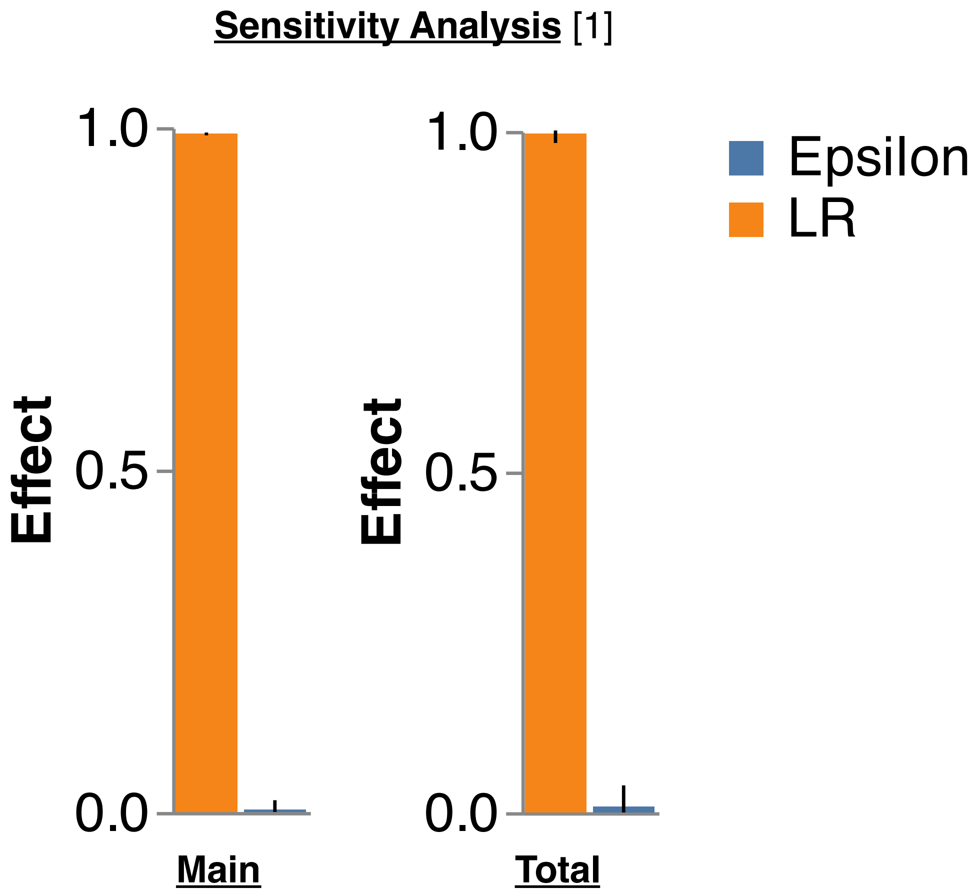


Epsilon
Learning rate
Difference in test accuracy
1
0
−1



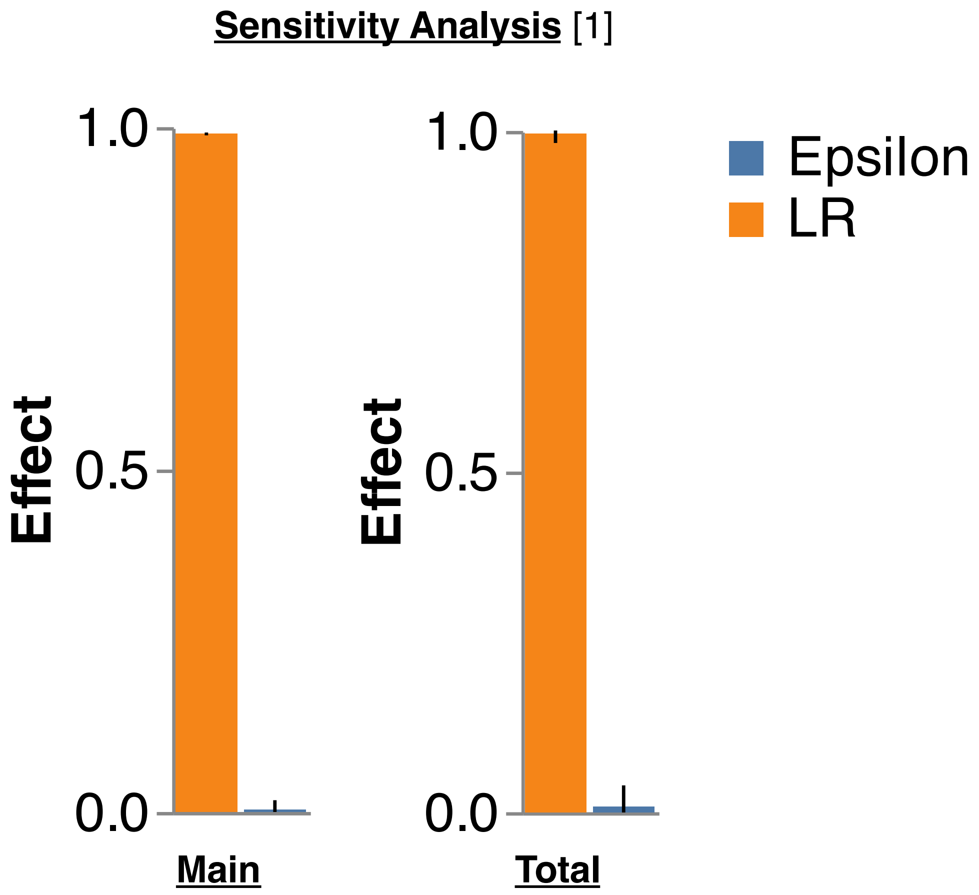
Variance
Epsilon
Learning rate
Variance
0.20
0.15
0.10
0.05
0.00

Adam vs. SGD



[1] Sobol (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation.*

Adam vs. SGD



1. No conclusive "best" optimizer
2. Conclusions vary by learning rate
3. No effect of ϵ

[1] Sobol (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation.*

Outline

1. Motivation: efficient machine learning
2. Introducing the multiverse analysis
3. Modeling the machine learning multiverse
4. Case study 1: adaptive optimizers
5. Case study 2: large-batch generalization gap
6. Future stuff, discussion

The large-batch generalization gap

- When training neural networks, we use *mini-batch* SGD
- But how large should the batch be?

The large-batch generalization gap

- When training neural networks, we use *mini-batch* SGD

- But how large should the batch be?

- Some evidence of a *generalization gap* at large batch sizes, e.g. [1]

- Some evidence against that, e.g. [2]

[1] Keskar et al. (2016). On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *ICLR.*
[2] Hoffer et al. (2017). Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *NeurIPS*.

The large-batch generalization gap

MV 2: Is there a large batch generalization gap?

ℓ: acc

X: LR $\times$ batch size $\times$ dataset $\times$ model

Interlude: multi-fidelity modeling

- How do we model categorical parameters?

Interlude: multi-fidelity modeling

- How do we model categorical parameters?
- Intrinsic Coregionalization Model [1, 2]
  - $K(X, X) = B \otimes k(X, X)$
  - $B_d = \boldsymbol{w}_d \boldsymbol{w}_d^{\top} + \mathrm{diag}(\boldsymbol{\kappa}_d)$

[1] Helterbrand & Cressie (1994). Universal cokriging under intrinsic coregionalization. Mathematical Geology.
[2] Alvarez et al. (2012). Kernels for vector-valued functions: A review. Foundations and Trends in Machine Learning.

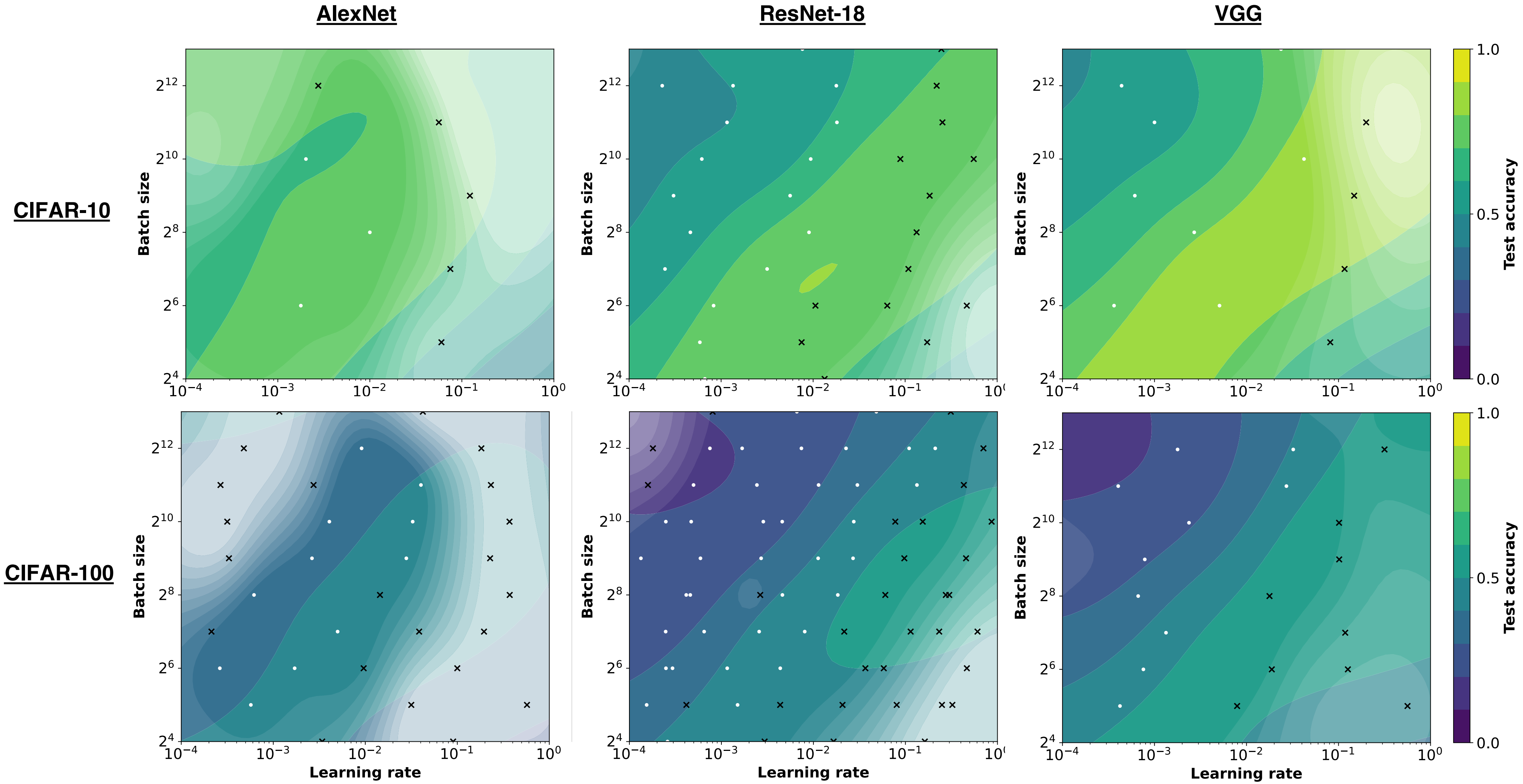
Interlude: multi-fidelity modeling

- How do we model categorical parameters?
- Intrinsic Coregionalization Model [1, 2]
  - $K(X, X) = B \otimes k(X, X)$
  - $B_d = \boldsymbol{w}_d \boldsymbol{w}_d^\top + \mathrm{diag}(\boldsymbol{\kappa}_d)$

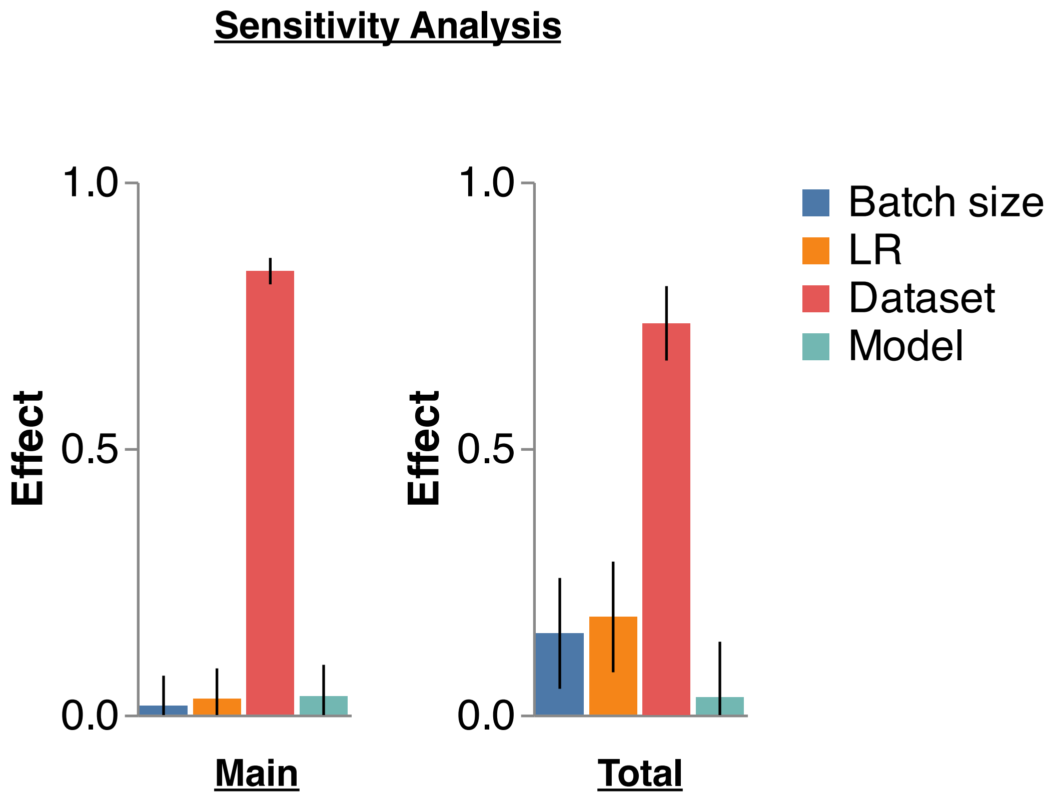
- Treat each dataset x model pair as a separate function
  - $K(X, X) = B_m \otimes B_d \otimes k(X, X)$
  - $B_m = \boldsymbol{w}_m \boldsymbol{w}_m^\top + \mathrm{diag}(\boldsymbol{\kappa}_m)$
  - $B_d = \boldsymbol{w}_d \boldsymbol{w}_d^\top + \mathrm{diag}(\boldsymbol{\kappa}_d)$

[1] Helterbrand & Cressie (1994). Universal cokriging under intrinsic coregionalization. Mathematical Geology.
[2] Alvarez et al. (2012). Kernels for vector-valued functions: A review. Foundations and Trends in Machine Learning.

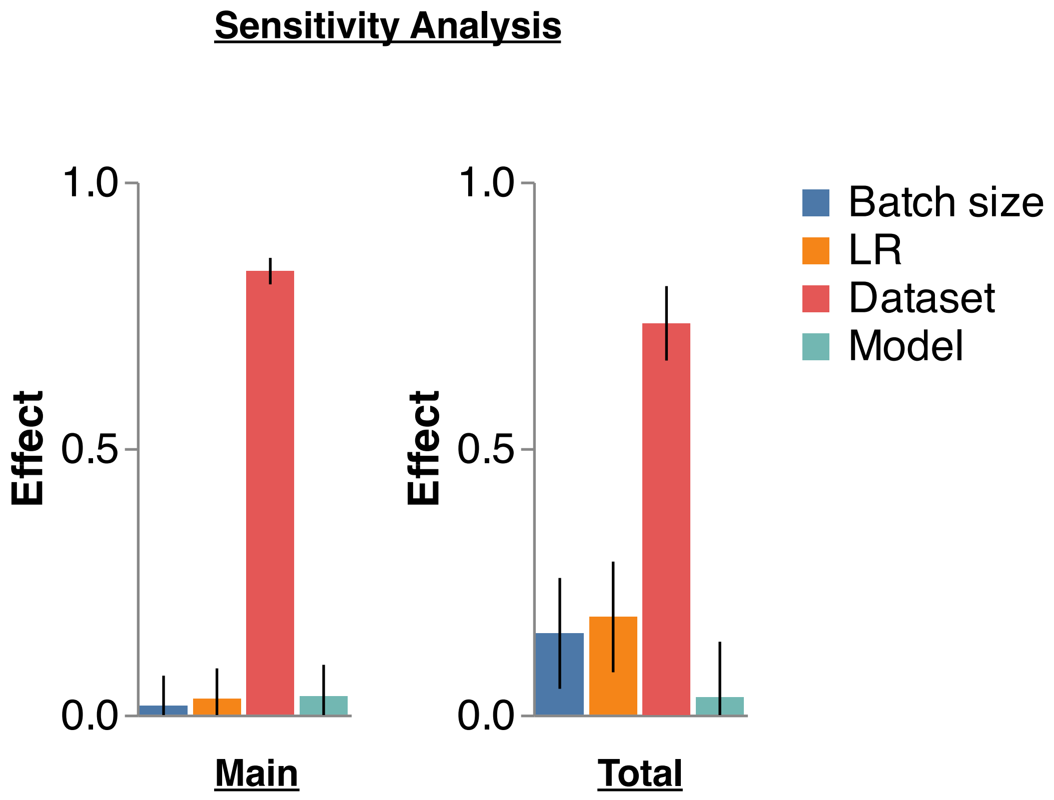
The large-batch generalization gap



The large-batch generalization gap



The large-batch generalization gap



1. Consistent across model/dataset
2. Batch size x LR interaction
3. No gap if scaled together

Outline

1. Motivation: efficient machine learning
2. Introducing the multiverse analysis
3. Modeling the machine learning multiverse
4. Case study 1: adaptive optimizers
5. Case study 2: large-batch generalization gap
6. Future stuff, discussion

Critique and open questions

- We still have to make choices about our search spaces too

Critique and open questions

- We still have to make choices about our search spaces too
- We still have to make choices about how to model the multiverse
  - GP kernel, hyperparameters, etc…

Critique and open questions

- We still have to make choices about our search spaces too
- We still have to make choices about how to model the multiverse
    - GP kernel, hyperparameters, etc…
- What about compute cost and climate impact?

Extensions and future ideas

- Making the multiverse bigger
  - more datasets, more models, termination criteria, …

Extensions and future ideas

- Making the multiverse bigger
  - more datasets, more models, termination criteria, …

- What other multiverse analyses could we run? What conclusions don’t you believe?

Extensions and future ideas

- Making the multiverse bigger
  - more datasets, more models, termination criteria, …
- What other multiverse analyses could we run? What conclusions don’t you believe?
- Multiverse analysis of the effect of “fairness” definitions

Extensions and future ideas

- Making the multiverse bigger
    - more datasets, more models, termination criteria, …
- What other multiverse analyses could we run? What conclusions don't you believe?
- Multiverse analysis of the effect of "fairness" definitions
- Accounting for heteroscedasticity and a principled tradeoff between replication and sampling a new point [1]

[1] Binois et al. (2019). Replication or exploration? Sequential design for stochastic simulation experiments. Technometrics.

Summary

- We want conclusions that are robust, general and useful
- The multiverse analysis is a framework for exploring the effect of choices on scientific conclusions

Summary

- We want conclusions that are robust, general and useful
- The multiverse analysis is a framework for exploring the effect of choices on scientific conclusions
- We make the multiverse tractable by modelling it with a GP
- And we explore it using Bayesian experimental design

Summary

- We want conclusions that are robust, general and useful
- The multiverse analysis is a framework for exploring the effect of choices on scientific conclusions
- We make the multiverse tractable by modelling it with a GP
- And we explore it using Bayesian experimental design
- Case study 1: Conclusions about best optimizer are sensitive to LR
- Case study 2: No generalization gap if batch size scaled with LR

Modeling the Machine Learning Multiverse

https://arxiv.org/abs/2206.05985

Samuel Bell

@neurosamuel

sjb326@cam.ac.uk

Onno Kampman

@KampmanOnno

Jesse Dodge

@JesseDodge

Neil Lawrence

@lawrennd