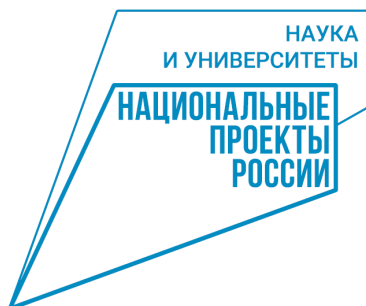


"AI Autumn School on Computational Optimization (ASCOMP 2024)"

07–09 октября 2024 г.,
г. Иннополис, Ресублика Татарстан,
Университет Иннополис и online



УНИВЕРСИТЕТ
ИННОПОЛИС



Институт
искусственного
интеллекта



SIMC

Steklov International Mathematical Center



Организации

Университет Иннополис, г. Иннополис, Республика Татарстан

Институт искусственного интеллекта,
Московский государственный университет имени М. В. Ломоносова, г. Москва

Математический институт им. В. А. Стеклова Российской академии наук, г. Москва

Математический центр мирового уровня
“Математический институт им. В. А. Стеклова Российской академии наук”
(МЦМУ МИАН), г. Москва

Конференция проводится при финансовой поддержке
Минобрнауки России (грант на создание и развитие МЦМУ МИАН,
соглашение № 075-15-2022-265).

Transformers in Computer Vision

Sergey Zagoruyko

October 7, 2024

Course contents:

1. Transformers in Computer Vision
2. Transformers in Object Detection
3. Transformers in Autonomous Driving

What is convolution?

Convolution Example

Input:

1	4	-1	0	2	-2	1	3	3	1
---	---	----	---	---	----	---	---	---	---

Filter:

1	2	0	-1
---	---	---	----

Output:

--	--	--	--	--	--	--

Source: François Fleuret's deep learning course

Convolution Example

Input:

1	4	-1	0	2	-2	1	3	3	1
---	---	----	---	---	----	---	---	---	---

Filter:

1	2	0	-1
---	---	---	----

Output:

9						
---	--	--	--	--	--	--

Source: François Fleuret's deep learning course

Convolution Example

Input:

1	4	-1	0	2	-2	1	3	3	1
---	---	----	---	---	----	---	---	---	---

Filter:

1	2	0	-1
---	---	---	----

Output:

9	0					
---	---	--	--	--	--	--

Source: François Fleuret's deep learning course

Convolution Example

Input:

1	4	-1	0	2	-2	1	3	3	1
---	---	----	---	---	----	---	---	---	---

Filter:

1	2	0	-1
---	---	---	----

Output:

9	0	1				
---	---	---	--	--	--	--

Source: François Fleuret's deep learning course

Convolution Example

Input:

1	4	-1	0	2	-2	1	3	3	1
---	---	----	---	---	----	---	---	---	---

Filter:

1	2	0	-1
---	---	---	----

Output:

9	0	1	3			
---	---	---	---	--	--	--

Source: François Fleuret's deep learning course

Convolution Example

Input:

1	4	-1	0	2	-2	1	3	3	1
---	---	----	---	---	----	---	---	---	---

Filter:

1	2	0	-1
---	---	---	----

Output:

9	0	1	3	-5		
---	---	---	---	----	--	--

Source: François Fleuret's deep learning course

Convolution Example

Input:

1	4	-1	0	2	-2	1	3	3	1
---	---	----	---	---	----	---	---	---	---

Filter:

1	2	0	-1
---	---	---	----

Output:

9	0	1	3	-5	-3	
---	---	---	---	----	----	--

Source: François Fleuret's deep learning course

Convolutions in 2D

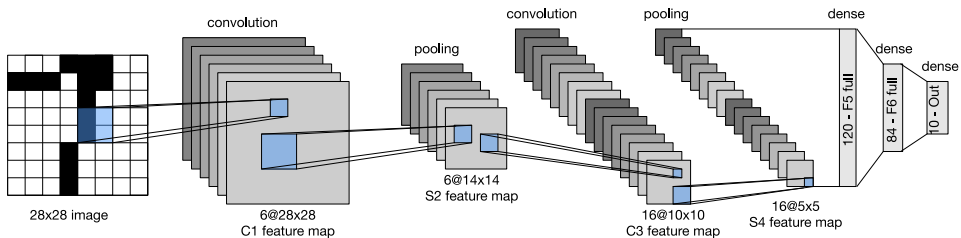
1 _{x1}	1 _{x0}	1 _{x1}	0	0
0 _{x0}	1 _{x1}	1 _{x0}	1	0
0 _{x1}	0 _{x0}	1 _{x1}	1	1
0	0	1	1	0
0	1	1	0	0

Image

4		

Convolved
Feature

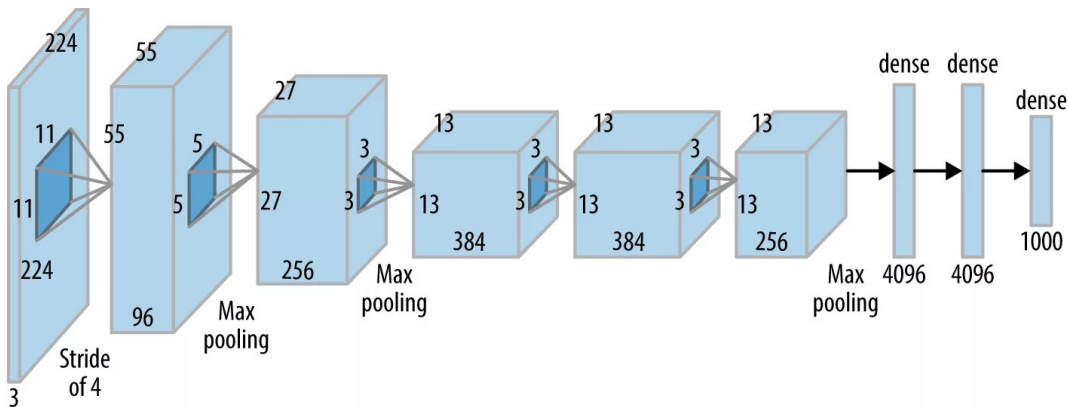
LeNet



LeNet-5 in 1998

<https://en.wikipedia.org/wiki/LeNet>

AlexNet



AlexNet in 2012

<https://machinelearningmastery.ru/the-w3h-of-alexnet-vggnet-resnet-and-inception-7baaaecccc96/>

Inductive Bias in Neural Networks

- ▶ **Definition:**

- ▶ Inductive bias refers to the set of assumptions a learning algorithm makes to generalize from the training data to unseen data.

Inductive Bias in Neural Networks

- ▶ **Definition:**

- ▶ Inductive bias refers to the set of assumptions a learning algorithm makes to generalize from the training data to unseen data.

- ▶ **Importance:**

- ▶ Helps in guiding the learning process and making predictions.
 - ▶ Determines the types of patterns a model can learn.

Inductive Bias in Neural Networks

- ▶ **Definition:**

- ▶ Inductive bias refers to the set of assumptions a learning algorithm makes to generalize from the training data to unseen data.

- ▶ **Importance:**

- ▶ Helps in guiding the learning process and making predictions.
 - ▶ Determines the types of patterns a model can learn.

- ▶ **Examples in Neural Networks:**

- ▶ *Convolutional Neural Networks (CNNs)*: Assumption of spatial hierarchies in images.
 - ▶ *Recurrent Neural Networks (RNNs)*: Assumption of sequential dependencies in time-series data.

Inductive Bias in Neural Networks

► Definition:

- Inductive bias refers to the set of assumptions a learning algorithm makes to generalize from the training data to unseen data.

► Importance:

- Helps in guiding the learning process and making predictions.
- Determines the types of patterns a model can learn.

► Examples in Neural Networks:

- *Convolutional Neural Networks (CNNs)*: Assumption of spatial hierarchies in images.
- *Recurrent Neural Networks (RNNs)*: Assumption of sequential dependencies in time-series data.

► Trade-offs:

- Stronger biases can lead to faster learning but may reduce flexibility.
- Weaker biases increase flexibility but may require more data.

Two pillars of deep learning:

Two pillars of deep learning:

- ▶ Large amounts of rich diverse data

Two pillars of deep learning:

- ▶ Large amounts of rich diverse data
- ▶ Large amounts of compute

AlexNet

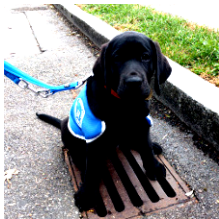
Visualizing parameters



Francois Fleuret's Deep Learning Course

AlexNet

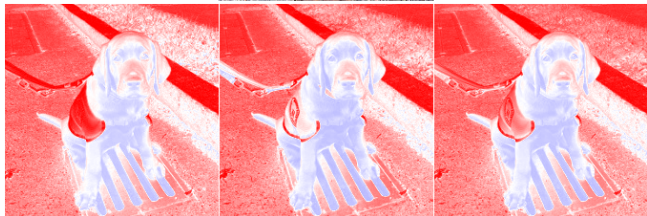
Visualizing activations



Francois Fleuret's Deep Learning Course

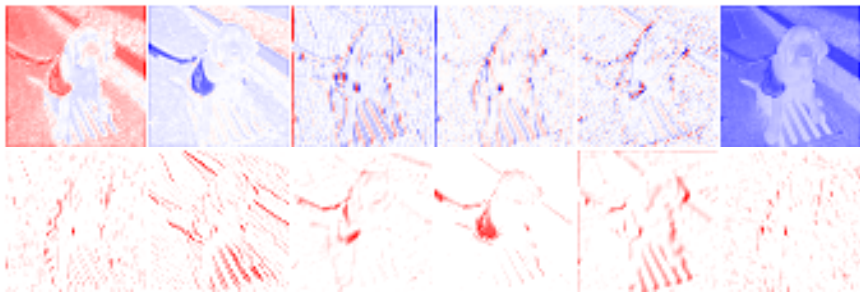
AlexNet

Visualizing activations



AlexNet

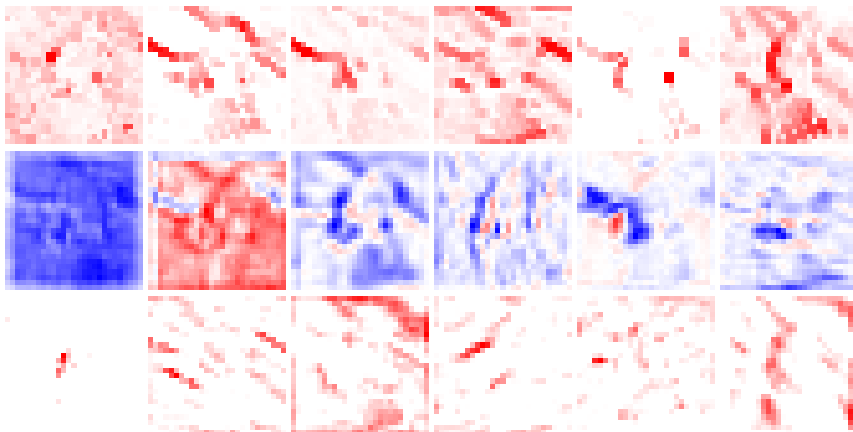
Visualizing activations



Francois Fleuret's Deep Learning Course

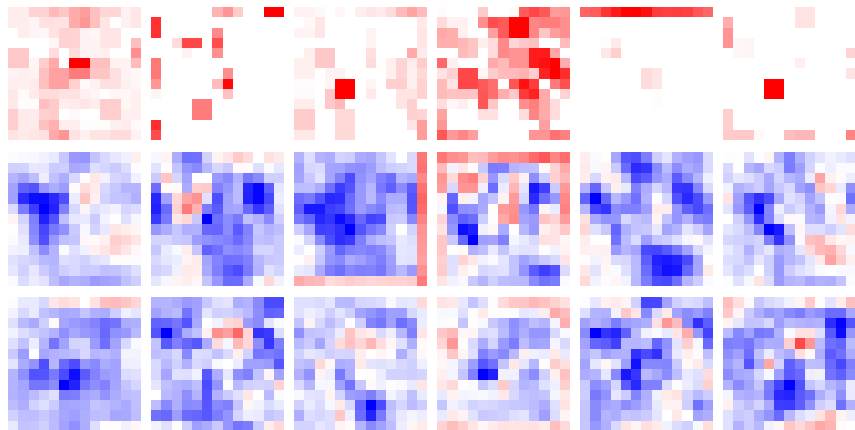
AlexNet

Visualizing activations



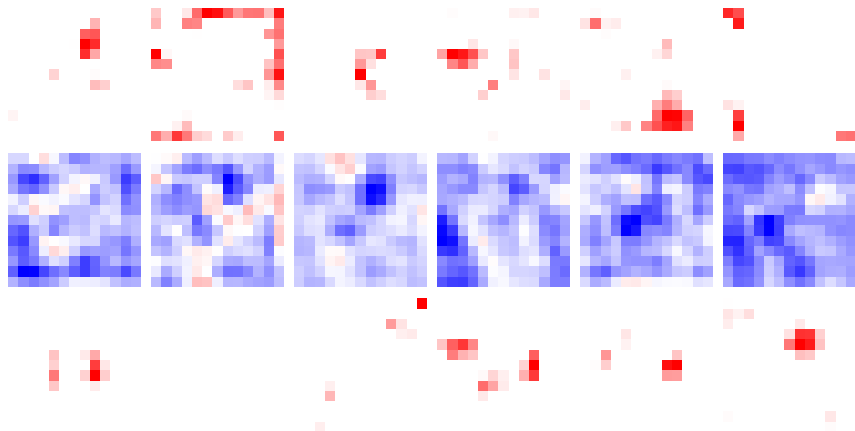
AlexNet

Visualizing activations

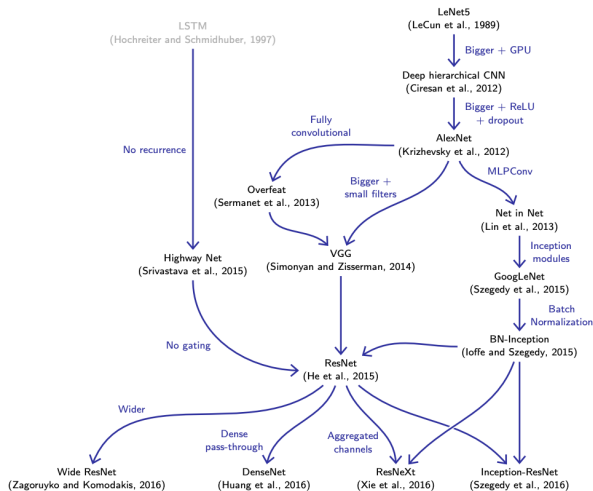


AlexNet

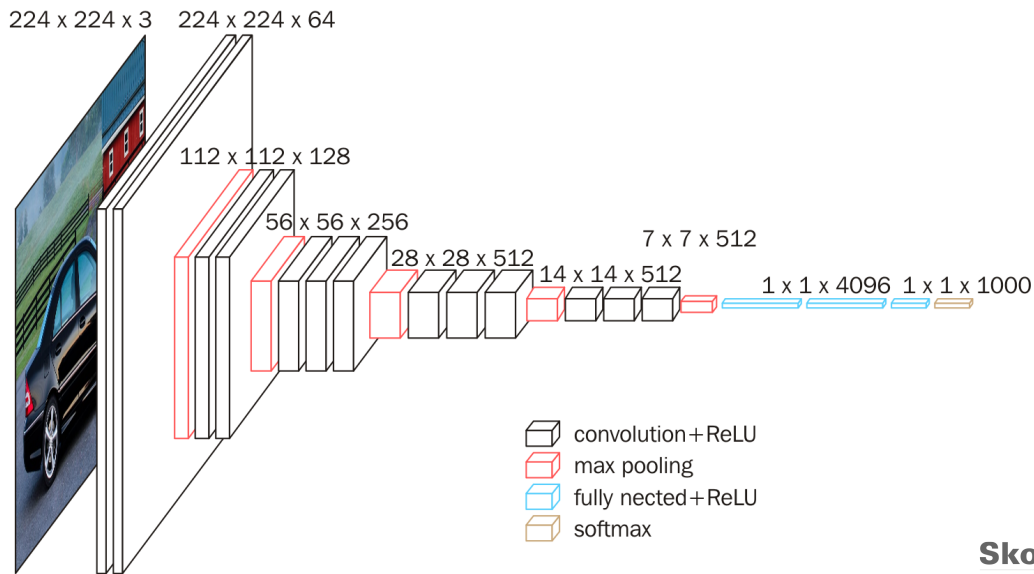
Visualizing activations



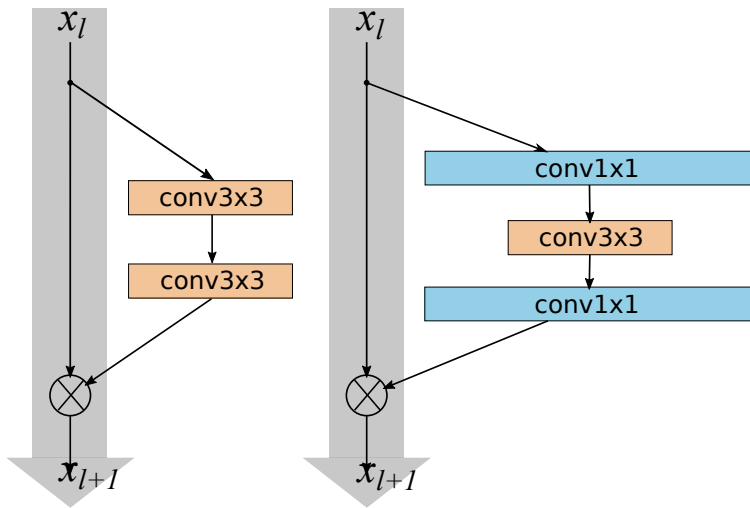
AlexNet



VGG



ResNet



(Zagoruyko and Komodakis 2016)

What is a Transformer?

Self-attention

Core operation in the Transformer:

$$Q = W_q X$$

$$K = W_k X$$

$$V = W_v X$$

$$Z = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V$$

Self-attention

Core operation in the Transformer:

$$Q = W_q X$$

$$K = W_k X$$

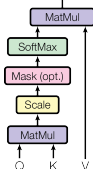
$$V = W_v X$$

$$Z = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V$$

- ▶ Quadratic cost
- ▶ Input order equivariant

Multi-head attention

Scaled Dot-Product Attention



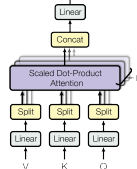
$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}}) V$$

$$MultiHead(Q, K, V) = Concat(H_1, ..., H_h) W^O$$

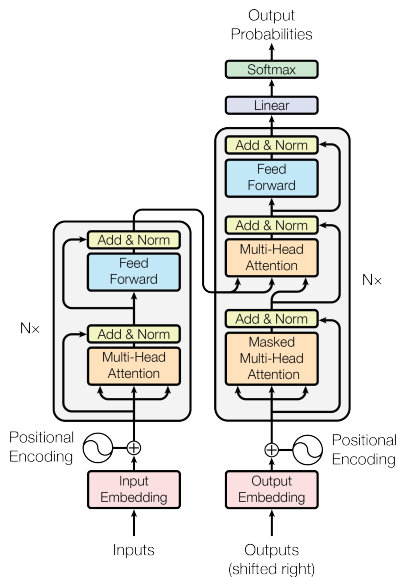
$$H_i = Attention(QW_i^Q, KW_i^K, VW_i^V), i = 1, ..., h$$

$$W_i^Q \in \mathbb{R}^{d_{model} \times d_k}, W_i^K \in \mathbb{R}^{d_{model} \times d_k}, W_i^V \in \mathbb{R}^{d_{model} \times d_v}, W_i^O \in \mathbb{R}^{hd_v \times d_{model}}$$

Multi-head Attention

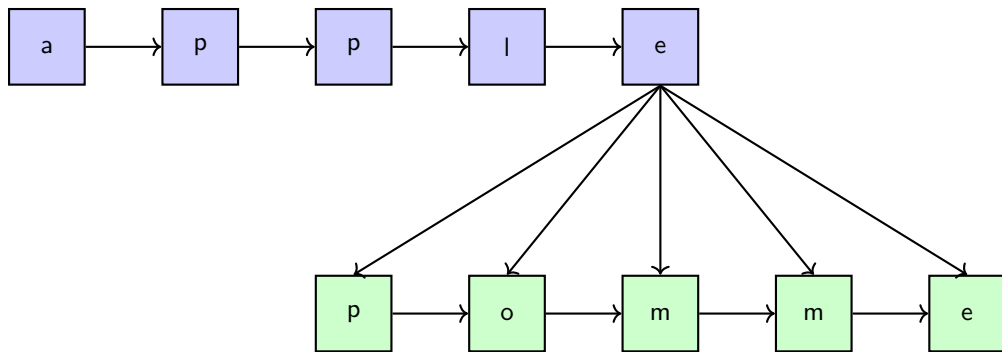


Original Transformer (Vaswani et al. 2017)



Seq2Seq Translation: Apple to Pomme

Encoder



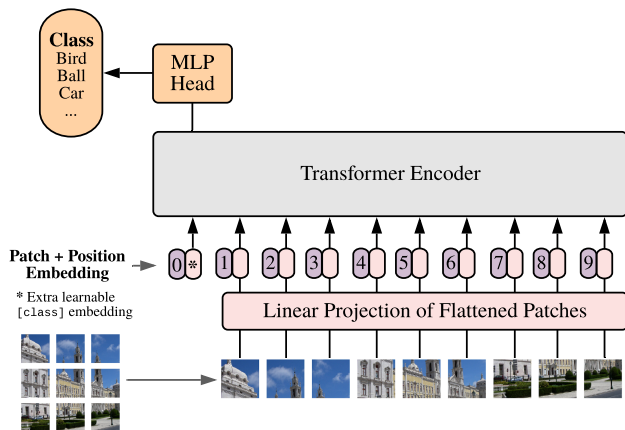
ViT: Vision Transformers

Vision Transformer, ViT (Dosovitskiy et al. 2020) -

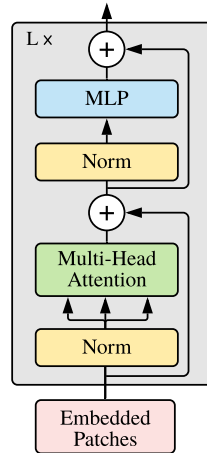
Inspired by the Transformer scaling successes in NLP, we experiment with applying a standard Transformer directly to images, with the fewest possible modifications. To do so, we split an image into patches and provide the sequence of linear embeddings of these patches as an input to a Transformer. Image patches are treated the same way as tokens (words) in an NLP application. We train the model on image classification in supervised fashion. (Dosovitskiy et al. 2020)

ViT: Vision Transformers

Vision Transformer (ViT)



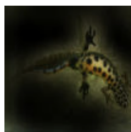
Transformer Encoder



(Dosovitskiy et al. 2020)

ViT attention

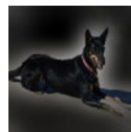
1



2



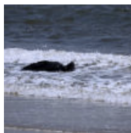
3



9



10

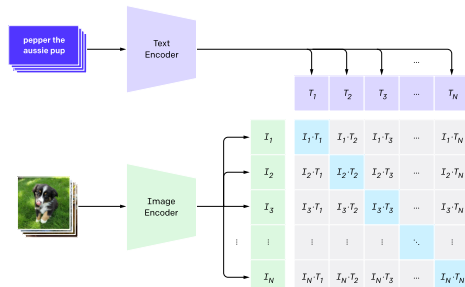


11



What is CLIP?

1. Contrastive pre-training

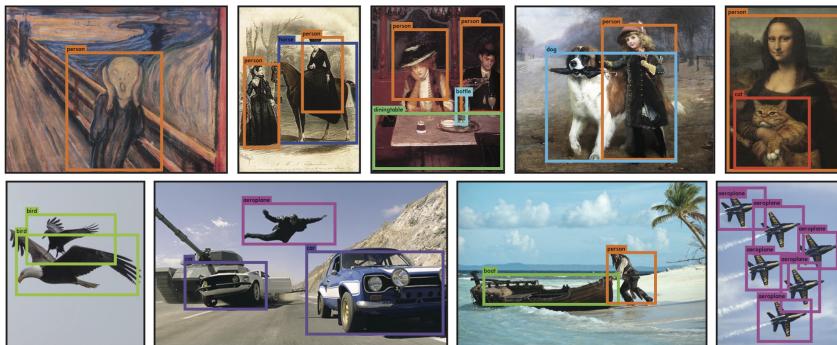


- ▶ **CLIP** stands for **C**ontrastive **L**anguage–**I**mage **P**retraining.
- ▶ Developed by OpenAI, it combines text and image understanding.
- ▶ Utilizes a large dataset of text-image pairs for training.

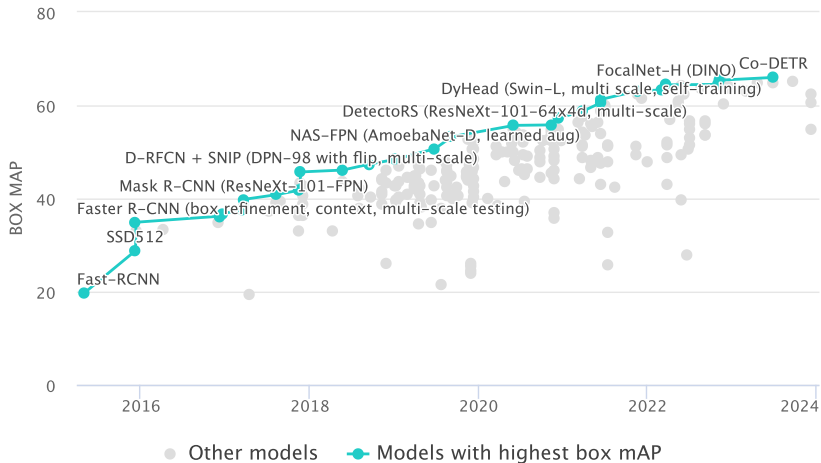
<https://openai.com/index/clip>

Object Detection

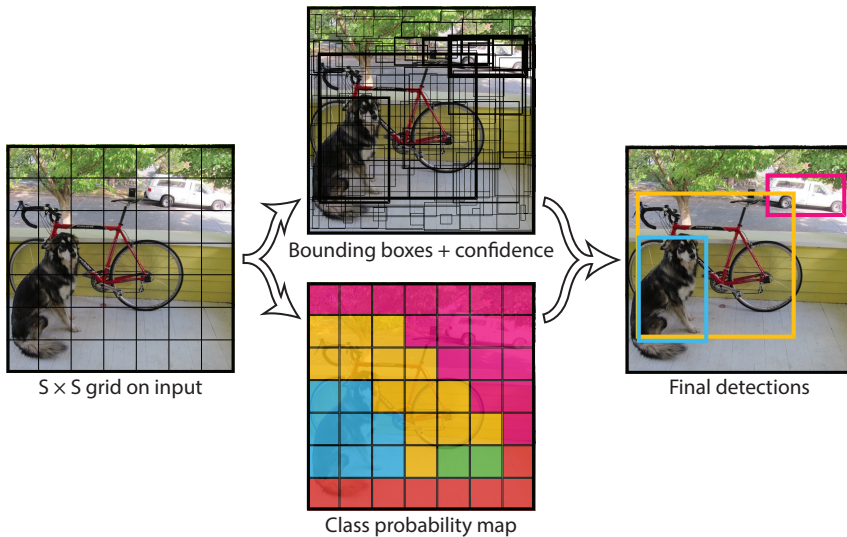
Object Detection



Object Detection



YOLO



Classical approach to detection

- Popular approach: detection $:=$ classification of boxes

Classical approach to detection

- ▶ Popular approach: detection $:=$ classification of boxes
- ▶ Requires selecting a subset of candidate boxes

Classical approach to detection

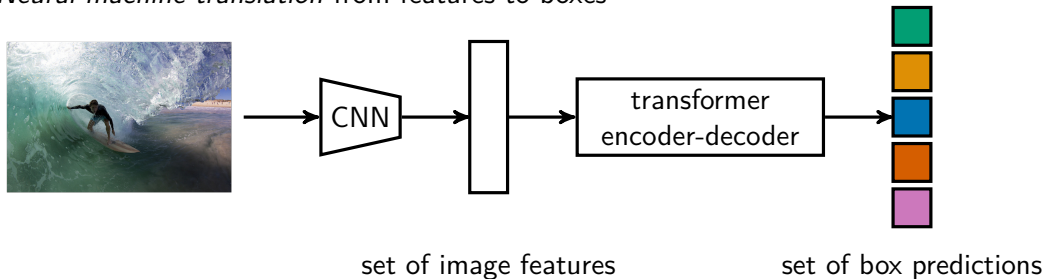
- ▶ Popular approach: detection $:=$ classification of boxes
- ▶ Requires selecting a subset of candidate boxes
- ▶ Regression step to refine the predictions

Classical approach to detection

- ▶ Popular approach: detection $:=$ classification of boxes
- ▶ Requires selecting a subset of candidate boxes
- ▶ Regression step to refine the predictions
- ▶ Typically non-differentiable

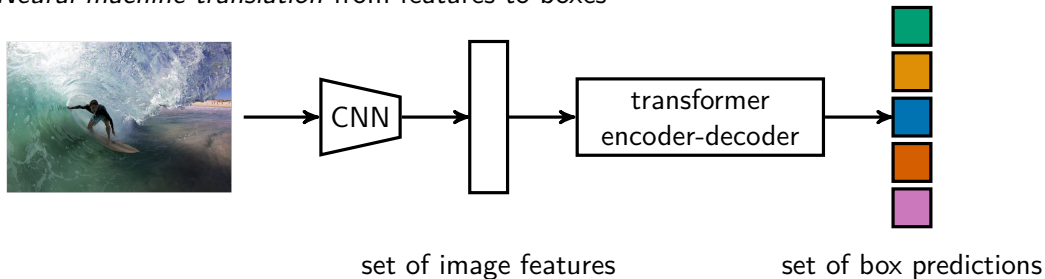
DETR: Rethinking object detection

- *Neural machine translation* from features to boxes



DETR: Rethinking object detection

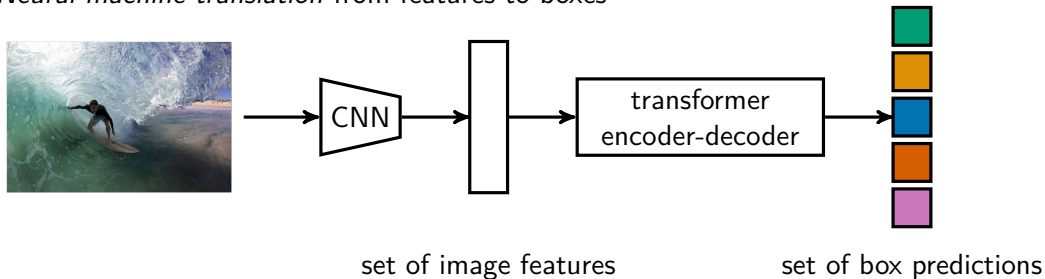
- *Neural machine translation* from features to boxes



- End-to-end parallel set prediction

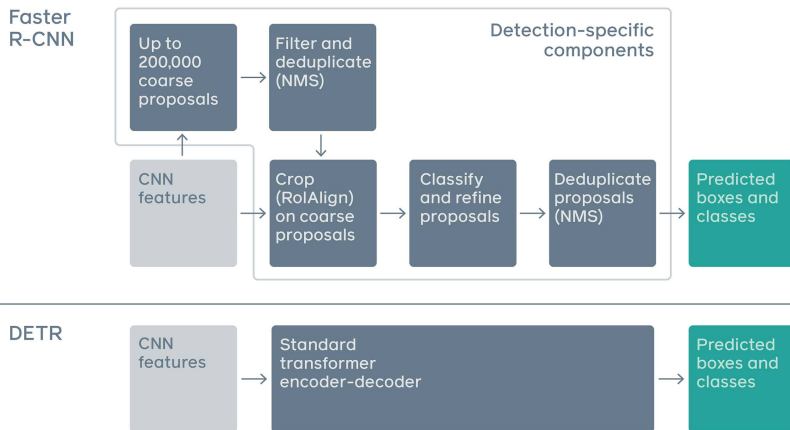
DETR: Rethinking object detection

- *Neural machine translation* from features to boxes

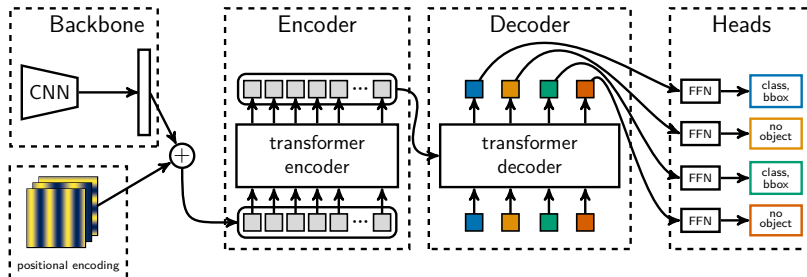


- End-to-end parallel set prediction
- Global scene reasoning

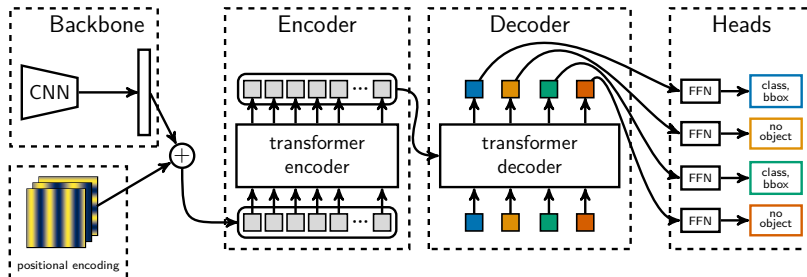
Streamlined detection pipeline



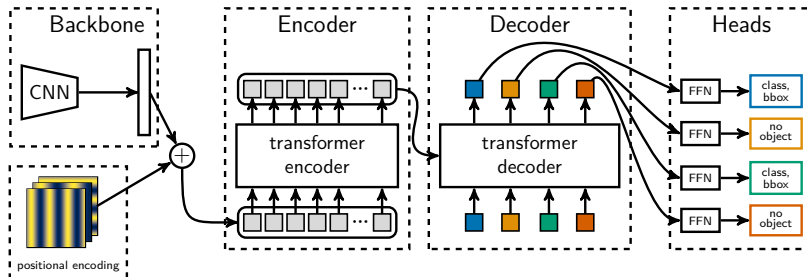
- We use standard ResNet from torchvision



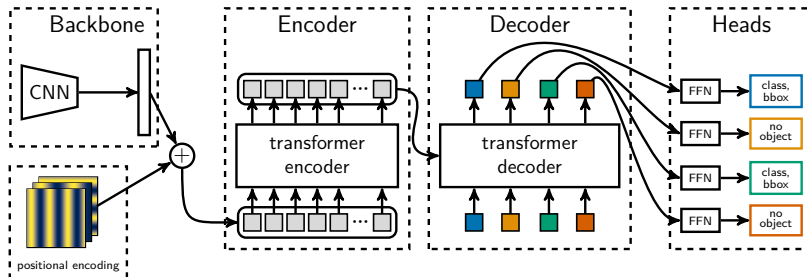
- We use standard ResNet from torchvision
- Pretraining on Imagenet is key (labels or SSL)



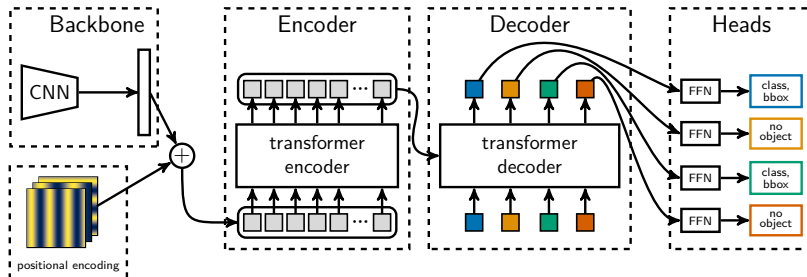
- We use 2D sine/cosine embeddings



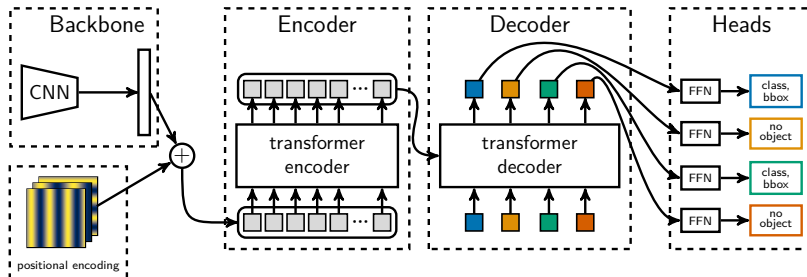
- We use 2D sine/cosine embeddings
- Embeddings are added to features



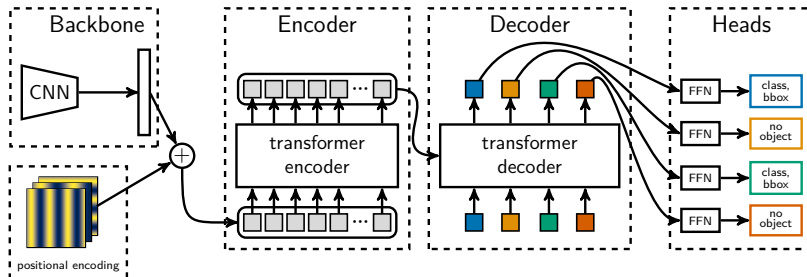
- We use 6 layers of transformer encoder



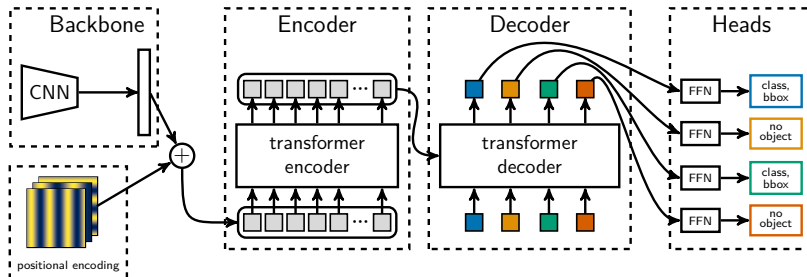
- We use 6 layers of transformer encoder
- Global reasoning through attention



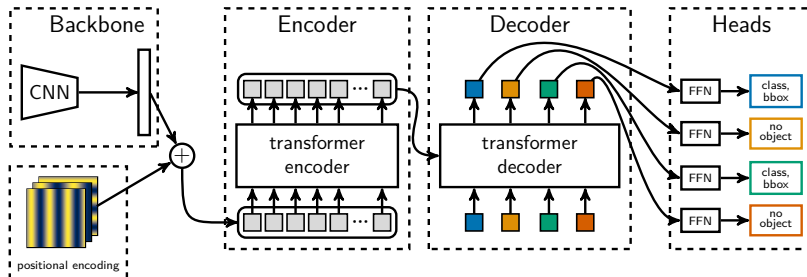
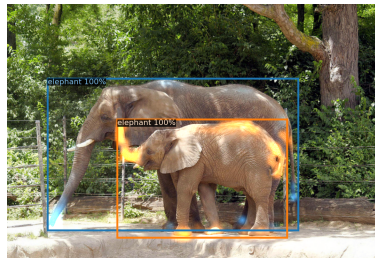
- We use 6 layers of transformer encoder
- Global reasoning through attention
- Starts separating instances



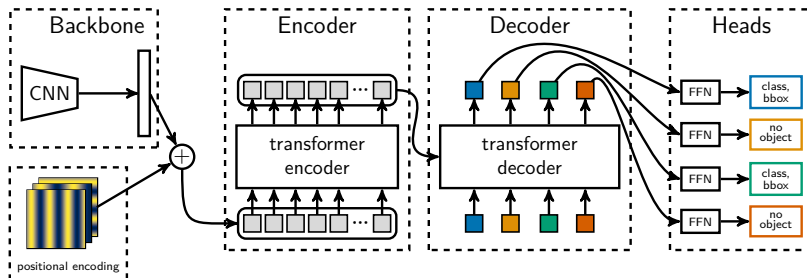
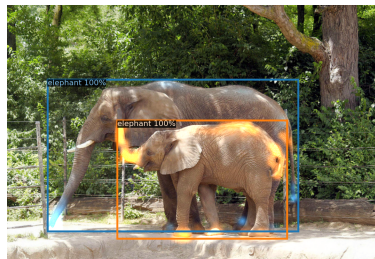
- We use 6 layers of transformer decoder;



- We use 6 layers of transformer decoder;
- Attention focuses on extremities



- We use 6 layers of transformer decoder;
- Attention focuses on extremities
- Predictions are refined at each layer **in parallel**



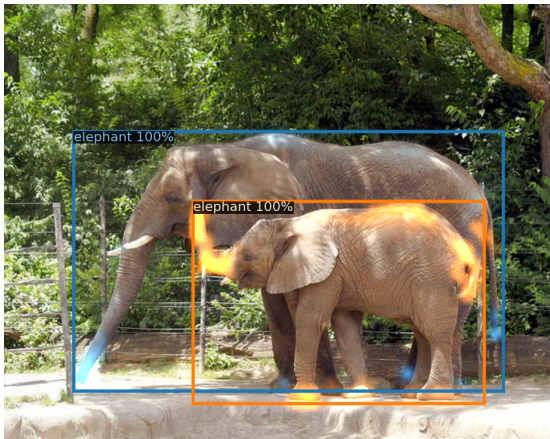
Decoder attention weights

6 decoder layers of:

- ▶ self-attention
- ▶ enc-dec attention
- ▶ FFN
- ▶ LayerNorm

All outputs are decoded *in parallel*

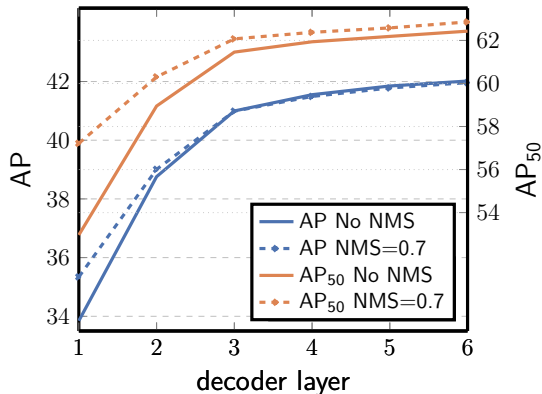
Attention focuses on extremities



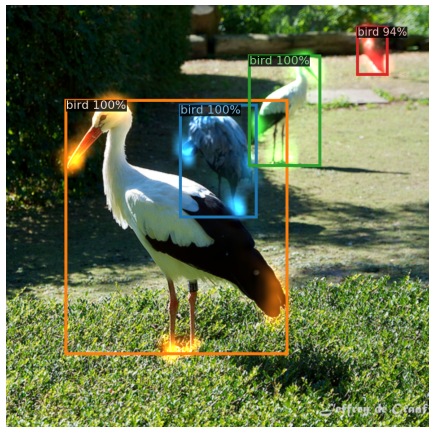
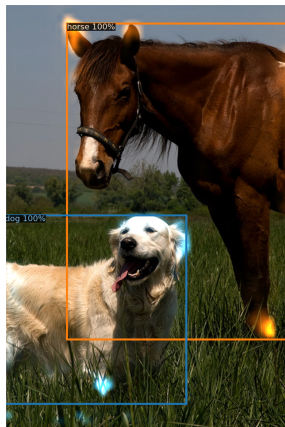
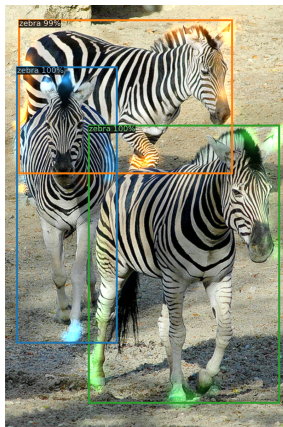
Decoder: no NMS needed

- ▶ AP/AP₅₀ goes up in lower layers (no communication)
- ▶ AP goes down in the last layers
- ▶ AP₅₀ goes up slightly

There is no need for NMS in DETR

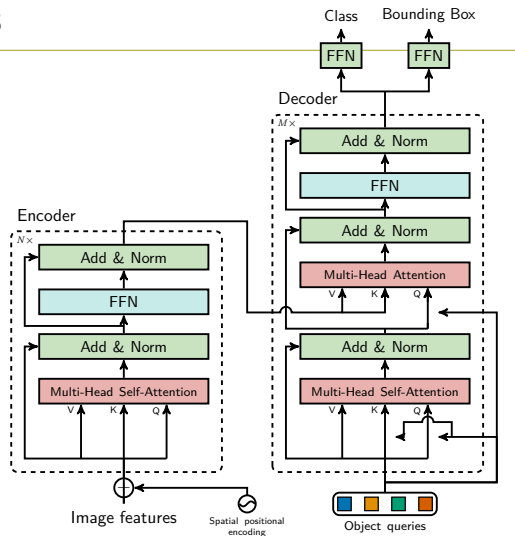


Decoder attention weights



Transformer architecture variations

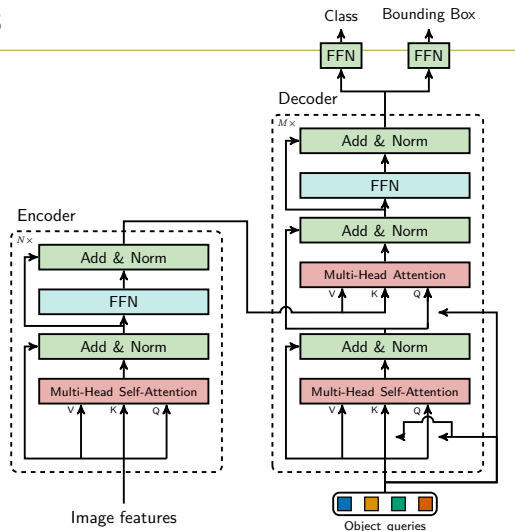
- Base transformer
39.2 AP^a



^aAshish Vaswani et al. (2017). "Attention is All you Need". In: *NeurIPS*.

Transformer architecture variations

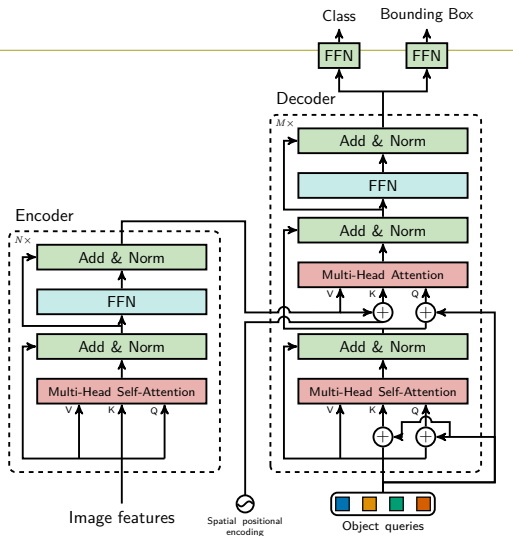
- Base transformer
39.2 AP^a
- No input positional encoding
32.8 AP



^aAshish Vaswani et al. (2017). "Attention is All you Need". In: *NeurIPS*.

Transformer architecture variations

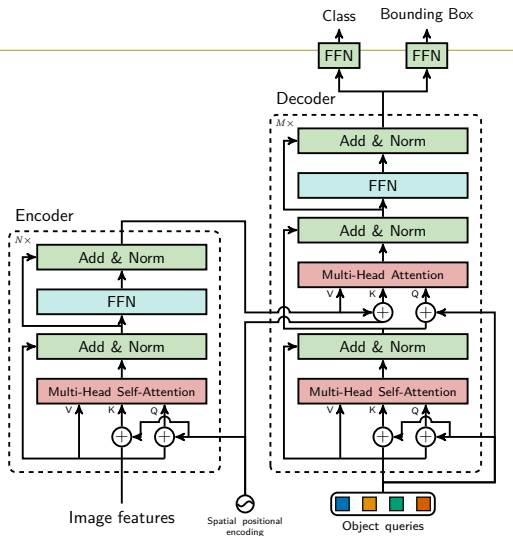
- Base transformer
39.2 AP^a
- No input positional encoding
32.8 AP
- Encodings in decoder attentions only
39.3 AP



^aAshish Vaswani et al. (2017). "Attention is All you Need". In: *NeurIPS*.

Transformer architecture variations

- Base transformer
39.2 AP^a
- No input positional encoding
32.8 AP
- Encodings in decoder attentions only
39.3 AP
- Encodings in all attentions
40.6 AP

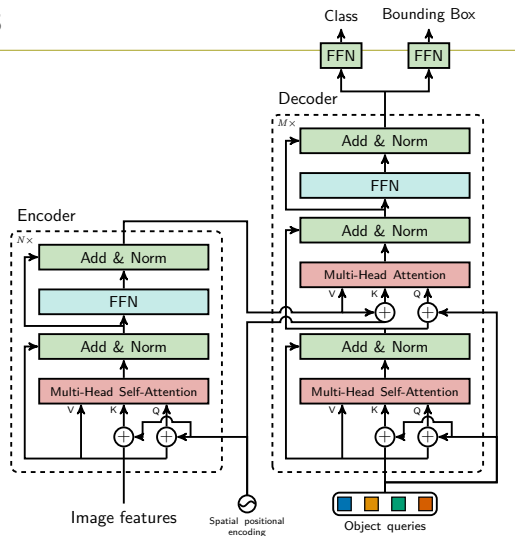


^aAshish Vaswani et al. (2017). "Attention is All you Need". In: *NeurIPS*.

Transformer architecture variations

- Base transformer
39.2 AP^a
- No input positional encoding
32.8 AP
- Encodings in decoder attentions only
39.3 AP
- Encodings in all attentions
40.6 AP

All transformer parts are contributing!



^aAshish Vaswani et al. (2017). "Attention is All you Need". In: *NeurIPS*.

Transformers in Autonomous driving

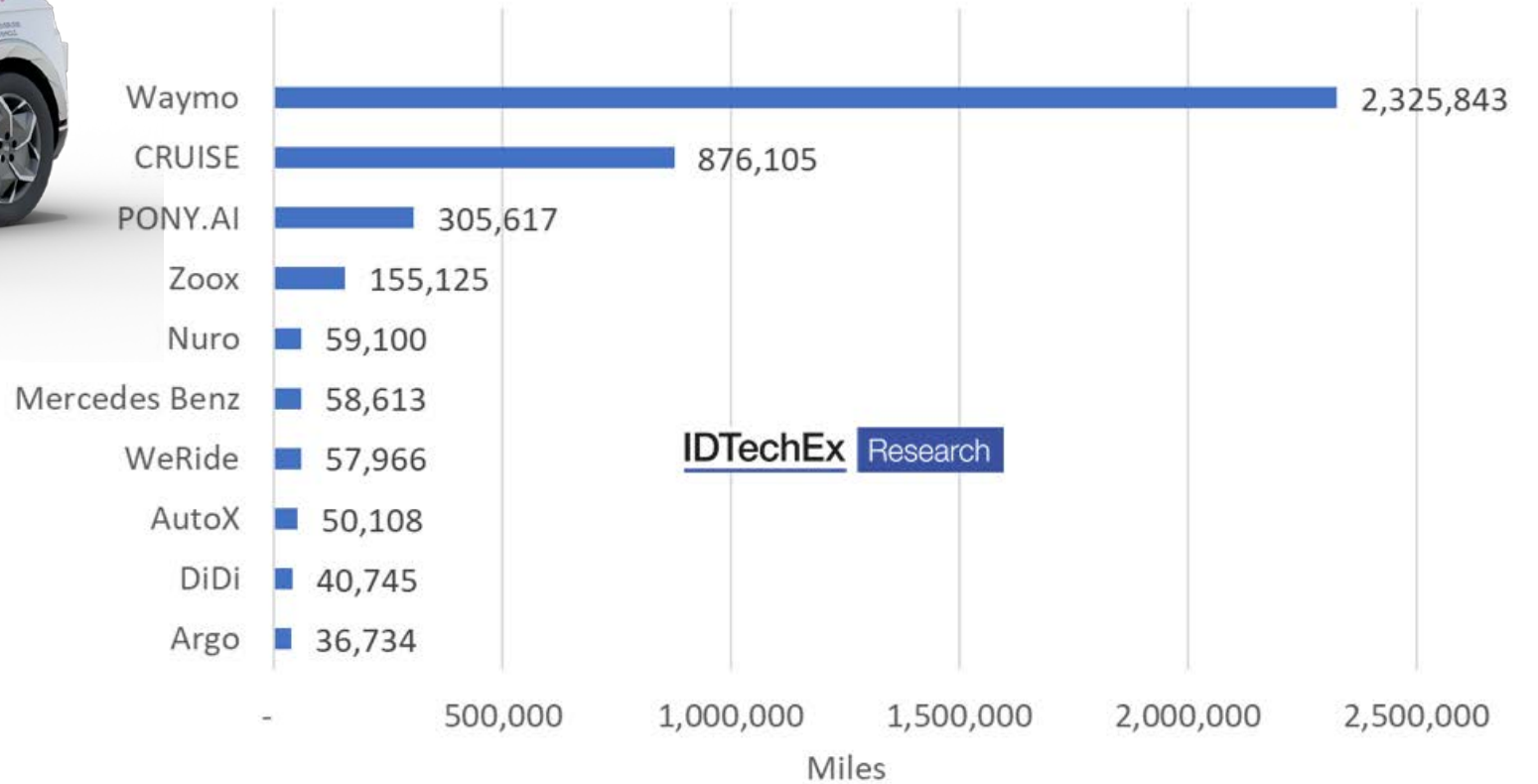


Sergey Zagoruyko

AV companies: the US



Top Players, No. Testing Miles, California DMV, 2021



AV companies: Russia

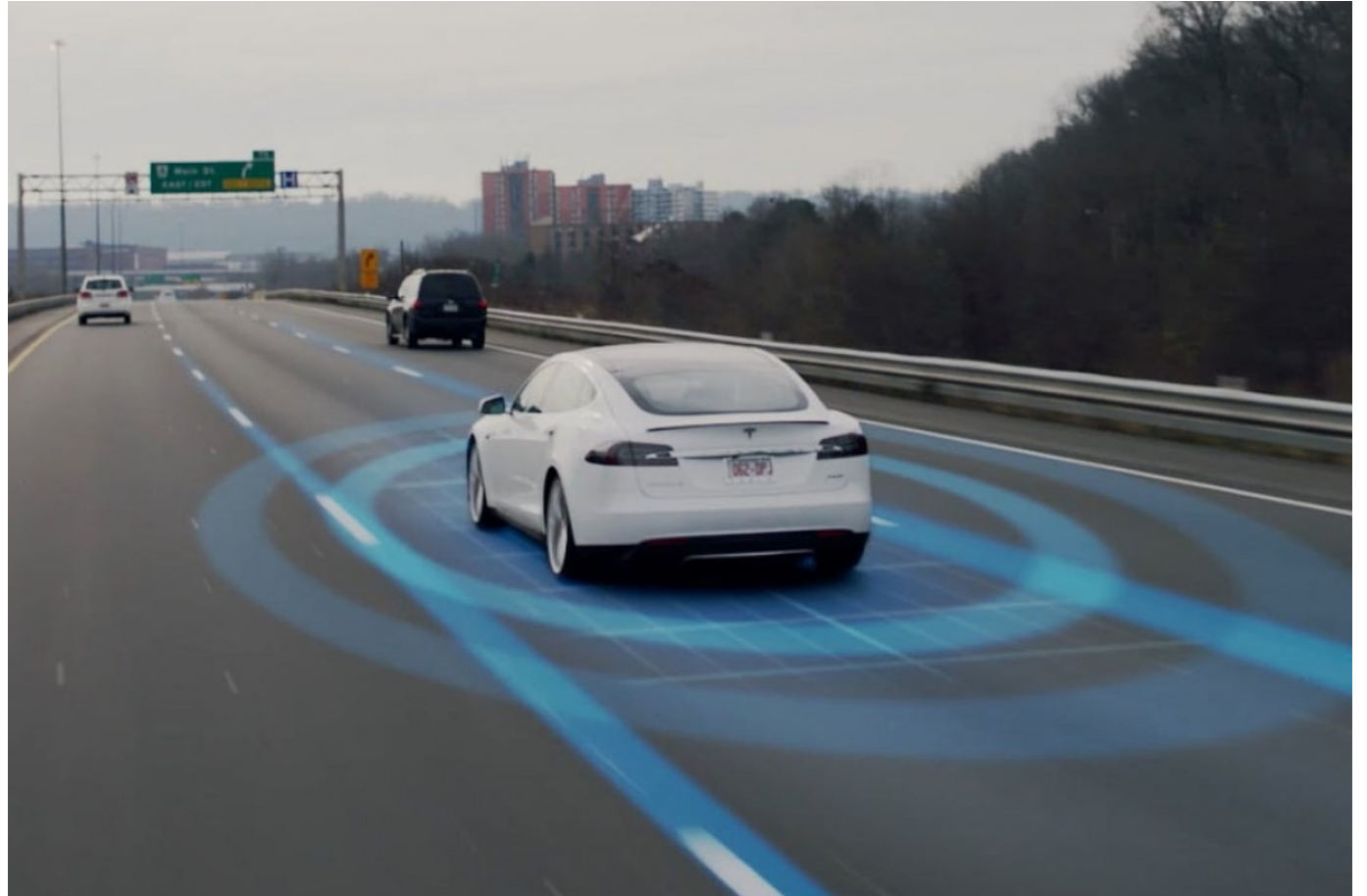
- [Yandex's autonomous cars have driven over six million miles in 'challenging conditions'](#)
- <https://www.youtube.com/watch?v=q-ekm7uqeq8>



- When did first autonomous vehicle was developed?

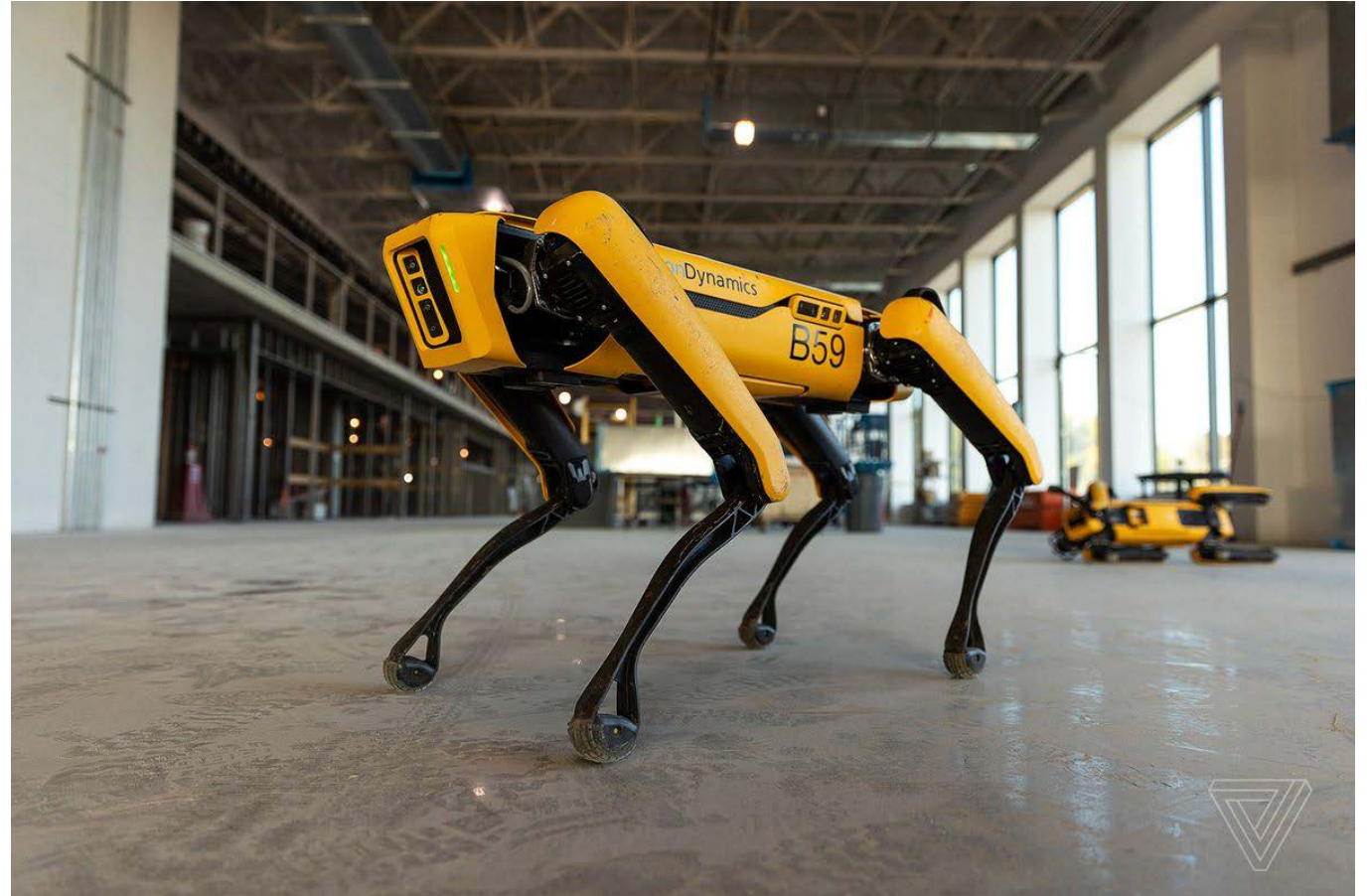
<https://youtu.be/YZ6nPhUG2i0>

In 1980s Ernst Dickmanns developed a system which was able to drive more than 1000 kms autonomously!



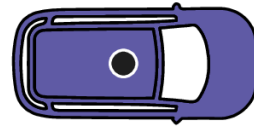
- What kind of machine learning is used in robots?

None.



Autonomous Driving stack

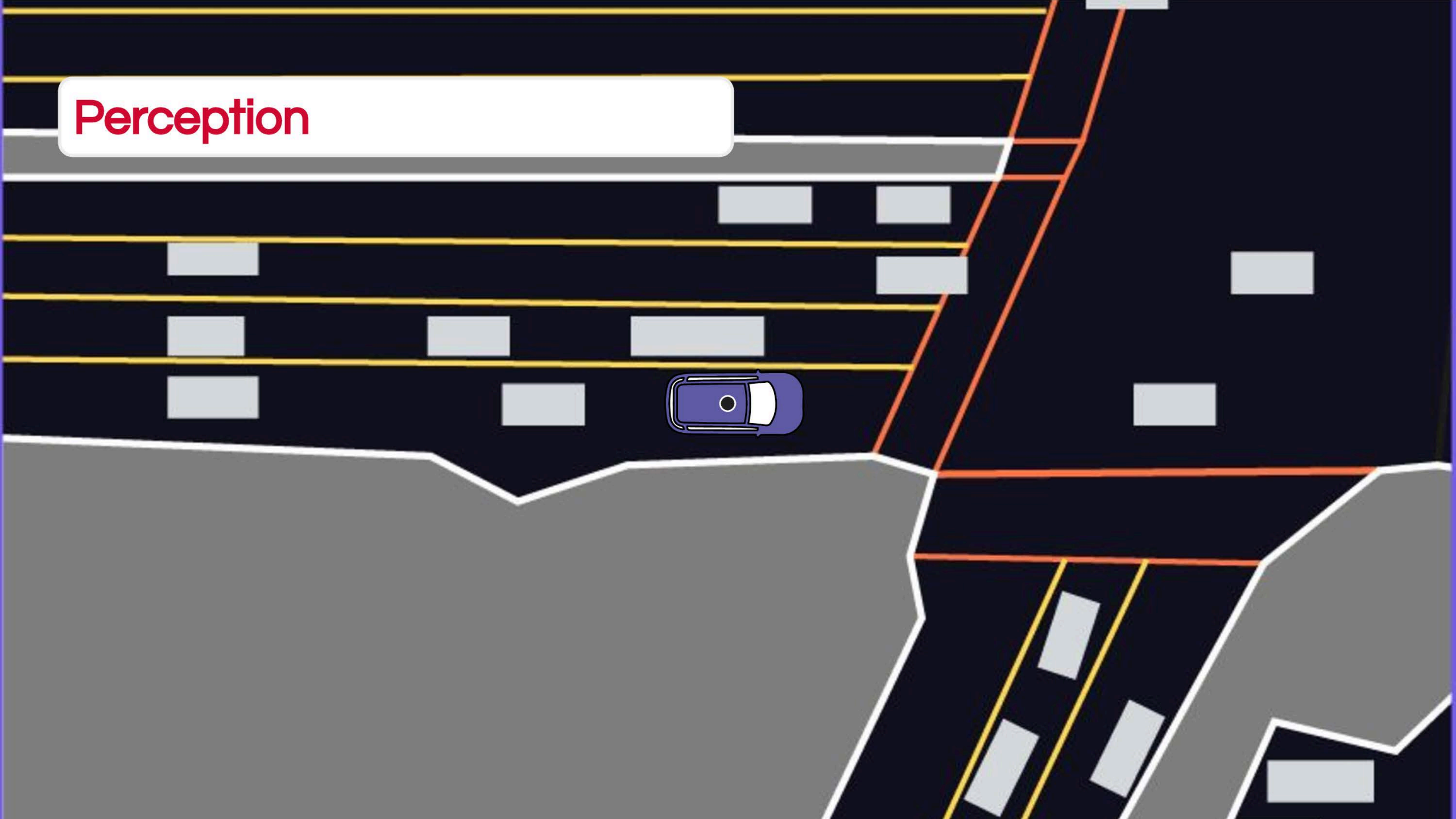
Traditional stack



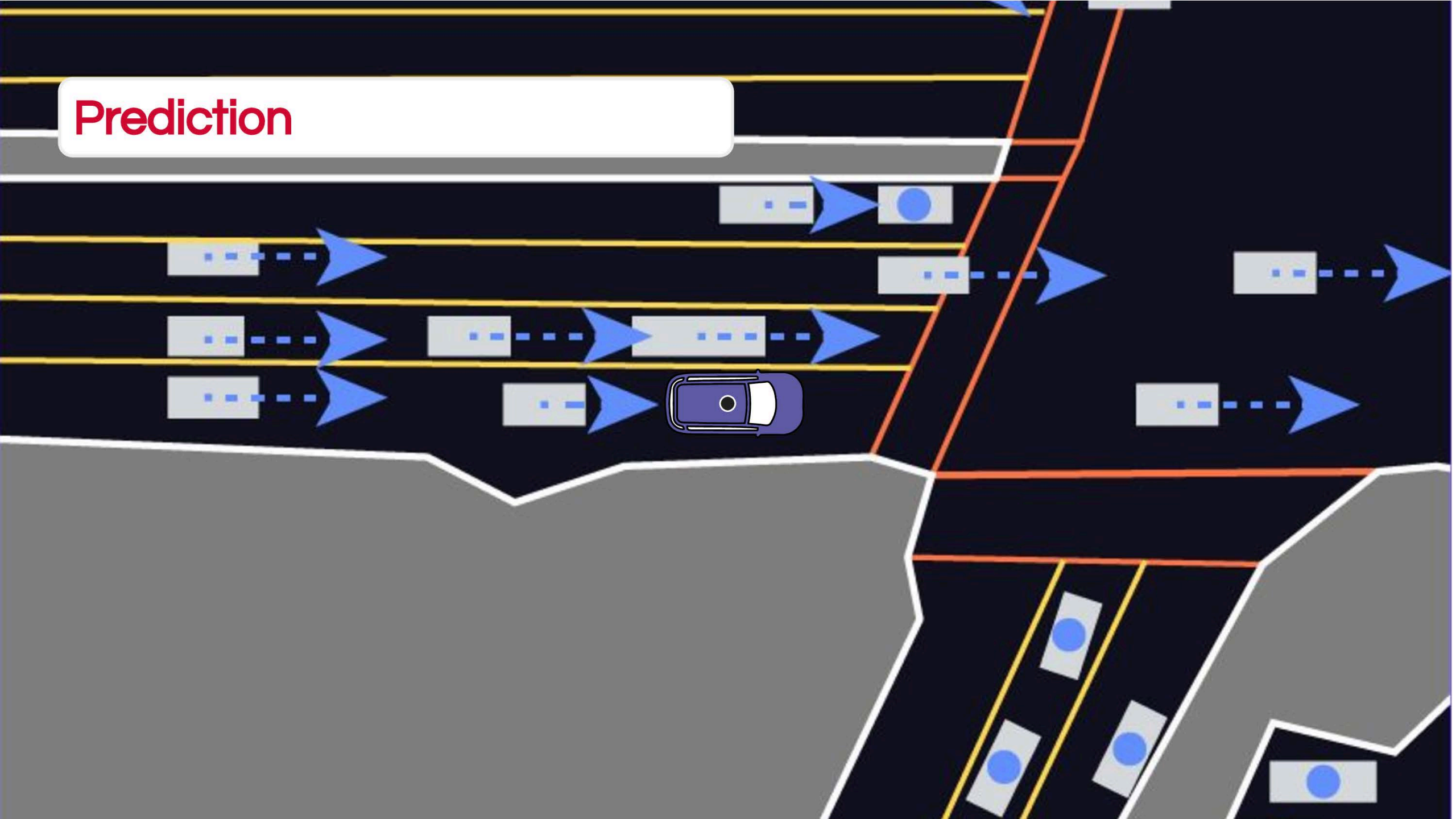
Sensors



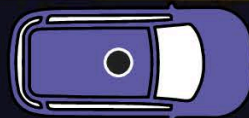
Perception



Prediction



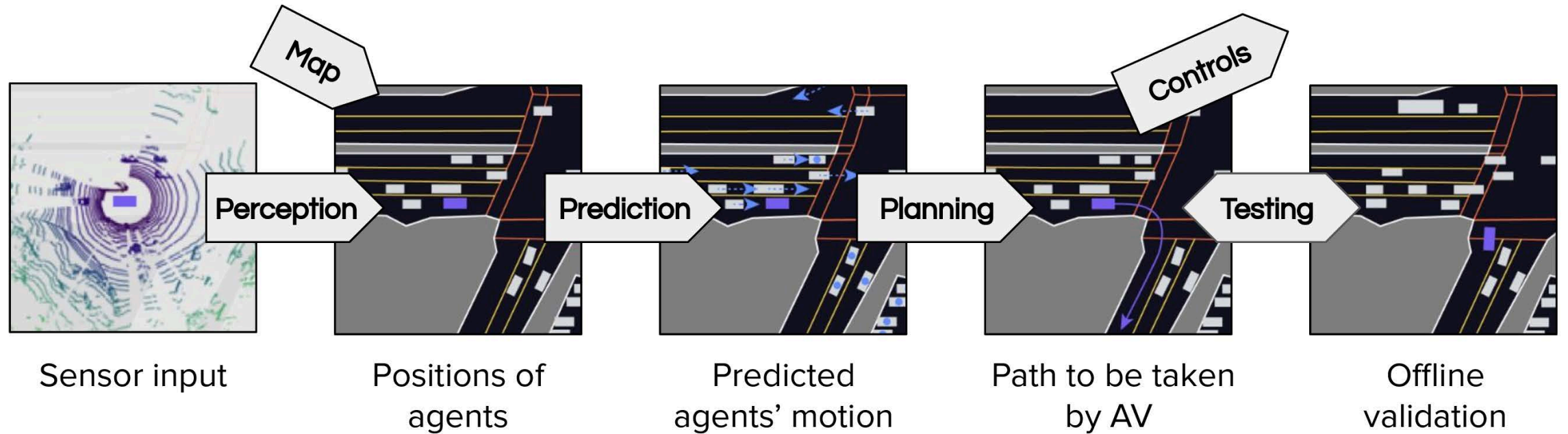
Planning



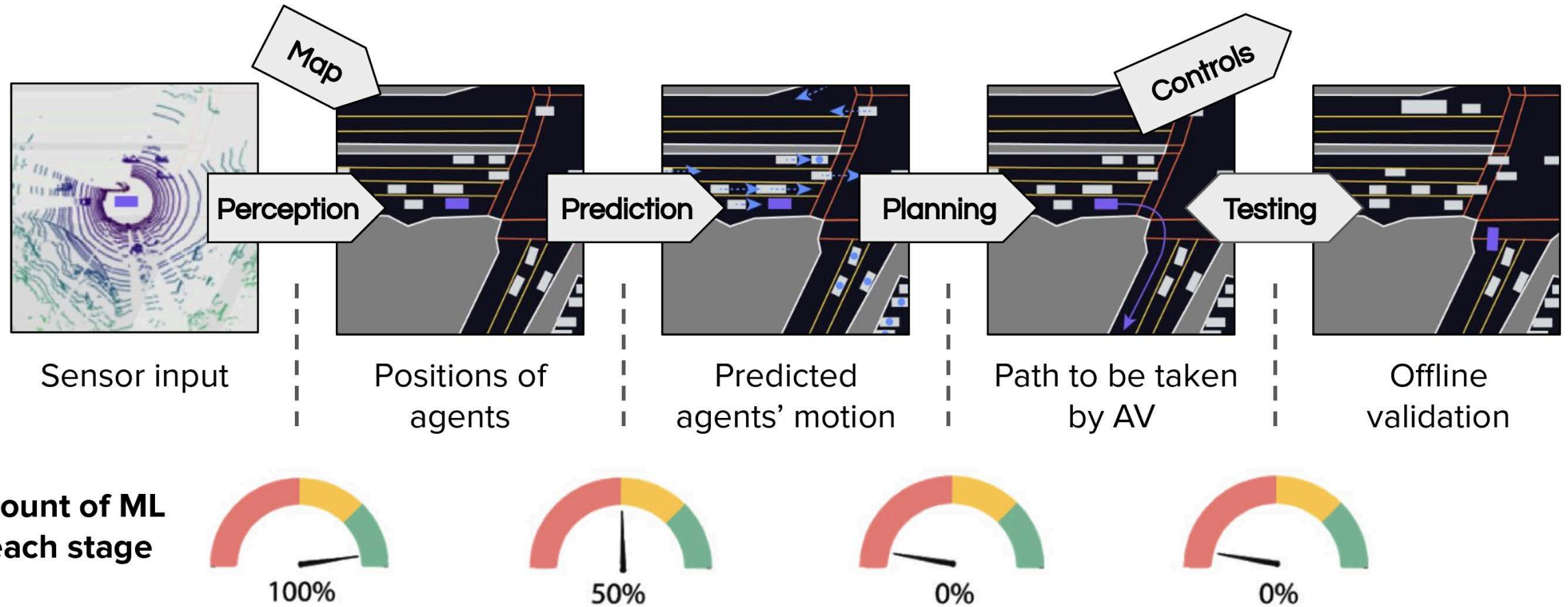
Control



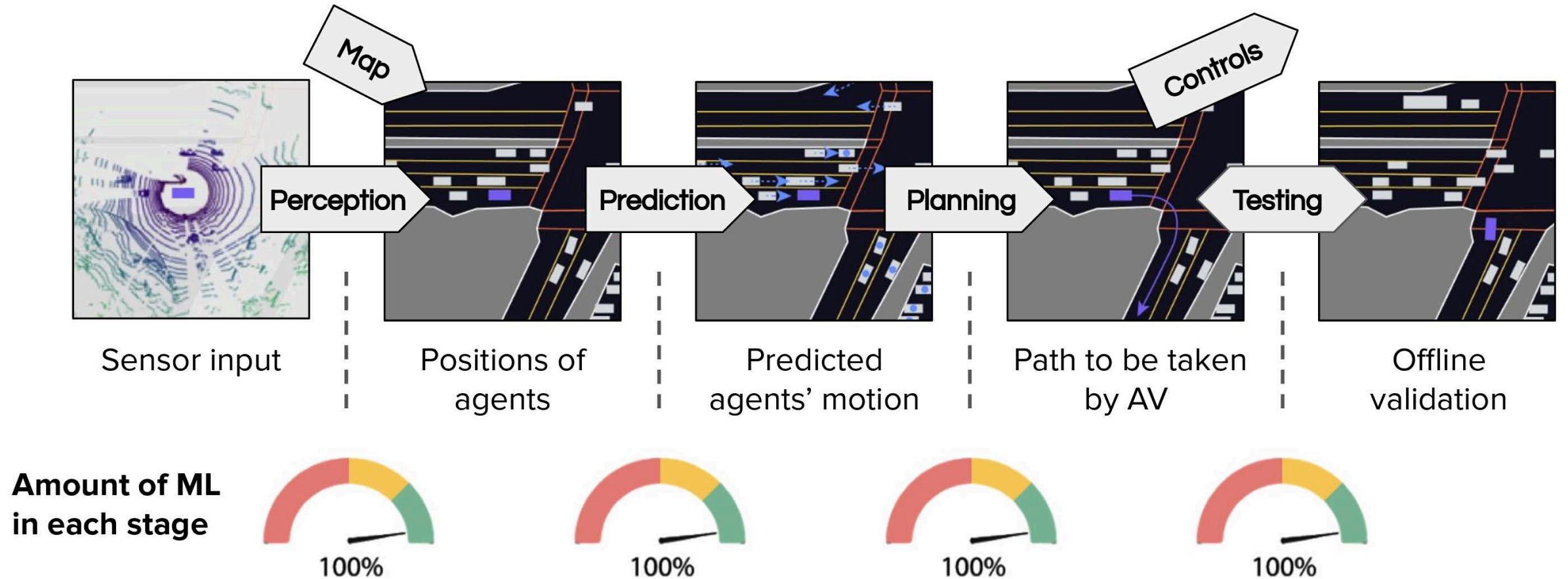
Traditional stack



Traditional stack



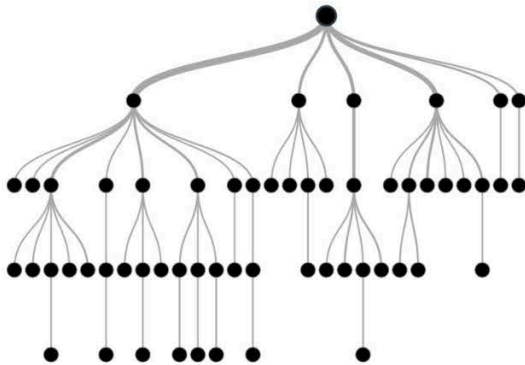
Towards data-driven decisions



Rule-based vs Data-Driven Systems

Rule-Based systems

- Hand-crafted by experts
- Difficult to scale
- More interpretable

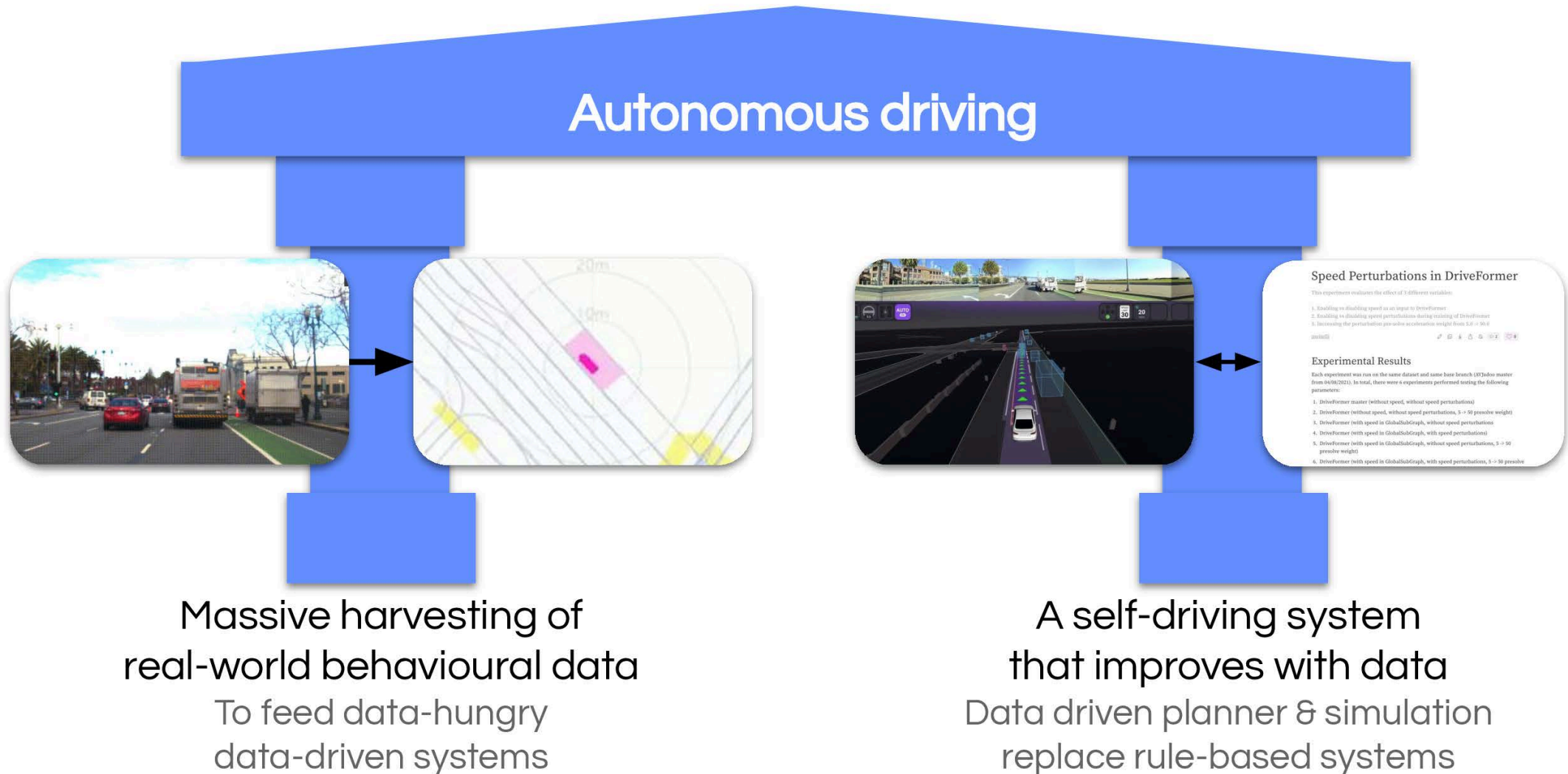


Data-driven systems

- Scales with data
- Easy to adapt
- Less interpretable



Two Pillars to Data Driven



Imitation learning

Decision Making

Sequential decision making is often formalized under the **MDP (Markov Decision Process)** framework.

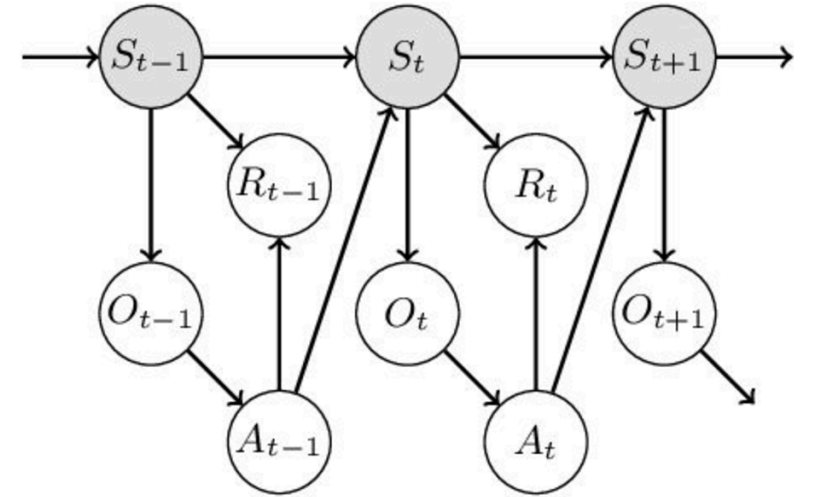
MDP is formalized as a 4-tuple $\langle \mathbf{S}, \mathbf{A}, \mathbf{P}, \mathbf{C} \rangle$, where:

S = State space $\bullet \longrightarrow s_t \in \mathcal{S}$

A = Action space $\bullet \longrightarrow a_t \in \mathcal{A}$

P = Transition dynamics $\bullet \longrightarrow P(S_{t+1} | s_t, a_t)$

C = Cost $\bullet \longrightarrow c(s_t, a_t)$



Source: Learning Partially Observable Markov Decision Processes Using Coupled Canonical Polyadic Decomposition. Kejun Huang et al. June 2019.

Decision Making

The goal is to find a policy π that will map states to actions:

$$\pi : s_t \mapsto a_t$$

such that:

$$\operatorname{argmin}_{\pi} J(\pi) = \mathbb{E}_{s_t, a_t \sim \pi} \sum_{t=1}^T c(s_t, a_t)$$

Decision Making

The goal is to find a policy π that will map states to actions:

$$\pi : s_t \mapsto a_t$$

such that:

$$\operatorname{argmin}_{\pi} J(\pi) = \mathbb{E}_{s_t, a_t \sim \pi} \sum_{t=1}^T c(s_t, a_t)$$

We actually
don't know this



Imitation Learning

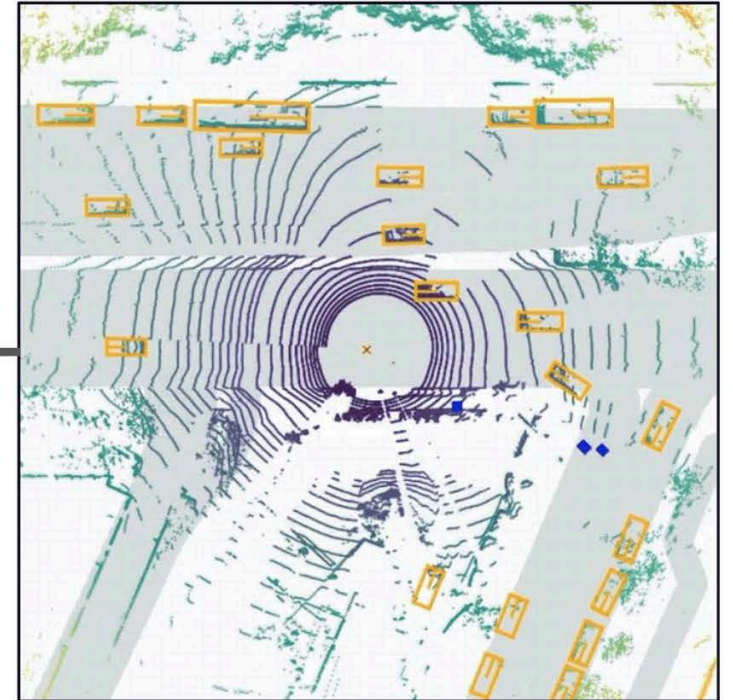
We can get away from knowing the cost by imitating experts (humans or systems). Let's say we collect a dataset:

$$D = \{(s, a)\}$$



Then, we use traditional empirical risk minimization (**ERM**) to learn to imitate:

$$\pi' = \operatorname{argmin}_{\pi} \mathbb{E}_{s, a \sim D} \mathcal{L}(a, \pi(s))$$



* This approach (and its limitations) dates back to 1989 ALVINN paper !

Imitation Learning

When you collect data: $D = \{(s_t, a_t)\} \mid \underbrace{s_t \sim d_{\pi^h}^t}_{\text{state distribution}}, \underbrace{a_t \sim \pi^h}_{\text{action distribution}}$

Mismatch

When you deploy:

$$s_t \sim d_{\pi_\theta}^t \quad a_t \sim \pi_\theta$$

Imitation Learning

- Does not work out of the box for AV planning
- One way to make it work is to add perturbations

ChauffeurNet: Learning to Drive by Imitating the Best and Synthesizing the Worst

Mayank Bansal, Alex Krizhevsky, Abhijit Ogale

Transformers in Perception

Autonomous Driving datasets

- KITTI
- Waymo Open Dataset
- nuScenes / nuPlan / nulmages
- Lyft prediction / perception datasets
- Yandex Vehicle Motion Prediction
- ...
- Many more

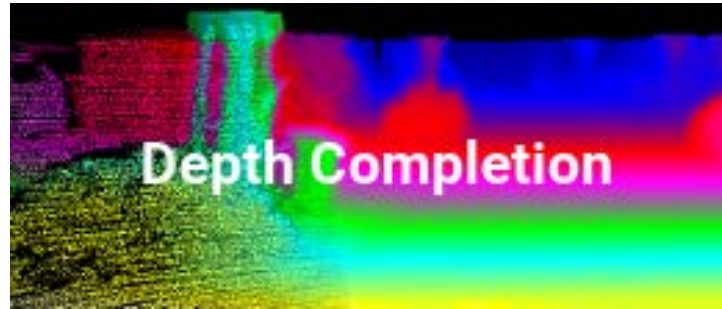
Autonomous Driving datasets

- KITTI: detection



Autonomous Driving datasets

- KITTI: stereo



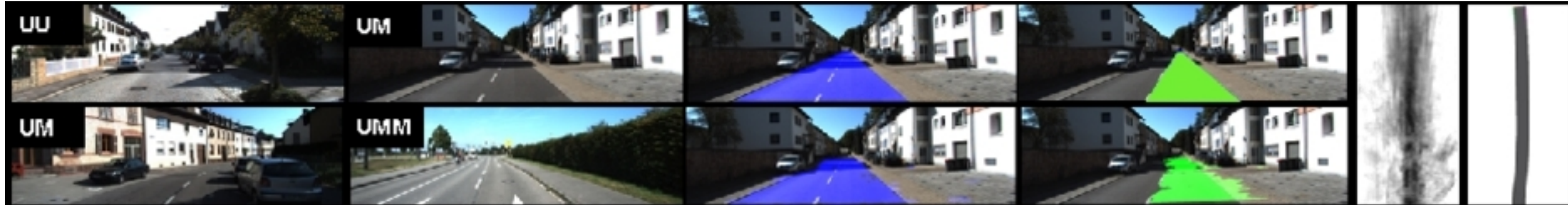
Autonomous Driving datasets

- KITTI: tracking



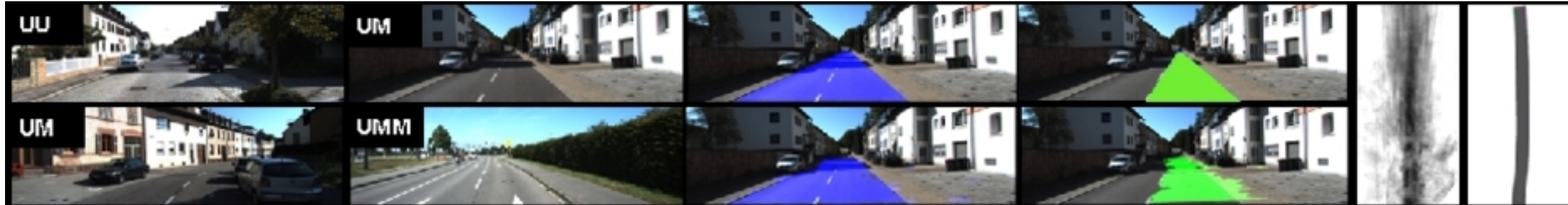
Autonomous Driving datasets

- KITTI: road lane detection



Autonomous Driving datasets

- KITTI: road lane detection



Autonomous Driving datasets

- KITTI: segmentation

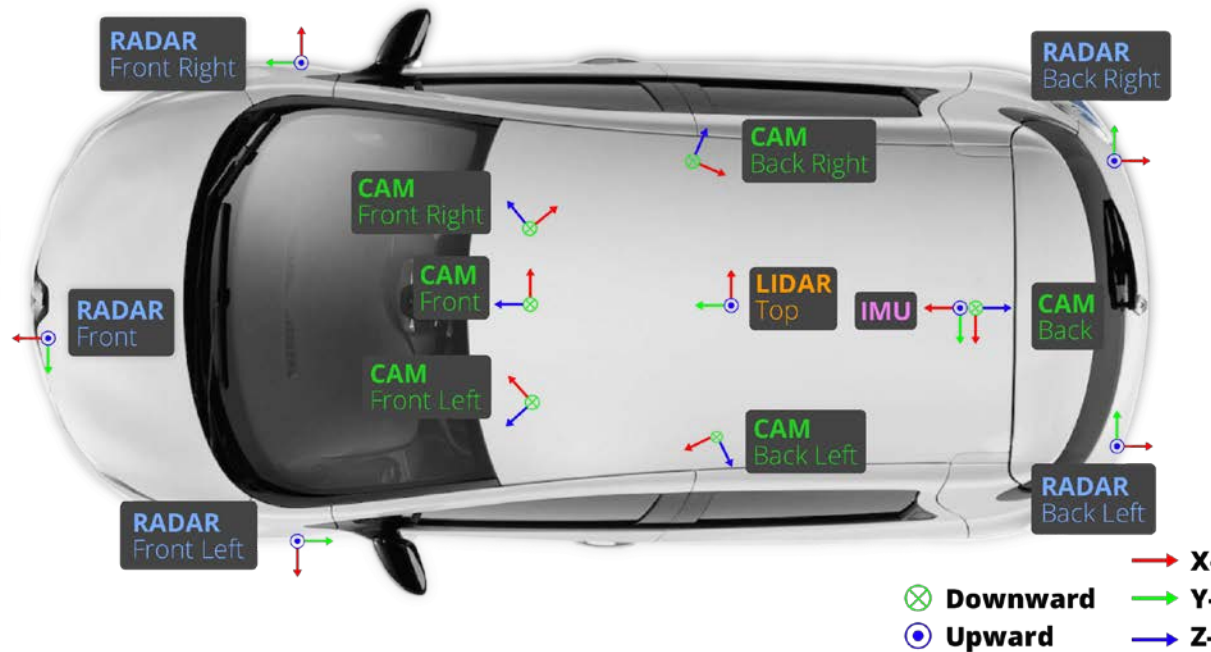
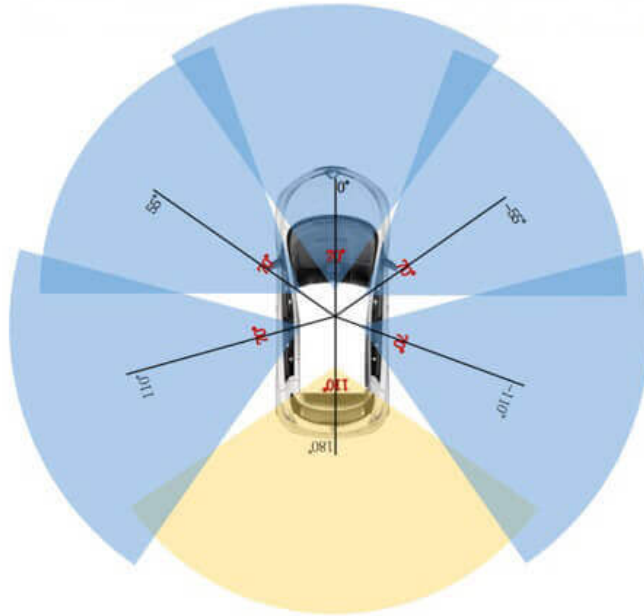


Autonomous Driving datasets

- nuScenes
- nuPlan
- nulmages

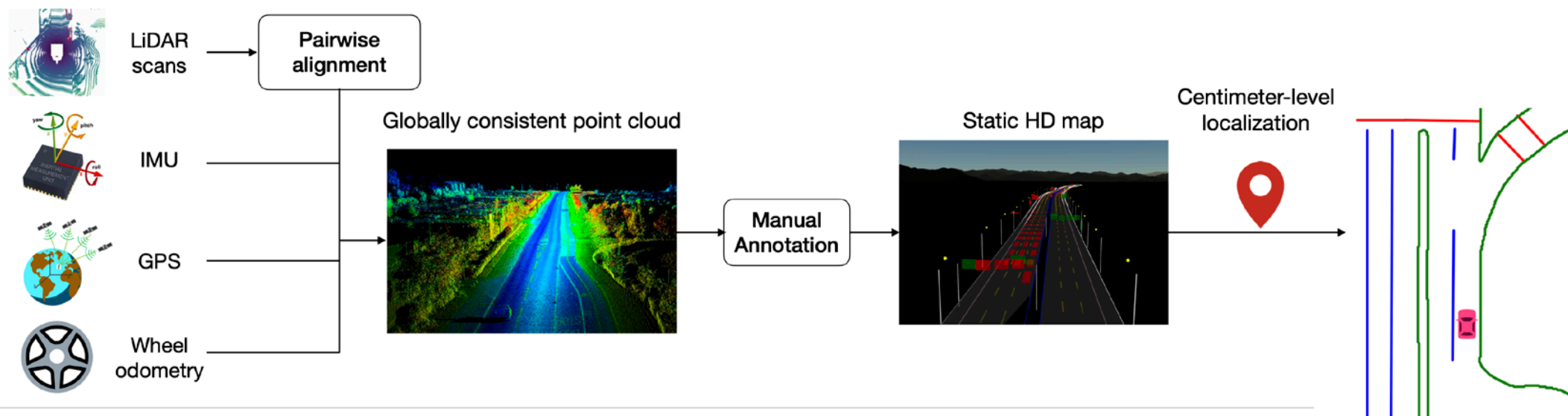


Singapore One North

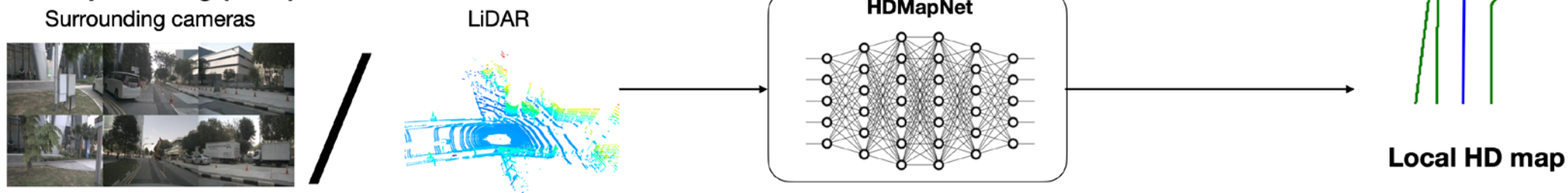


HDMapNet

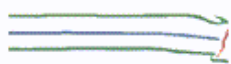
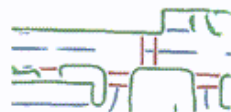
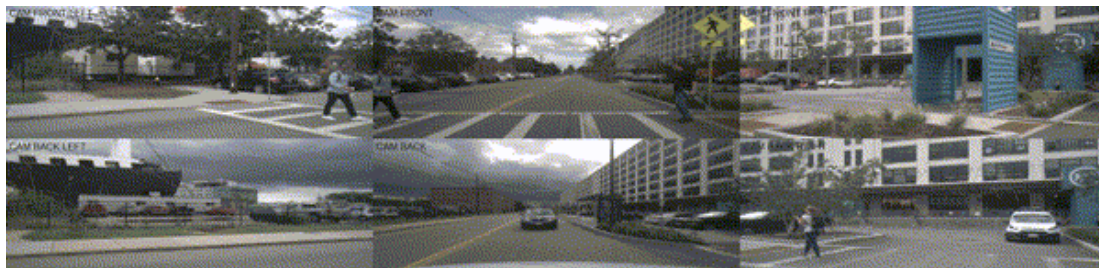
Traditional mapping pipeline



Online map learning (Ours)

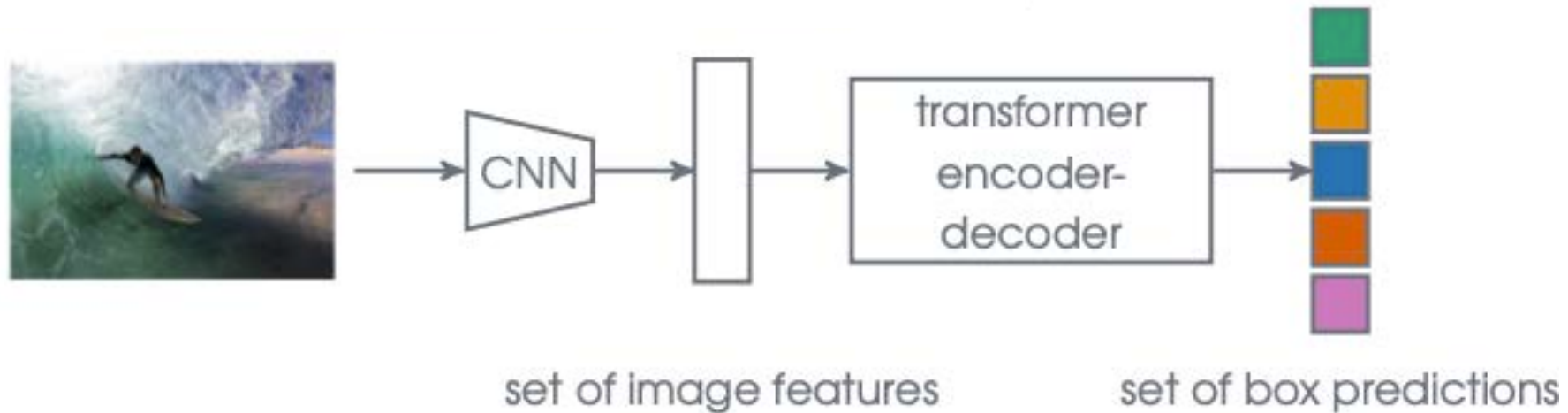


HDMapNet



DETR: End-to-end Object Detection

Applying NMT to translate CNN features to detections:



[End-to-end object detection with transformers](#)

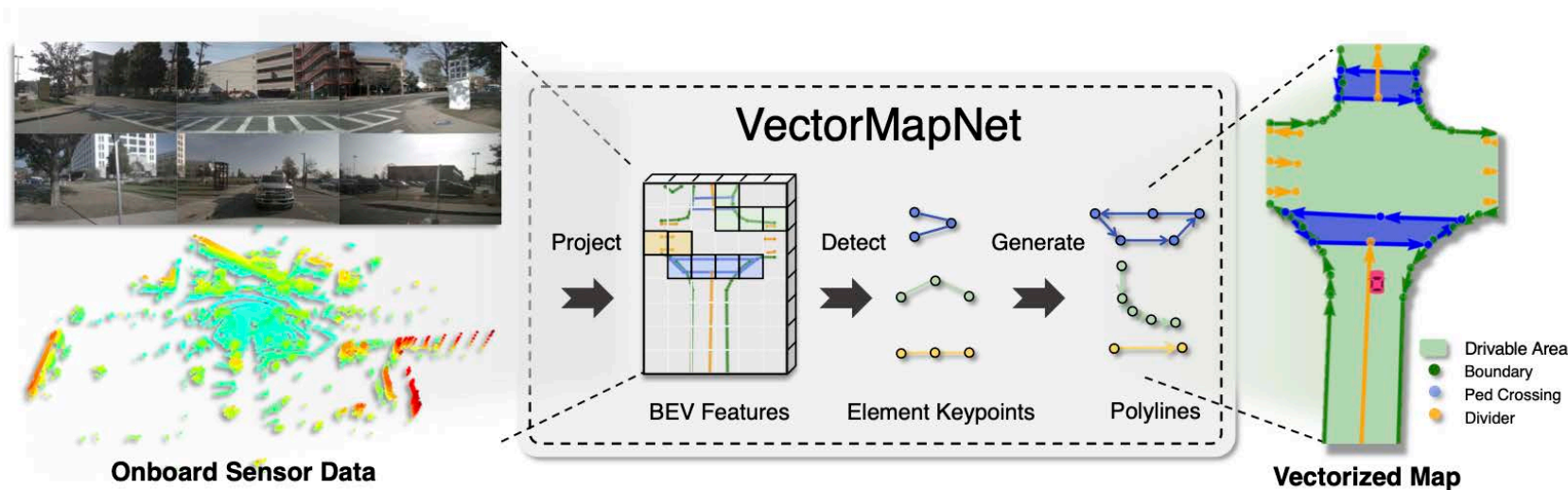
N Carion, F Massa, G Synnaeve, N Usunier, A Kirillov, S Zagoruyko

Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23 ...

4872

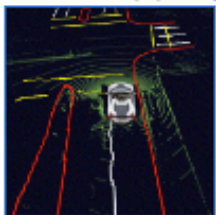
2020

VectorMapNet and MapTR



Related Projects on **VCAD (Vision-Centric Autonomous Driving)**

BEV Mapping



HDMapNet

BEV Detection



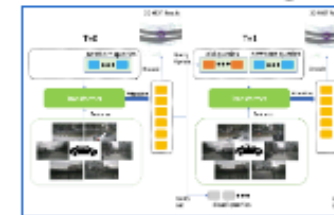
DETR3D

BEV Fusion



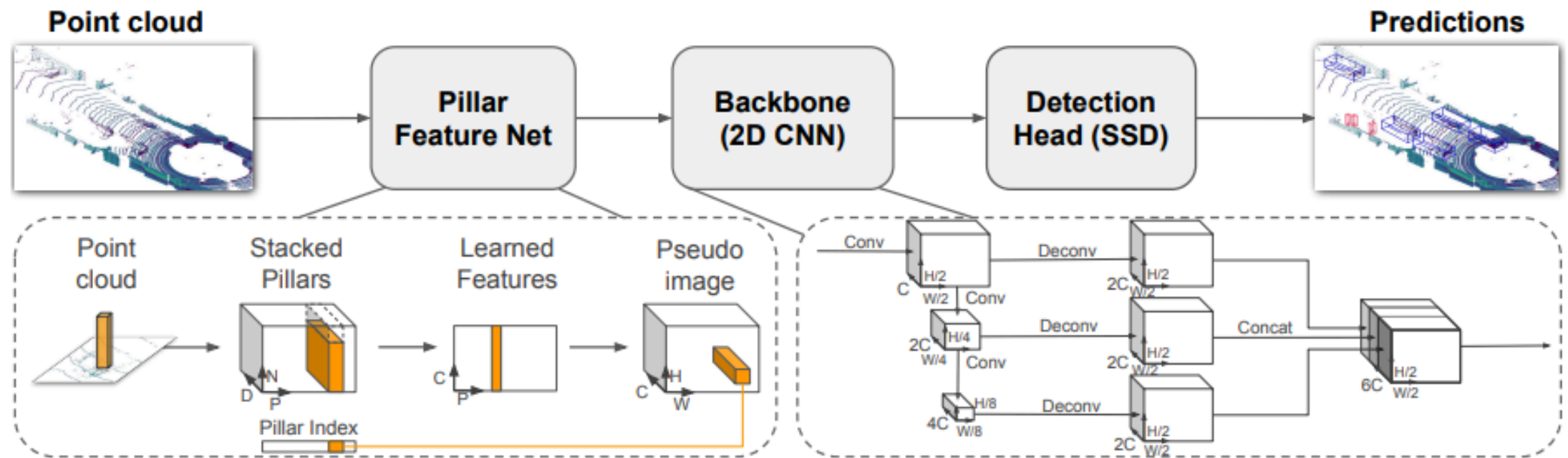
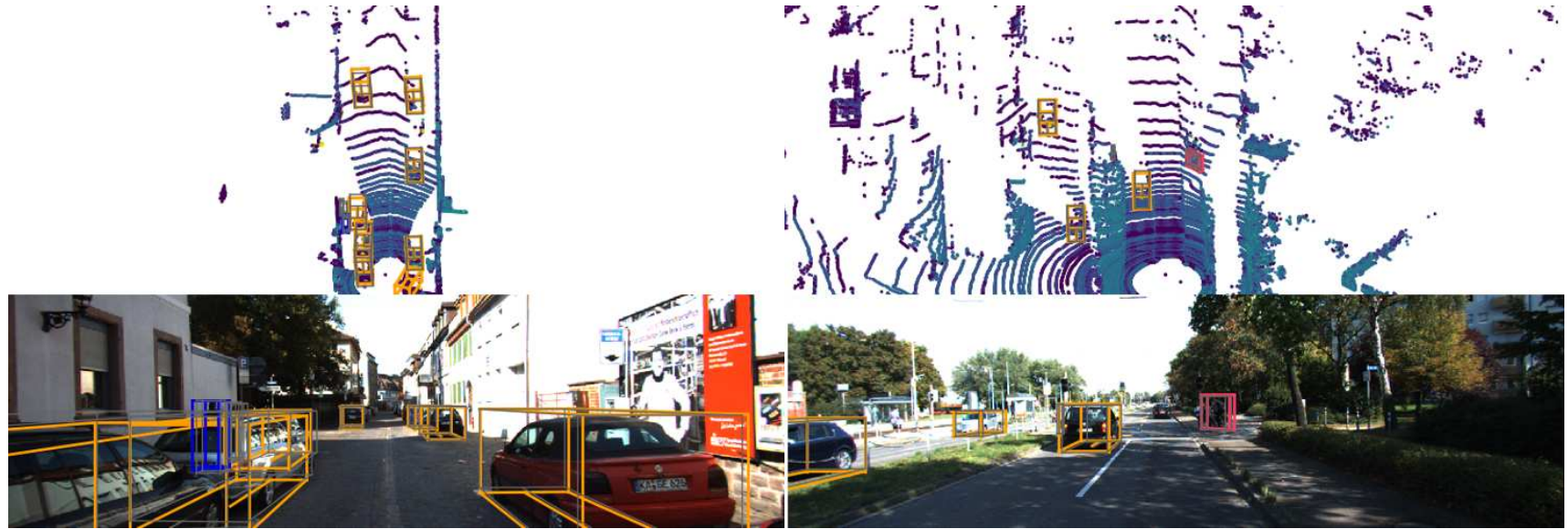
FUTR3D

BEV Tracking

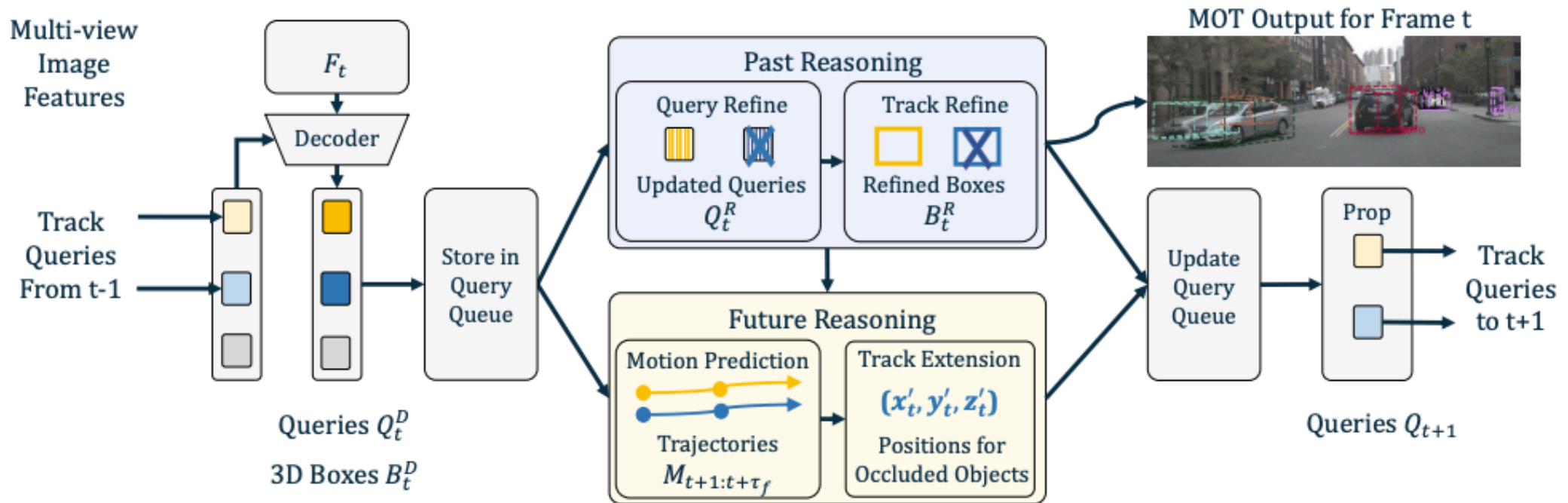


MUTR3D

PointPillars



End-to-end multi-camera 3D MOT

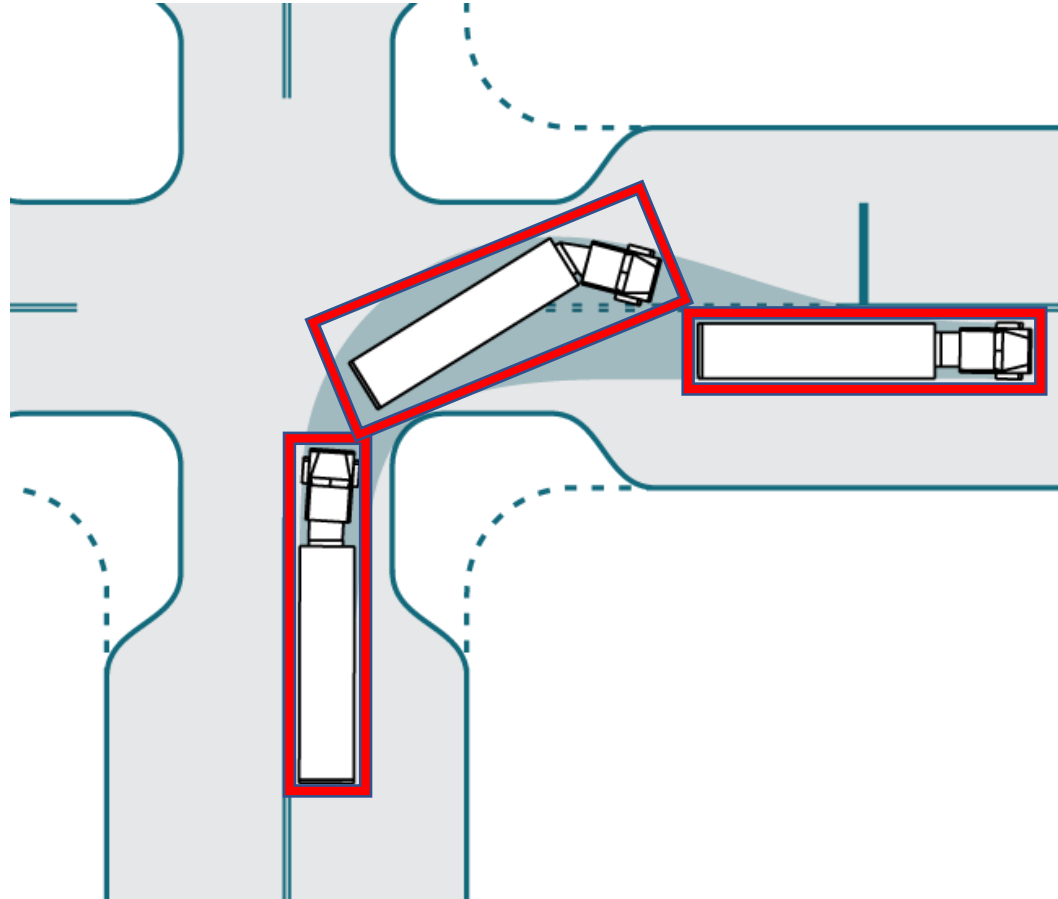


Perception problems

Planning and simulation uncover perception problems:

- 1. Long tail**
2. Data collection and annotation are expensive
3. Offline perception is computationally heavy
4. Requires storing large amounts of data
5. Creating high-definition maps is difficult, become obsolete with time

Anecdotal cases: trucks



Problem for simulation
and planning

Anecdotal cases: trucks

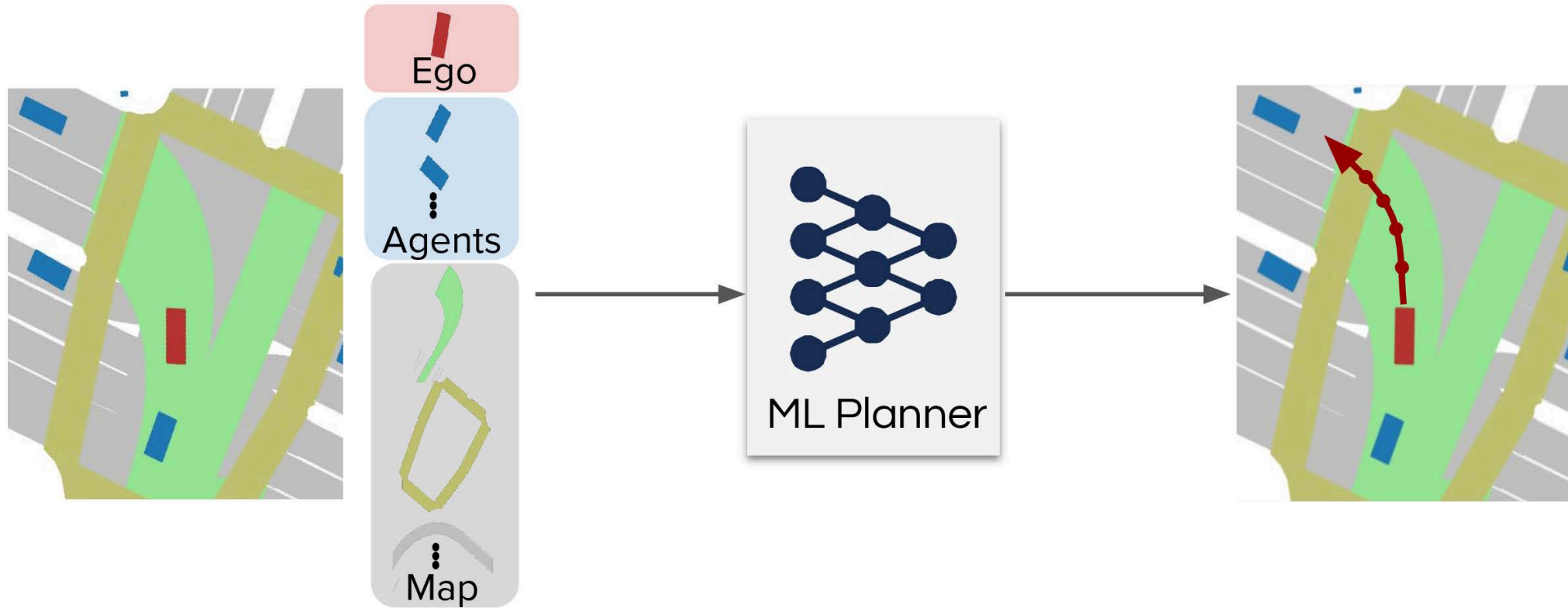


Anecdotal cases: animals



Transformers in Motion Planning

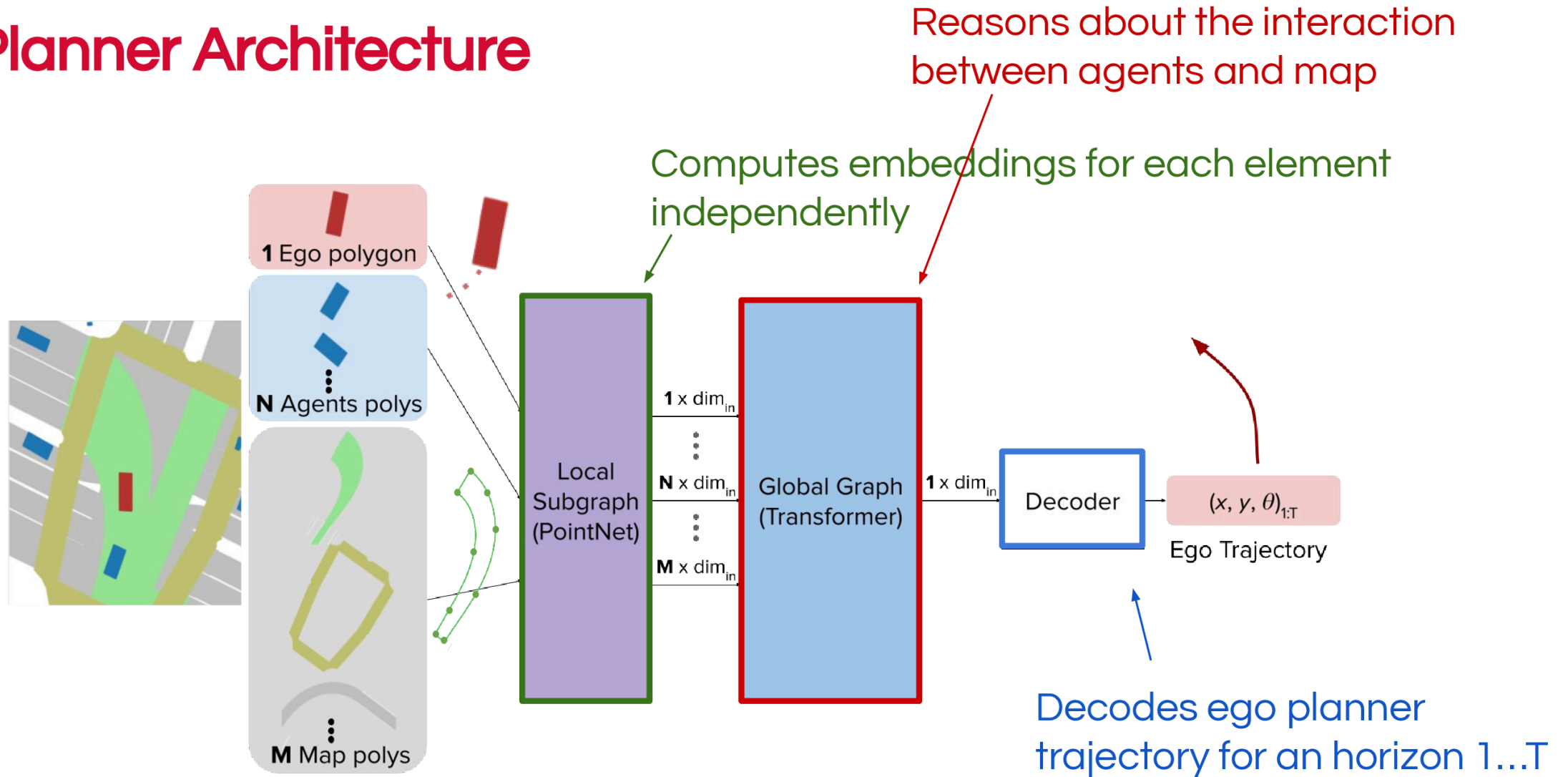
ML Planner



Inputs: localization, perception agents, traffic light status, & HD map

Output: ML planned trajectory

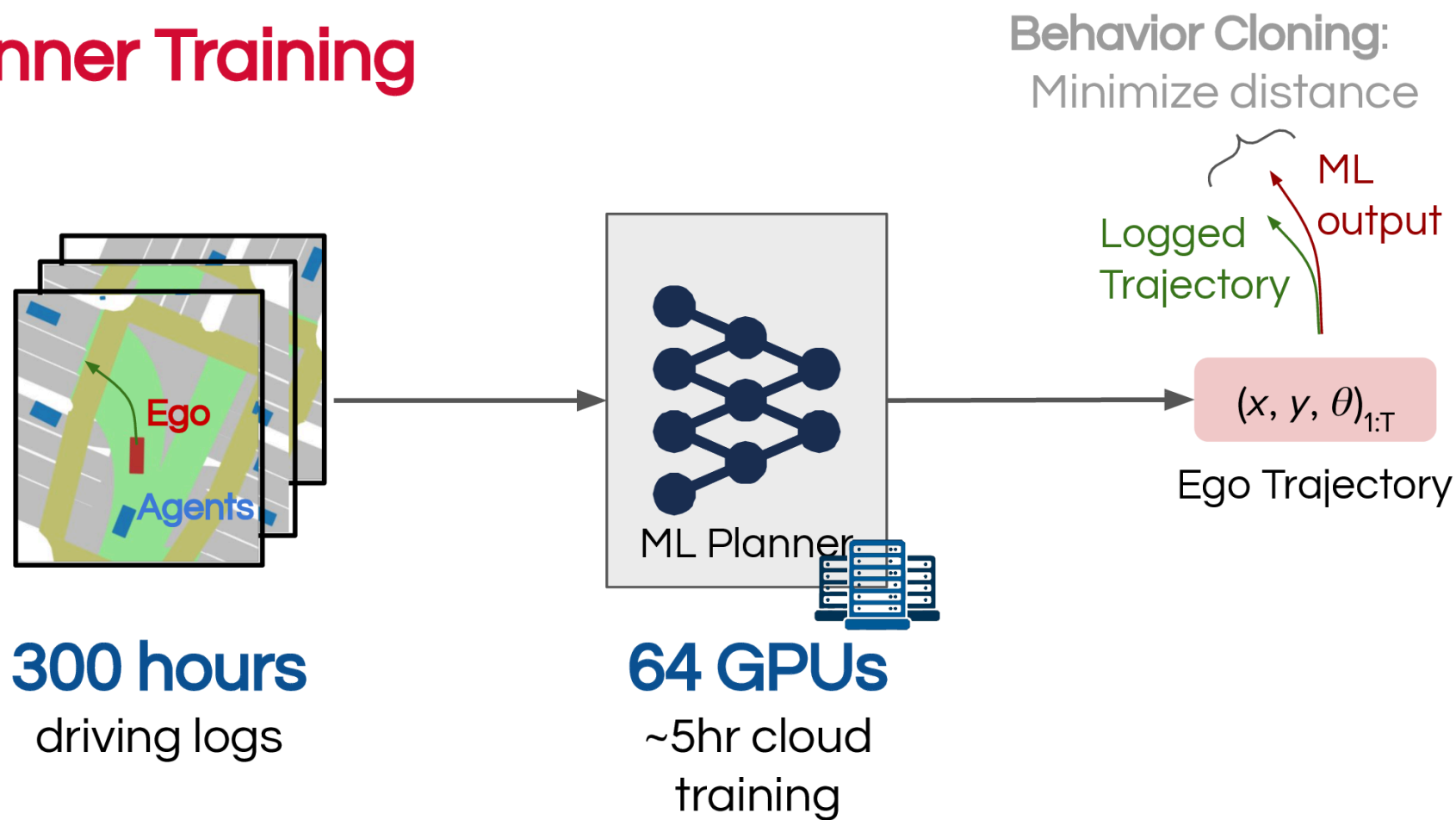
ML Planner Architecture



¹PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation (Qi, 20

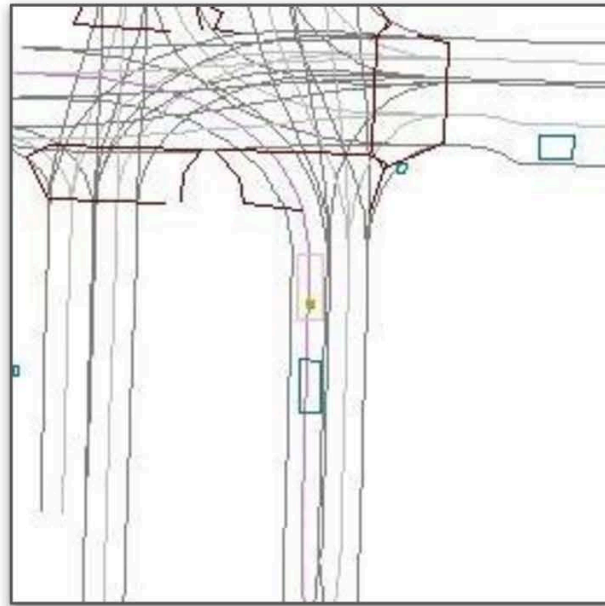
²Attention Is All You Need (Vaswani, 20

ML Planner Training



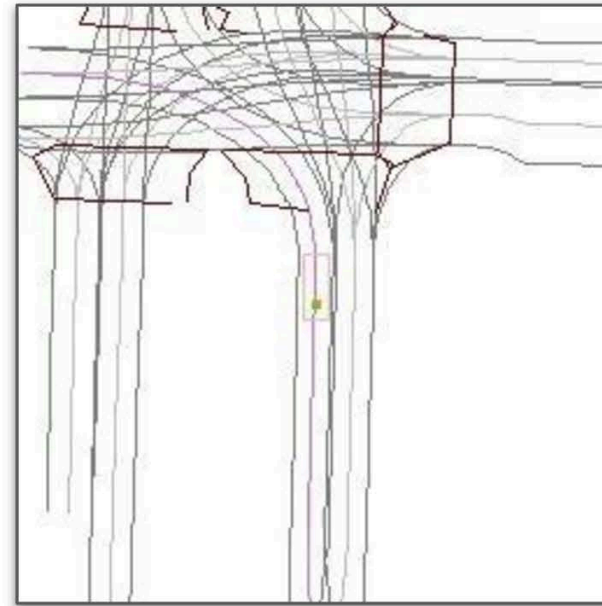
Causal stability

Example: at a red light, the ML Planner learned “**chase vehicle movement means go**” instead of “**green light means go**”



Agents from log

Same driving
experience

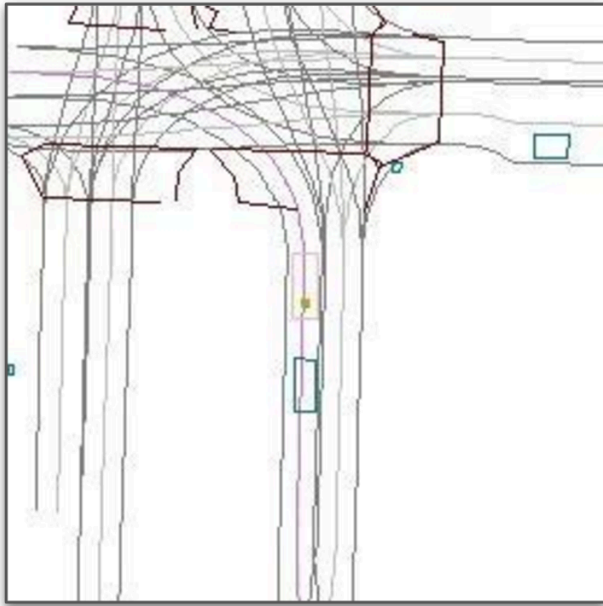


Agents removed
“Ghost town”

— Logged Ego
— ML planner

Causal stability

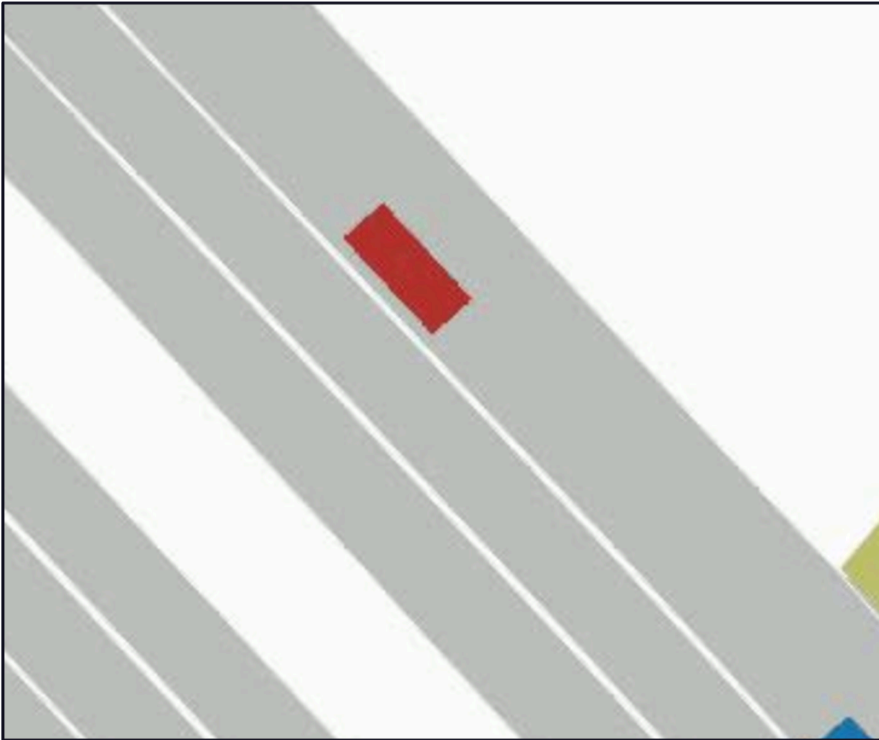
Example: at a red light, the ML Planner learned “**chase vehicle movement means go**” instead of “**green light means go**”



Agents from log

- Model is doing the **right thing**, but for the **wrong reasons**, important having means to **uncover it**
- **Causal instability** if naively trained, common problem in ML
- **Hint:** what would be the cause of these issues ? Would that happen if the model is trained with data that breaks this **spurious correlation** ?

Simulation



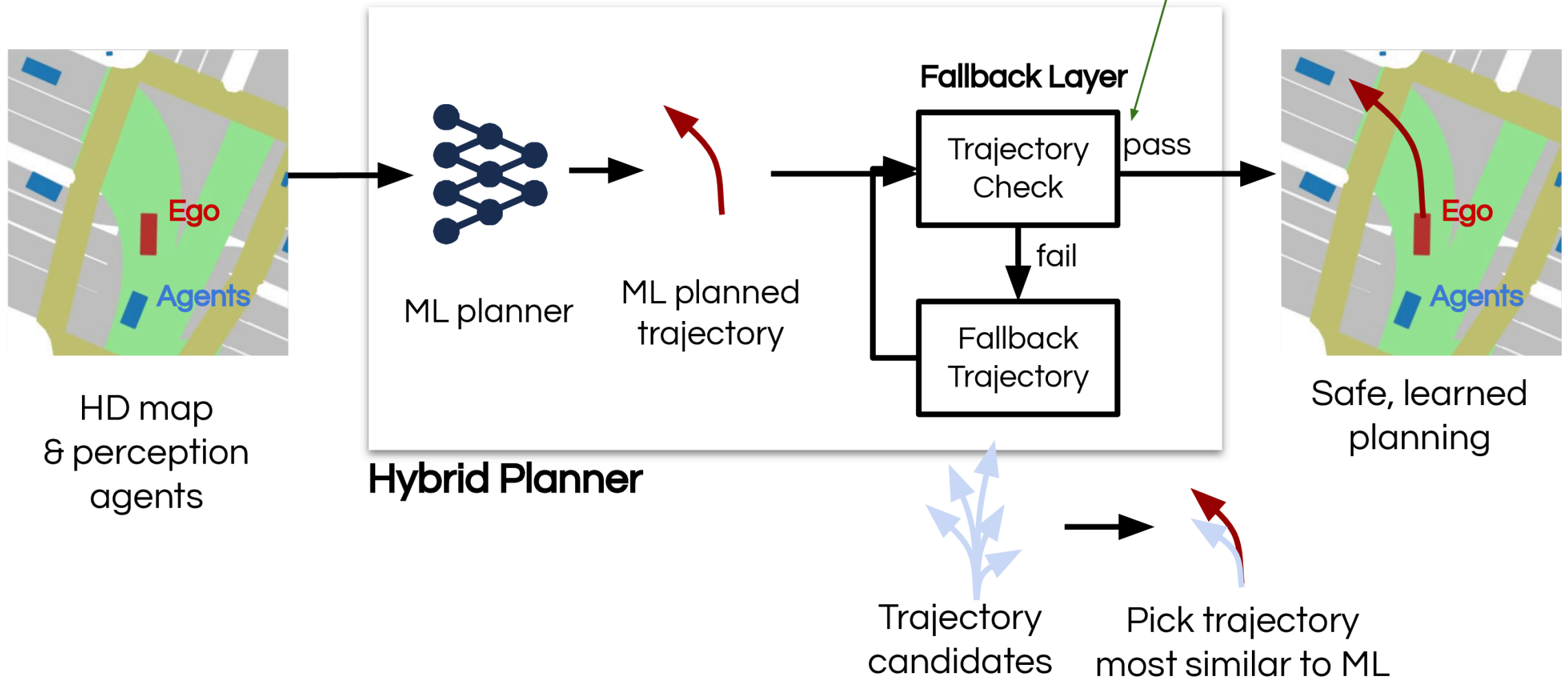
- Road lane
- Agent
- Self driving vehicle

- Agent replay during evaluation is **non-reactive**
- To uncover diverse issues, you need **diverse scenarios**
- Simulating diverse scenes can **uncover failure modes and causal instability** of the policy

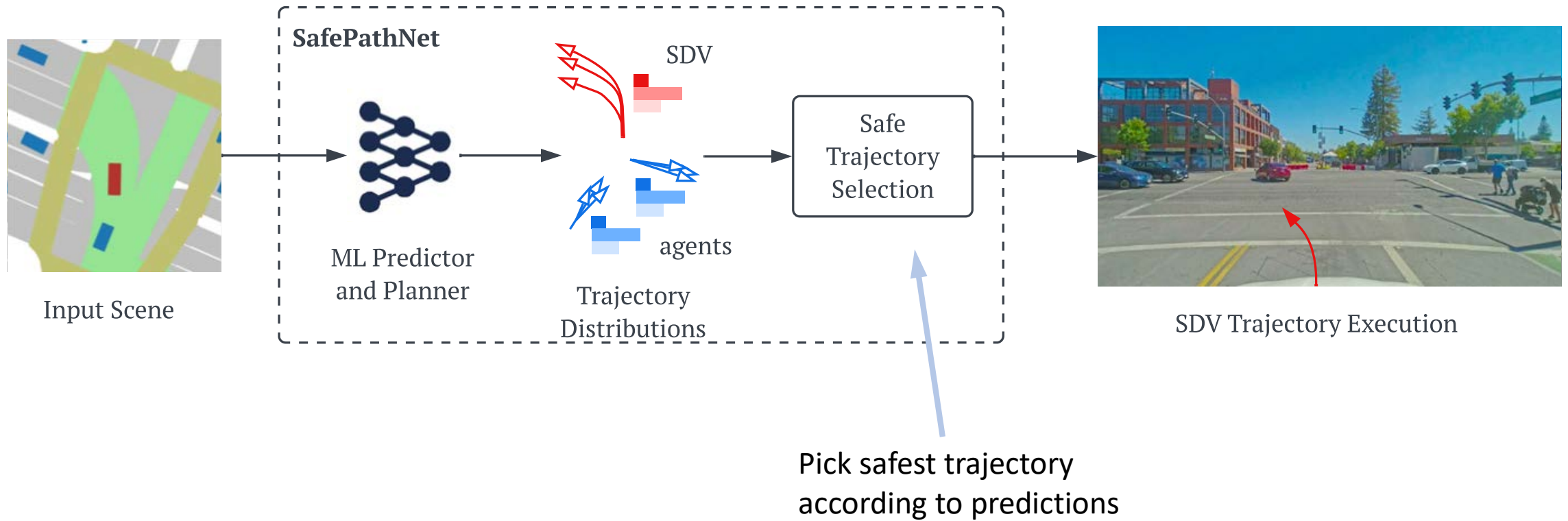
Ensuring safety

SafetyNet - Hybrid Planner

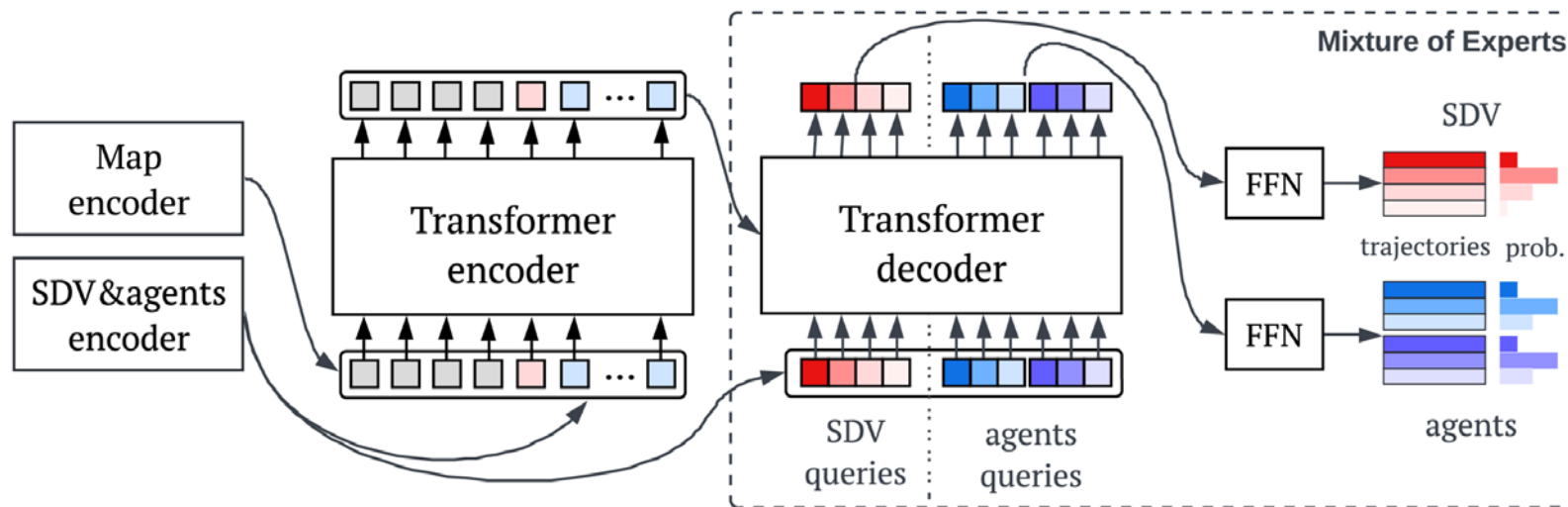
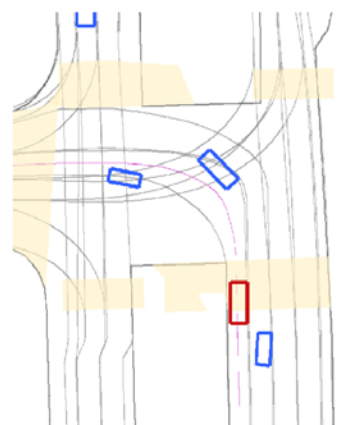
- ✓ Comfort
- ✓ Safety
- ✓ Legality



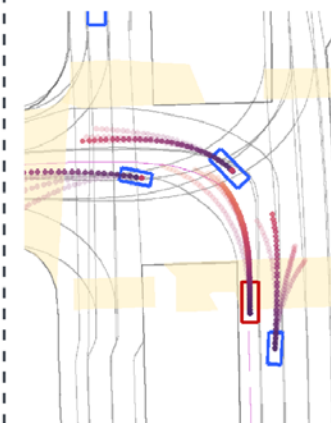
SafePathNet



SafePathNet

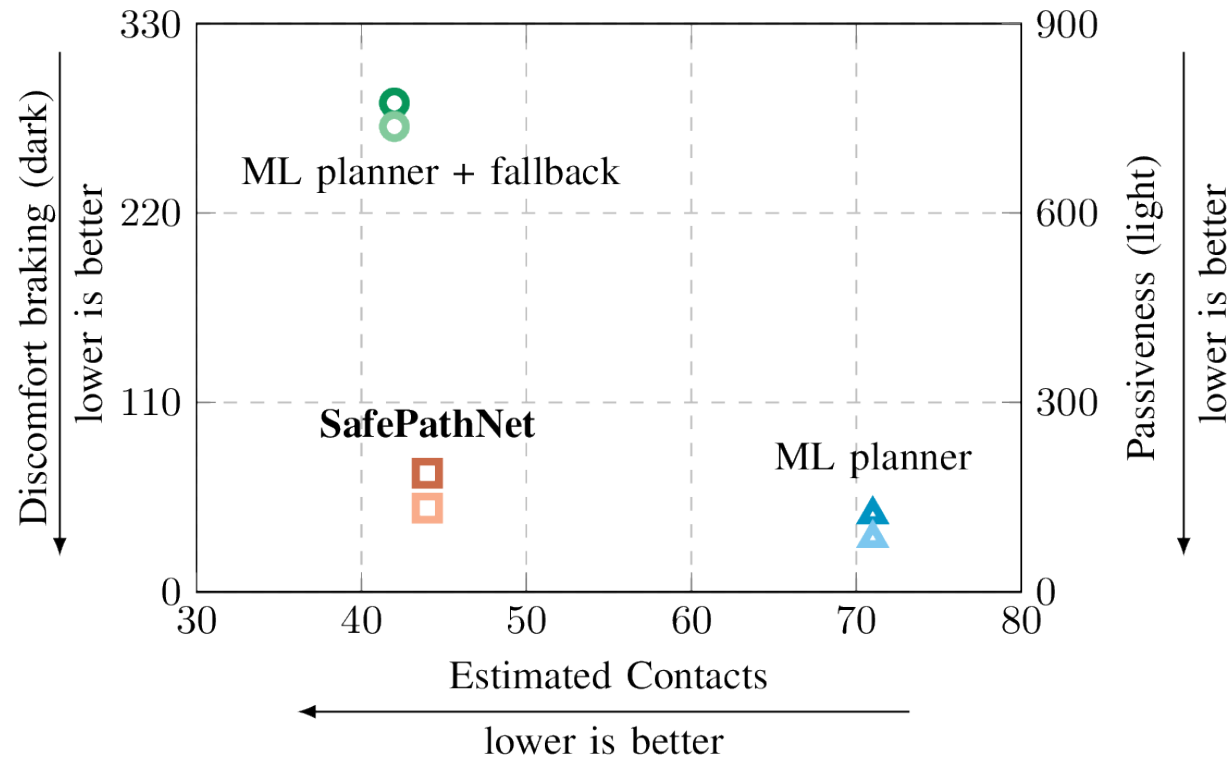
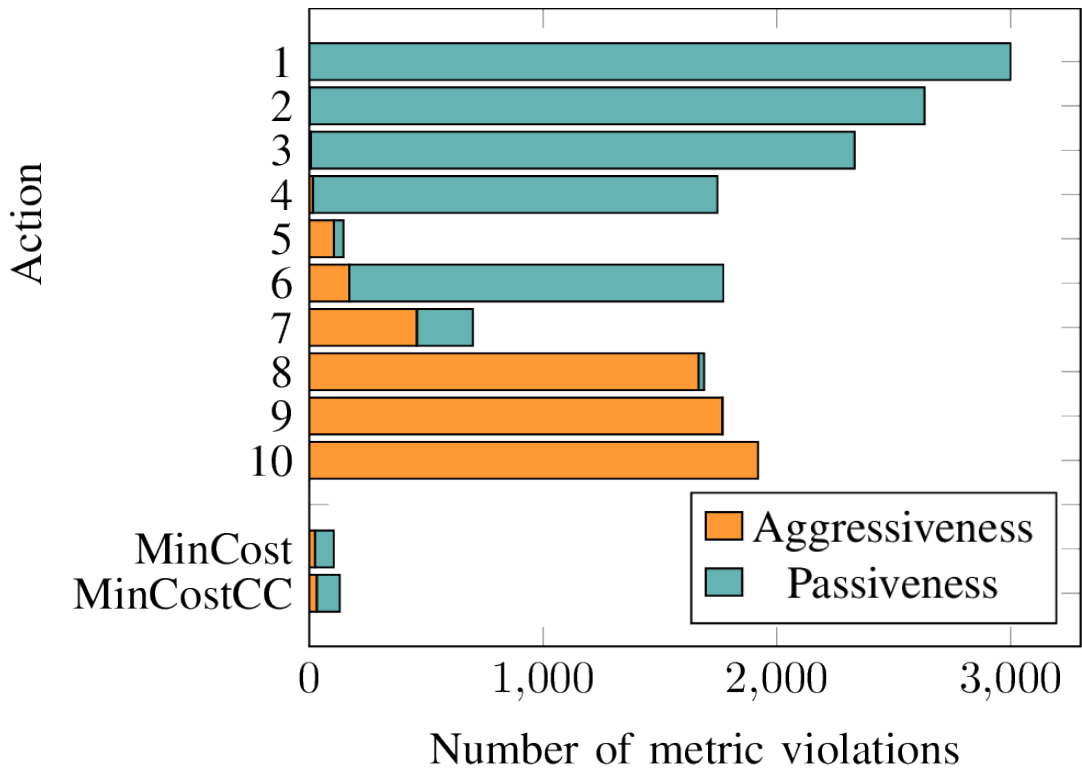


Learns multiple futures
in unsupervised way



Similar to DETR

SafePathNet



Safe Real-world Autonomous Driving by Learning to Predict and Plan with a Mixture of Experts

Stefano Pini, Christian S. Perone, Aayush Ahuja,
Ana Sofia Rufino Ferreira, Moritz Niendorf, Sergey Zagoruyko

Tabular Deep Learning

Lecturer: Artem Babenko



ASCOMP 2024

Lecturer

- Artem Babenko, Research Lead @ Yandex Research
- Publications on deep/machine learning for tabular data by Yandex Research
 - (NeurIPS 2018) CatBoost: unbiased boosting with categorical features
 - (ICLR 2020) Neural Oblivious Decision Ensembles for Deep Learning on Tabular Data
 - (NeurIPS 2021) Revisiting Deep Learning Models for Tabular Data
 - (NeurIPS 2022) On Embeddings for Numerical Features in Tabular Deep Learning
 - (arXiv 2022) Revisiting Pretraining Objectives for Tabular Deep Learning
 - (ICML 2023) TabDDPM: Modelling Tabular Data with Diffusion Models
 - (ICLR 2024) TabR: Tabular Deep Learning Meets Nearest Neighbors
 - (2024) Several projects under submission
- Tabular DL projects by Yandex Research: github.com/yandex-research/rtdl
(RTDL = Research on Tabular Deep Learning)

YR Tabular DL team



Artem Babenko



Yura Gorishniy



Nikolay Kartashev



Akim Kotelnikov



Ivan Rubachev

Outline

- Introduction
- The pre-deep learning era of Tabular ML
- Modern Tabular Deep Learning
- Real-world impact

Outline

- Introduction
- The pre-deep learning era of Tabular ML
- Modern Tabular Deep Learning
- Real-world impact

Tabular data

Tabular data — two-dimensional tables

- rows ~ objects
- columns ~ features

Today we focus on

- supervised regression
- supervised classification

Applications

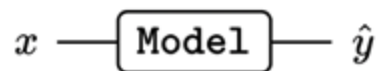
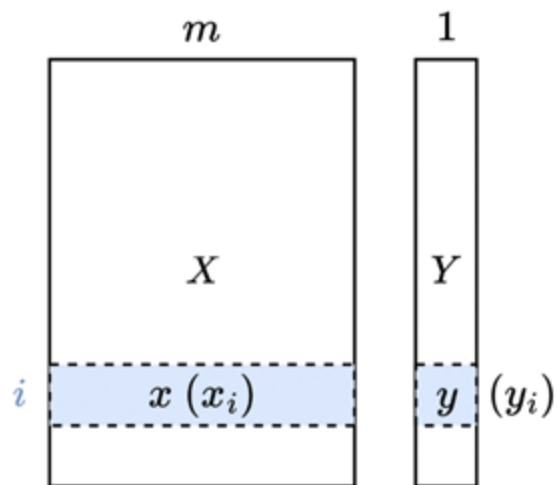
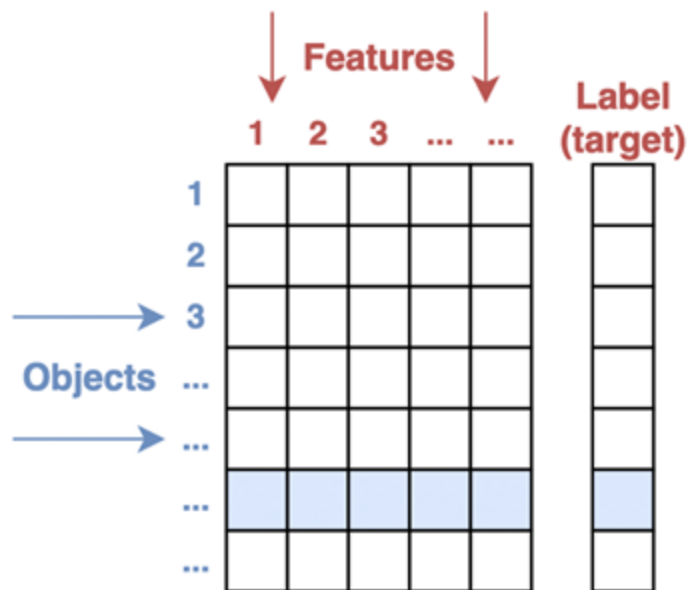
- everyday tasks...
- ...and many others

A	B	C
...	...	...
...	...	...
...	...	...

X y

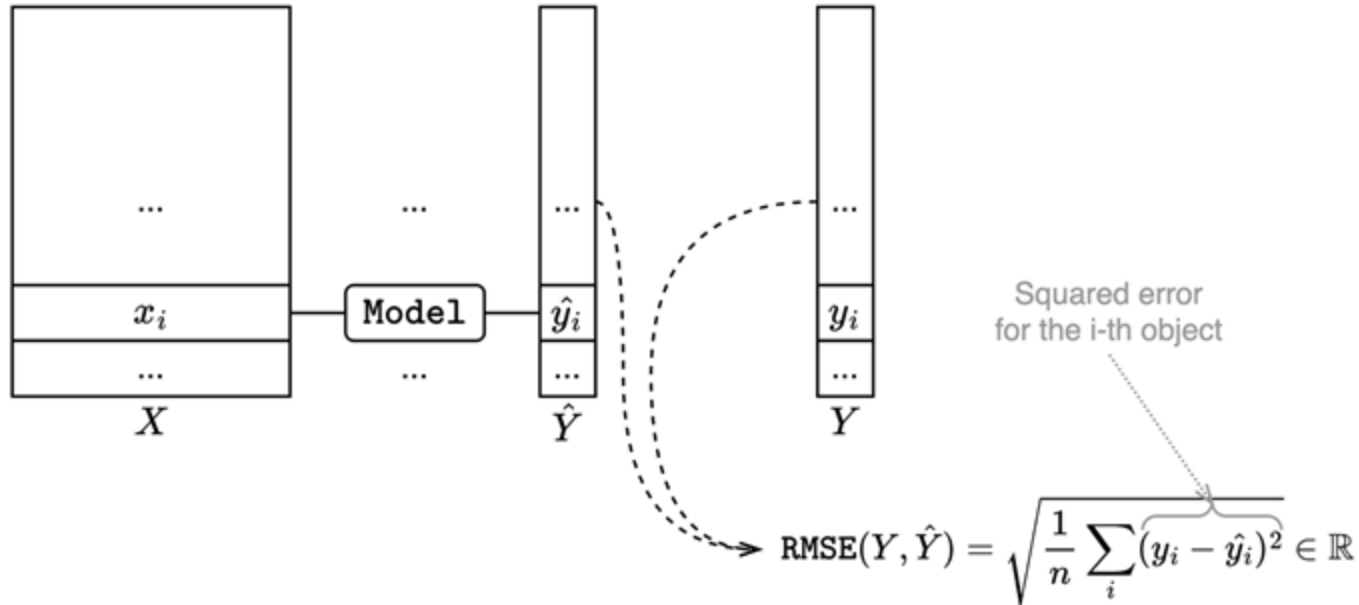


Notation

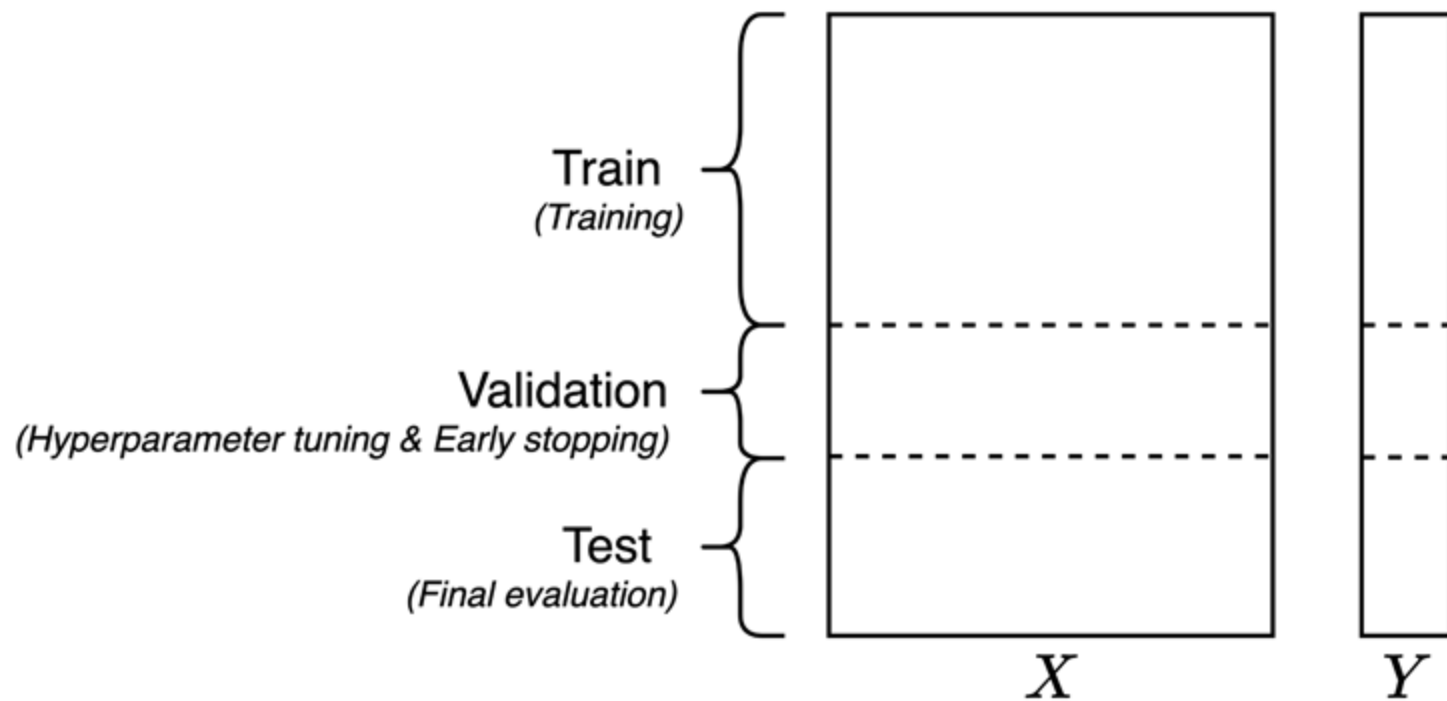


Metrics

Metrics are used to evaluate how well predictions approximate labels.
Example: Root Mean Squared Error (RMSE)



Dataset splitting



Data preprocessing

Continuous features

- [QuantileTransformer](#)
- QuantileTransformer with noise ([example](#))
- [StandardScaler](#)
- Missing data: $x \rightarrow (0, 1)$ if x is NaN else $(x, 0)$

Categorical features

- One-hot encoding
(typically used when the number of distinct values is not too high)
- Embeddings
- Missing data: make NaN a new category

Binary features

- Just encode as $\{0, 1\}$
- Missing data: any reasonable strategy (see “Continuous” and “Categorical”)

Ordinal features

- [OrdinalEncoder](#)
- Thermometer encoding
- Cumulative embeddings

P.S. Standardize regression labels

Specifics of Tabular ML problems

- Limited dataset sizes
- Heterogeneous and mixed-type features
- Each problem has its own nature
- Target dependencies are often “ill-behaved”

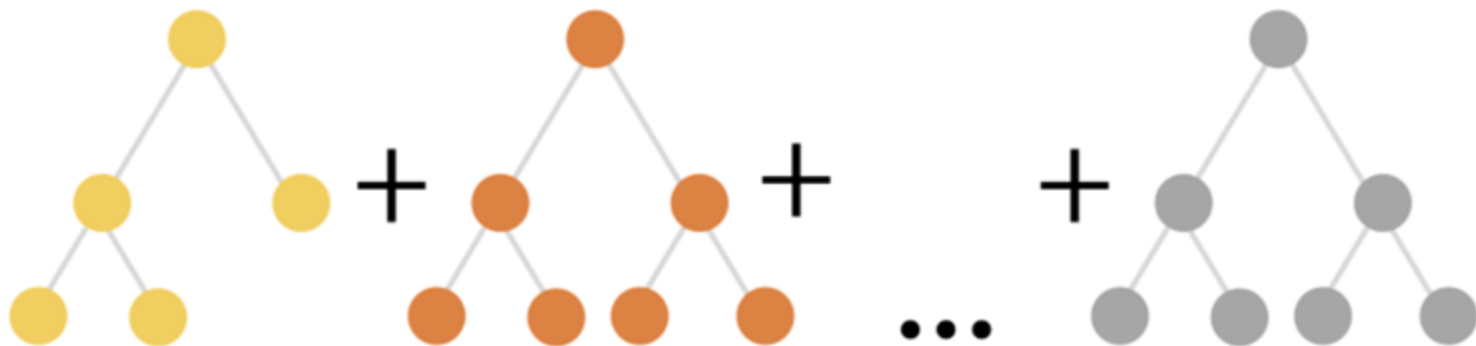
Outline

- Introduction
- The pre-deep learning era of Tabular ML
- Modern Tabular Deep Learning
- Real-world impact

Classic machine learning algorithms

- K-Nearest neighbors
- Linear model (Linear regression, Logistic regression, ...)
- Support vector machine (SVM)
- Decision tree
- Random forest
- Gradient-boosted decision tree (GBDT)

Gradient Boosting Decision Trees (GBDT)



dmlc
XGBoost



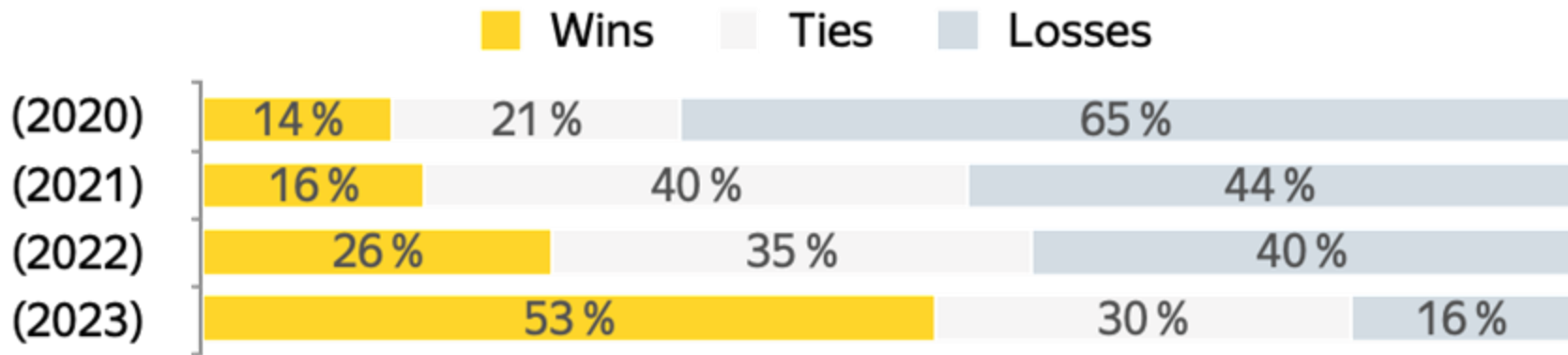
Yandex
CatBoost



LightGBM

GBDT is a strong baseline for Tabular ML

- Efficient
- Easy-to-use
- Effective



Best DL model vs XGBoost on the academic benchmark of ~40 datasets

Outline

- Introduction
- The pre-deep learning era of Tabular ML
- Modern Tabular Deep Learning
- Real-world impact

Chaos in Tabular DL before 2021

Differentiable trees

- NODE (Popov et al., 2020)

“Attention”-based models

- AutoInt (Song et al., 2019)
- TabNet (Arik and Pfister, 2020)

Multiplicative feature interactions

- DCN2 (Wang et al., 2020)

Specific activation functions

- SNN (Klambauer et al., 2017)

Boosting-like models

- GrowNet (Badirli et al., 2020)

And many others

- ...

2021: Are we really making progress in Tabular DL? [1,2,3]

- Tuning protocols and evaluation are often unfair
- GBDT is still superior to DL
- Sophisticated DL models are often inferior to simple ones

[1] Revisiting Deep Learning Models for Tabular Data, Gorishniy et al., 2021

[2] Tabular Data: Deep Learning is not all you need, Schwartz-Ziv et al., 2021

[3] Regularization is all you need: simple neural nets can excel on tabular data, Kadra et al., 2021

Outline

- Introduction
- The pre-deep learning era of Tabular ML
- Modern Tabular Deep Learning
 - MLP, Resnet, FT-Transformer
- Real-world impact

MLP

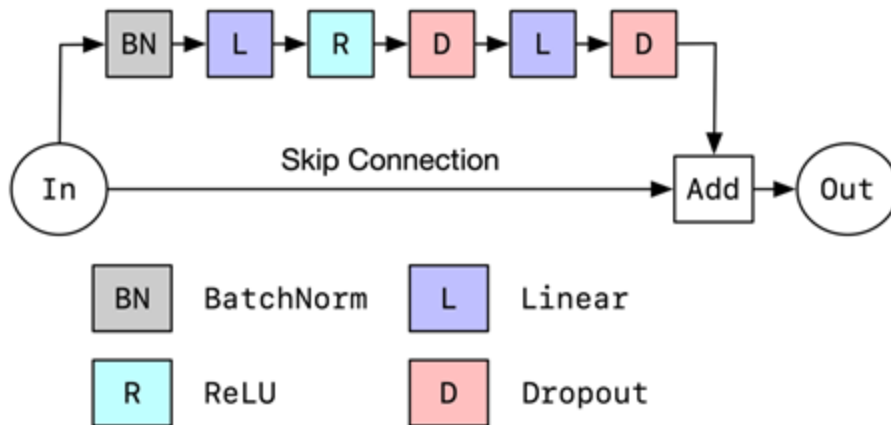
- Simple and fast
- Average performance



One MLP block

ResNet for Tabular Data

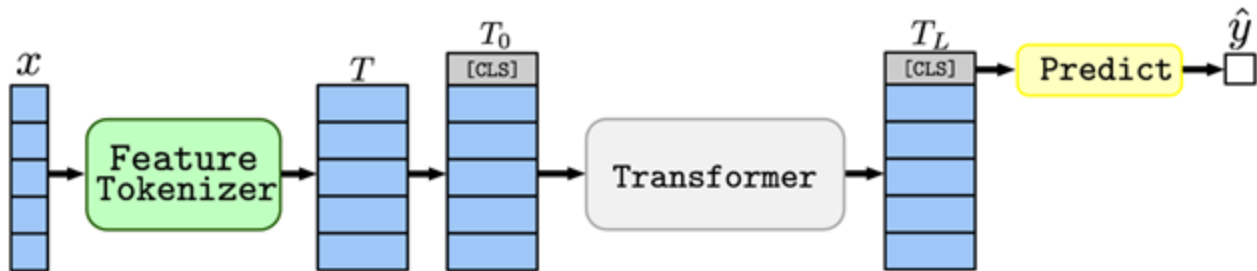
- Inspired by ResNet (He et al., 2015)
- Quite simple and relatively fast
- Hopefully, more powerful than MLP



One ResNet block

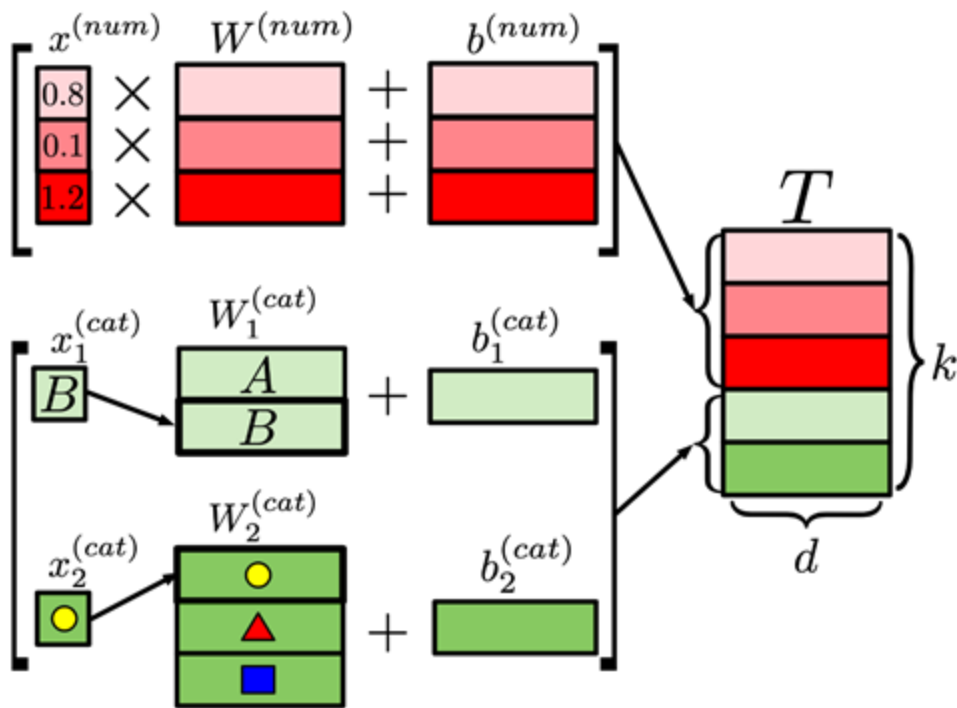
FT-Transformer (Ours)

- Based on Transformer (Vaswani et al., 2017)
- Slower than ResNet
- Hopefully, more powerful than MLP and ResNet

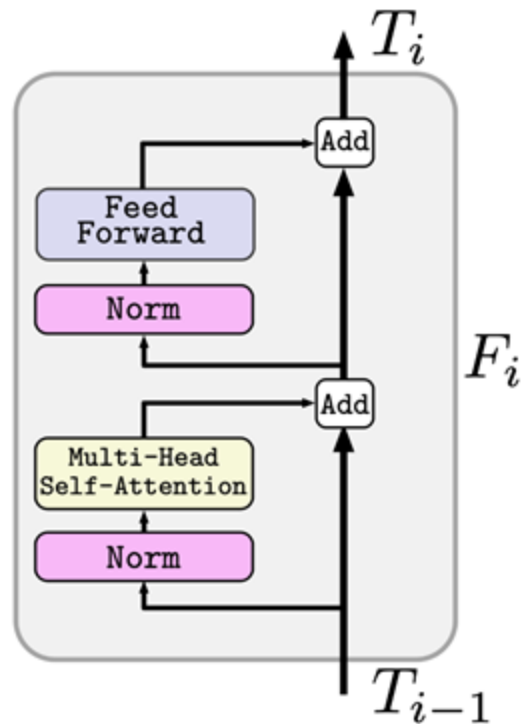


FT-Transformer

FT-Transformer (Ours)



Feature Tokenizer



One Transformer block

Experiments

Experiments: datasets and protocol

Dataset	N	K	Metric
California Housing	21K	8	RMSE
Adult	49K	14	Accuracy (B)
Helena	66K	27	Accuracy (M)
Jannis	84K	54	Accuracy (M)
Higgs (small)	99K	28	Accuracy (B)
ALOI	108K	128	Accuracy (M)
Epsilon	500K	2000	Accuracy (B)
Year	516K	90	RMSE
Covtype	582K	54	Accuracy (M)
Yahoo	710K	699	RMSE
Microsoft	1201K	136	RMSE

$N \sim$ dataset size

$K \sim$ number of features

$B \sim$ binary

$M \sim$ multiclass

- Tuning
 - mostly Optuna (Akiba et al., 2019) (50-100 iterations)
 - grid search from original papers
- Evaluation
 - 15 random seeds
 - ensembles: three ensembles (each consists of five single models)
- No DL tricks
 - no augmentation
 - no lr scheduling
 - no pretraining
 - etc.

Experiments: Neural Networks

Model	Average rank (std)
TabNet	7.5 (2.0)
SNN	6.4 (1.4)
AutoInt	5.7 (2.3)
GrowNet	5.7 (2.2)
MLP	4.8 (1.9)
DCN V2	4.7 (2.0)
NODE	3.9 (2.8)
ResNet	3.3 (1.8)
FT-Transformer	1.8 (1.2)

Takeaways

- MLP is still a good sanity check
- ResNet is a strong baseline
- FT-Transformer outperforms existing solutions on most of the tasks
- Tuning matters

Experiments: FT-Transformer vs GBDT (ensembles)

Dataset	CA 	AD 	HE 	JA 	HI 	AL 	EP 	YE 	CO 	YA 	MI 
#objects	20K	49K	65K	84K	98K	108K	500K	515K	581K	710K	1200K
XGBoost (d)	0.462	0.874	0.348	0.711	0.717	0.924	0.88	9.192	0.964	0.761	0.751
CatBoost (d)	0.428	0.873	0.386	0.724	0.728	0.948	0.889	8.885	0.91	0.749	0.744
FT-Transformer (d)	0.454	0.86	0.395	0.734	0.731	0.966	0.897	8.727	0.973	0.747	0.742
FT-Transformer*	0.448	0.86	0.398	0.739	0.731	0.967	0.898	8.751	0.973	0.747	0.743

(d) ~ default configuration *out of competition

 Accuracy  RMSE **Best**

Takeaways

- ensemble of default FT-Transformers is a powerful thing

Experiments: ResNet & FT-Transformer vs GBDT (ensembles)

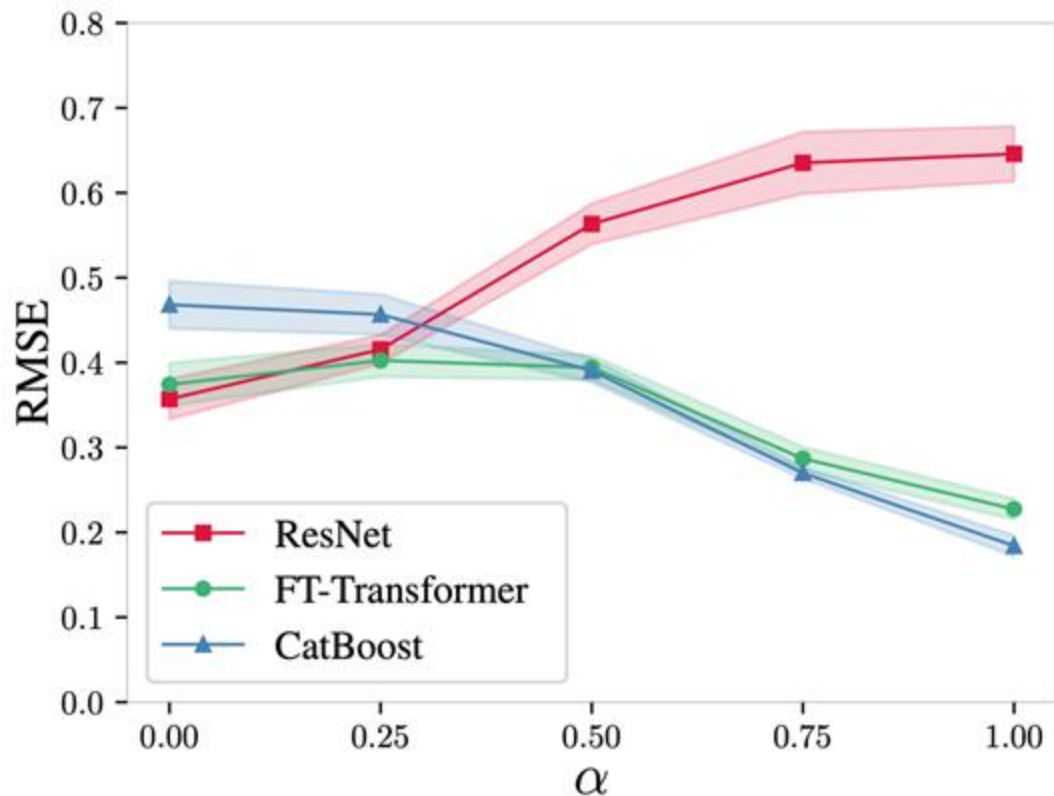
<i>Dataset</i>	CA 	AD 	HE 	JA 	HI 	AL 	EP 	YE 	CO 	YA 	MI 
<i>#objects</i>	20K	49K	65K	84K	98K	108K	500K	515K	581K	710K	1200K
XGBoost	0.431	0.872	0.377	0.724	0.728	-	0.886	8.819	0.969	0.732	0.742
CatBoost	0.423	0.874	0.388	0.727	0.729	-	0.89	8.837	0.968	0.74	0.741
ResNet	0.478	0.857	0.398	0.734	0.731	0.966	0.898	8.77	0.967	0.751	0.745
FT-Transformer	0.448	0.86	0.398	0.739	0.731	0.967	0.898	8.751	0.973	0.747	0.743

 Accuracy  RMSE
Best

Takeaways

- “DL vs GBDT” is an open problem
- **FT-Transformer reduces the gap between ResNet and GBDT**

An intriguing property of FT-Transformer



$$x \sim \mathcal{N}(0, I_k),$$

$$y = \alpha \cdot f_{GBDT}(x) + (1 - \alpha) \cdot f_{DNN}(x).$$

$f_{GBDT} \sim$ easy for GBDT

$f_{DNN} \sim$ easy for ResNet

Takeaways

- FT-Transformer is a more universal architecture for Tabular Data
- Further research is needed to understand this phenomenon

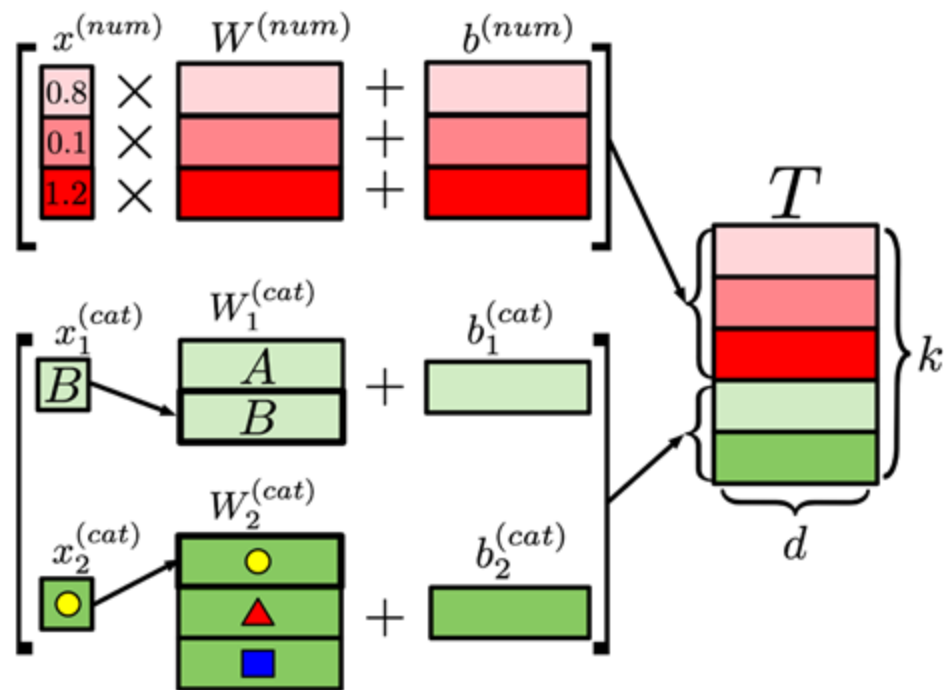
Conclusion

- **MLP and ResNet**
 - fast and strong baselines
- **FT-Transformer**
 - slower
 - can yield even better performance
- FT-Transformer is a more universal architecture for Tabular Data
- Python package with the new models:
`pip install rtdl`
- Source code:
<https://github.com/yandex-research/rtdl>

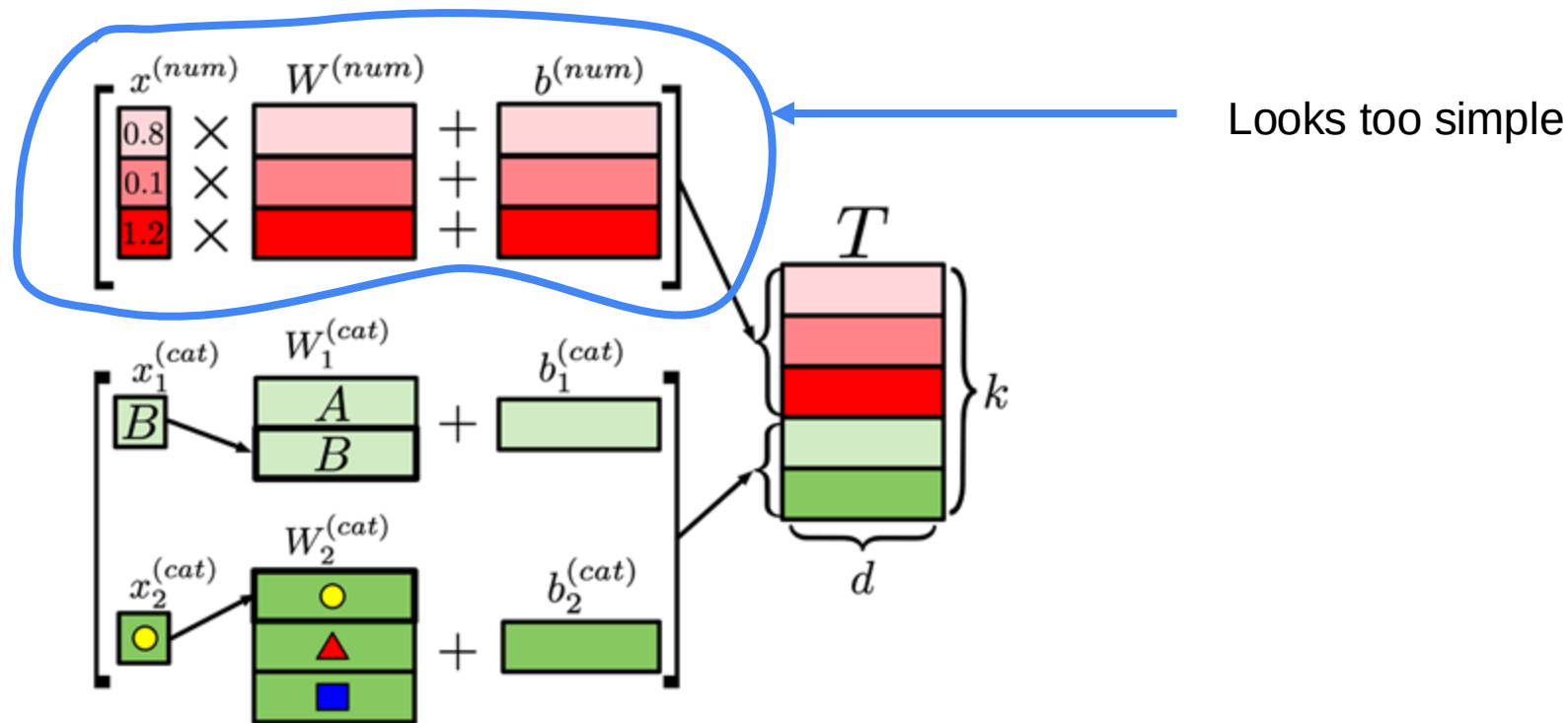
Outline

- Introduction
- The pre-deep learning era of Tabular ML
- Modern Tabular Deep Learning
 - Embeddings for Numerical Features
- Real-world impact

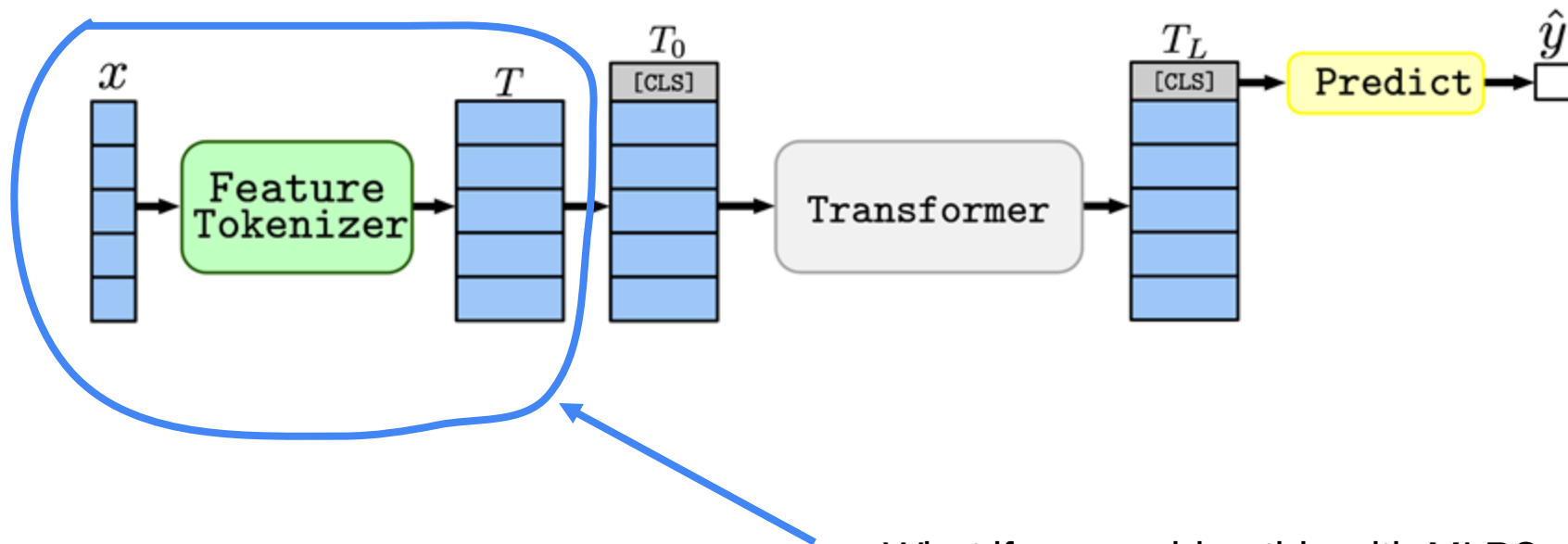
How can we improve FT-Transformer?



How can we improve FT-Transformer?



But wait...



What if we combine this with MLP?

Moreover...

- Transformers perform well
 - The only model with embeddings for **numerical features**

Moreover...

- Transformers perform well
 - The only model with embeddings for **numerical features**
- GBDTs process **numerical features** via thresholds

Moreover...

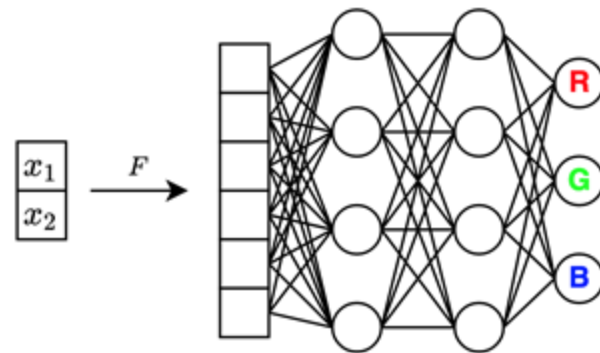
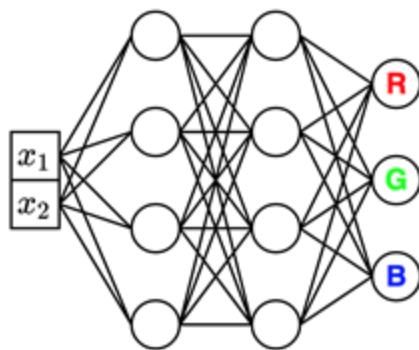
- Transformers perform well
 - The only model with embeddings for **numerical features**
- GBDTs process **numerical features** via thresholds
- MLP is a universal approximator in theory...

Moreover...

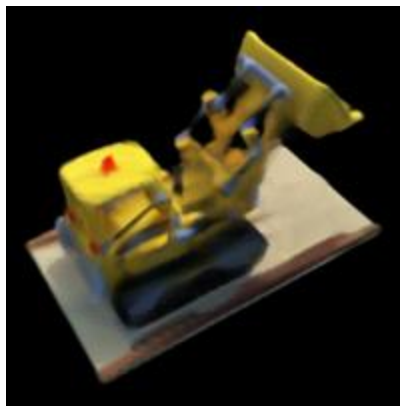
- Transformers perform well
 - The only model with embeddings for **numerical features**
- GBDTs process **numerical features** via thresholds
- MLP is a universal approximator in theory...
- ... but not in practice. Though, **changing the input space can help**
 - “Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains” (Matthew Tancik et al., 2020)
 - “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis” (Ben Mildenhall et al., 2020)

Input representation matters

Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains (Tancik et al., NeurIPS 2020)



The original image



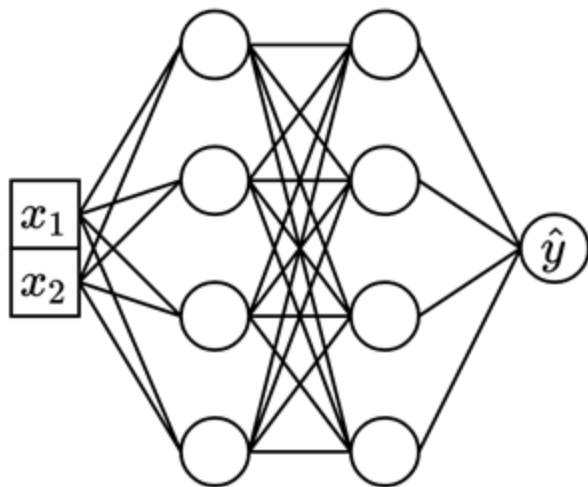
Moreover...

- Transformers perform well
 - The only model with embeddings for **numerical features**
- GBDTs process **numerical features** via thresholds
- MLP is a universal approximator in theory...
- ... but not in practice. Though, **changing the input space can help**
 - “Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains” (Matthew Tancik et al., 2020)
 - “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis” (Ben Mildenhall et al., 2020)
- Little work on **numerical features** processing

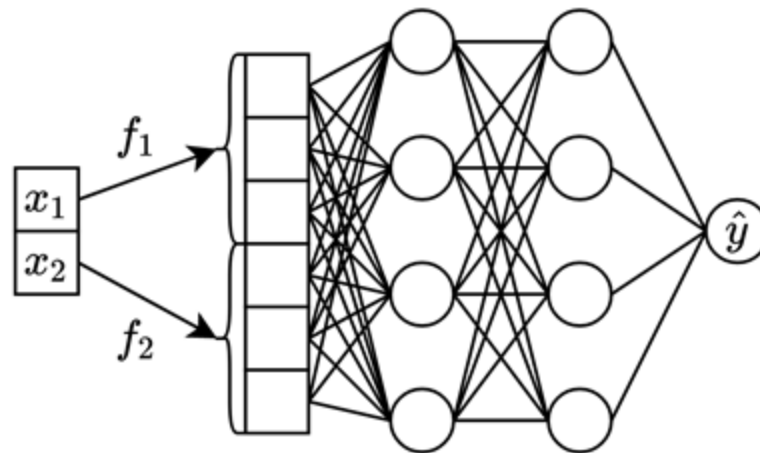
Questions

- Can we improve the way numerical features are processed?
- Can MLP-like models benefit from embeddings for numerical features?

MLP with embeddings

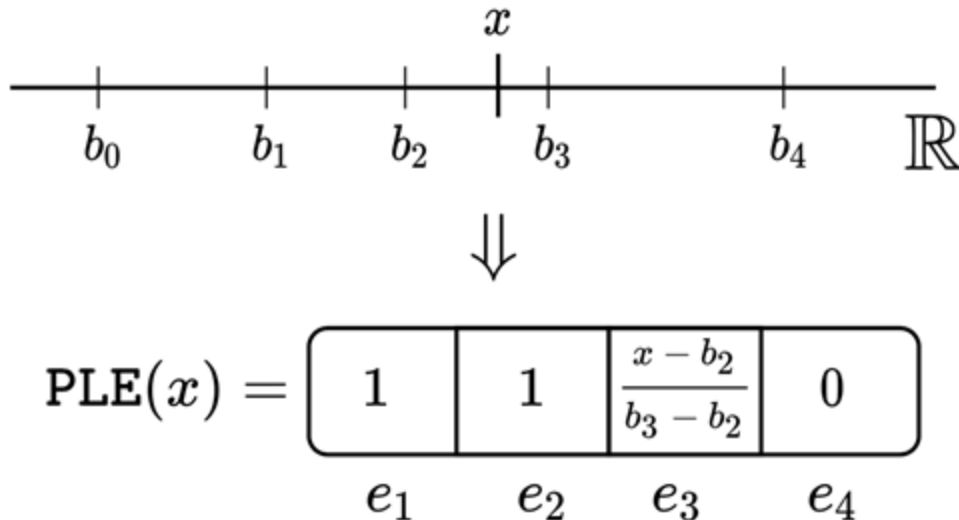


Without embeddings

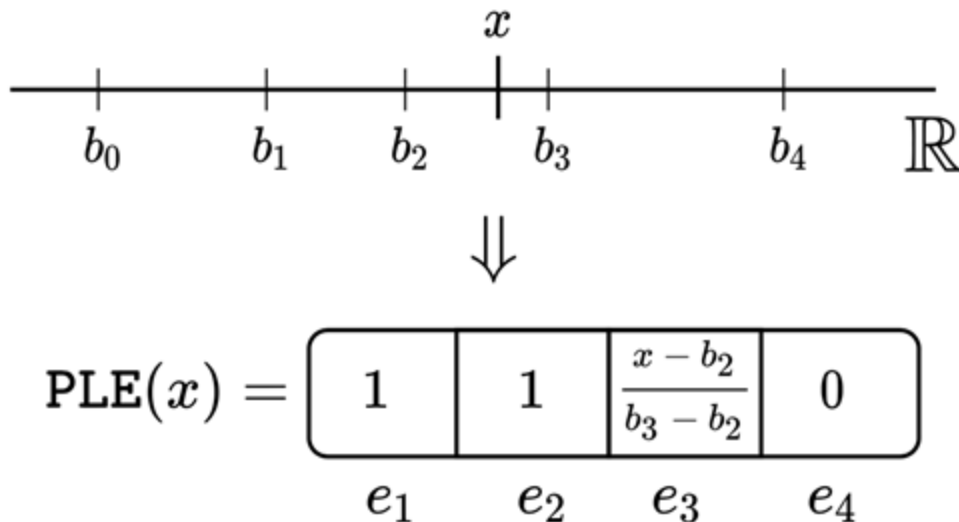


With embeddings

Piecewise-linear encoding



Piecewise-linear encoding



For Transformer-based models:

- v_t - the embedding of the t -th bin

$$f_i(x) = v_0 + \sum_{t=1}^T e_t \cdot v_t = \text{Linear}(\text{PLE}(x))$$

(2022) *On Embeddings for Numerical Features in Tabular Deep Learning*

Piecewise-linear encoding

Quantile binning

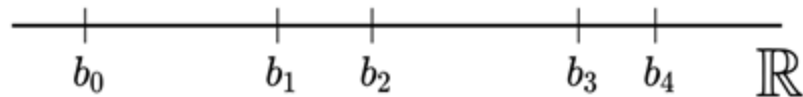
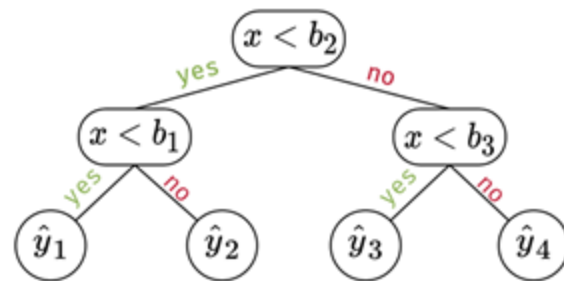
$$b_t = \mathcal{Q}_{\frac{t}{T}} \left(\{x_i^{j(num)}\}_{j \in J_{train}} \right)$$

Piecewise-linear encoding

Quantile binning

$$b_t = Q_{\frac{t}{T}} \left(\{x_i^{j(num)}\}_{j \in J_{train}} \right)$$

Target-aware binning



Periodic activation functions

- (this approach is unrelated to PLE)
- Inspired by the success of periodic functions in other fields

$$f_i(x) = \text{Periodic}(x) = \text{concat}[\sin(v), \cos(v)]$$

$$v = [2\pi c_1 x, \dots, 2\pi c_k x]$$

Other approaches

- Stacking “conventional” layers (linear, ReLU, SoftMax, ...)
- Stacking “conventional” layers on top of PLE or Periodic

Model names

Embedding name	Embedding function f_i	Comment
L	Linear(x)	
LR	ReLU(Linear(x))	
Q-LR	ReLU(Linear(PLE(x)))	quantile-based PLE
T-LR	ReLU(Linear(PLE(x)))	target-based PLE
PLR	ReLU(Linear(Periodic(x)))	The “LR” addition is more important, than for PLE

Model name = <Backbone-Embedding>

Examples:

- Transformer-L (== FT-Transformer)
- MLP-PLR

(2022) On Embeddings for Numerical Features in Tabular Deep Learning

Experiments: datasets and protocol

Dataset	N	K	Metric
Gesture	10K	32	Accuracy (M)
Churn modelling	10K	11	Accuracy (B)
Eye movements	11K	26	Accuracy (M)
California Housing	21K	8	RMSE
House pricing	23K	16	RMSE
Adult income	49K	14	Accuracy (B)
Otto products	62K	93	Accuracy (M)
Higgs (small)	98K	28	Accuracy (B)
FB comments	197K	51	RMSE
Santander	200K	200	Accuracy (M)
Covertypes	581K	54	Accuracy (M)
Microsoft	1201K	136	RMSE

- Tuning
 - mostly Optuna (Akiba et al., 2019) (50-100 iterations)
- Evaluation
 - 15 random seeds
 - ensembles: three ensembles (each consists of five single models)
- No DL tricks
 - no augmentation
 - no lr scheduling
 - no pretraining
 - etc.

N ~ dataset size

K ~ number of features

B ~ binary

M ~ multiclass

Experiments: results

Model	Average rank (std.)
CatBoost	6.8 (4.9)
XGBoost	9.0 (5.7)
MLP	15.6 (2.4)
MLP-LR	10.2 (4.4)
MLP-Q-LR	10.7 (4.6)
MLP-T-LR	10.3 (3.8)
MLP-PLR	4.9 (4.8)
Transformer-L	10.6 (3.3)
Transformer-LR	9.4 (4.1)
Transformer-Q-LR	8.5 (5.5)
Transformer-T-LR	7.2 (4.6)
Transformer-PLR	6.0 (4.5)

- The benchmark is biased towards GBDT-friendly problems
- MLP-LR is consistently better than MLP

Experiments: results

Model	Average rank (std.)
CatBoost	6.8 (4.9)
XGBoost	9.0 (5.7)
MLP	15.6 (2.4)
MLP-LR	10.2 (4.4)
MLP-Q-LR	10.7 (4.6)
MLP-T-LR	10.3 (3.8)
MLP-PLR	4.9 (4.8)
Transformer-L	10.6 (3.3)
Transformer-LR	9.4 (4.1)
Transformer-Q-LR	8.5 (5.5)
Transformer-T-LR	7.2 (4.6)
Transformer-PLR	6.0 (4.5)

- The benchmark is biased towards GBDT-friendly problems
- MLP-LR is consistently better than MLP

Embeddings for numerical features:

- **can provide significant boost**
- **are applicable to MLP-like models**
 - See MLP vs MLP-PLR!
- **allow MLP-like models to compete with Transformer**

Conclusion

- Backbones
 - MLP is a great backbone for researchers and practitioners
 - ResNet may (or may not) provide an extra bit of performance
 - Transformers are competitive, but slow (unclear if it is worth it)
- Embeddings for numerical features
 - can provide significant performance boost
 - Linear + ReLU
 - low risk & low reward
 - Periodic + Linear + ReLU
 - tune sigma: [0.01, 0.02, 0.05, 0.1, 0.5, 1.0, ...]
 - for other hyperparameters, take inspiration from the official repository
 - PLE-based solutions can also provide good performance

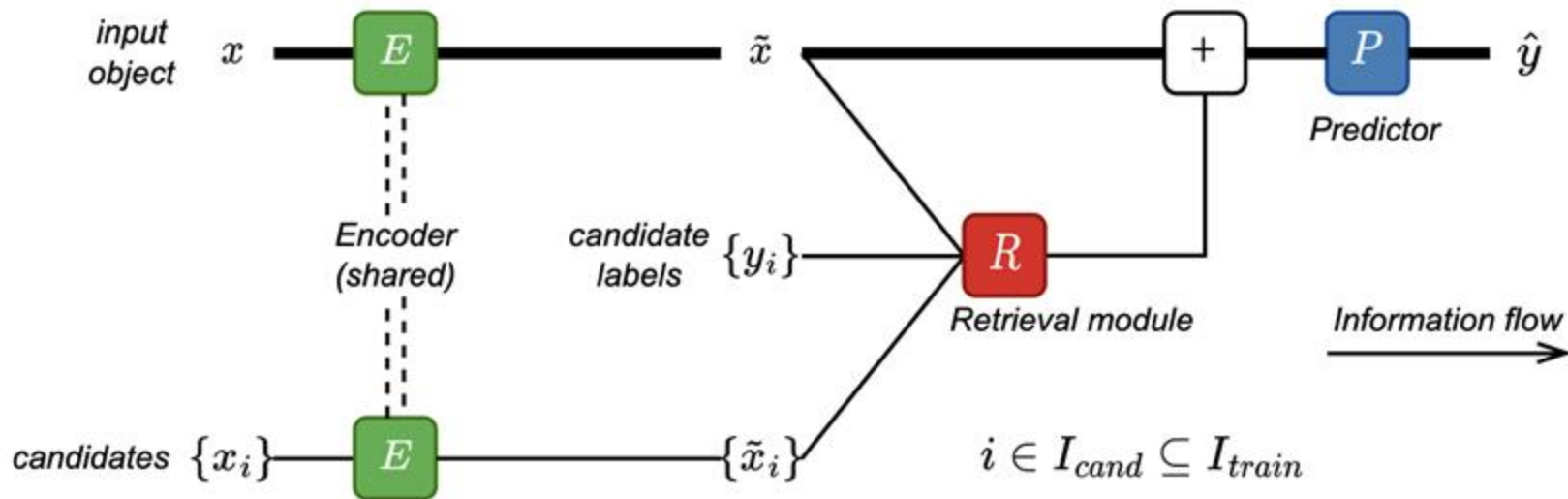
Outline

- Introduction
- The pre-deep learning era of Tabular ML
- Modern Tabular Deep Learning
 - TabR
- Real-world impact

Retrieval-Augmented Learning

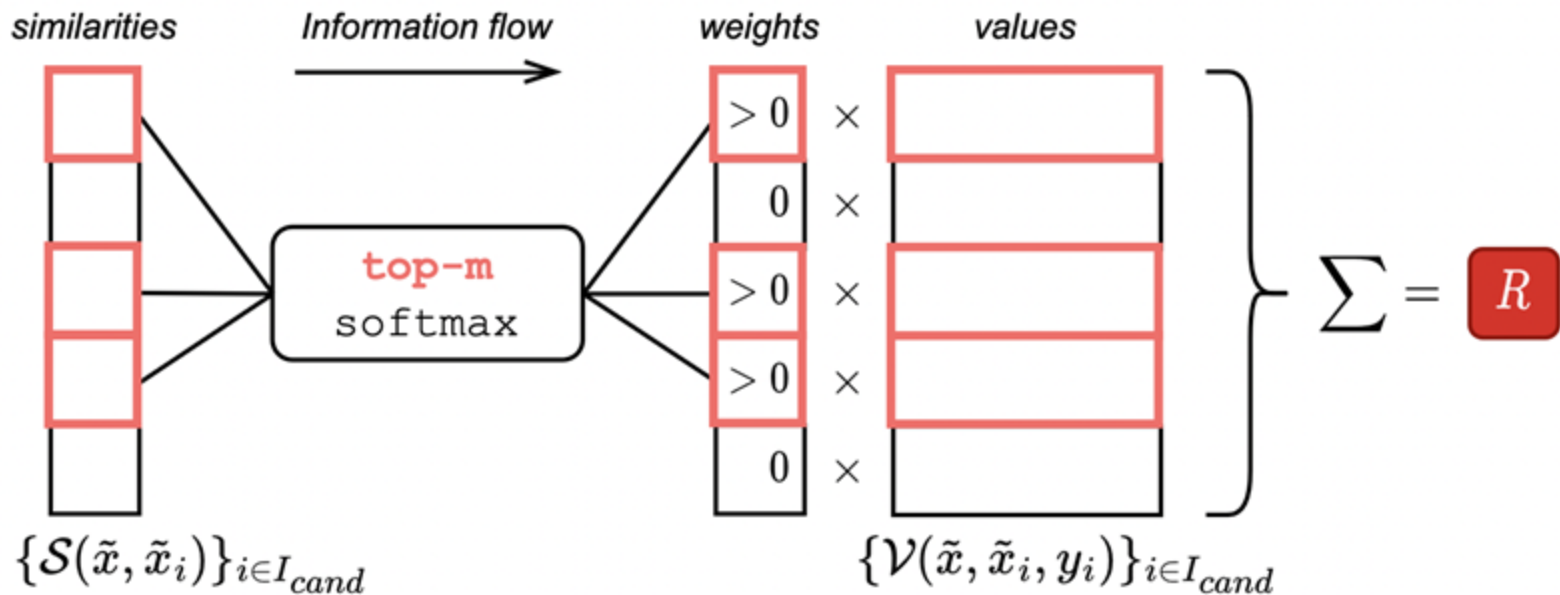
- Is originally motivated by the local learning paradigm (Vapnik et al. 1992)
- Demonstrates success in NLP and computer vision tasks
- Provides higher interpretability and robustness

TabR



TabR: Tabular Deep Learning Meets Nearest Neighbors (ICLR 2024)

TabR



Technical insights

The retrieval module R

- Linear complexity w.r.t. the number of candidates
- The inter-object communication happens only once

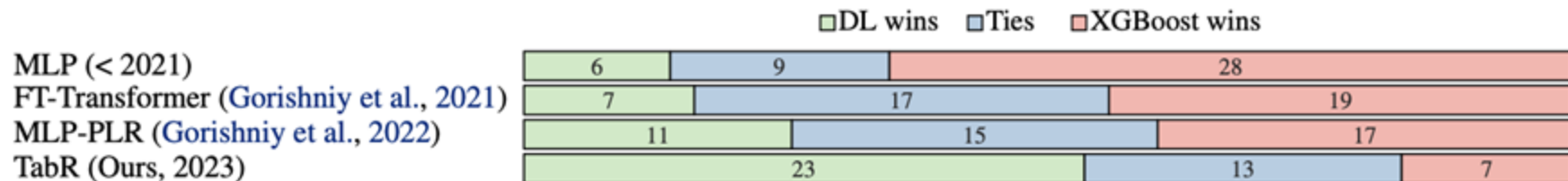
The similarity module S

- By default, the L2 distance is recommended (important!)

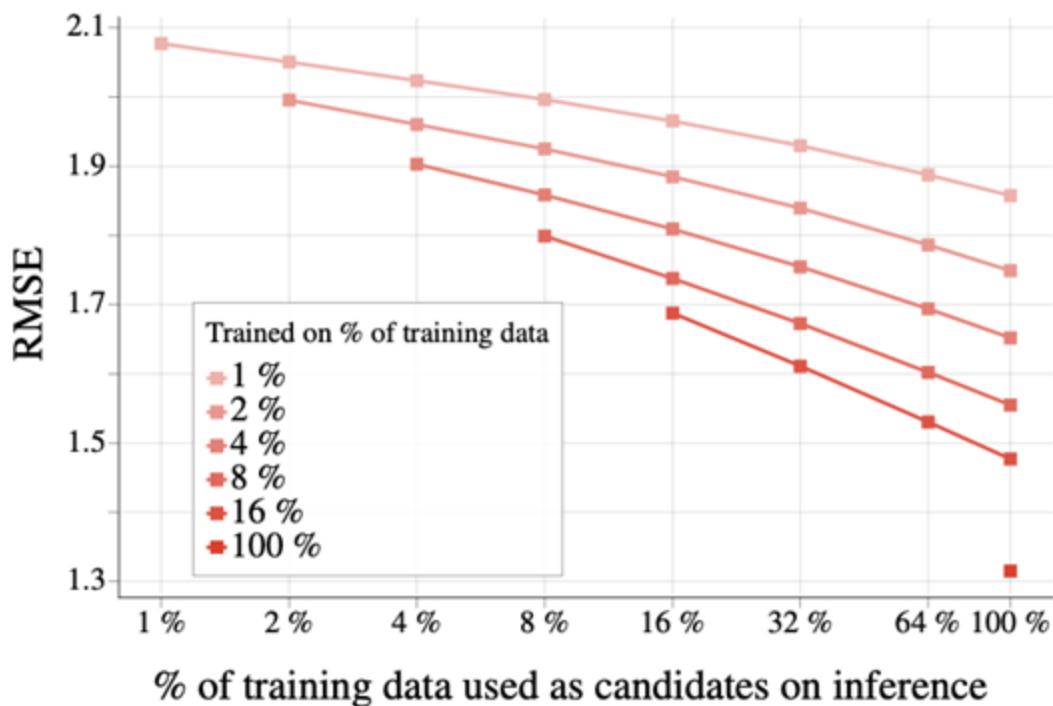
The value module V

- Can depend on objects and their interactions

TabR results



Training on a subset of data



Limitations

- Reminder: simple ML models suffer from distributions shifts in features and/or labels of individual objects.
- Retrieval-based models also suffer from distribution shifts in interactions between objects.
- To prevent such problems, one has to think how to configure the retrieval behavior in each individual use case.

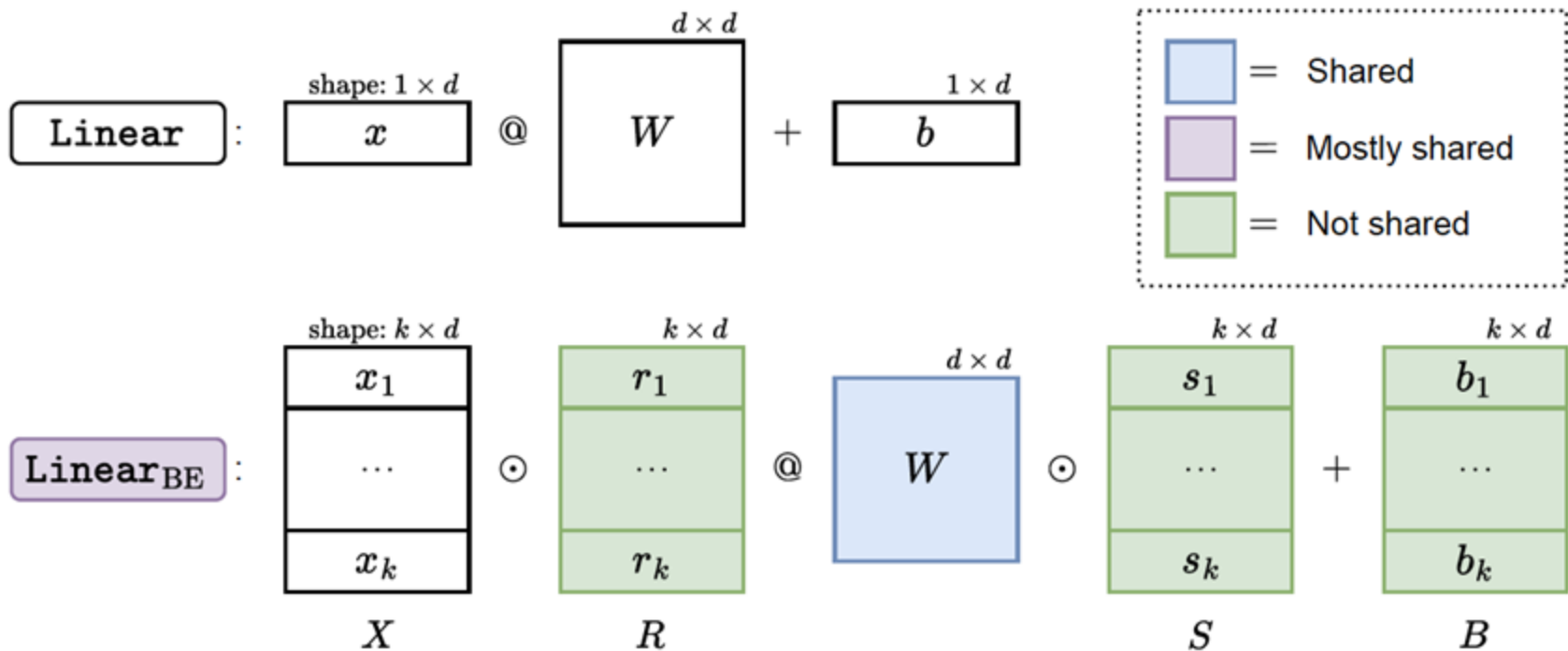
Outline

- Introduction
- The pre-deep learning era of Tabular ML
- Modern Tabular Deep Learning
 - TabM
- Real-world impact

Ensembles of Models in Machine Learning

- Main idea: train several models and combine predictions from them
- GBDT are essentially an ensemble
- Go-to recipe in DL: train several *independent* models and average the predictions
 - Can be used for any model
 - Often improves accuracy
 - Higher memory and runtime costs

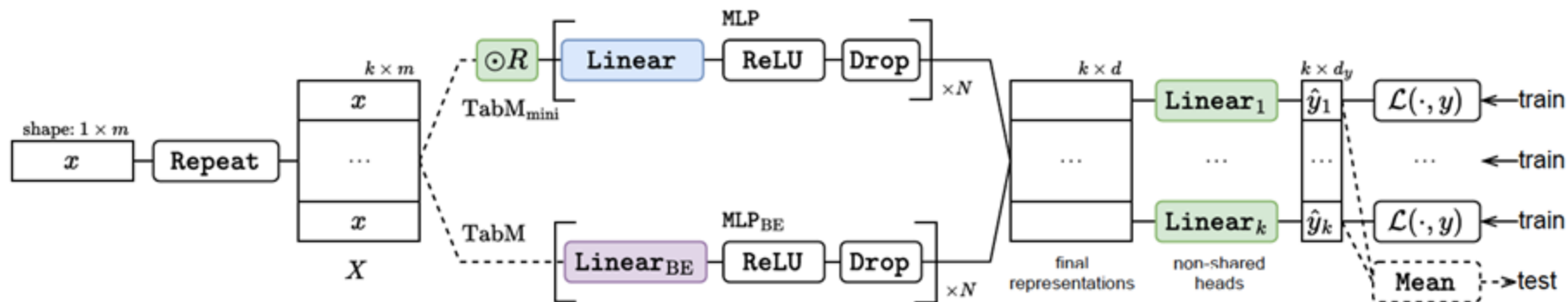
BatchEnsemble (Wen et al., 2020): main idea



R, S, B - *adapters*

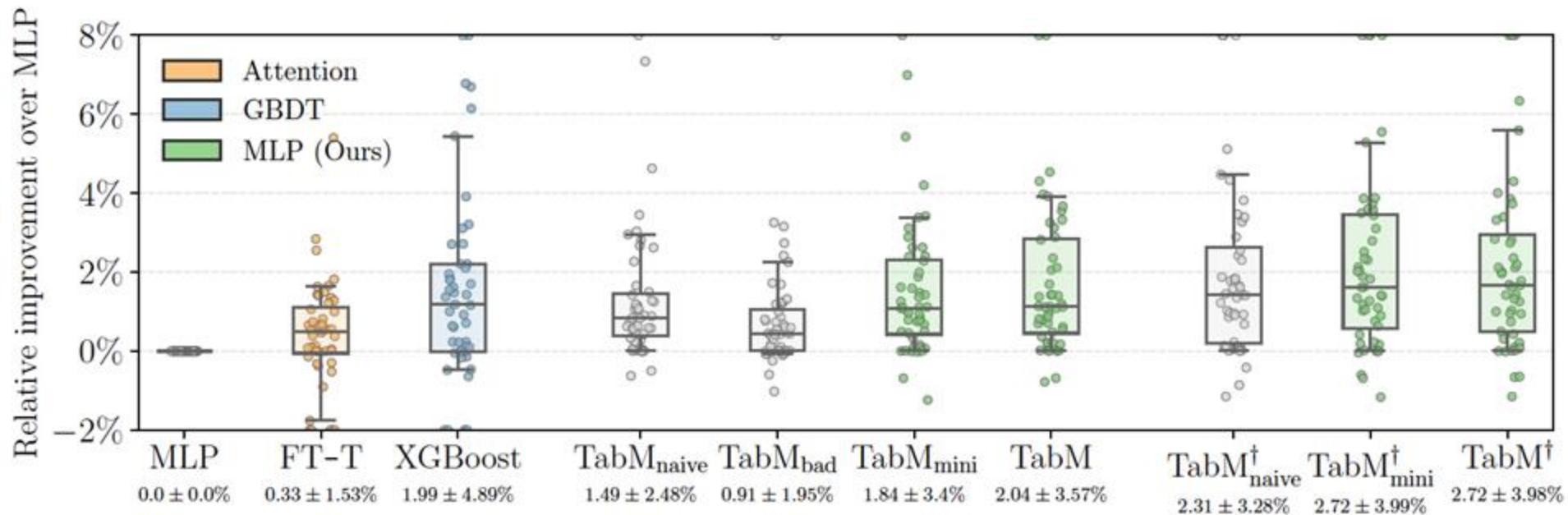
Since $k \ll d$, runtime and memory overhead are tolerable!

TabM: BatchEnsemble meets Tabular DL

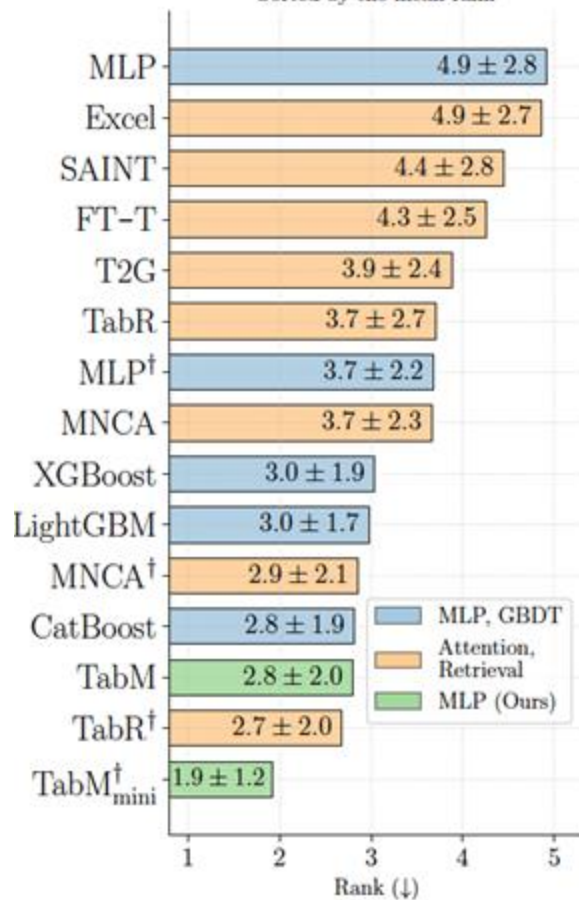


- TabM with $k = 1$ is equivalent to MLP
- Specific initialization of adapters is needed
- Can be combined with non-linear feature embeddings

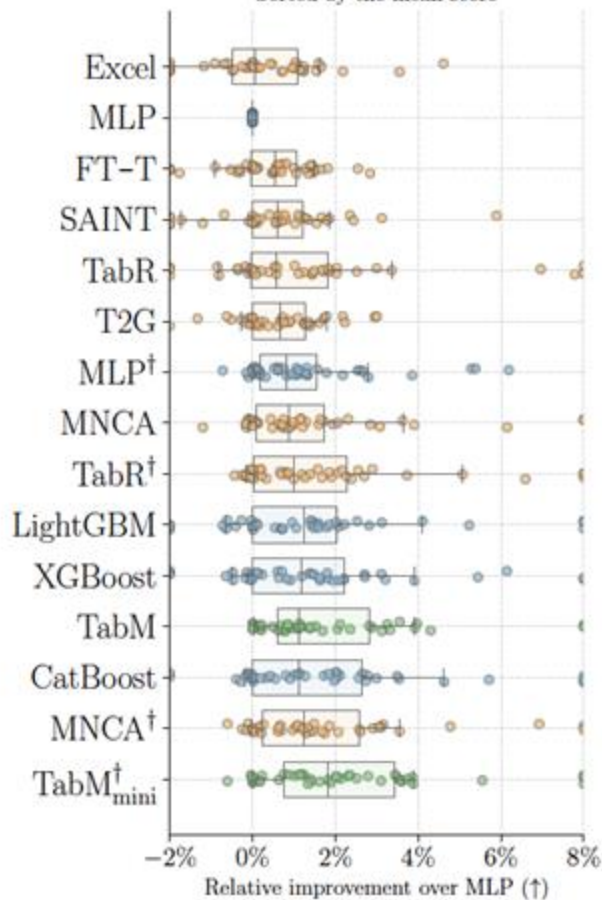
TabM: results



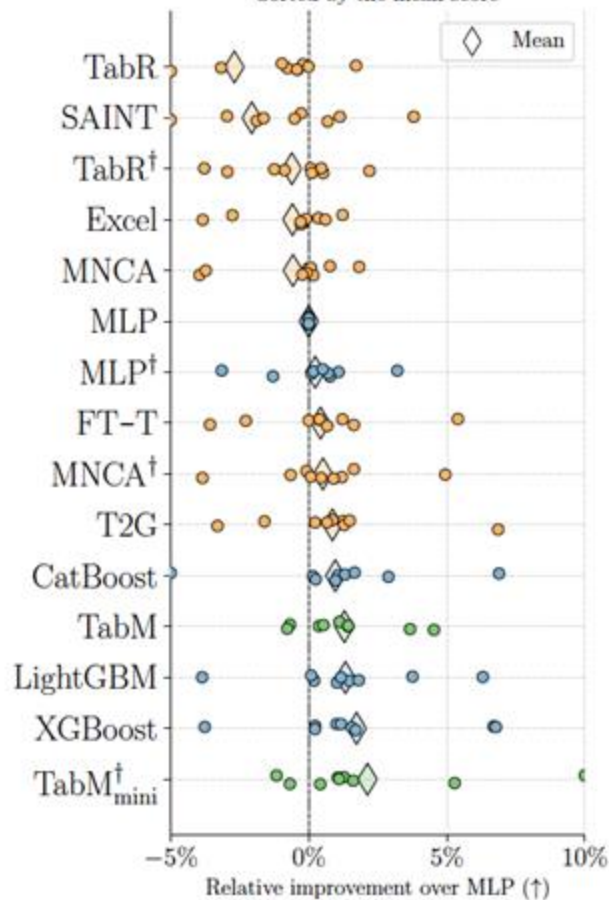
Performance ranks with std. dev.
On all datasets
Sorted by the mean rank



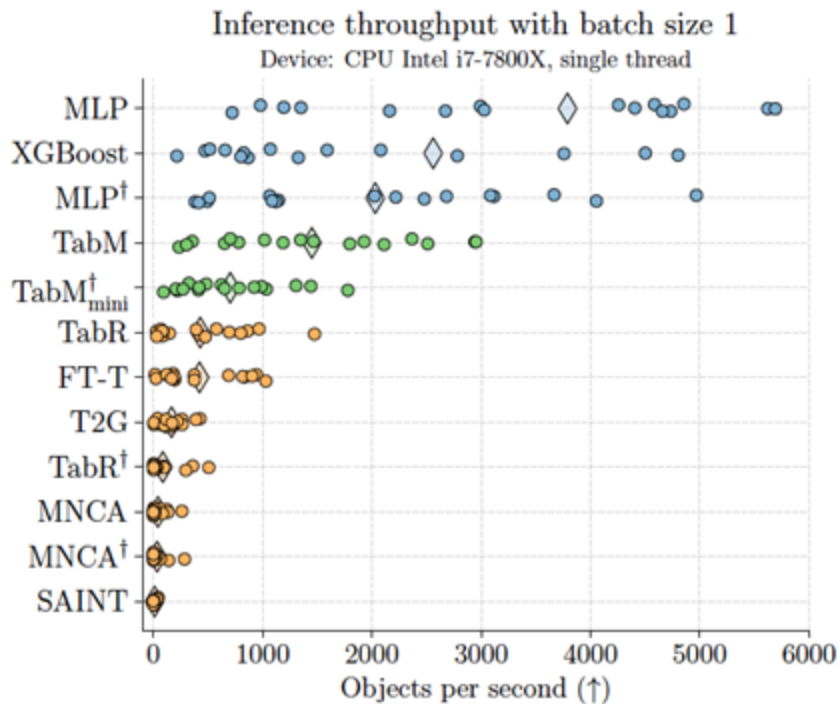
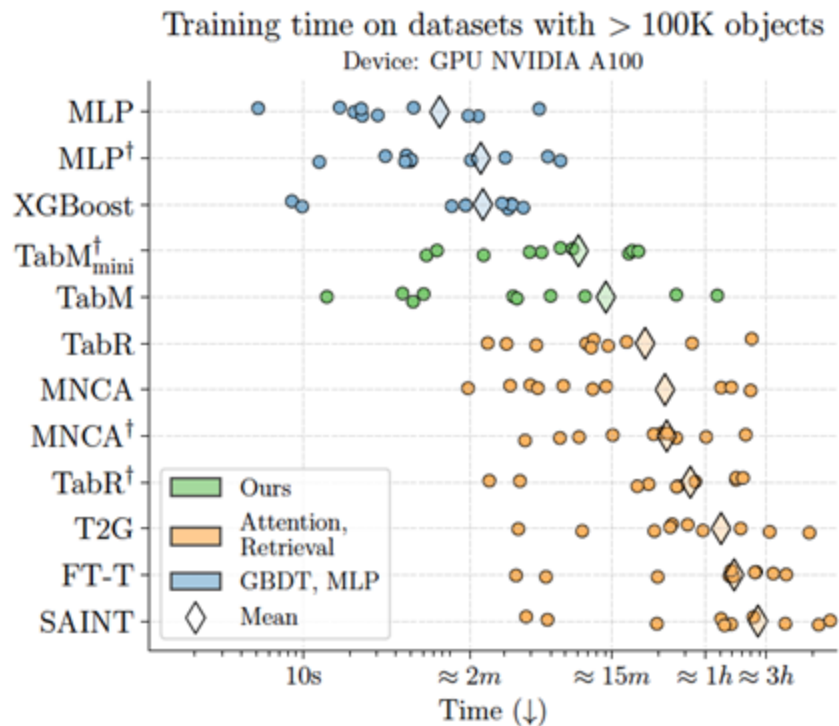
Performance scores
On 41 datasets with random split
Sorted by the mean score



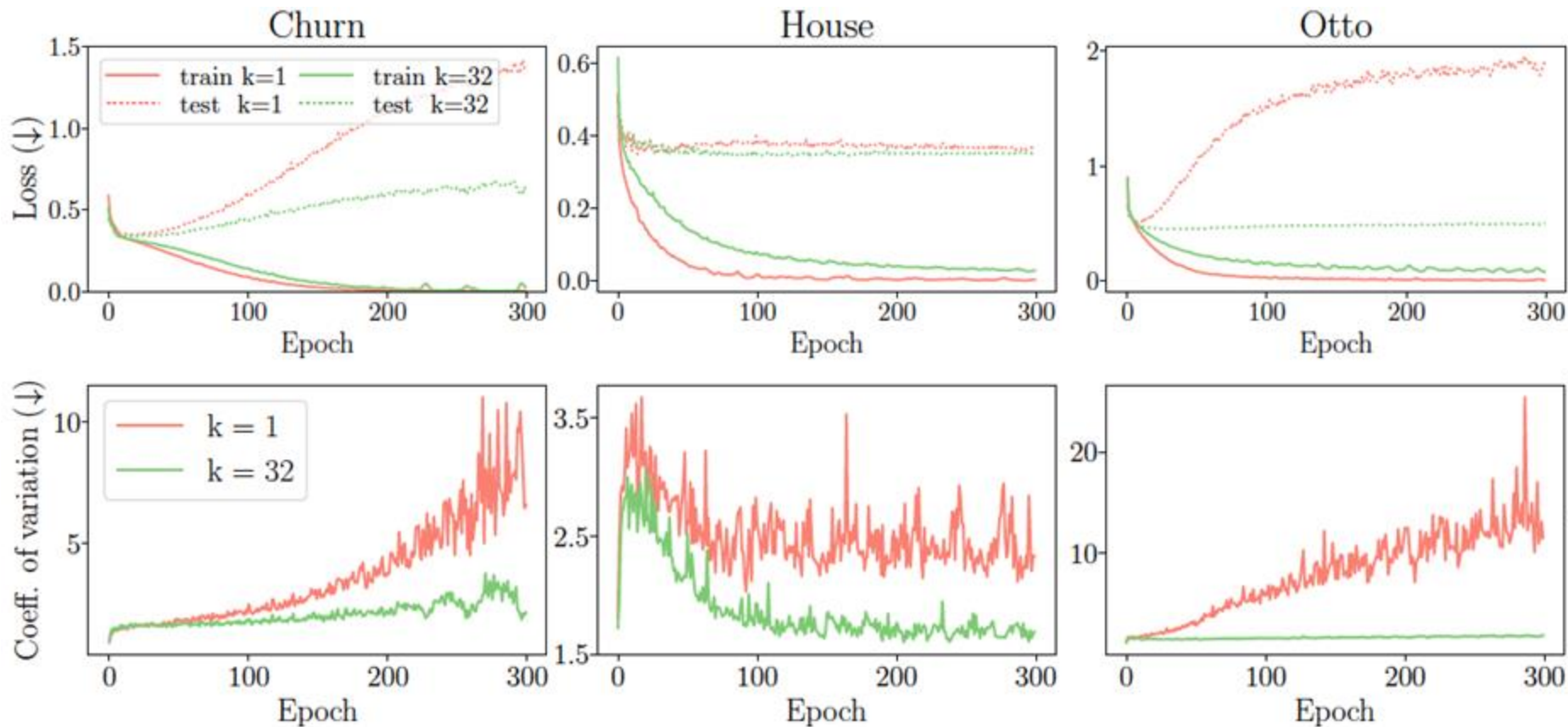
Performance scores
On 9 datasets with domain-aware split
Sorted by the mean score



Efficiency



Optimization properties of TabM



Conclusion

- TabM with non-linear feature embeddings are currently the state-of-the-art
- TabM typically outperforms GBDT on existing benchmarks
- TabM exhibits stable optimization and less overfitting

Outline

- Introduction
- The pre-deep learning era of Tabular ML
- Modern Tabular Deep Learning
- Real-world impact

Tabular DL in our lives

DeepETA: How Uber Predicts Arrival Times Using Deep Learning

February 10, 2022 / Global

Rides



Freight



Eats



How we built it: Stripe Radar

Our most recent architecture evolution occurred in mid-2022 when we migrated from an ensemble “Wide & Deep model,” composed of an XGBoost model and a deep neural network (DNN), to a pure DNN-only model. The result was a model that trains faster, scales better, and is more adaptable to the most cutting-edge ML techniques.

В ННГУ усовершенствовали нейросеть для диагностики скорости старения

Ученые Университета Лобачевского усовершенствовали нейросеть для диагностики скорости старения. Новая модель иммунологических часов получила название SImAge (Small Immuno Age). Она построена на основе глубокой нейронной сети FT-Transformer. Нейросеть оценивает состояние организма по 10 биомаркерам, которые отражают ...

Challenging Gradient Boosted Decision Trees with Tabular Transformers for Fraud Detection at Booking.com

Conclusion

- Tabular DL is extremely impactful research field with many unresolved questions
- New models are being developed and the progress has not converged
- GBDTs are still in wide use but their primacy has been challenged
- Tomorrow: Advanced topics in Tabular DL

Questions?

Advanced Topics in Tabular Deep Learning

Lecturer: Artem Babenko



ASCOMP 2024

Outline

- Quick recap
- Tabular Benchmarks
- Pretraining in Tabular DL
- Cross-domain learning
- Generative tabular models
- Future directions

Outline

- Quick recap
- Tabular Benchmarks
- Pretraining in Tabular DL
- Cross-domain learning
- Generative tabular models
- Future directions

Recap from yesterday

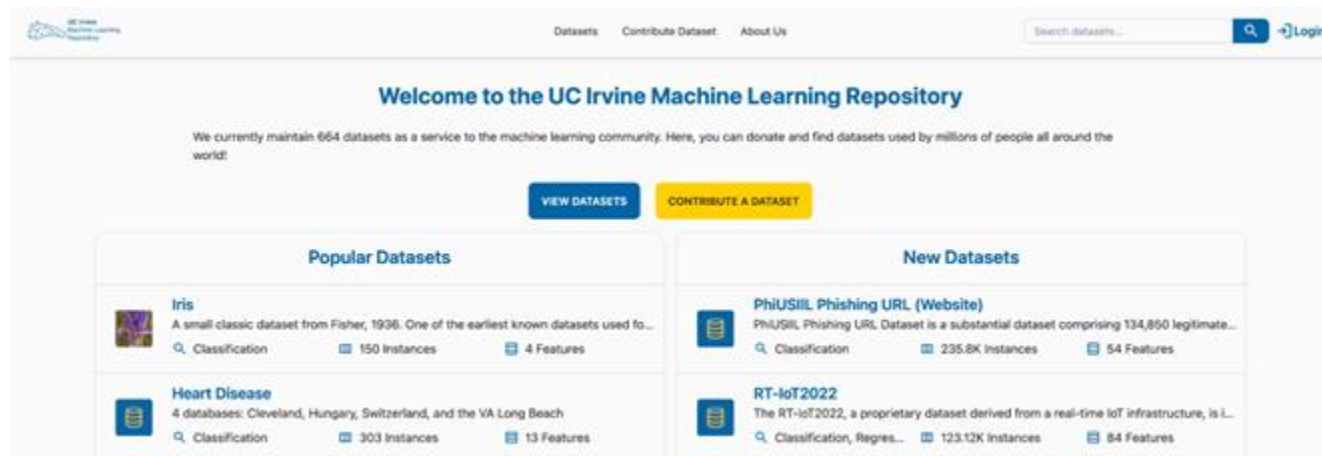
- Tabular problems are everywhere
- “Shallow” GBDT models are still a popular choice
- Tabular DL architectures are actively developed
- Are new DL architectures the only research direction?
 - No!

Outline

- Quick recap
- **Tabular Benchmarks**
- Pretraining in Tabular DL
- Cross-domain learning
- Generative tabular models
- Future directions

Where do tabular DL researchers get datasets?

- openml.org
- archive.ics.uci.edu
- kaggle.com/datasets
- `from sklearn.datasets import *`
- Do we care to examine those 10-20-100 datasets? - Rarely!



Let's Look at the Academic Benchmarks

	A	B	C	D	E	F	G	H	I	J	K
1	Dataset	#Samples	#Features	Citation	Not a Real Word Task	Time split impossible, or not used (if needed)	Leak	Not Tabular	Synthetic Unknown origin	GOV Records	Questions
2	lymph	148	19	https://www.openml.org/search?type=data&status=active&id=10	1	0	0	0	0	0	
3	qsar-biodeg	155	42	https://www.openml.org/search?type=data&status=active&id=1494	0	0	0	1	0	0	
4	audiology	226	70	https://archive.ics.uci.edu/dataset/3/audiology+standardized	1	0	0	0	0	0	
5	heart-h	294	14	https://openml.org/search?type=data&status=active&id=51	0	0	0	0	0	0	
6	colic	368	27	https://www.openml.org/search?type=data&status=active&id=25	1	0	0	0	1	0	
7	monks-problems-2	601	7	https://www.openml.org/search?type=data&status=active&id=334	1	0	0	0	1	0	
8	balance-scale	625	5	http://archive.ics.uci.edu/dataset/12/balance+scale	1	0	0	0	1	0	
9	prob	672	10	https://www.openml.org/search?type=data&status=active&id=470	1	0	0	0	0	0	
10	Australian	690	15	https://archive.ics.uci.edu/dataset/143/statlog+australian+credit+approval	0	1	0	0	0	0	
11	credit-approval	690	16	https://archive.ics.uci.edu/dataset/27/credit+approval	0	1	0	0	1	0	
12	vehicle	846	19	https://www.openml.org/search?type=data&status=active&id=54	0	0	0	1	1	0	
13	cnae-9	1080	857	https://www.openml.org/search?type=data&status=active&id=1468	0	0	0	1	0	0	
14	socmob	1156	6	https://www.openml.org/search?type=data&status=active&id=44987	0	1	0	0	1	1	
15	100-plants-texture	1599	65	https://archive.ics.uci.edu/dataset/241/one+hundred+plant+species+leaves+data+set	1	0	0	1	0	0	

What did we find?

Problems:

Data Leakage (10 datasets). data-leaks stemming from data preparation errors, or inappropriate data splits being used in papers using the datasets.

No time data available (most datasets). either represent a fixed snapshot of some real-world phenomena, or don't have a way to construct a time-based validation/test sets

Dataset Duplication (California Housing, House 16H, house_sales, kdd_ipums_la_97-small, houses) - all datasets are from 1990 census data

Dataset Size. 19/100 less than 10k samples.

Synthetic data. (or from an unknown source). datasets for which the original data source is untraceable.

Not Tabular. datasets where underlying data is not tabular like images, audio, text or graphs

>50%

Datasets don't handle time properly

38%

"Problematic" Datasets

~20

Features available

<1kk

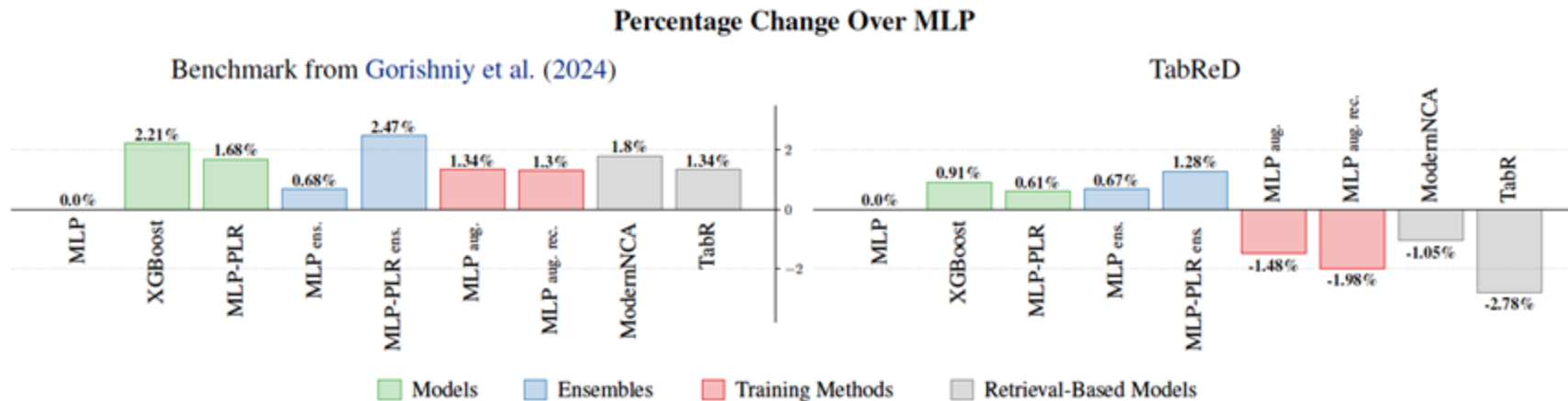
Small sample sizes
Majority is below 100k samples

TabRed: focus on temporal-shift based evaluation

Benchmark	Dataset Sizes (Q ₅₀)		Issues (#Issues / #Datasets)			Time-split		
	#Samples	#Features	Data-Leakage	Synthetic or Untraceable	Non-Tabular	Needed	Possible	Used
Grinsztajn et al. [22]	16,679	13	7 / 44	1 / 44	7 / 44	22	5	
Tabzilla [40]	3,087	23	3 / 36	6 / 36	12 / 36	12	0	
WildTab [35]	546,543	10	1* / 3	1 / 3	0 / 3	1	1	✗
TableShift [18]	840,582	23	0 / 15	0 / 15	0 / 15	15	8	
Gorishniy et al. [21]	57,909	20	1* / 10	1 / 10	0 / 10	7	1	
TabRed (ours)	7,163,150	261	✗	✗	✗	✓	✓	✓

Methods	Classification (ROC AUC $\uparrow$)			Regression (RMSE $\downarrow$)					Average Rank
	Homesite Insurance	Ecom Offers	HomeCredit Default	Sberbank Housing	Cooking Time	Delivery ETA	Maps Routing	Weather	
Classical ML Baselines									
XGBoost	0.9601	0.5763	0.8670	<u>0.2419</u>	0.4823	<u>0.5468</u>	<u>0.1616</u>	<u>1.4671</u>	2.6 $\pm$ 1.2
LightGBM	0.9603	0.5758	<u>0.8664</u>	0.2468	0.4826	<u>0.5468</u>	0.1618	1.4625	2.9 $\pm$ 1.2
CatBoost	0.9606	0.5596	0.8621	0.2482	0.4823	0.5465	0.1619	<u>1.4688</u>	3.1 $\pm$ 1.4
RandomForest	0.9570	0.5764	0.8269	0.2640	0.4884	0.5959	0.1653	1.5838	7.1 $\pm$ 2.0
Linear	0.9290	0.5665	0.8168	0.2509	0.4882	0.5579	0.1709	1.7679	8.1 $\pm$ 2.5
Tabular DL Models									
MLP	0.9500	<u>0.6015</u>	0.8545	0.2508	0.4820	0.5504	0.1622	1.5470	4.8 $\pm$ 1.7
SNN	0.9492	0.5996	0.8551	0.2858	0.4838	0.5544	0.1651	1.5649	6.4 $\pm$ 1.9
DCNv2	0.9392	0.5955	0.8466	0.2770	0.4842	0.5532	0.1672	1.5782	7.4 $\pm$ 2.3
ResNet	0.9469	0.5998	0.8493	0.2743	0.4825	0.5527	0.1625	1.5021	5.5 $\pm$ 2.1
FT-Transformer	<u>0.9622</u>	0.5775	0.8571	0.2440	0.4820	0.5542	0.1625	1.5104	4.4 $\pm$ 1.4
MLP-PLR	<u>0.9621</u>	0.5957	0.8568	0.2438	<u>0.4812</u>	0.5527	<u>0.1616</u>	1.5177	3.6 $\pm$ 1.5
Trompt	0.9546	0.5792	0.8381	0.2596	<u>0.4834</u>	0.5563	0.1652	1.5722	6.8 $\pm$ 2.0
Retrieval Augmented Tabular DL									
TabR-S	0.9487	0.5943	0.8501	0.2820	0.4828	0.5514	0.1639	<u>1.4666</u>	5.8 $\pm$ 2.2
ModernNCA	0.9514	0.5765	0.8531	0.2593	0.4825	0.5498	0.1625	1.5062	5.0 $\pm$ 1.3

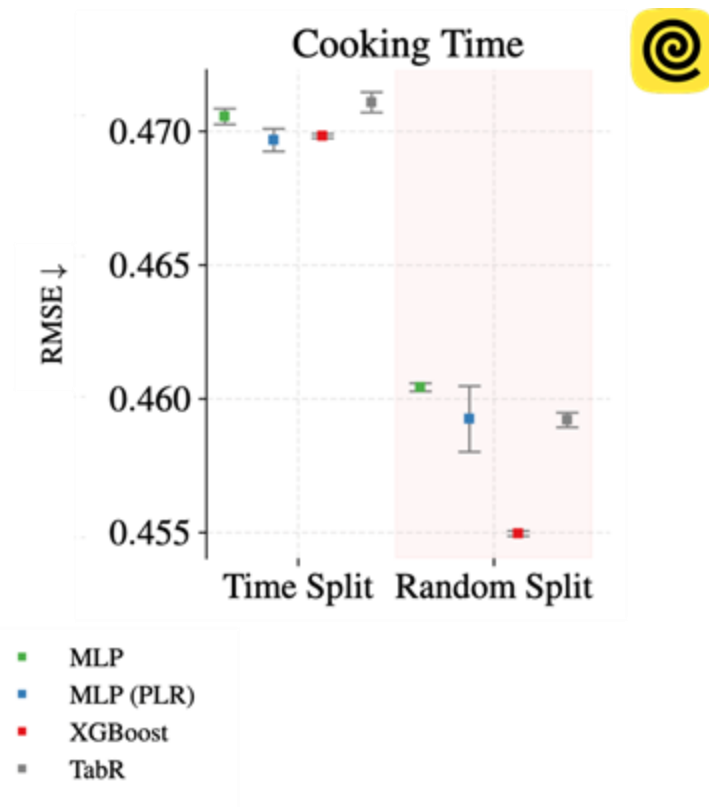
Findings on TabReD



- Performance differences are less pronounced (feature engineering)
- Non-linear feature embeddings and ensembles are helpful
- FT-Transformer is not justified
- Retrieval-augmented models are generally less performant

Temporal shift

- GBDTs are less robust to temporal shift
- Realistic evaluation setups are important for healthy progress



Summary

- A new benchmark with datasets, closer resembling real-world scenarios
- Sources: Kaggle and Yandex Eats, Maps, Weather, Lavka
- Datasets with 10M samples and feature-engineering (*with up-to 1000s of features*)
- All datasets have timestamps

Outline

- Quick recap
- Tabular Benchmarks
- **Pretraining in Tabular DL**
- Cross-domain learning
- Generative tabular models
- Future directions

Pretraining in DL: main idea

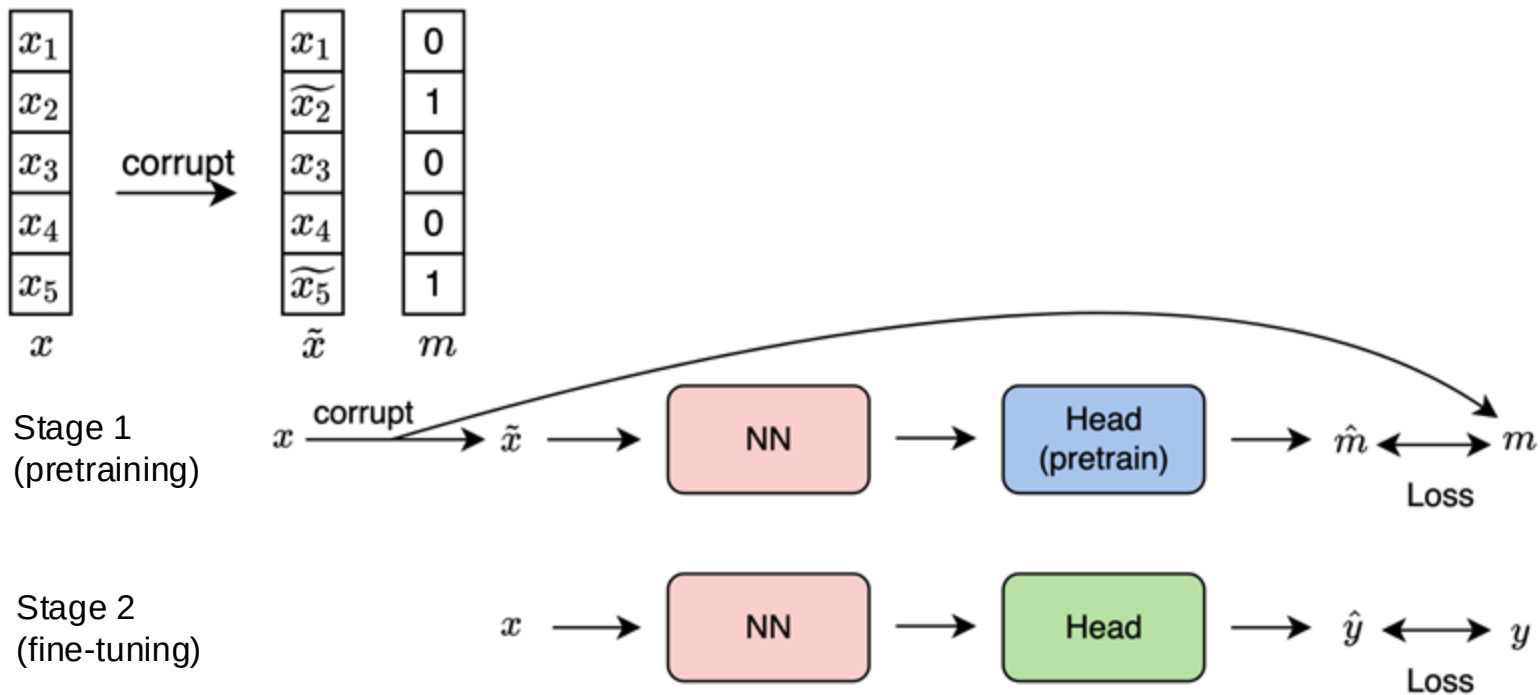
- To train the model to solve a related problem before the main learning process
 - Same data but different tasks (e.g. with cheaper labels)
 - "Extra" data from the same or a similar domain
- Inner logic of the pretrained model can be helpful for the target problem
- Provides better than random initialization for subsequent gradient optimization
- De facto standard for typical pipelines in NLP and CV
 - Contrastive learning
 - Self-prediction

Pretraining in Tabular DL: specifics

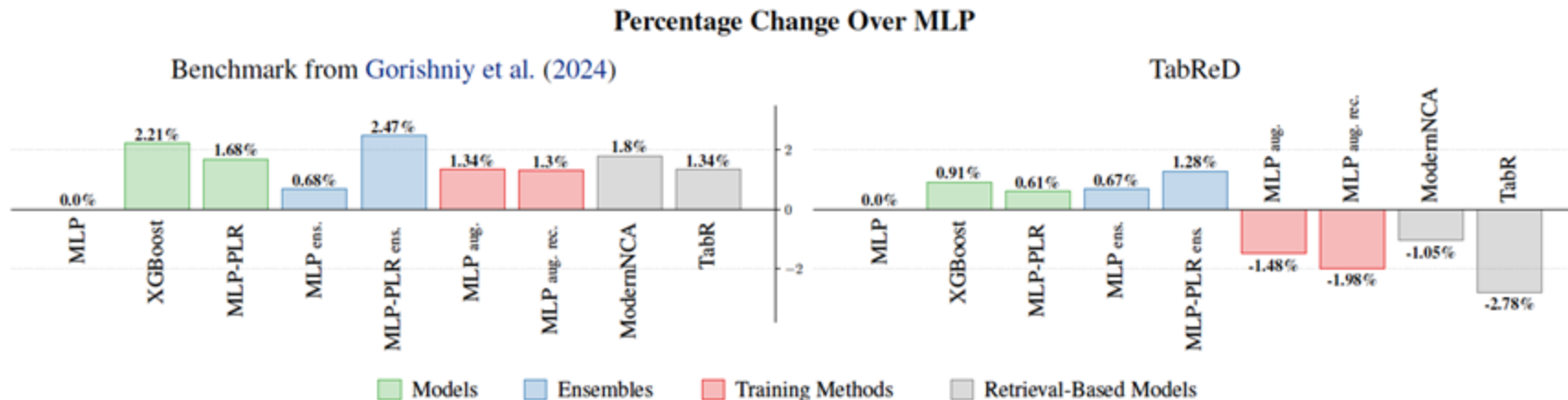
- No "extra" data
 - Need to pretrain on the main train set
- Lack of "valid" data augmentations
 - Any augmentation can TODO the data distribution
 - Pretraining can be harmful
- Problems from a large number of domains
 - Need of the universal pretraining recipe

Unsupervised pretraining for tabular data

Mask prediction

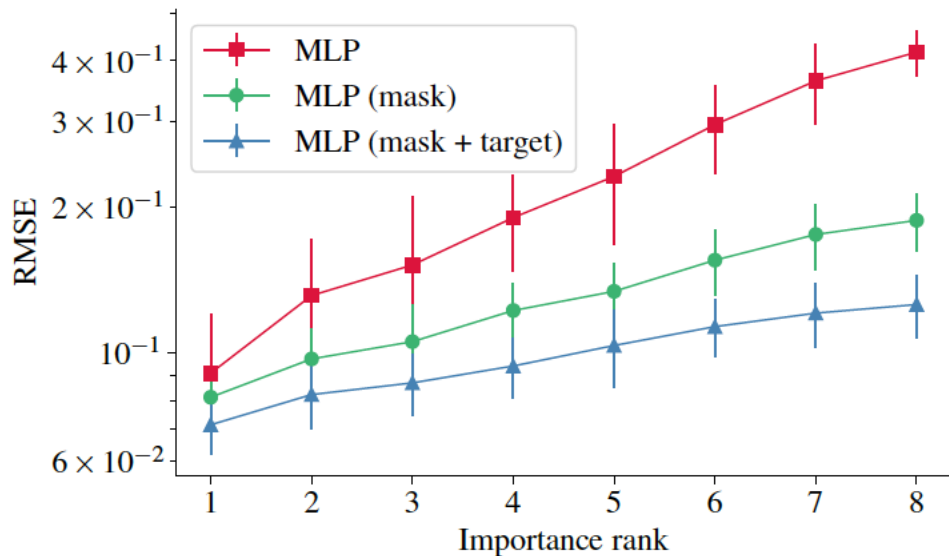


Experiments with pretraining



- All pretraining strategies perform on par to each other
- Pretraining is beneficial for both simple and advanced tabular DL models
- In temporal-shift based evaluation, pretraining can be harmful

When and why pretraining helps?



- An experiment on synthetic data with controllable feature importances
- For different models, we measure the reconstruction quality of different features from the inner model representations
- Pretrained models capture less important (but still significant!) features better

Conclusion

- Pretraining does have some potential in Tabular DL
- The choice of pretraining objective does not matter much
- The pretraining effect depends on the distribution shift between train and test
 - Effect is often negative when the shift is noticeable
 - The universal pretraining recipe is yet to discover

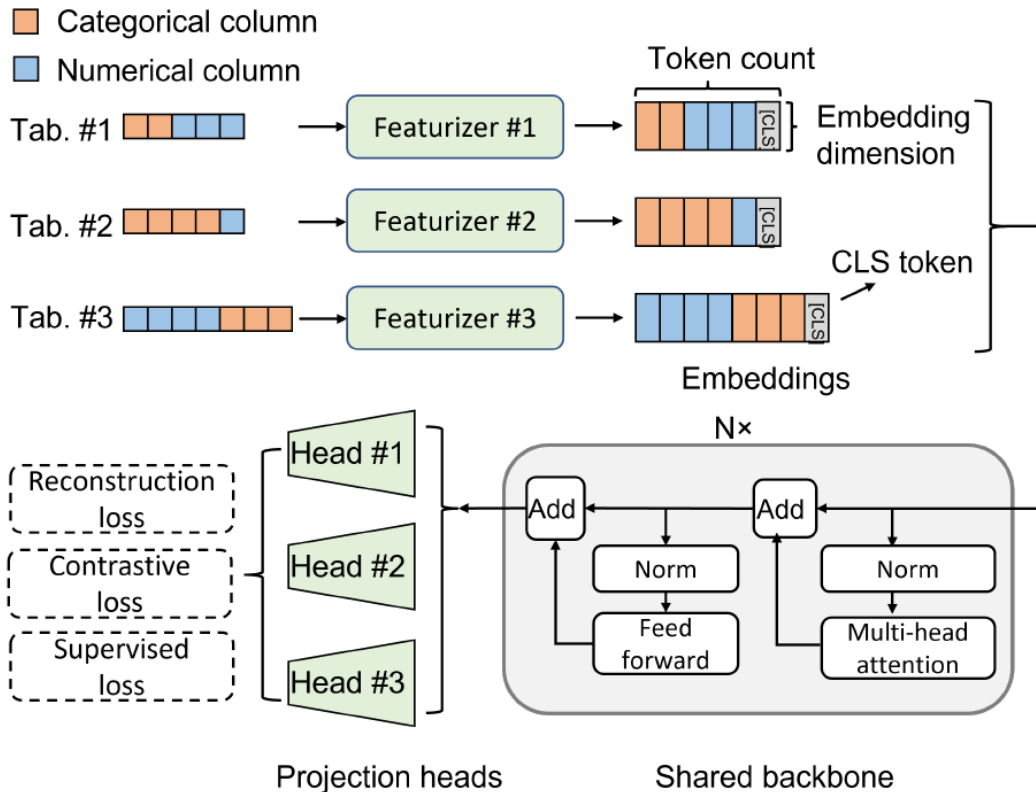
Outline

- Quick recap
- Tabular Benchmarks
- Pretraining in Tabular DL
- Cross-domain learning
- Generative tabular models
- Future directions

Main idea of cross-domain Tabular DL

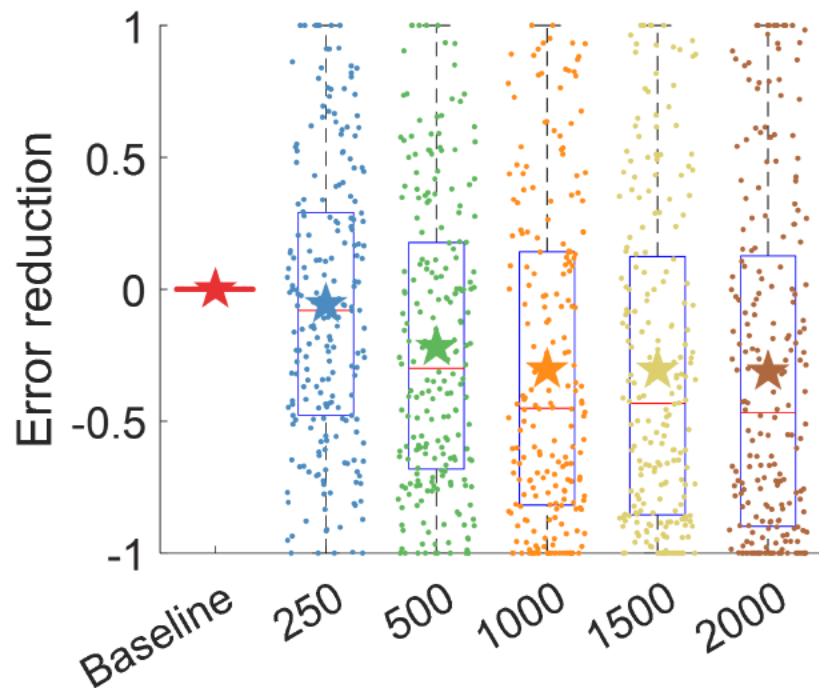
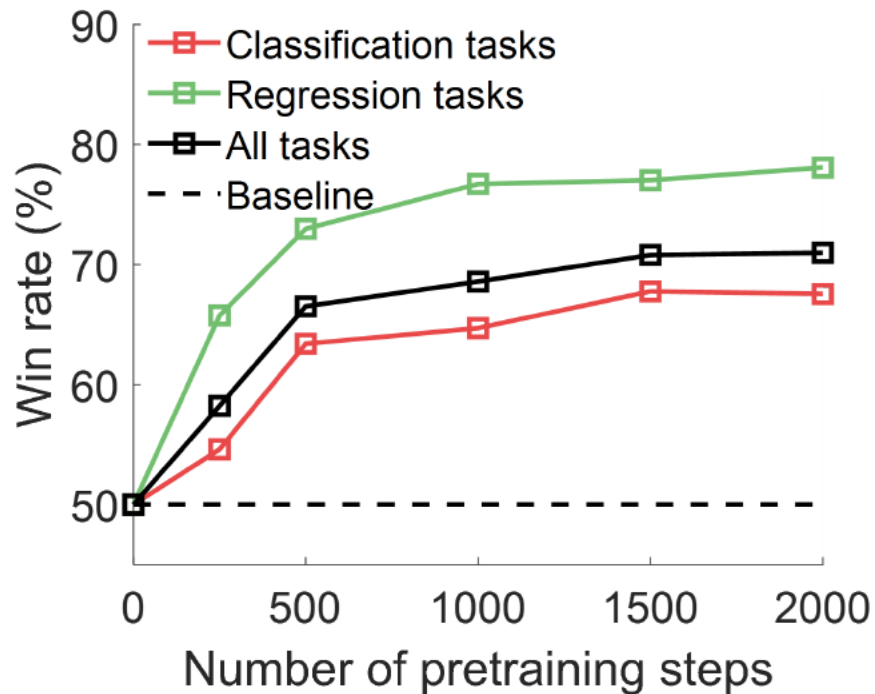
- Leverage knowledge from one domain to improve predictions in another one
- Sounds like magic for tabular DL but ...
- Sometimes does make sense (and even works)

XTAB (Zhu et al., ICML'2023)

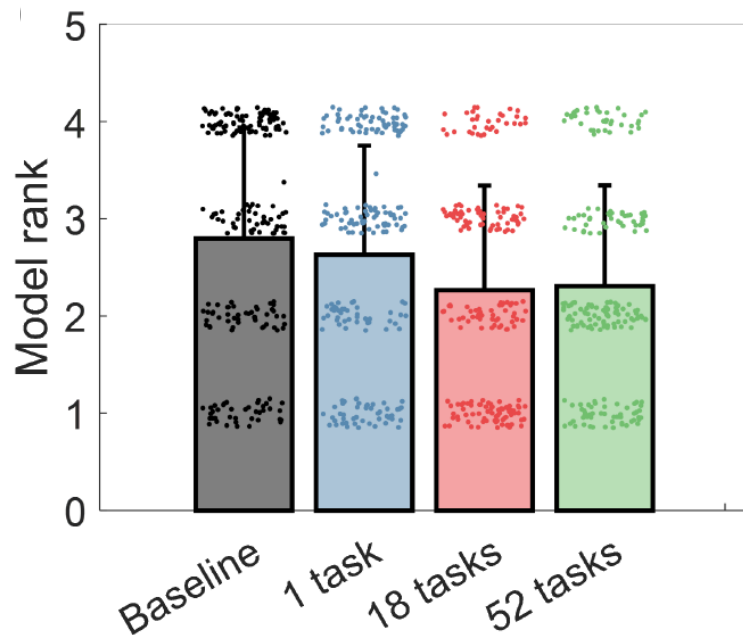
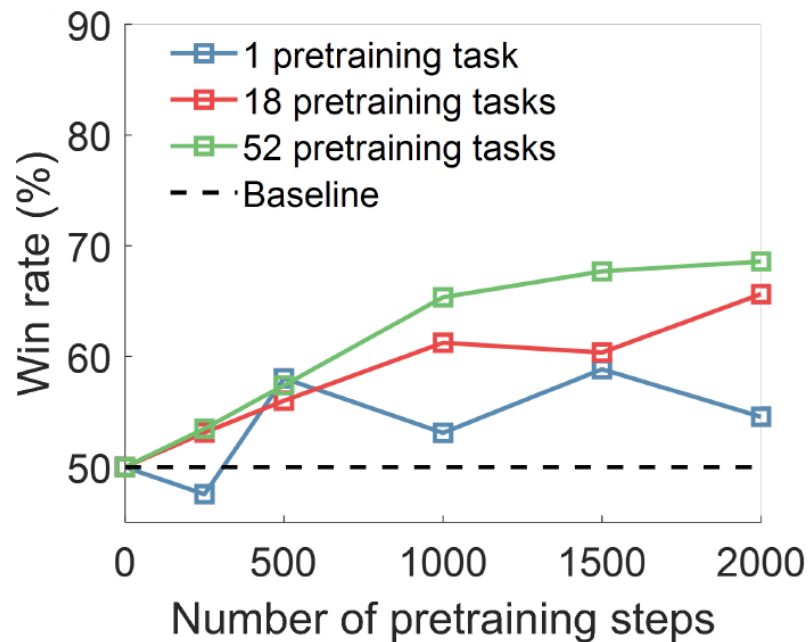


- Pretrains a shared FT-Transformer backbone on many tabular tasks
- Feature tokenizers and final "heads" are not shared
- Can be used as a starting point for a new tabular task

XTAB: results



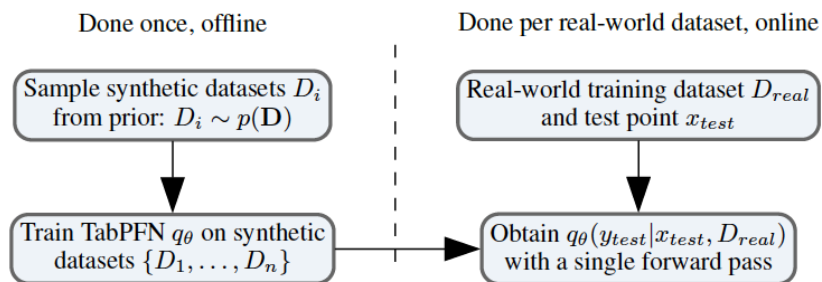
XTAB: dependence on the train size



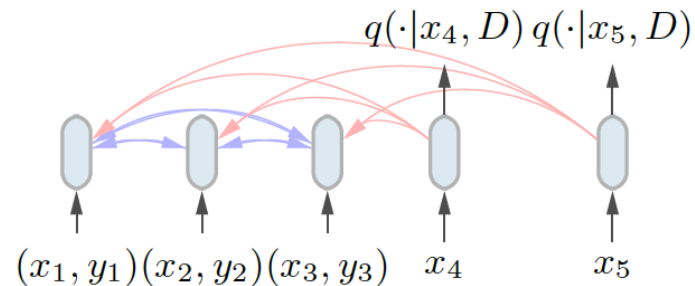
XTAB: conclusion

- Does provide some profit but ...
- Is limited to Transformer-based architectures
 - Can be slow
 - Can be suboptimal
- Typical improvements are moderate

TabPFN (Hollmann et al., ICLR'2023)



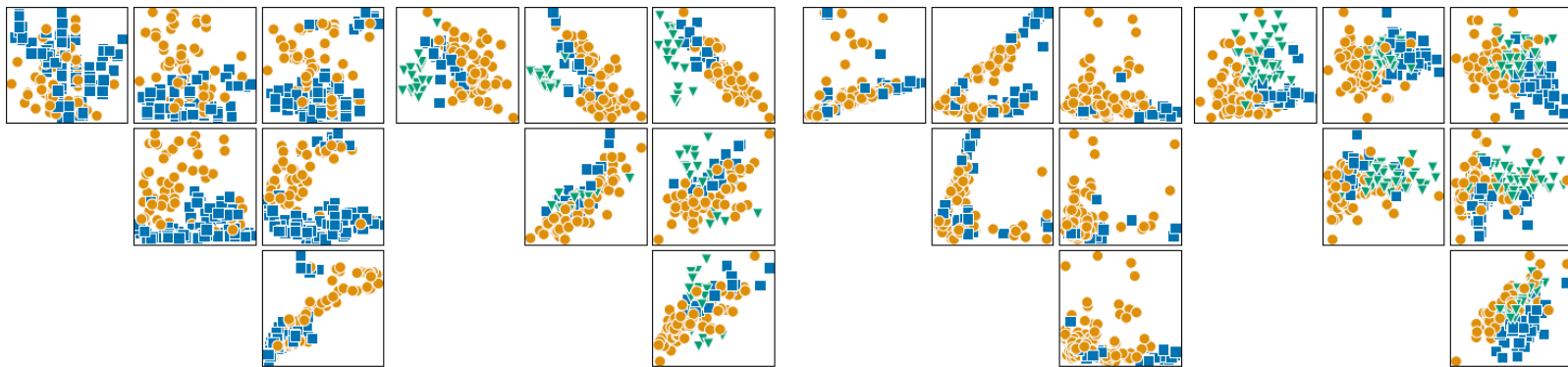
(a) Prior-fitting and inference



(b) Architecture and attention mechanism

TabPFN: synthetics

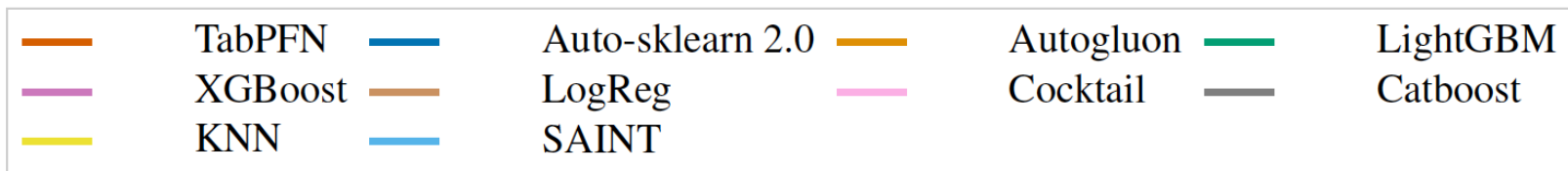
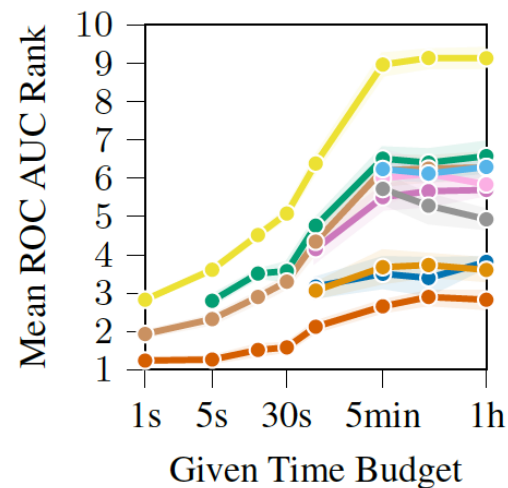
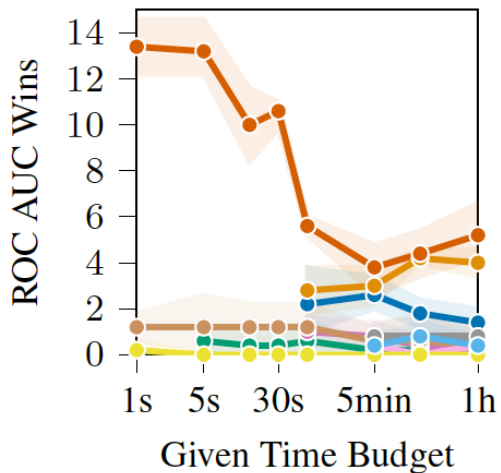
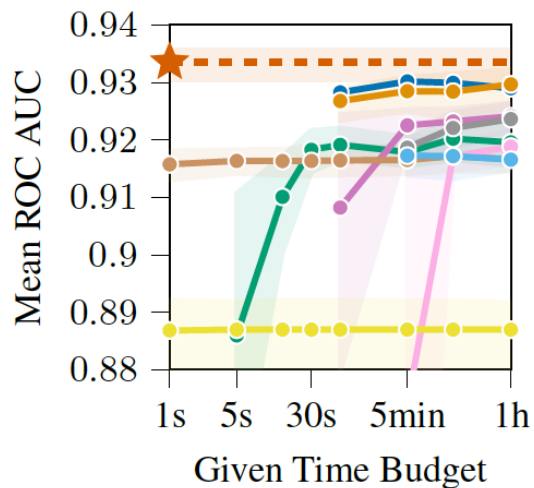
- Synthetic datasets are sampled from an accurately designed prior



(a) Synthetic datasets

(b) Actual datasets

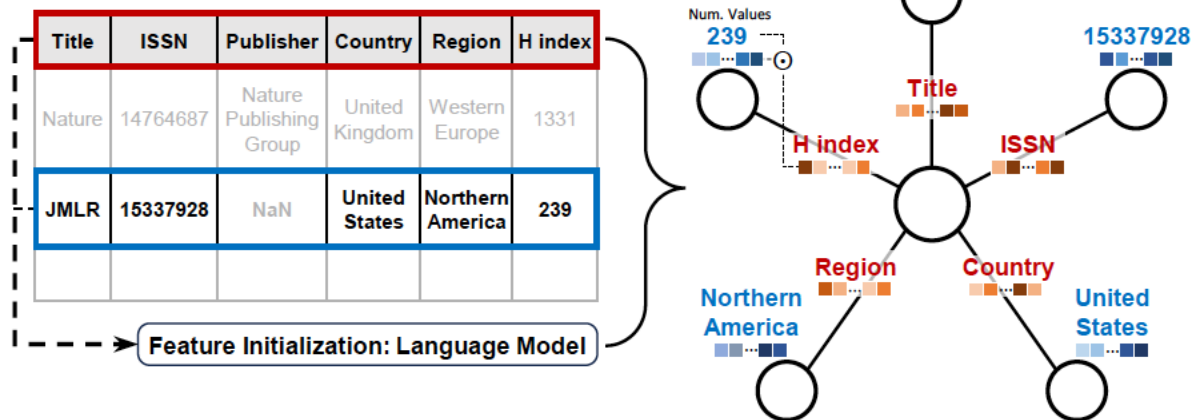
TabPFN: results



TabPFN: conclusion

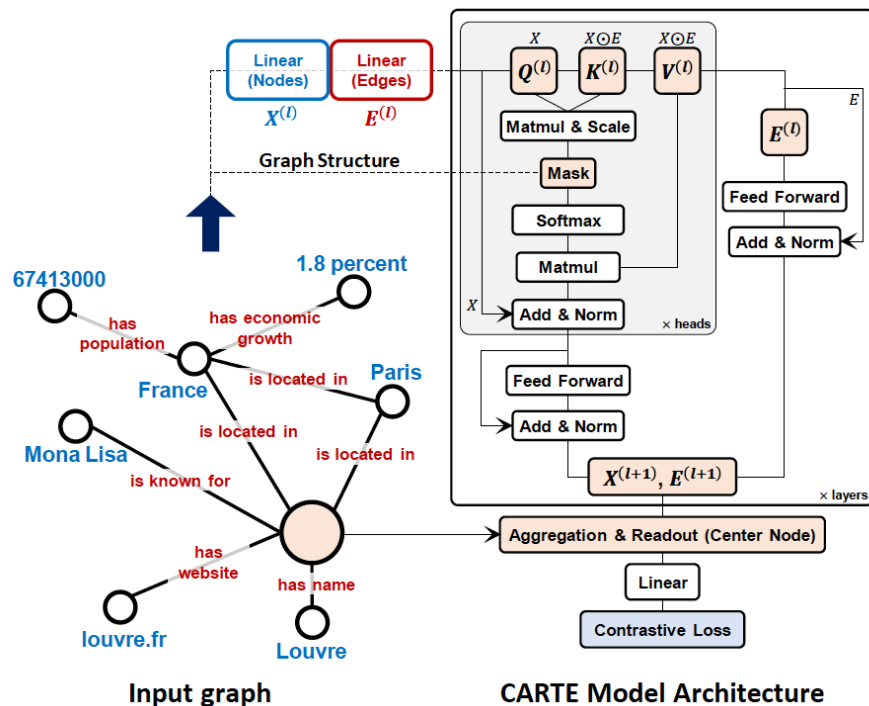
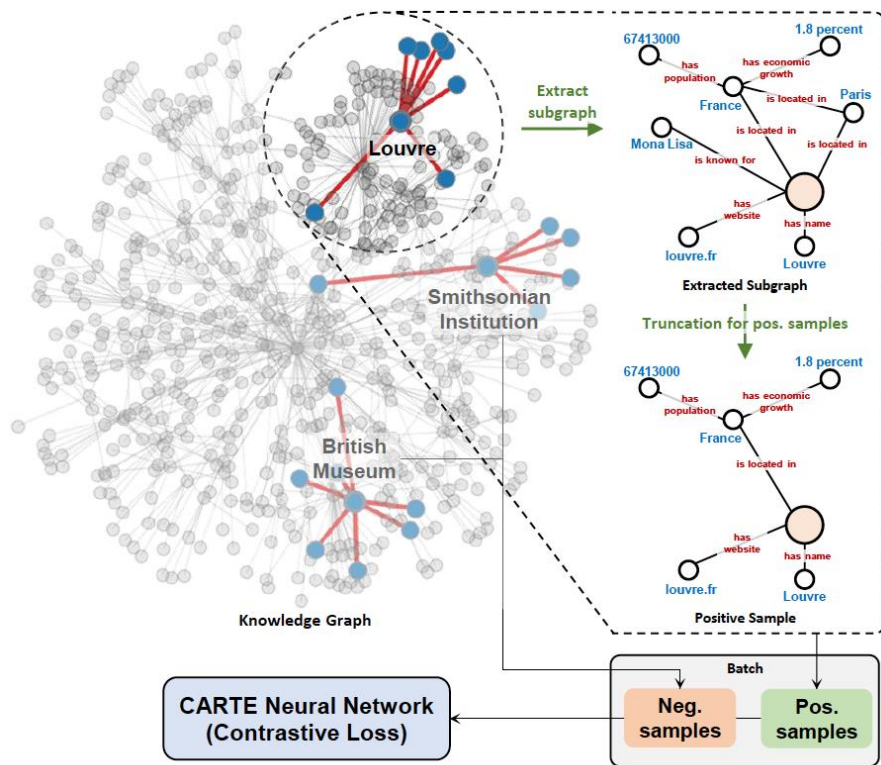
- Very interesting and novel idea but ...
- Is limited to Transformer-based architectures
- Is limited to small-scale problems
 - A lot of current research aims to scale TabPFN
- Focuses on a low-runtime-budget niche
 - In many applications, performance cannot be traded off against runtime

CARTE (Kim et al., ICML'2024)

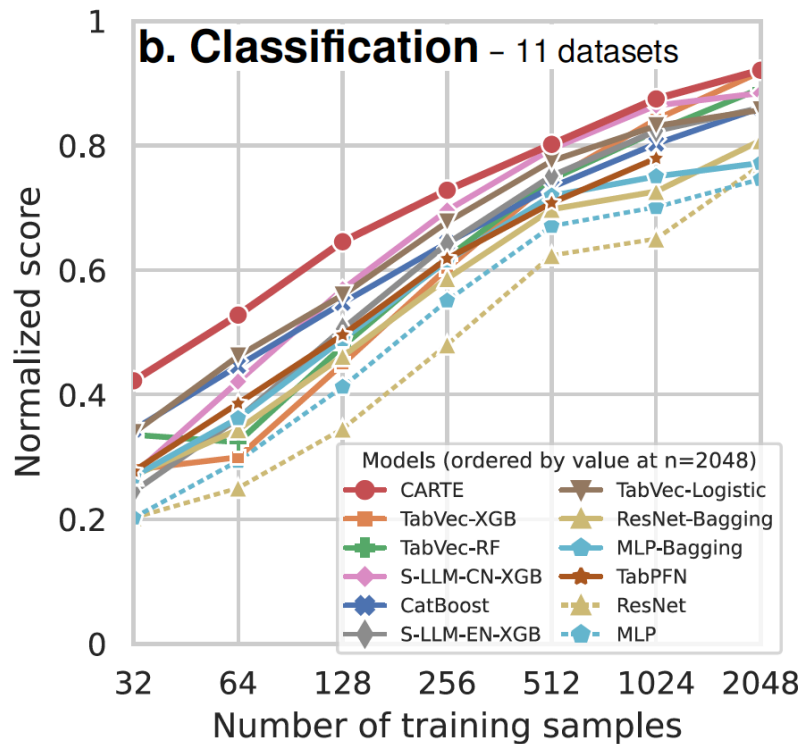
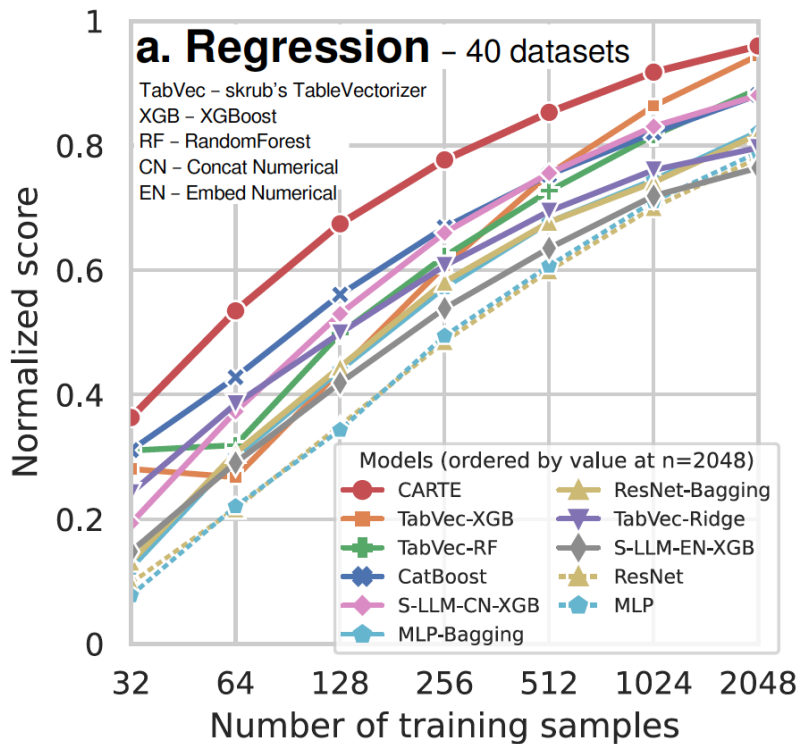


- Each datapoint is represented by a “star”-shaped graph
- “Textual” features are initialized based on LLM
- Special initialization of numerical features and the central node

CARTE: pretraining from the external knowledge graph



CARTE: results



CARTE: conclusion

- The method does work but ...
- The success is shown only for Transformer-based architectures
- The success is shown only for small-scale problems (up to a few thousand objects)
- The method needs meaningful column names
- For certain domains there could be a lack of external knowledge graphs

Outline

- Quick recap
- Tabular Benchmarks
- Pretraining in Tabular DL
- Cross-domain learning
- Generative tabular models
- Future directions

Generative Modeling in ML

- Goal: to approximate the data distribution by a probabilistic model
- One of potential applications: to produce useful synthetic data
- Several families of methods exist: GAN, VAE, NF, DDPM

Our work: TabDDPM

- Diffusion models were shown to outperform GAN/VAE/NF for images
- GAN/VAE were used for tabular data but without much success
- Let's use diffusion models for tabular data!

What are diffusion models?

- Forward process gradually adds noise to an initial sample with the predefined distributions $q(x_t|x_{t-1})$.
- Reverse process gradually denoises a latent variable with distributions $p(x_{t-1}|x_t)$ that are approximated by a neural network
- For example, Gaussian distribution for continuous data and categorical distributions for categorical data

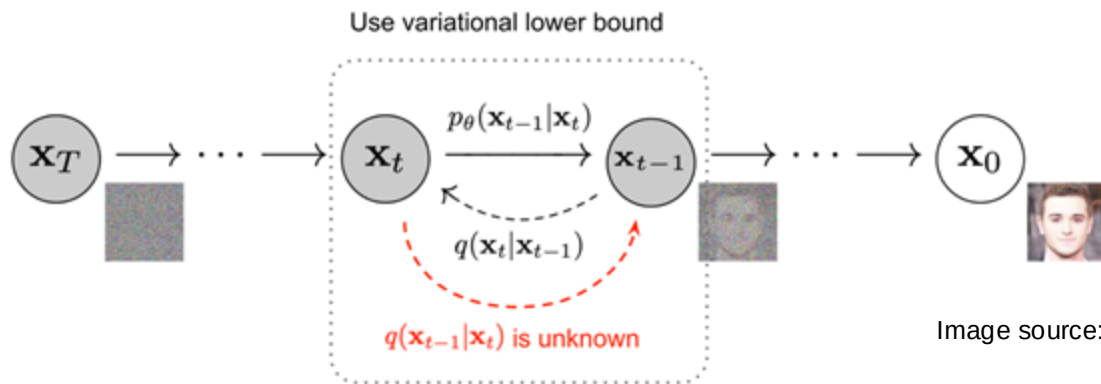
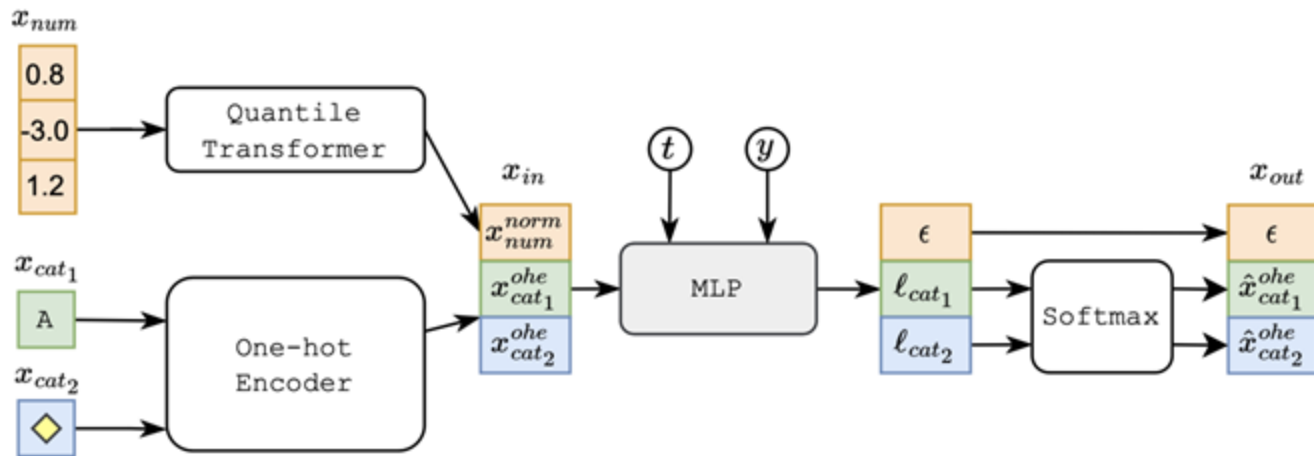


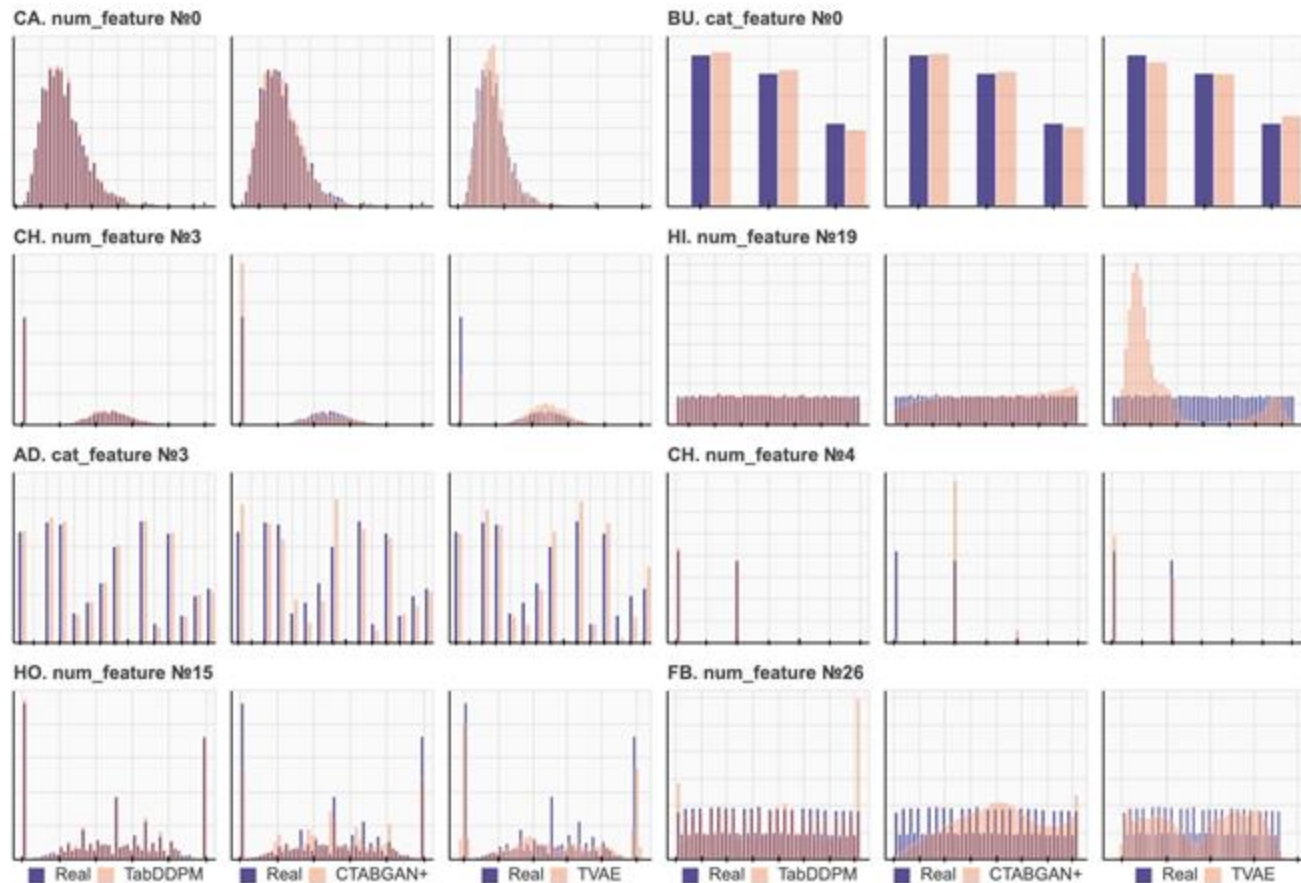
Image source: <https://lilianweng.github.io>

TabDDPM

- Gaussian diffusion for numerical features
- Multinomial diffusion (Hooeboom et al., 2021) for categorical features
- TabDDPM models joint distribution since MLP takes both numerical and categorical features to approximate reverse process
- Consider regression target as an additional feature
- Final loss is *sum* of gaussian DDPM and categorical DDPM losses

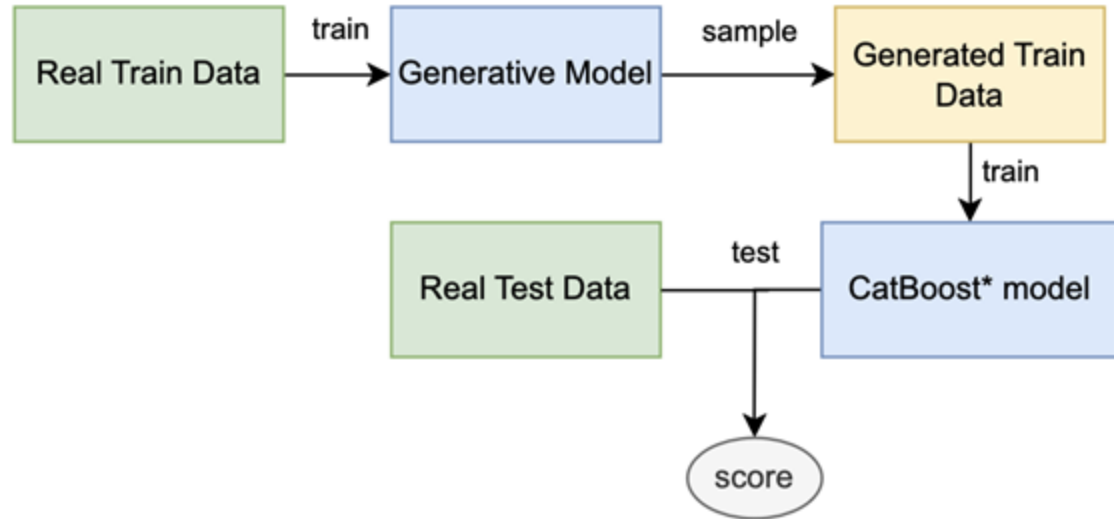


Individual feature distributions



Evaluation

- Machine Learning utility (Xu et al., 2019)
- Privacy metrics



*(Prokhorenkova et al., 2018)

Compare this score
with the real one

ML utility with CatBoost model

Average rank (over 16 datasets) with std in terms of ML utility of synthetic data

1 – the best

5 – the worst

Model	Avg. rank	Std of rank
CTGAN	4.25	1.06
TVAE	3.81	0.83
CTABGAN+	3.63	1.02
SMOTE	1.75	0.84
TabDDPM	1.56	0.60

SMOTE (Chawla et al., 2002) –
linear interpolation of two
random samples from train

ML utility with CatBoost model

Average rank (over 16 datasets) with std in terms of ML utility of synthetic data

1 – the best

5 – the worst

Model	Avg. rank	Std of rank
CTGAN	4.25	1.06
TVAE	3.81	0.83
CTABGAN+	3.63	1.02
SMOTE	1.75	0.84
TabDDPM	1.56	0.60

SMOTE (Chawla et al., 2002) –
linear interpolation of two
random samples from train

Main Conclusions:

- TabDDPM outperforms GAN/VAE-based baselines
- SMOTE is a simple and strong baseline

ML utility with Catboost. Numbers.

- TabDDPM performs on par with SMOTE
- Real score is almost always the highest one

Table 5. The values of machine learning efficiency computed w.r.t. the state-of-the-art tuned CatBoost model.

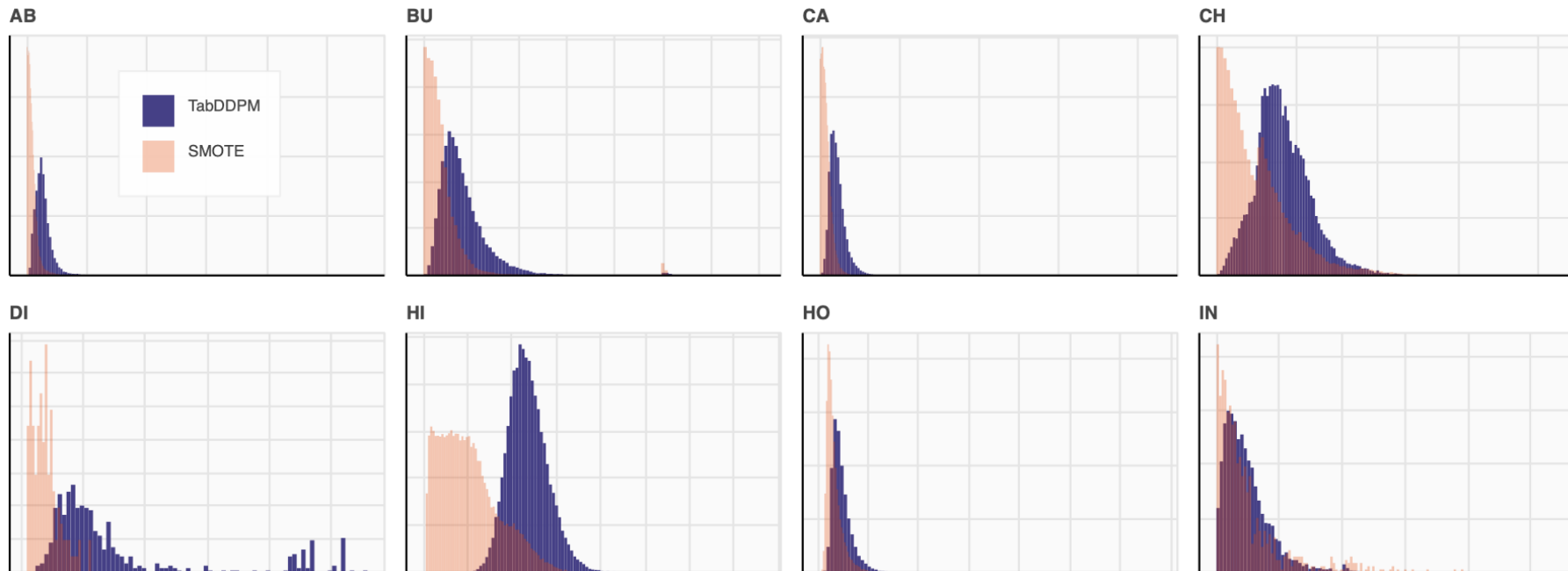
	AB (R_2)	AD (F_1)	BU (F_1)	CA (R_2)	CAR (F_1)	CH (F_1)	DE (F_1)	DI (F_1)
CTGAN	0.420 $\pm$.004	0.789 $\pm$.001	0.867 $\pm$.003	0.686 $\pm$.003	0.730 $\pm$.001	0.723 $\pm$.006	0.699$\pm$.002	0.459 $\pm$.096
TVAE	0.433 $\pm$.008	0.781 $\pm$.002	0.864 $\pm$.005	0.752 $\pm$.001	0.717 $\pm$.001	0.732 $\pm$.006	0.656 $\pm$.007	0.714$\pm$.039
CTABGAN	–	0.783 $\pm$.002	0.855 $\pm$.005	–	0.717 $\pm$.001	0.688 $\pm$.006	0.644 $\pm$.011	0.731$\pm$.022
CTABGAN+	0.467 $\pm$.004	0.772 $\pm$.003	0.884 $\pm$.005	0.525 $\pm$.004	0.733 $\pm$.001	0.702 $\pm$.012	0.686 $\pm$.004	0.734$\pm$.020
SMOTE	0.549$\pm$.005	0.791 $\pm$.002	0.891 $\pm$.003	0.840$\pm$.001	0.732 $\pm$.001	0.743 $\pm$.005	0.693 $\pm$.003	0.683 $\pm$.037
TabDDPM	0.550$\pm$.010	0.795$\pm$.001	0.906$\pm$.003	0.836 $\pm$.002	0.737$\pm$.001	0.755$\pm$.006	0.691 $\pm$.004	0.740$\pm$.020
Real	0.556 $\pm$.004	0.815 $\pm$.002	0.906 $\pm$.002	0.857 $\pm$.001	0.738 $\pm$.001	0.740 $\pm$.009	0.688 $\pm$.003	0.785 $\pm$.013

	FB (R_2)	GE (F_1)	HI (F_1)	HO (R_2)	IN (R_2)	KI (R_2)	MI (F_1)	WI (F_1)
CTGAN	0.443 $\pm$.005	0.333 $\pm$.013	0.575 $\pm$.006	0.433 $\pm$.005	0.745 $\pm$.009	0.772 $\pm$.005	0.783 $\pm$.005	0.749 $\pm$.015
TVAE	0.685 $\pm$.003	0.434 $\pm$.006	0.638 $\pm$.003	0.493 $\pm$.006	0.784 $\pm$.010	0.824 $\pm$.003	0.912 $\pm$.001	0.501 $\pm$.012
CTABGAN	–	0.392 $\pm$.006	0.575 $\pm$.004	–	–	–	0.889 $\pm$.002	0.906$\pm$.019
CTABGAN+	0.509 $\pm$.011	0.406 $\pm$.009	0.664 $\pm$.002	0.504 $\pm$.005	0.797 $\pm$.005	0.444 $\pm$.014	0.892 $\pm$.002	0.798 $\pm$.021
SMOTE	0.803$\pm$.002	0.658$\pm$.007	0.722$\pm$.001	0.662 $\pm$.004	0.812$\pm$.002	0.842$\pm$.004	0.932 $\pm$.001	0.913$\pm$.007
TabDDPM	0.713 $\pm$.002	0.597 $\pm$.006	0.722$\pm$.001	0.677$\pm$.010	0.809 $\pm$.002	0.833$\pm$.014	0.936$\pm$.001	0.904$\pm$.009
Real	0.837 $\pm$.001	0.636 $\pm$.007	0.724 $\pm$.001	0.662 $\pm$.003	0.814 $\pm$.001	0.907 $\pm$.002	0.934 $\pm$.000	0.898 $\pm$.006

Privacy. Distance to closest record (DCR)

- For each synthetic sample, we find the minimum distance to real datapoints and take the mean of these distances
- Low DCR values = all synthetic samples are essentially copies of some real datapoints
- Larger DCR values = generative model can produce something “new” rather than just copies of real data

Histograms of DCR values for TabDDPM and SMOTE



DCR comparison

Average rank (over 16 datasets) with std in terms of DCR

1 – the best

4 – the worst

Model	Avg. rank	Std of rank
TVAE	2.31	0.95
CTABGAN+	1.56	0.81
SMOTE	3.44	1.09
TabDDPM	2.69	0.79

Main Conclusions:

- TabDDPM outperforms SMOTE
- GAN/VAE methods show high DCR but generate useless (in terms of ML utility) samples

Conclusion

- Diffusion models generate tabular data of higher quality than GAN/VAE data
 - But still not enough for usage as "useful" synthetics
- "Old-school" SMOTE is a strong baseline that should not be overlooked
- TabDDPM is a step forward towards strong yet private method

Outline

- Quick recap
- Tabular Benchmarks
- Pretraining in Tabular DL
- Cross-domain learning
- Generative tabular models
- Future directions

Future of Tabular DL research

- More theory and understanding
 - Optimization dynamics
 - Dealing with 'high-frequencies'
- Synergy with Graph ML and GNNs
 - For graphs with tabular features in the nodes/edges
 - For multi-table problems with relations between tables
- Exploit LLM for Tabular problems
 - Use textual metadata about features
 - Multi-modal datasets
- Usability
 - Tooling

Questions?



Weierstraß-Institut für
Angewandte Analysis und Stochastik



Linearly perturbed optimization: theory and applications

Vladimir Spokoiny ,
WIAS, HU Berlin

7. Oktober 2024

1 Introduction

2 Linearly perturbed optimization

- Quadratic case
- 2S expansions
- Linear perturbation under third order smoothness
- Uniform smoothness

3 Fourth order approximation

4 Quadratic penalization

Let $f(\mathbf{v})$ be a smooth concave function,

$$\mathbf{v}^* = \operatorname{argmax}_{\mathbf{v}} f(\mathbf{v}), \quad \mathbb{F} = -\nabla^2 f(\mathbf{v}^*).$$

Let another function $g(\mathbf{v})$ satisfy for some vector $\mathbf{A}$

$$g(\mathbf{v}) - g(\mathbf{v}^*) = \langle \mathbf{v} - \mathbf{v}^*, \mathbf{A} \rangle + f(\mathbf{v}) - f(\mathbf{v}^*). \quad (1)$$

Define

$$\mathbf{v}^\circ \stackrel{\text{def}}{=} \operatorname{argmax}_{\mathbf{v}} g(\mathbf{v}), \quad g(\mathbf{v}^\circ) = \max_{\mathbf{v}} g(\mathbf{v}). \quad (2)$$

Aim: evaluate the quantities $\mathbf{v}^\circ - \mathbf{v}^*$ and $g(\mathbf{v}^\circ) - g(\mathbf{v}^*)$.

Let $L(\boldsymbol{v})$ be a **log-likelihood** function. Consider the MLE

$$\tilde{\boldsymbol{v}} = \operatorname{argmax}_{\boldsymbol{v}} L(\boldsymbol{v})$$

and the **background truth**

$$\boldsymbol{v}^* = \operatorname{argmax}_{\boldsymbol{v}} \mathbb{E}L(\boldsymbol{v}).$$

Stochastically linear smooth (SLS) models: $\mathbb{E}L(\boldsymbol{v})$ is **smooth** and **concave** in $\boldsymbol{v}$ and $\zeta(\boldsymbol{v}) = L(\boldsymbol{v}) - \mathbb{E}L(\boldsymbol{v})$ is **linear** in $\boldsymbol{v}$:

$$\boldsymbol{A} = \nabla \zeta(\boldsymbol{v}) = \nabla \zeta.$$

Outcome: Fisher theorem and Wilks phenomenon in statistics.

Let $h(\cdot)$ be **concave** and

$$\mathbf{v}^* = \operatorname{argmax} h(\mathbf{v}).$$

Consider

$$g(\mathbf{v}) = h(\mathbf{v}) - \|G\mathbf{v}\|^2/2,$$

$$f(\mathbf{v}) = h(\mathbf{v}) - \|G\mathbf{v}\|^2/2 + \langle G^2\mathbf{v}^*, \mathbf{v} \rangle, \quad .$$

Then $\nabla f(\mathbf{v}^*) = 0$ and $\mathbf{v}^* = \operatorname{argmax} f(\mathbf{v})$.

g is a linear perturbation of f with $\mathbf{A} = -G^2\mathbf{v}^*$.

Outcome: roughness penalty, effective dimension, critical dimension.

Let f be a **concave** function and

$$\mathbf{v}^* = \operatorname{argmax} f(\mathbf{v}).$$

Let also $\mathbf{v}^\circ$ be a **current guess**. Define

$$g(\mathbf{v}) = f(\mathbf{v}) - \langle \nabla f(\mathbf{v}^\circ), \mathbf{v} - \mathbf{v}^\circ \rangle.$$

Then $\nabla g(\mathbf{v}^\circ) = 0$ and hence,

$$\mathbf{v}^\circ = \operatorname{argmax} g(\mathbf{v}).$$

g is a linear perturbation of f with $\mathbf{A} = \nabla f(\mathbf{v}^\circ)$.

Outcome: Newton – Kantorovich – Nemirovskii-Nesterov theorem on **quadratic convergence** of **strongly convex** optimization.

Let $\mathbb{P}_f \sim \exp f(\mathbf{x})$. Denote by $\mathbb{N}_{\mathbf{x}, \mathbb{Z}}$ the Gaussian measure with the mean $\mathbf{x}$ and covariance $\mathbb{Z}^{-1}$, i.e. $\mathbb{N}_{\mathbf{x}, \mathbb{Z}} \stackrel{\text{def}}{=} \mathcal{N}(\mathbf{x}, \mathbb{Z}^{-1})$.

$$\text{Gauss VI: } (\mathbf{x}_{\text{VI}}, \mathbb{Z}_{\text{VI}}) = \underset{\mathbf{x}, \mathbb{Z}}{\operatorname{arginf}} \mathcal{K}(\mathbb{N}_{\mathbf{x}, \mathbb{Z}} \parallel \mathbb{P}_f).$$

Natural candidates:

1. **Laplace:** $\mathbf{x}_{\text{VI}} \approx \operatorname{argmax} f(\mathbf{x})$, $\mathbb{Z}_{\text{VI}} \approx -\nabla^2 f(\mathbf{x}^*)$;
2. **Moments:** $\mathbf{x}_{\text{VI}} \approx \mathbb{E}_f \mathbf{X}$, $\mathbb{Z}_{\text{VI}}^{-1} \approx \operatorname{Var}_f(\mathbf{X})$.

Let $\mathbb{P}_f \sim \exp f(\mathbf{x})$. Denote by $\mathbb{N}_{\mathbf{x}, \mathbb{Z}}$ the Gaussian measure with the mean $\mathbf{x}$ and covariance $\mathbb{Z}^{-1}$, i.e. $\mathbb{N}_{\mathbf{x}, \mathbb{Z}} \stackrel{\text{def}}{=} \mathcal{N}(\mathbf{x}, \mathbb{Z}^{-1})$.

$$\text{Gauss VI: } (\mathbf{x}_{\text{VI}}, \mathbb{Z}_{\text{VI}}) = \underset{\mathbf{x}, \mathbb{Z}}{\operatorname{arginf}} \mathcal{H}(\mathbb{N}_{\mathbf{x}, \mathbb{Z}} \parallel \mathbb{P}_f).$$

Natural candidates:

1. **Laplace:** $\mathbf{x}_{\text{VI}} \approx \operatorname{argmax} f(\mathbf{x})$, $\mathbb{Z}_{\text{VI}} \approx -\nabla^2 f(\mathbf{x}^*)$;
2. **Moments:** $\mathbf{x}_{\text{VI}} \approx \mathbb{E}_f \mathbf{X}$, $\mathbb{Z}_{\text{VI}}^{-1} \approx \operatorname{Var}_f(\mathbf{X})$.

[Katsevich and Rigollet, 2023] argued for (2).

1 Introduction

2 Linearly perturbed optimization

- Quadratic case
- 2S expansions
- Linear perturbation under third order smoothness
- Uniform smoothness

3 Fourth order approximation

4 Quadratic penalization

Lemma

Let $f(\mathbf{v})$ be *quadratic* with $\nabla^2 f(\mathbf{v}) \equiv -\mathbb{F}$. If $g(\mathbf{v})$ satisfy (1), then

$$\mathbf{v}^\circ - \mathbf{v}^* = \mathbb{F}^{-1} \mathbf{A}, \quad g(\mathbf{v}^\circ) - g(\mathbf{v}^*) = \frac{1}{2} \|\mathbb{F}^{-1/2} \mathbf{A}\|^2.$$

Proof. Clearly $-\nabla^2 g(\mathbf{v}) \equiv -\mathbb{F}$ and

$$\nabla g(\mathbf{v}^*) - \nabla g(\mathbf{v}^\circ) = \mathbb{F}(\mathbf{v}^\circ - \mathbf{v}^*).$$

Further, (1) and $\nabla f(\mathbf{v}^*) = 0$ yield $\nabla g(\mathbf{v}^*) = \mathbf{A}$. Together with $\nabla g(\mathbf{v}^\circ) = 0$, this implies $\mathbf{v}^\circ - \mathbf{v}^* = \mathbb{F}^{-1} \mathbf{A}$.

Taylor expansion of g at $\mathbf{v}^\circ$ yields by $\nabla g(\mathbf{v}^\circ) = 0$

$$g(\mathbf{v}^*) - g(\mathbf{v}^\circ) = -\frac{1}{2}(\mathbf{v}^\circ - \mathbf{v}^*)^\top \mathbb{F}(\mathbf{v}^\circ - \mathbf{v}^*) = -\frac{1}{2} \|\mathbb{F}^{-1/2} \mathbf{A}\|^2.$$

Define

$$\delta_3(\mathbf{v}, \mathbf{u}) = f(\mathbf{v} + \mathbf{u}) - f(\mathbf{v}) - \langle \nabla f(\mathbf{v}), \mathbf{u} \rangle - \frac{1}{2} \langle \nabla^2 f(\mathbf{v}), \mathbf{u}^{\otimes 2} \rangle,$$

$$\delta'_3(\mathbf{v}, \mathbf{u}) = \langle \nabla f(\mathbf{v} + \mathbf{u}), \mathbf{u} \rangle - \langle \nabla f(\mathbf{v}), \mathbf{u} \rangle - \langle \nabla^2 f(\mathbf{v}), \mathbf{u}^{\otimes 2} \rangle.$$

For $\mathbb{D}^2 \subseteq \mathbb{F}(\mathbf{v}) = -\nabla^2 f(\mathbf{v})$, define

$$\begin{aligned}\omega(\mathbf{v}) &\stackrel{\text{def}}{=} \sup_{\mathbf{u}: \|\mathbb{D}\mathbf{u}\| \leq r} \frac{2|\delta_3(\mathbf{v}, \mathbf{u})|}{\|\mathbb{D}\mathbf{u}\|^2}, \\ \omega'(\mathbf{v}) &\stackrel{\text{def}}{=} \sup_{\mathbf{u}: \|\mathbb{D}\mathbf{u}\| \leq r} \frac{|\delta'_3(\mathbf{v}, \mathbf{u})|}{\|\mathbb{D}\mathbf{u}\|^2}.\end{aligned}\tag{3}$$

Proposition

Fix $\nu \leq 2/3$ and r such that $\|\mathbb{F}^{-1/2} \mathbf{A}\| \leq \nu r$. Suppose now that $f(\mathbf{v})$ satisfy (3) for $\mathbf{v} = \mathbf{v}^*$, $\mathbb{D} = \mathbb{F}^{1/2}$, and ω' such that

$$1 - \nu - \omega' \geq 0. \quad (4)$$

Then for $\mathbf{v}^\circ$ from (2), it holds

$$\|\mathbb{F}^{1/2}(\mathbf{v}^\circ - \mathbf{v}^*)\| \leq r.$$

With $\mathbb{D} = \mathbb{F}^{1/2}$, the bound $\|\mathbb{D}^{-1}\mathbf{A}\| \leq \nu r$ implies for any $\mathbf{u}$

$$|\langle \mathbf{A}, \mathbf{u} \rangle| = |\langle \mathbb{D}^{-1}\mathbf{A}, \mathbb{D}\mathbf{u} \rangle| \leq \nu r \|\mathbb{D}\mathbf{u}\|.$$

If $\|\mathbb{D}\mathbf{u}\| > r$, then $r\|\mathbb{D}\mathbf{u}\| \leq \|\mathbb{D}\mathbf{u}\|^2$. Therefore,

$$|\langle \mathbf{A}, \mathbf{u} \rangle| \leq \nu \|\mathbb{D}\mathbf{u}\|^2, \quad \|\mathbb{D}\mathbf{u}\| > r. \quad (5)$$

Let $\mathbf{v}$ satisfy $\|\mathbb{D}(\mathbf{v} - \mathbf{v}^*)\| = r$. Denote $\mathbf{u} = \mathbf{v} - \mathbf{v}^*$. The idea is to show that the derivative $\frac{d}{dt}g(\mathbf{v}^* + t\mathbf{u}) < 0$ is negative for $t > 1$. Then all the extreme points of $g(\mathbf{v})$ are within $\mathcal{A}(r)$. We use the decomposition

$$g(\mathbf{v}^* + t\mathbf{u}) - g(\mathbf{v}^*) = \langle \mathbf{A}, \mathbf{u} \rangle t + f(\mathbf{v}^* + t\mathbf{u}) - f(\mathbf{v}^*).$$

With $h(t) = f(\mathbf{v}^* + t\mathbf{u}) - f(\mathbf{v}^*) - \langle \mathbf{A}, \mathbf{u} \rangle t$, it holds

$$\frac{d}{dt}f(\mathbf{v}^* + t\mathbf{u}) = \langle \mathbf{A}, \mathbf{u} \rangle + h'(t). \quad (6)$$

By definition of $\mathbf{v}^*$, it also holds $h'(0) = -\langle \mathbf{A}, \mathbf{u} \rangle$. The identity $\nabla^2 f(\mathbf{v}^*) = -\mathbb{D}^2$ yields $h''(0) = -\|\mathbb{D}\mathbf{u}\|^2$. Bound (3) implies for $|t| \leq 1$

$$|h'(t) - h'(0) - th''(0)| \leq t^2 |h''(0)| \omega'.$$

For $t = 1$, we obtain by (4) and (5)

$$h'(1) \leq -\langle \mathbf{A}, \mathbf{u} \rangle + h''(0) - h''(0) \omega' \leq -|h''(0)| (1 - \omega' - \nu) < 0.$$

Moreover, concavity of $h(t)$ imply that $h'(t) - h'(0)$ decreases in t for $t > 1$. Further, summing up the above derivation yields

$$\left. \frac{d}{dt} h(\mathbf{v}^* + t\mathbf{u}) \right|_{t=1} \leq -\|\mathbb{D}\mathbf{u}\|^2 (1 - \nu - \omega') < 0.$$

As $\frac{d}{dt} h(\mathbf{v}^* + t\mathbf{u})$ decreases with $t \geq 1$ together with $h'(t)$ due to (6), the same applies to all such t . This implies the assertion.

Proposition

Under the conditions of Proposition 1, with $\xi = \mathbb{D}^{-1}\mathbf{A} = \mathbb{F}^{-1/2}\mathbf{A}$

$$-\frac{\omega}{1+\omega}\|\xi\|^2 \leq 2g(\mathbf{v}^\circ) - 2g(\mathbf{v}^*) - \|\xi\|^2 \leq \frac{\omega}{1-\omega}\|\xi\|^2. \quad (7)$$

Also

$$\begin{aligned} \|\mathbb{D}(\mathbf{v}^\circ - \mathbf{v}^* - \mathbb{F}^{-1}\mathbf{A})\|^2 &\leq \frac{3\omega}{(1-\omega)^2} \|\mathbb{F}^{-1/2}\mathbf{A}\|^2, \\ \|\mathbb{D}(\mathbf{v}^\circ - \mathbf{v}^*)\| &\leq \frac{1 + \sqrt{2\omega}}{1-\omega} \|\mathbb{F}^{-1/2}\mathbf{A}\|. \end{aligned} \quad (8)$$

By (3), for any $\mathbf{v} \in \mathcal{A}(\mathbf{r})$

$$\left| f(\mathbf{v}^*) - f(\mathbf{v}) - \frac{1}{2} \|\mathbb{D}(\mathbf{v} - \mathbf{v}^*)\|^2 \right| \leq \frac{\omega}{2} \|\mathbb{D}(\mathbf{v} - \mathbf{v}^*)\|^2. \quad (9)$$

Further,

$$\begin{aligned} g(\mathbf{v}) - g(\mathbf{v}^*) - \frac{1}{2} \|\mathbb{D}^{-1} \mathbf{A}\|^2 &= \langle \mathbf{v} - \mathbf{v}^*, \mathbf{A} \rangle + f(\mathbf{v}) - f(\mathbf{v}^*) - \frac{1}{2} \|\mathbb{D}^{-1} \mathbf{A}\|^2 \\ &= -\frac{1}{2} \|\mathbb{D}(\mathbf{v} - \mathbf{v}^*) - \mathbb{D}^{-1} \mathbf{A}\|^2 + f(\mathbf{v}) - f(\mathbf{v}^*) + \frac{1}{2} \|\mathbb{D}(\mathbf{v} - \mathbf{v}^*)\|^2. \end{aligned} \quad (10)$$

As $\mathbf{v}^\circ \in \mathcal{A}(\mathbf{r})$ and it maximizes $g(\mathbf{v})$, we derive by (9)

$$\begin{aligned} g(\mathbf{v}^\circ) - g(\mathbf{v}^*) - \frac{1}{2} \|\mathbb{D}^{-1} \mathbf{A}\|^2 &= \max_{\mathbf{v} \in \mathcal{A}(\mathbf{r})} \left\{ g(\mathbf{v}) - g(\mathbf{v}^*) - \frac{1}{2} \|\mathbb{D}^{-1} \mathbf{A}\|^2 \right\} \\ &\leq \max_{\mathbf{v} \in \mathcal{A}(\mathbf{r})} \left\{ -\frac{1}{2} \|\mathbb{D}(\mathbf{v} - \mathbf{v}^*) - \mathbb{D}^{-1} \mathbf{A}\|^2 + \frac{\omega}{2} \|\mathbb{D}(\mathbf{v} - \mathbf{v}^*)\|^2 \right\}. \end{aligned}$$

Further, $\max_{\mathbf{u}} \{\omega \|\mathbf{u}\|^2 - \|\mathbf{u} - \boldsymbol{\xi}\|^2\} = \frac{\omega}{1-\omega} \|\boldsymbol{\xi}\|^2$ for $\omega \in [0, 1)$ and $\boldsymbol{\xi} \in \mathbb{R}^p$, yielding

$$g(\mathbf{v}^\circ) - g(\mathbf{v}^*) - \frac{1}{2} \|\mathbb{D}^{-1} \mathbf{A}\|^2 \leq \frac{\omega}{2(1-\omega)} \|\mathbb{D}^{-1} \mathbf{A}\|^2.$$

Similarly

$$\begin{aligned} & g(\mathbf{v}^\circ) - g(\mathbf{v}^*) - \frac{1}{2} \|\mathbb{D}^{-1} \mathbf{A}\|^2 \\ & \geq \max_{\mathbf{v} \in \mathcal{A}(\mathbf{r})} \left\{ -\frac{1}{2} \|\mathbb{D}(\mathbf{v} - \mathbf{v}^*) - \mathbb{D}^{-1} \mathbf{A}\|^2 - \frac{\omega}{2} \|\mathbb{D}(\mathbf{v} - \mathbf{v}^*)\|^2 \right\} \\ & = -\frac{\omega}{2(1+\omega)} \|\mathbb{D}^{-1} \mathbf{A}\|^2. \end{aligned} \tag{11}$$

These bounds imply (7).

Now we derive similarly to (10) that for $\mathbf{v} \in \mathcal{A}(\mathbf{r})$

$$g(\mathbf{v}) - g(\mathbf{v}^*) \leq \langle \mathbf{v} - \mathbf{v}^*, \mathbf{A} \rangle - \frac{1 - \omega}{2} \|\mathbb{D}(\mathbf{v} - \mathbf{v}^*)\|^2.$$

A particular choice $\mathbf{v} = \mathbf{v}^\circ$ yields

$$g(\mathbf{v}^\circ) - g(\mathbf{v}^*) \leq \langle \mathbf{v}^\circ - \mathbf{v}^*, \mathbf{A} \rangle - \frac{1 - \omega}{2} \|\mathbb{D}(\mathbf{v}^\circ - \mathbf{v}^*)\|^2.$$

Combining this result with (11) allows to bound

$$\langle \mathbf{v}^\circ - \mathbf{v}^*, \mathbf{A} \rangle - \frac{1 - \omega}{2} \|\mathbb{D}(\mathbf{v}^\circ - \mathbf{v}^*)\|^2 - \frac{1}{2} \|\mathbb{D}^{-1} \mathbf{A}\|^2 \geq -\frac{\omega}{2(1 + \omega)} \|\mathbb{D}^{-1} \mathbf{A}\|^2.$$

For $\xi = \mathbb{D}^{-1} \mathbf{A}$, $\mathbf{u} = \mathbb{D}(\mathbf{v}^\circ - \mathbf{v}^*)$, and $\omega \in [0, 1/3]$, the inequality

$$2\langle \mathbf{u}, \xi \rangle - (1 - \omega)\|\mathbf{u}\|^2 - \|\xi\|^2 \geq -\frac{\omega}{1 + \omega}\|\xi\|^2$$

implies

$$\left\| \mathbf{u} - \frac{1}{1 - \omega} \xi \right\|^2 \leq \frac{2\omega}{(1 + \omega)(1 - \omega)^2} \|\xi\|^2$$

yielding for $\omega \leq 1/3$

$$\begin{aligned} \|\mathbf{u} - \xi\| &\leq \left(\omega + \sqrt{\frac{2\omega}{1 + \omega}} \right) \frac{\|\xi\|}{1 - \omega} \leq \frac{\sqrt{3\omega}\|\xi\|}{1 - \omega}, \\ \|\mathbf{u}\| &\leq \left(1 + \sqrt{\frac{2\omega}{1 + \omega}} \right) \frac{\|\xi\|}{1 - \omega} \leq \frac{1 + \sqrt{2\omega}\|\xi\|}{1 - \omega}, \end{aligned}$$

and (8) follows.

Lemma

It holds

$$\max_{\mathbf{u}} \{ \omega \|\mathbf{u}\|^2 - \|\mathbf{u} - \boldsymbol{\xi}\|^2 \} = \frac{\omega}{1 - \omega} \|\boldsymbol{\xi}\|^2.$$

If $\omega \leq 1/3$, then the inequality

$$2\langle \mathbf{u}, \boldsymbol{\xi} \rangle - (1 - \omega) \|\mathbf{u}\|^2 - \|\boldsymbol{\xi}\|^2 \geq -\frac{\omega}{1 + \omega} \|\boldsymbol{\xi}\|^2$$

implies

$$\begin{aligned} \left\| \mathbf{u} - \frac{1}{1 - \omega} \boldsymbol{\xi} \right\|^2 &\leq \frac{2\omega}{(1 + \omega)(1 - \omega)^2} \|\boldsymbol{\xi}\|^2 \\ \|\mathbf{u} - \boldsymbol{\xi}\| &\leq \left(\omega + \sqrt{\frac{2\omega}{1 + \omega}} \right) \frac{\|\boldsymbol{\xi}\|}{1 - \omega} \leq \frac{\sqrt{3\omega} \|\boldsymbol{\xi}\|}{1 - \omega}. \end{aligned}$$

($\mathcal{T}_3$) *There exists τ_3 such that for all $\mathbf{u}$ with $\|\mathbb{D}\mathbf{u}\| \leq r$*

$$|\delta_3(\mathbf{v}, \mathbf{u})| \leq \frac{\tau_3}{6} \|\mathbb{D}\mathbf{u}\|^3, \quad |\delta'_3(\mathbf{v}, \mathbf{u})| \leq \frac{\tau_3}{2} \|\mathbb{D}\mathbf{u}\|^3.$$

($\mathcal{T}_4$) *There exists τ_4 such that for all $\mathbf{u}$ with $\|\mathbb{D}\mathbf{u}\| \leq r$*

$$|\delta_4(\mathbf{v}, \mathbf{u})| \leq \frac{\tau_4}{24} \|\mathbb{D}\mathbf{u}\|^4.$$

$(\mathcal{T}_3^*)$ $f(\mathbf{v})$ is strongly concave, $\mathbb{D}^2 \leq \nabla^2 f(\mathbf{v})$, and

$$\sup_{\mathbf{u}: \|\mathbb{D}\mathbf{u}\| \leq r} \sup_{\mathbf{z} \in \mathbb{R}^p} \frac{|\langle \nabla^3 f(\mathbf{v} + \mathbf{u}), \mathbf{z}^{\otimes 3} \rangle|}{\|\mathbb{D}\mathbf{z}\|^3} \leq \tau_3.$$

$(\mathcal{T}_4^*)$ $f(\mathbf{v})$ is strongly concave, $\mathbb{D}^2 \leq \nabla^2 f(\mathbf{v})$, and

$$\sup_{\mathbf{u}: \|\mathbb{D}\mathbf{u}\| \leq r} \sup_{\mathbf{z} \in \mathbb{R}^p} \frac{|\langle \nabla^4 f(\mathbf{v} + \mathbf{u}), \mathbf{z}^{\otimes 4} \rangle|}{\|\mathbb{D}\mathbf{z}\|^4} \leq \tau_4.$$

Banach's characterization [Banach, 1938] yields under $(\mathcal{T}_3^*)$ (resp $(\mathcal{T}_4^*)$)

$$|\langle \nabla^3 f(\mathbf{v} + \mathbf{u}), \mathbf{z}_1 \otimes \mathbf{z}_2 \otimes \mathbf{z}_3 \rangle| \leq \tau_3 \|\mathbb{D}\mathbf{z}_1\| \|\mathbb{D}\mathbf{z}_2\| \|\mathbb{D}\mathbf{z}_3\|; \quad (12)$$

$$|\langle \nabla^4 f(\mathbf{v} + \mathbf{u}), \mathbf{z}_1 \otimes \mathbf{z}_2 \otimes \mathbf{z}_3 \otimes \mathbf{z}_4 \rangle| \leq \tau_4 \prod_{k=1}^4 \|\mathbb{D}\mathbf{z}_k\|. \quad (13)$$

Proposition

Let $f(\mathbf{v})$ be a strongly concave function with $f(\mathbf{v}^*) = \max_{\mathbf{v}} f(\mathbf{v})$ and $\mathbb{F} = -\nabla^2 f(\mathbf{v}^*)$. Let $g(\mathbf{v})$ fulfill (1) with some vector $\mathbf{A}$. Suppose that $f(\mathbf{v})$ follows $(\mathcal{T}_3)$ with $\mathbf{r} = \nu^{-1} \|\mathbb{F}^{-1/2} \mathbf{A}\|$ for $\nu < 1$ and some $\tau_3 \geq 0$. Let

$$\tau_3 \|\mathbb{F}^{-1/2} \mathbf{A}\| < 2\nu(1 - \nu).$$

Then $\mathbf{v}^\circ = \operatorname{argmax}_{\mathbf{v}} g(\mathbf{v})$ satisfies

$$\|\mathbb{F}^{1/2}(\mathbf{v}^\circ - \mathbf{v}^*)\| \leq \nu^{-1} \|\mathbb{F}^{-1/2} \mathbf{A}\|.$$

Proposition

Under the conditions of Proposition 3

$$\begin{aligned}
 -\frac{2\tau_3}{3} \|\mathbb{F}^{-1/2} \mathbf{A}\|^3 &\leq 2g(\mathbf{v}^\circ) - 2g(\mathbf{v}^*) - \|\mathbb{F}^{-1/2} \mathbf{A}\|^2 \\
 &\leq \tau_3 \|\mathbb{F}^{-1/2} \mathbf{A}\|^3.
 \end{aligned} \tag{14}$$

Moreover, under $(\mathcal{T}_3^)$*

$$\begin{aligned}
 \|\mathbb{F}^{1/2}(\mathbf{v}^\circ - \mathbf{v}^*) - \mathbb{F}^{-1/2} \mathbf{A}\| &\leq \frac{3\tau_3}{4} \|\mathbb{F}^{-1/2} \mathbf{A}\|^2, \\
 \|\mathbb{F}^{1/2}(\mathbf{v}^\circ - \mathbf{v}^*)\| &\leq \|\mathbb{F}^{-1/2} \mathbf{A}\| + \frac{3\tau_3}{4} \|\mathbb{F}^{-1/2} \mathbf{A}\|^2.
 \end{aligned} \tag{15}$$

By $(\mathcal{T}_3)$ and $\nabla f(\mathbf{v}^*) = 0$, for any $\mathbf{v} \in \mathcal{A}(\mathbf{r})$

$$\begin{aligned} \left| f(\mathbf{v}^*) - f(\mathbf{v}) - \frac{1}{2} \|\mathbb{F}^{1/2}(\mathbf{v} - \mathbf{v}^*)\|^2 \right| &\leq \frac{\tau_3}{6} \|\mathbb{D}(\mathbf{v} - \mathbf{v}^*)\|^3 \\ &\leq \frac{\tau_3}{6} \|\mathbb{F}^{1/2}(\mathbf{v} - \mathbf{v}^*)\|^3. \end{aligned} \quad (16)$$

Further,

$$\begin{aligned} g(\mathbf{v}) - g(\mathbf{v}^*) - \frac{1}{2} \|\mathbb{F}^{-1/2} \mathbf{A}\|^2 \\ &= \langle \mathbf{v} - \mathbf{v}^*, \mathbf{A} \rangle + f(\mathbf{v}) - f(\mathbf{v}^*) - \frac{1}{2} \|\mathbb{F}^{-1/2} \mathbf{A}\|^2 \\ &= -\frac{1}{2} \|\mathbb{F}^{1/2}(\mathbf{v} - \mathbf{v}^*) - \mathbb{F}^{-1/2} \mathbf{A}\|^2 + f(\mathbf{v}) - f(\mathbf{v}^*) + \frac{1}{2} \|\mathbb{F}^{1/2}(\mathbf{v} - \mathbf{v}^*)\|^2. \end{aligned}$$

As $\mathbf{v}^\circ \in \mathcal{A}(\mathbf{r})$ and it maximizes $g(\mathbf{v})$, we derive by (16) and Lemma 3

$$\begin{aligned}
 g(\mathbf{v}^\circ) - g(\mathbf{v}^*) - \frac{1}{2} \|\mathbb{F}^{-1/2} \mathbf{A}\|^2 &= \max_{\mathbf{v} \in \mathcal{A}(\mathbf{r})} \left\{ g(\mathbf{v}) - g(\mathbf{v}^*) - \frac{1}{2} \|\mathbb{F}^{-1/2} \mathbf{A}\|^2 \right\} \\
 &\leq \max_{\mathbf{v} \in \mathcal{A}(\mathbf{r})} \left\{ -\frac{1}{2} \|\mathbb{F}^{1/2}(\mathbf{v} - \mathbf{v}^*) - \mathbb{F}^{-1/2} \mathbf{A}\|^2 + \frac{\tau_3}{6} \|\mathbb{F}^{1/2}(\mathbf{v} - \mathbf{v}^*)\|^3 \right\} \\
 &\leq \frac{\tau_3}{2} \|\mathbb{F}^{-1/2} \mathbf{A}\|^3.
 \end{aligned}$$

Now (14) follows from this and

$$\begin{aligned}
 g(\mathbf{v}^\circ) - g(\mathbf{v}^*) - \frac{1}{2} \|\mathbb{F}^{-1/2} \mathbf{A}\|^2 &\geq \max_{\mathbf{v} \in \mathcal{A}(\mathbf{r})} \left\{ -\frac{1}{2} \|\mathbb{F}^{1/2}(\mathbf{v} - \mathbf{v}^*) - \mathbb{F}^{-1/2} \mathbf{A}\|^2 - \frac{\tau_3}{6} \|\mathbb{F}^{1/2}(\mathbf{v} - \mathbf{v}^*)\|^3 \right\} \\
 &\geq -\frac{\tau_3}{3} \|\mathbb{F}^{-1/2} \mathbf{A}\|^3.
 \end{aligned}$$

For proving (21) use that $\nabla f(\mathbf{v}^*) = 0$, $\nabla g(\mathbf{v}^\circ) = 0$,
 $\nabla f(\mathbf{v}^\circ) = \nabla g(\mathbf{v}^\circ) - \mathbf{A} = -\mathbf{A}$, and $-\nabla^2 f(\mathbf{v}^*) = \mathbb{F}$. By Lemma 4 with
 $\mathbf{u} = \mathbb{F}^{-1}\mathbf{A}$

$$\|\mathbb{F}^{-1/2}\{\nabla f(\mathbf{v}^* + \mathbb{F}^{-1}\mathbf{A}) + \mathbf{A}\}\| \leq \frac{\tau_3}{2}\|\mathbb{F}^{-1/2}\mathbf{A}\|^2. \quad (17)$$

Further, by (1)

$$\begin{aligned} \|\mathbb{F}^{-1/2}\nabla g(\mathbf{v}^* + \mathbb{F}^{-1}\mathbf{A})\| &= \|\mathbb{F}^{-1/2}\{\nabla g(\mathbf{v}^* + \mathbb{F}^{-1}\mathbf{A}) - \mathbf{A} + \mathbf{A}\}\| \\ &\leq \|\mathbb{F}^{-1/2}\{\nabla f(\mathbf{v}^* + \mathbb{F}^{-1}\mathbf{A}) + \mathbf{A}\}\| \leq \frac{\tau_3}{2}\|\mathbb{F}^{-1/2}\mathbf{A}\|^2. \end{aligned}$$

By definition $\nabla g(\mathbf{v}^\circ) = 0$. This yields

$$\|\mathbb{F}^{-1/2}\{\nabla g(\mathbf{v}^* + \mathbb{F}^{-1}\mathbf{A}) - \nabla g(\mathbf{v}^\circ)\}\| \leq \frac{\tau_3}{2}\|\mathbb{F}^{-1/2}\mathbf{A}\|^2. \quad (18)$$

Now we can use with $\Delta = \mathbf{v}^* + \mathbb{F}^{-1} \mathbf{A} - \mathbf{v}^\circ$

$$\begin{aligned} & \mathbb{F}^{-1/2} \{ \nabla g(\mathbf{v}^* + \mathbb{F}^{-1} \mathbf{A}) - \nabla g(\mathbf{v}^\circ) \} \\ &= \left(\int_0^1 \mathbb{F}^{-1/2} \nabla^2 g(\mathbf{v}^\circ + t\Delta) \mathbb{F}^{-1/2} dt \right) \mathbb{F}^{1/2} \Delta. \end{aligned}$$

By (1) $\nabla^2 g(\mathbf{v}) = \nabla^2 f(\mathbf{v})$ for all $\mathbf{v}$. If $\|\mathbb{F}^{1/2}(\mathbf{v} - \mathbf{v}^*)\| \leq \mathbf{r}$, then $(\mathcal{T}_3^*)$ implies $\|\mathbb{F}^{-1/2} \nabla^2 f(\mathbf{v}) \mathbb{F}^{-1/2} + \mathbb{I}_p\| \leq \omega^+ \leq \tau_3 \mathbf{r} \leq 1/3$. Hence,

$$\|\mathbb{F}^{-1/2} \{ \nabla g(\mathbf{v}^* + \mathbb{F}^{-1} \mathbf{A}) - \nabla g(\mathbf{v}^\circ) \}\| \geq (1 - \omega^+) \|\mathbb{F}^{1/2}(\mathbf{v}^\circ - \mathbf{v}^* - \mathbb{F}^{-1} \mathbf{A})\|.$$

This and (26) yield

$$\|\mathbb{F}^{1/2}(\mathbf{v}^\circ - \mathbf{v}^* - \mathbb{F}^{-1} \mathbf{A})\| \leq \frac{\tau_3}{2(1 - \omega^+)} \|\mathbb{F}^{-1/2} \mathbf{A}\|^2 \leq \frac{3\tau_3}{4} \|\mathbb{F}^{-1/2} \mathbf{A}\|^2,$$

and (21) follows.

Lemma

For any $\xi \in \mathbb{R}^p$ with $\|\xi\| \leq 2r/3$ and τ with $\tau r \leq 1/2$, it holds

$$\max_{\|u\| \leq r} \left(\frac{\tau}{3} \|u\|^3 - \|u - \xi\|^2 \right) \leq \frac{\tau}{2} \|\xi\|^3, \quad (19)$$

$$\min_{\|u\| \leq r} \left(\frac{\tau}{3} \|u\|^3 + \|u - \xi\|^2 \right) \leq \frac{\tau}{3} \|\xi\|^3. \quad (20)$$

Any maximizer $\mathbf{u}$ of the left hand-side of (19) satisfies

$$\tau \|\mathbf{u}\|^{1/2} \mathbf{u} - 2(\mathbf{u} - \boldsymbol{\xi}) = 0.$$

Therefore, $\mathbf{u} = \rho \boldsymbol{\xi}$ for some ρ , reducing the problem to the univariate case:

$$\max_{\|\mathbf{u}\| \leq r} \left(\frac{\tau}{3} \|\mathbf{u}\|^3 - \|\mathbf{u} - \boldsymbol{\xi}\|^2 \right) = \|\boldsymbol{\xi}\|^2 \max_{\rho: \|\rho \boldsymbol{\xi}\| \leq r} \left(\frac{\tau \|\boldsymbol{\xi}\|}{3} \rho^3 - (\rho - 1)^2 \right).$$

Define $a = \tau \|\boldsymbol{\xi}\|$. The conditions $\|\boldsymbol{\xi}\| \leq 2r/3$ and $\tau r \leq 1/2$ imply $a \leq 1/3$ and $\|\rho \boldsymbol{\xi}\| \leq r$ implies $|\rho| \leq 3/2$. The function $a\rho^3/3 - (\rho - 1)^2$ is concave on the interval $|\rho| \leq 3/2$ and hence, the maximizer ρ fulfills $a\rho^2 - 2\rho + 2 = 0$ yielding

$$\rho = \frac{1 \pm \sqrt{1 - 2a}}{a}, \quad |\rho| \leq 3/2.$$

As $a \in [0, 1/3]$, we can only use

$$\rho_a = \frac{1 - \sqrt{1 - 2a}}{a} = \frac{2}{1 + \sqrt{1 - 2a}}, \quad \rho_a - 1 = \frac{2a}{(1 + \sqrt{1 - 2a})^2}.$$

Therefore,

$$\begin{aligned} \max_{\|u\| \leq r} \left(\frac{\tau}{3} \|u\|^3 - \|u - \xi\|^2 \right) &= \frac{\tau \|\xi\|^3 \rho_a^3}{3} - \|\xi\|^2 (\rho_a - 1)^2 \\ &= \frac{\tau \|\xi\|^3}{3} \frac{8(1 + \sqrt{1 - 2a}) - 12a}{(1 + \sqrt{1 - 2a})^4} \leq \frac{\tau \|\xi\|^3}{3} \max_{a \in [0, 1/3]} \frac{8(1 + \sqrt{1 - 2a}) - 12a}{(1 + \sqrt{1 - 2a})^4} \leq \frac{\tau \|\xi\|^3}{2} \end{aligned}$$

With $y = 1 + \sqrt{1 - 2a}$ or $-2a = (y - 1)^2 - 1 = y^2 - 2y$, represent

$$\phi(a) \stackrel{\text{def}}{=} \frac{8(1 + \sqrt{1 - 2a}) - 12a}{(1 + \sqrt{1 - 2a})^4} = \frac{8y + 6y^2 - 12y}{y^4} = \frac{6y - 4}{y^3},$$

and the latter decreases with $y \geq 1$. As $\phi(1/3) \leq 3/2$, (19) follows.

The proof of (20) is similar. The general case can be reduced to the univariate one by using $\mathbf{u} = \rho \boldsymbol{\xi}$. With $a = \tau \|\boldsymbol{\xi}\|$, the minimizer ρ_a reads as

$$\rho_a = \frac{2}{1 + \sqrt{1 + 2a}}, \quad 1 - \rho_a = \frac{\sqrt{1 + 2a} - 1}{\sqrt{1 + 2a} + 1} = \frac{2a}{(\sqrt{1 + 2a} + 1)^2},$$

yielding for $a \in [0, 1/3]$

$$\begin{aligned} \min_{\|\mathbf{u}\| \leq r} \left(\frac{\tau}{3} \|\mathbf{u}\|^3 + \|\mathbf{u} - \boldsymbol{\xi}\|^2 \right) &= \frac{\tau \|\boldsymbol{\xi}\|^3 \rho_a^3}{3} + \|\boldsymbol{\xi}\|^2 (\rho_a - 1)^2 \\ &\leq \frac{\tau \|\boldsymbol{\xi}\|^3}{3} \max_{a \in [0, 1/3]} \frac{8(1 + \sqrt{1 + 2a}) + 12a}{(1 + \sqrt{1 + 2a})^4}, \end{aligned}$$

and with $y = 1 + \sqrt{1 + 2a}$ or $2a = y^2 - 2y$,

$$\max_{a \in [0, 1/3]} \frac{8(1 + \sqrt{1 + 2a}) + 12a}{(1 + \sqrt{1 + 2a})^4} \leq \max_{y \geq 2} \frac{8y + 6y^2 - 12y}{y^4} = \max_{y \geq 2} \frac{6y - 4}{y^3} = 1.$$

Proposition

Let $f(\mathbf{v})$ be a strongly concave function with $f(\mathbf{v}^*) = \max_{\mathbf{v}} f(\mathbf{v})$ and $\mathbb{F} = -\nabla^2 f(\mathbf{v}^*)$. Assume $(\mathcal{T}_3^*)$ at $\mathbf{v}^*$ with $\mathbb{D}^2$, $\mathbf{r}$, and τ_3 such that

$$\mathbb{D}^2 \leq \mathbb{F}, \quad \mathbf{r} \geq \frac{3}{2} \|\mathbb{D} \mathbb{F}^{-1} \mathbf{A}\|, \quad \tau_3 \|\mathbb{D} \mathbb{F}^{-1} \mathbf{A}\| < \frac{4}{9}.$$

Then $\|\mathbb{D}(\mathbf{v}^\circ - \mathbf{v}^*)\| \leq (3/2) \|\mathbb{D} \mathbb{F}^{-1} \mathbf{A}\|$ and moreover,

$$\|\mathbb{D}^{-1} \mathbb{F}(\mathbf{v}^\circ - \mathbf{v}^* - \mathbb{F}^{-1} \mathbf{A})\| \leq \frac{3\tau_3}{4} \|\mathbb{D} \mathbb{F}^{-1} \mathbf{A}\|^2. \quad (21)$$

If the function f is quadratic and concave with the maximum at $\mathbf{v}^*$ then the linearly perturbed function g is also quadratic and concave with the maximum at $\check{\mathbf{v}} = \mathbf{v}^* + \mathbb{F}^{-1}\mathbf{A}$.

In general, the point $\check{\mathbf{v}}$ is not the maximizer of g , however, it is very close to $\mathbf{v}^*$. We use that $\nabla f(\mathbf{v}^*) = 0$ and $-\nabla^2 f(\mathbf{v}^*) = \mathbb{F}$. Then (27) of Lemma 4 yields

$$\begin{aligned}\|\mathbb{D}^{-1}\nabla g(\check{\mathbf{v}})\| &= \|\mathbb{D}^{-1}\{\nabla f(\mathbf{v}^* + \mathbb{F}^{-1}\mathbf{A}) - \nabla f(\mathbf{v}^*) + \mathbf{A}\}\| \\ &\leq \frac{\tau_3}{2}\|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\|^2.\end{aligned}\tag{22}$$

As $\|\mathbb{D} \mathbb{F}^{-1} \mathbf{A}\| \leq 2r/3$, condition $(\mathcal{T}_3^*)$ can be applied in the $r/3$ -vicinity of $\check{\mathbf{v}}$. Fix any $\mathbf{v}$ with $\|\mathbb{D}(\mathbf{v} - \check{\mathbf{v}})\| \leq r/3$ and define $\Delta = \mathbf{v} - \check{\mathbf{v}}$. By (29) of Lemma 4

$$\begin{aligned} \|\mathbb{D}^{-1}\{\nabla g(\mathbf{v}) - \nabla g(\check{\mathbf{v}}) + \mathbb{F}\Delta\}\| &= \|\mathbb{D}^{-1}\{\nabla f(\mathbf{v}) - \nabla f(\check{\mathbf{v}}) + \mathbb{F}\Delta\}\| \\ &\leq \frac{3\tau_3}{2} \|\mathbb{D}\Delta\|^2. \end{aligned}$$

In particular, this and (22) yield

$$\|\mathbb{D}^{-1}\{\nabla g(\check{\mathbf{v}} + \Delta) + \mathbb{F}\Delta\}\| \leq 2\tau_3 \|\mathbb{D}\Delta\|^2.$$

For any $\mathbf{u}$ with $\|\mathbf{u}\| = 1$, this implies

$$|\langle \nabla g(\check{\mathbf{v}} + \Delta) + \mathbb{F}\Delta, \mathbb{D}^{-1}\mathbf{u} \rangle| \leq 2\tau_3 \|\mathbb{D}\Delta\|^2. \quad (23)$$

Suppose now that $\|\mathbb{D}\Delta\| = r/3$ and consider the function $h(t) = g(\check{\mathbf{v}} + t\Delta)$. Then $h'(t) = \langle \nabla g(\check{\mathbf{v}} + t\Delta), \Delta \rangle$ and (23) implies with $\mathbf{u} = \mathbb{D}\Delta/\|\mathbb{D}\Delta\|$

$$|\langle \nabla g(\check{\mathbf{v}} + \Delta), \Delta \rangle + \|\mathbb{F}^{1/2}\Delta\|^2| \leq 2\tau_3 \|\mathbb{D}\Delta\|^3.$$

As $\mathbb{F} \geq \mathbb{D}^2$, this yields

$$h'(1) \leq 2\tau_3 \|\mathbb{D}\Delta\|^3 - \|\mathbb{D}\Delta\|^2. \quad (24)$$

Similarly, (22) yields by $\|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\| = 2r/3$

$$|h'(0)| = |\langle \nabla g(\check{\mathbf{v}}), \Delta \rangle| \leq \frac{\tau_3}{2} \|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\|^2 \|\mathbb{D}\Delta\| = \frac{2\tau_3}{9} r^2 \|\mathbb{D}\Delta\|. \quad (25)$$

Concavity of $g(\cdot)$ ensures that $t^* = \operatorname{argmax}_t h(t)$ satisfies $|t^*| \leq 1$ if

$$h'(1) < -|h'(0)|, \quad h'(-1) < |h'(0)|.$$

Due to (24), (25), and $\|\mathbb{D}\Delta\| = r/3$, the latter condition reads

$$\frac{2\tau_3}{9}r^2\|\mathbb{D}\Delta\| + 2\tau_3\|\mathbb{D}\Delta\|^3 - \|\mathbb{D}\Delta\|^2 = \|\mathbb{D}\Delta\|r\left(\frac{2\tau_3r}{9} + \frac{2\tau_3r}{9} - \frac{1}{3}\right) < 0.$$

which is fulfilled because of $\tau_3\|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\| \leq 4/9$ and $\|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\| = 2r/3$. We summarize that $\mathbf{v}^\circ = \operatorname{argmax}_{\mathbf{v}} g(\mathbf{v})$ satisfies $\|\mathbb{D}(\mathbf{v}^\circ - \check{\mathbf{v}})\| \leq r/3$ while $\|\mathbb{D}(\check{\mathbf{v}} - \mathbf{v}^*)\| = \|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\| = 2r/3$. Therefore,

$$\|\mathbb{D}(\mathbf{v}^\circ - \mathbf{v}^*)\| \leq r.$$

This allows us to use $(\mathcal{T}_3^*)$ at this point for establishing (21). By definition $\nabla g(\mathbf{v}^\circ) = 0$ and hence,

$$\|\mathbb{D}^{-1}\{\nabla g(\mathbf{v}^* + \mathbb{F}^{-1}\mathbf{A}) - \nabla g(\mathbf{v}^\circ)\}\| \leq \frac{\tau_3}{2}\|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\|^2. \quad (26)$$

By (29) of Lemma 4, it holds with $\Delta = \mathbf{v}^* + \mathbb{F}^{-1}\mathbf{A} - \mathbf{v}^\circ$

$$\|\mathbb{D}^{-1}\{\nabla g(\mathbf{v}^* + \mathbb{F}^{-1}\mathbf{A}) - \nabla g(\mathbf{v}^\circ) - \nabla^2 g(\mathbf{v}^*)\Delta\}\| \leq \frac{3\tau_3}{2}\|\mathbb{D}\Delta\|^2.$$

Combining with (26) yields

$$\|\mathbb{D}^{-1}\mathbb{F}\Delta\| \leq \frac{3\tau_3}{2}\|\mathbb{D}\Delta\|^2 + \frac{\tau_3}{2}\|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\|^2 \leq \frac{3\tau_3}{2}\|\mathbb{D}^{-1}\mathbb{F}\Delta\|^2 + \frac{\tau_3}{2}\|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\|^2.$$

As $2x \leq \alpha x^2 + \beta$ with $\alpha = 3\tau_3$, $\beta = \tau_3\|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\|^2$, and $x = \|\mathbb{D}^{-1}\mathbb{F}\Delta\| \in (0, 1/\alpha)$ implies $x \leq \beta/(2 - \alpha\beta)$, this yields

$$\|\mathbb{D}^{-1}\mathbb{F}(\mathbf{v}^\circ - \mathbf{v}^* - \mathbb{F}^{-1}\mathbf{A})\| \leq \frac{\tau_3}{2 - 3\tau_3^2\|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\|^2}\|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\|^2$$

and (21) follows by $\tau_3\|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\| \leq 4/9$.

Lemma

Assume $(\mathcal{T}_3^*)$ at $\mathbf{v}$. Let $\mathcal{U}_r = \{\mathbf{u}: \|\mathbb{D}\mathbf{u}\| \leq r\}$. Then

$$\|\mathbb{D}^{-1}\{\nabla f(\mathbf{v} + \mathbf{u}) - \nabla f(\mathbf{v}) - \langle \nabla^2 f(\mathbf{v}), \mathbf{u} \rangle\}\| \leq \frac{\tau_3}{2} \|\mathbb{D}\mathbf{u}\|^2, \quad \mathbf{u} \in \mathcal{U}_r. \quad (27)$$

Also for all $\mathbf{u}, \mathbf{u}_1 \in \mathcal{U}_r$

$$\|\mathbb{D}^{-1}\{\nabla^2 f(\mathbf{v} + \mathbf{u}_1) - \nabla^2 f(\mathbf{v} + \mathbf{u})\}\mathbb{D}^{-1}\| \leq \tau_3 \|\mathbb{D}(\mathbf{u}_1 - \mathbf{u})\| \quad (28)$$

$$\|\mathbb{D}^{-1}\{\nabla f(\mathbf{v} + \mathbf{u}_1) - \nabla f(\mathbf{v} + \mathbf{u}) - \nabla^2 f(\mathbf{v})(\mathbf{u}_1 - \mathbf{u})\}\| \leq \frac{3\tau_3}{2} \|\mathbb{D}(\mathbf{u}_1 - \mathbf{u})\|^2. \quad (29)$$

Moreover, under $(\mathcal{T}_4^*)$, for any $\mathbf{u} \in \mathcal{U}_r$,

$$\|\mathbb{D}^{-1}\{\nabla f(\mathbf{v} + \mathbf{u}) - \nabla f(\mathbf{v}) - \langle \nabla^2 f(\mathbf{v}), \mathbf{u} \rangle - \frac{1}{2}\langle \nabla^3 f(\mathbf{v}), \mathbf{u}^{\otimes 2} \rangle\}\| \leq \frac{\tau_4}{6} \|\mathbb{D}\mathbf{u}\|^3. \quad (30)$$

Denote

$$\mathbf{A} \stackrel{\text{def}}{=} \nabla f(\mathbf{v} + \mathbf{u}) - \nabla f(\mathbf{v}) - \langle \nabla^2 f(\mathbf{v}), \mathbf{u} \rangle.$$

For any vector $\mathbf{w} \in \mathbb{R}^p$, $(\mathcal{T}_3^*)$ and (12) imply

$$|\langle \mathbf{A}, \mathbf{w} \rangle| \leq \frac{\tau_3}{2} \|\mathbb{D}\mathbf{u}\|^2 \|\mathbb{D}\mathbf{w}\|.$$

Therefore,

$$\|\mathbb{D}^{-1}\mathbf{A}\| = \sup_{\|\mathbf{w}\|=1} |\langle \mathbb{D}^{-1}\mathbf{A}, \mathbf{w} \rangle| = \sup_{\|\mathbf{w}\|=1} |\langle \mathbf{A}, \mathbb{D}^{-1}\mathbf{w} \rangle| \leq \frac{\tau_3}{2} \|\mathbb{D}\mathbf{u}\|^2$$

which yields the first statement.

For (30), apply

$$\mathbf{A} \stackrel{\text{def}}{=} \nabla f(\mathbf{v} + \mathbf{u}) - \nabla f(\mathbf{v}) - \langle \nabla^2 f(\mathbf{v}), \mathbf{u} \rangle - \frac{1}{2} \langle \nabla^3 f(\mathbf{v}), \mathbf{u}^{\otimes 2} \rangle$$

and use $(\mathcal{T}_4^*)$ and (13) instead of $(\mathcal{T}_3^*)$ and (12).

Further, with $\mathcal{B}_1 \stackrel{\text{def}}{=} \nabla^2 f(\mathbf{v} + \mathbf{u}_1) - \nabla^2 f(\mathbf{v} + \mathbf{u})$ and $\Delta = \mathbf{u}_1 - \mathbf{u}$, by $(\mathcal{T}_3^*)$, for any $\mathbf{w} \in \mathbb{R}^p$ and some $t \in [0, 1]$,

$$\begin{aligned} |\langle \mathbb{D}^{-1} \{ \nabla^2 f(\mathbf{v} + \mathbf{u}_1) - \nabla^2 f(\mathbf{v} + \mathbf{u}) \} \mathbb{D}^{-1}, \mathbf{w}^{\otimes 2} \rangle| &= |\langle \mathcal{B}_1, (\mathbb{D}^{-1} \mathbf{w})^{\otimes 2} \rangle| \\ &= |\langle \nabla^3 f(\mathbf{v} + \mathbf{u} + t\Delta), \Delta \otimes (\mathbb{D}^{-1} \mathbf{w})^{\otimes 2} \rangle| \leq \tau_3 \|\mathbb{D} \Delta\| \|\mathbf{w}\|^2. \end{aligned}$$

This proves (28). Similarly, for some $t \in [0, 1]$

$$\begin{aligned} |\langle \mathbb{D}^{-1} \{ \nabla f(\mathbf{v} + \mathbf{u}_1) - \nabla f(\mathbf{v} + \mathbf{u}) \} - \nabla^2 f(\mathbf{v} + \mathbf{u}) \Delta, \mathbf{w} \rangle| \\ = \frac{1}{2} |\langle \nabla^3 f(\mathbf{v} + \mathbf{u} + t\Delta), \Delta \otimes \Delta \otimes \mathbb{D}^{-1} \mathbf{w} \rangle| \leq \frac{\tau_3}{2} \|\mathbb{D} \Delta\|^2 \|\mathbf{w}\| \end{aligned}$$

and with $\mathcal{B} = \nabla^2 f(\mathbf{v} + \mathbf{u}) - \nabla^2 f(\mathbf{v})$, by (28),

$$\|\mathbb{D}^{-1} \mathcal{B} \Delta\| \leq \|\mathbb{D}^{-1} \mathcal{B} \mathbb{D}^{-1}\| \|\mathbb{D} \Delta\| \leq \tau_3 \|\mathbb{D} \Delta\|^2.$$

This completes the proof of (29).

Lemma

If $2x \leq \alpha x^2 + \beta$ and $x \in (0, 1/\alpha)$ for $\alpha\beta \leq 1$, then

$$x \leq \frac{\beta}{2 - \alpha\beta}.$$

The roots of $\alpha x^2 + \beta = 2x$ satisfy

$$x = \frac{1 \pm \sqrt{1 - \alpha\beta}}{\alpha}$$

As $x \leq 1/\alpha$, we only consider

$$x \leq \frac{1 - \sqrt{1 - \alpha\beta}}{\alpha} = \frac{\alpha\beta}{\alpha(1 + \sqrt{1 - \alpha\beta})} \leq \frac{\beta}{1 + 1 - \alpha\beta}.$$

Lemma

Assume $\mathbb{D}^2 \leq \mathbb{F}$ and let some other matrix $\mathbb{F}_1 \in \mathfrak{M}_p$ satisfy

$$\|\mathbb{D}^{-1} (\mathbb{F}_1 - \mathbb{F}) \mathbb{D}^{-1}\| \leq \omega \quad (31)$$

with $\omega < 1$. Then for any vector $\mathbf{u}$

$$\|\mathbb{F}^{-1/2} (\mathbb{F}_1 - \mathbb{F}) \mathbb{F}^{-1/2}\| \leq \omega, \quad (32)$$

$$\|\mathbb{F}^{1/2} (\mathbb{F}_1^{-1} - \mathbb{F}^{-1}) \mathbb{F}^{1/2}\| \leq \frac{\omega}{1 - \omega}, \quad (33)$$

$$\frac{1}{1 + \omega} \|\mathbb{D} \mathbb{F}^{-1} \mathbb{D}\| \leq \|\mathbb{D} \mathbb{F}_1^{-1} \mathbb{D}\| \leq \frac{1}{1 - \omega} \|\mathbb{D} \mathbb{F}^{-1} \mathbb{D}\|, \quad (34)$$

$$(1 - \omega) \|\mathbb{D}^{-1} \mathbb{F} \mathbf{u}\| \leq \|\mathbb{D}^{-1} \mathbb{F}_1 \mathbf{u}\| \leq (1 + \omega) \|\mathbb{D}^{-1} \mathbb{F} \mathbf{u}\|, \quad (35)$$

$$\frac{1 - 2\omega}{1 - \omega} \|\mathbb{D} \mathbb{F}^{-1} \mathbf{u}\| \leq \|\mathbb{D} \mathbb{F}_1^{-1} \mathbf{u}\| \leq \frac{1}{1 - \omega} \|\mathbb{D} \mathbb{F}^{-1} \mathbf{u}\|. \quad (36)$$

Statement (32) follows from (31) because of $\mathbb{F}^{-1} \leq \mathbb{D}^{-2}$. Define now $\mathbf{U} \stackrel{\text{def}}{=} \mathbb{F}^{-1/2} (\mathbb{F}_1 - \mathbb{F}) \mathbb{F}^{-1/2}$. Then $\|\mathbf{U}\| \leq \omega$ and

$$\|\mathbb{F}^{1/2} (\mathbb{F}_1^{-1} - \mathbb{F}^{-1}) \mathbb{F}^{1/2}\| = \|(\mathbf{I} + \mathbf{U})^{-1} - \mathbf{I}\| \leq \frac{1}{1 - \omega} \|\mathbf{U}\|$$

yielding (33). Further,

$$\begin{aligned} \|\mathbb{D} (\mathbb{F}_1^{-1} - \mathbb{F}^{-1}) \mathbb{D}\| &= \|\mathbb{D} \mathbb{F}_1^{-1} \mathbb{F}_1 (\mathbb{F}_1^{-1} - \mathbb{F}^{-1}) \mathbb{F} \mathbb{F}^{-1} \mathbb{D}\| \\ &= \|\mathbb{D} \mathbb{F}_1^{-1} \mathbb{D} \mathbb{D}^{-1} (\mathbb{F}_1 - \mathbb{F}) \mathbb{D}^{-1} \mathbb{D} \mathbb{F}^{-1} \mathbb{D}\| \\ &\leq \|\mathbb{D} \mathbb{F}_1^{-1} \mathbb{D}\| \|\mathbb{D} \mathbb{F}^{-1} \mathbb{D}\| \|\mathbb{D}^{-1} (\mathbb{F}_1 - \mathbb{F}) \mathbb{D}^{-1}\| \leq \omega \|\mathbb{D} \mathbb{F}_1^{-1} \mathbb{D}\|. \end{aligned}$$

This implies (34).

Also, by $\mathbb{D}^2 \leq \mathbb{F}$

$$\begin{aligned}\|\mathbb{D}^{-1}\mathbb{F}_1\mathbf{u}\| &\leq \|\mathbb{D}^{-1}\mathbb{F}\mathbf{u}\| + \|\mathbb{D}^{-1}(\mathbb{F}_1 - \mathbb{F})\mathbb{D}^{-1}\mathbb{D}\mathbf{u}\| \leq \|\mathbb{D}^{-1}\mathbb{F}\mathbf{u}\| + \omega \|\mathbb{D}\mathbf{u}\| \\ &\leq \|\mathbb{D}^{-1}\mathbb{F}\mathbf{u}\| + \omega \|\mathbb{D}^{-1}\mathbb{F}\mathbf{u}\| \leq (1 + \omega) \|\mathbb{D}^{-1}\mathbb{F}\mathbf{u}\|,\end{aligned}$$

and (35) follows. Similarly

$$\begin{aligned}\|\mathbb{D}(\mathbb{F}_1^{-1} - \mathbb{F}^{-1})\mathbf{u}\| &= \|\mathbb{D}\mathbb{F}_1^{-1}(\mathbb{F}_1 - \mathbb{F})\mathbb{F}^{-1}\mathbf{u}\| \\ &= \|\mathbb{D}\mathbb{F}_1^{-1}\mathbb{D}\mathbb{D}^{-1}(\mathbb{F}_1 - \mathbb{F})\mathbb{D}^{-1}\mathbb{D}\mathbb{F}^{-1}\mathbf{u}\| \\ &\leq \|\mathbb{D}^{-1}(\mathbb{F}_1 - \mathbb{F})\mathbb{D}^{-1}\| \|\mathbb{D}\mathbb{F}_1^{-1}\mathbb{D}\| \|\mathbb{D}\mathbb{F}^{-1}\mathbf{u}\| \\ &\leq \frac{\omega}{1 - \omega} \|\mathbb{D}\mathbb{F}^{-1}\mathbf{u}\|\end{aligned}$$

and (36) follows as well.

1 Introduction

2 Linearly perturbed optimization

- Quadratic case
- 2S expansions
- Linear perturbation under third order smoothness
- Uniform smoothness

3 Fourth order approximation

4 Quadratic penalization

Proposition

Let $f(\mathbf{v})$ be a strongly concave function with $f(\mathbf{v}^*) = \max_{\mathbf{v}} f(\mathbf{v})$ and $\mathbb{F} = -\nabla^2 f(\mathbf{v}^*)$, and let $f(\mathbf{v})$ follow $(\mathcal{T}_3^*)$ and $(\mathcal{T}_4^*)$ with some $\mathbb{D}^2$, τ_3 , τ_4 , and $\mathbf{r}$ satisfying

$$\mathbb{D}^2 \leq \mathbb{F}, \quad \mathbf{r} = \frac{3}{2} \|\mathbb{D} \mathbb{F}^{-1} \mathbf{A}\|, \quad \tau_3 \|\mathbb{D} \mathbb{F}^{-1} \mathbf{A}\| < \frac{4}{9}, \quad \tau_4 \|\mathbb{D} \mathbb{F}^{-1} \mathbf{A}\|^2 < \frac{1}{3}. \quad (37)$$

Let $g(\mathbf{v})$ fulfill (1) with some vector $\mathbf{A}$ and $g(\mathbf{v}^\circ) = \max_{\mathbf{v}} g(\mathbf{v})$. Then $\|\mathbb{D}(\mathbf{v}^\circ - \mathbf{v}^*)\| \leq (3/2) \|\mathbb{D} \mathbb{F}^{-1} \mathbf{A}\|$. Further, define

$$\mathbf{a} = \mathbb{F}^{-1} \{ \mathbf{A} + \nabla \mathcal{T}(\mathbb{F}^{-1} \mathbf{A}) \}, \quad (38)$$

where $\mathcal{T}(\mathbf{u}) = \frac{1}{6} \langle \nabla^3 f(\mathbf{v}^*), \mathbf{u}^{\otimes 3} \rangle$ for $\mathbf{u} \in \mathbb{R}^p$. Then

$$\|\mathbb{D}^{-1} \mathbb{F}(\mathbf{v}^\circ - \mathbf{v}^* - \mathbf{a})\| \leq (\tau_4/2 + \tau_3^2) \|\mathbb{D} \mathbb{F}^{-1} \mathbf{A}\|^3. \quad (39)$$

Proposition 5 yields (21). By $(\mathcal{T}_3^*)$

$$\begin{aligned} \|\mathbb{D}^{-1} \mathbb{F}(\mathbf{a} - \mathbb{F}^{-1} \mathbf{A})\| &= \|\mathbb{D}^{-1} \nabla \mathcal{T}(\mathbb{F}^{-1} \mathbf{A})\| \\ &= \sup_{\|\mathbf{u}\|=1} 3 \left| \langle \mathcal{T}, \mathbb{F}^{-1} \mathbf{A} \otimes \mathbb{F}^{-1} \mathbf{A} \otimes \mathbb{D}^{-1} \mathbf{u} \rangle \right| \leq \frac{\tau_3}{2} \|\mathbb{D} \mathbb{F}^{-1} \mathbf{A}\|^2. \end{aligned} \quad (40)$$

As $\mathbb{D}^{-1} \mathbb{F} \geq \mathbb{F}^{1/2} \geq \mathbb{D}$, this implies by $\tau_3 \|\mathbb{D} \mathbb{F}^{-1} \mathbf{A}\| \leq 4/9$

$$\begin{aligned} \|\mathbb{D} \mathbf{a}\| &\leq \|\mathbb{D} \mathbb{F}^{-1} \mathbf{A}\| + \|\mathbb{D} \mathbb{F}^{-1} \nabla \mathcal{T}(\mathbb{F}^{-1} \mathbf{A})\| \\ &\leq \left(1 + \frac{\tau_3}{2} \|\mathbb{D} \mathbb{F}^{-1} \mathbf{A}\|\right) \|\mathbb{D} \mathbb{F}^{-1} \mathbf{A}\| \leq \frac{11}{9} \|\mathbb{D} \mathbb{F}^{-1} \mathbf{A}\| \end{aligned} \quad (41)$$

and

$$\|\mathbb{F}^{1/2} \mathbf{a} - \mathbb{F}^{-1/2} \mathbf{A}\| \leq \frac{\tau_3}{2} \|\mathbb{D} \mathbb{F}^{-1} \mathbf{A}\|^2.$$

Next, again by $(\mathcal{T}_3^*)$, for any $\mathbf{w}$

$$\|\mathbb{D}^{-1} \nabla^2 \mathcal{T}(\mathbf{w}) \mathbb{D}^{-1}\| = \sup_{\|\mathbf{u}\|=1} 6 |\langle \mathcal{T}, \mathbf{w} \otimes (\mathbb{D}^{-1} \mathbf{u})^{\otimes 2} \rangle| \leq \tau_3 \|\mathbb{D} \mathbf{w}\|.$$

The tensor $\nabla^2 \mathcal{T}(\mathbf{u})$ is linear in $\mathbf{u}$, hence for any $t \in [0, 1]$

$$\begin{aligned} & \|\mathbb{D}^{-1} \nabla^2 \mathcal{T}(t\mathbf{a} + (1-t)\mathbb{F}^{-1}\mathbf{A}) \mathbb{D}^{-1}\| \\ & \leq \max\{\|\mathbb{D}^{-1} \nabla^2 \mathcal{T}(\mathbb{F}^{-1}\mathbf{A}) \mathbb{D}^{-1}\|, \|\mathbb{D}^{-1} \nabla^2 \mathcal{T}(\mathbf{a}) \mathbb{D}^{-1}\|\} \\ & \leq \tau_3 \max\{\|\mathbb{D} \mathbb{F}^{-1}\mathbf{A}\|, \|\mathbb{D} \mathbf{a}\|\}. \end{aligned}$$

Based on (41), assume $\|\mathbb{D} \mathbb{F}^{-1}\mathbf{A}\| \leq \|\mathbb{D} \mathbf{a}\| \leq (11/9)\|\mathbb{D} \mathbb{F}^{-1}\mathbf{A}\|$. Then (40) yield

$$\begin{aligned} & \|\mathbb{D}^{-1} \nabla \mathcal{T}(\mathbf{a}) - \mathbb{D}^{-1} \nabla \mathcal{T}(\mathbb{F}^{-1}\mathbf{A})\| \\ & = \mathbb{D}^{-1} \nabla^2 \mathcal{T}(t\mathbf{a} + (1-t)\mathbb{F}^{-1}\mathbf{A}) \mathbb{D}^{-1} \|\mathbb{D} \mathbb{F}^{-1}(\mathbf{a} - \mathbb{F}^{-1}\mathbf{A})\| \\ & \leq \frac{\tau_3^2}{2} \|\mathbb{D} \mathbb{F}^{-1}\mathbf{A}\|^2 \|\mathbb{D} \mathbf{a}\| \leq \frac{2\tau_3^2}{3} \|\mathbb{D} \mathbb{F}^{-1}\mathbf{A}\|^3. \end{aligned}$$

Further, $-\nabla^2 f(0) = \mathbb{F}$, $\nabla \mathcal{T}(\mathbf{a}) = \frac{1}{2} \langle \nabla^3 f(0), \mathbf{a} \otimes \mathbf{a} \rangle$. By (30) and (41)

$$\begin{aligned} \|\mathbb{D}^{-1}\{\nabla f(\mathbf{a}) + \mathbb{F}\mathbf{a} - \nabla \mathcal{T}(\mathbf{a})\}\| &\leq \frac{\tau_4}{6} \|\mathbb{D}\mathbf{a}\|^3 \leq \frac{(11/9)^3 \tau_4}{6} \|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\|^3 \\ &\leq \frac{\tau_4}{3} \|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\|^3. \end{aligned}$$

Next we bound $\|\mathbb{D}^{-1}\{\nabla g(\mathbf{a}) - \nabla g(\mathbf{v}^\circ)\}\|$. As $\nabla g(\mathbf{v}^\circ) = 0$, (1) and (38) imply

$$\begin{aligned} \|\mathbb{D}^{-1}\{\nabla g(\mathbf{a}) - \nabla g(\mathbf{v}^\circ)\}\| &= \|\mathbb{D}^{-1}\nabla g(\mathbf{a})\| = \|\mathbb{D}^{-1}\{\nabla g(\mathbf{a}) + \mathbb{F}\mathbf{a} - \nabla \mathcal{T}(\mathbf{A}) - \mathbf{A}\}\| \\ &\leq \|\mathbb{D}^{-1}\{\nabla f(\mathbf{a}) + \mathbb{F}\mathbf{a} - \nabla \mathcal{T}(\mathbf{a})\}\| + \|\mathbb{D}^{-1}\{\nabla \mathcal{T}(\mathbf{a}) - \nabla \mathcal{T}(\mathbf{A})\}\| \leq \diamond_1, \end{aligned} \quad (42)$$

where $\diamond_1 \stackrel{\text{def}}{=} \frac{\tau_4 + 2\tau_3^2}{3} \|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\|^3$, and by (37)

$$3\tau_3 \diamond_1 = \tau_3 \|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\| \tau_4 \|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\|^2 + 2\tau_3^3 \|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\|^3 < \frac{1}{3}. \quad (43)$$

Further, $\nabla^2 g(0) = \nabla^2 f(0) = -\mathbb{F}$, and (29) of Lemma 4 implies

$$\begin{aligned} & \|\mathbb{D}^{-1}\{\nabla g(\mathbf{a}) - \nabla g(\mathbf{v}^\circ) + \mathbb{F}(\mathbf{a} - \mathbf{v}^\circ)\}\| \\ &= \|\mathbb{D}^{-1}\{\nabla f(\mathbf{a}) - \nabla f(\mathbf{v}^\circ) + \mathbb{F}(\mathbf{a} - \mathbf{v}^\circ)\}\| \leq \frac{3\tau_3}{2} \|\mathbb{D}(\mathbf{a} - \mathbf{v}^\circ)\|^2. \end{aligned}$$

Combining with (42) yields in view of $\mathbb{D}^2 \leq \mathbb{F}$

$$\|\mathbb{D}^{-1}\mathbb{F}(\mathbf{a} - \mathbf{v}^\circ)\| \leq \frac{3\tau_3}{2} \|\mathbb{D}(\mathbf{a} - \mathbf{v}^\circ)\|^2 + \diamond_1 \leq \frac{3\tau_3}{2} \|\mathbb{D}^{-1}\mathbb{F}(\mathbf{a} - \mathbf{v}^\circ)\|^2 + \diamond_1.$$

As $2x \leq \alpha x^2 + \beta$ with $\alpha = 3\tau_3$, $\beta = 2\diamond_1$, and $x \in (0, 1/\alpha)$ implies $x \leq \beta/(2 - \alpha\beta)$, we conclude by (43)

$$\|\mathbb{D}^{-1}\mathbb{F}(\mathbf{a} - \mathbf{v}^\circ)\| \leq \frac{\diamond_1}{1 - 3\tau_3 \diamond_1} \leq \frac{\tau_4 + 2\tau_3^2}{2} \|\mathbb{D}\mathbb{F}^{-1}\mathbf{A}\|^3,$$

and (39) follows.

1 Introduction

2 Linearly perturbed optimization

- Quadratic case
- 2S expansions
- Linear perturbation under third order smoothness
- Uniform smoothness

3 Fourth order approximation

4 Quadratic penalization

Here we discuss the case when $g(\mathbf{v}) - f(\mathbf{v})$ is **quadratic**.

The general case can be reduced to the situation with $g(\mathbf{v}) = f(\mathbf{v}) - \|G\mathbf{v}\|^2/2$. To make the dependence of G more explicit, denote $f_G(\mathbf{v}) \stackrel{\text{def}}{=} f(\mathbf{v}) - \|G\mathbf{v}\|^2/2$,

$$\mathbf{v}^* = \operatorname{argmax}_{\mathbf{v}} f(\mathbf{v}),$$

$$\mathbf{v}_G^* = \operatorname{argmax}_{\mathbf{v}} f_G(\mathbf{v}) = \operatorname{argmax}_{\mathbf{v}} \{f(\mathbf{v}) - \|G\mathbf{v}\|^2/2\}.$$

We study the bias $\mathbf{v}_G^* - \mathbf{v}^*$ induced by this penalization.

Lemma

Let $f(\mathbf{v})$ be *quadratic* with $\mathbb{F} \equiv -\nabla^2 f(\mathbf{v})$. Define

$$\mathbf{S}_G \equiv -G^2 \mathbf{v}^*.$$

Then it holds with $\mathbb{F}_G = \mathbb{F} + G^2$

$$\mathbf{v}^* - \mathbf{v}_G^* = \mathbb{F}_G^{-1} \mathbf{S}_G = -\mathbb{F}_G^{-1} G^2 \mathbf{v}^*,$$

$$f_G(\mathbf{v}_G^*) - f_G(\mathbf{v}^*) = \frac{1}{2} \|\mathbb{F}_G^{-1/2} \mathbf{S}_G\|^2 = \frac{1}{2} \|\mathbb{F}_G^{-1/2} G^2 \mathbf{v}^*\|^2.$$

Quadraticity of $f(\mathbf{v})$ implies quadraticity of $f_G(\mathbf{v})$ with $\nabla^2 f_G(\mathbf{v}) \equiv -\mathbb{F}_G$ and

$$\nabla f_G(\mathbf{v}^*) - \nabla f_G(\mathbf{v}_G^*) = \mathbb{F}_G (\mathbf{v}_G^* - \mathbf{v}^*).$$

Further, $\nabla f(\mathbf{v}^*) = 0$ yielding $\nabla f_G(\mathbf{v}^*) = \mathbf{S}_G = -G^2 \mathbf{v}^*$. Together with $\nabla f_G(\mathbf{v}_G^*) = 0$, this implies

$$\mathbf{v}^* - \mathbf{v}_G^* = \mathbb{F}_G^{-1} \mathbf{S}_G.$$

The Taylor expansion of f_G at $\mathbf{v}_G^*$ yields

$$f_G(\mathbf{v}^*) - f_G(\mathbf{v}_G^*) = -\frac{1}{2} \|\mathbb{F}_G^{1/2} (\mathbf{v}^* - \mathbf{v}_G^*)\|^2 = -\frac{1}{2} \|\mathbb{F}_G^{-1/2} \mathbf{S}_G\|^2$$

and the assertion follows.

Proposition

Let $f_G(\mathbf{v}) = f(\mathbf{v}) - \|G\mathbf{v}\|^2/2$ be concave and follow $(\mathcal{T}_3^*)$ with some $\mathbb{D}^2$, τ_3 , and $\mathbf{r}$ satisfying

$$\mathbb{D}^2 \leq \mathbb{F}_G, \quad \mathbf{r} \geq 3\mathbf{b}_G/2, \quad \tau_3 \mathbf{b}_G < 4/9,$$

where $\mathbf{b}_G = \|\mathbb{D} \mathbb{F}_G^{-1} G^2 \mathbf{v}^*\|$. Then

$$\|\mathbb{D}(\mathbf{v}_G^* - \mathbf{v}^*)\| \leq 3\mathbf{b}_G/2.$$

Moreover,

$$\|\mathbb{D}^{-1} \mathbb{F}_G(\mathbf{v}_G^* - \mathbf{v}^* + \mathbb{F}_G^{-1} G^2 \mathbf{v}^*)\| \leq \frac{3\tau_3}{4} \mathbf{b}_G^2,$$

$$\left| 2f_G(\mathbf{v}_G^*) - 2f_G(\mathbf{v}^*) - \frac{1}{2} \|\mathbb{F}_G^{-1/2} G^2 \mathbf{v}^*\|^2 \right| \leq \frac{\tau_3}{2} \mathbf{b}_G^3.$$

Define $g_G(\mathbf{v})$ by

$$g_G(\mathbf{v}) - g_G(\mathbf{v}_G^*) = f_G(\mathbf{v}) - f_G(\mathbf{v}_G^*) + \langle G^2 \mathbf{v}^*, \mathbf{v} - \mathbf{v}_G^* \rangle. \quad (44)$$

The function f_G is concave, the same holds for g_G from (44).

Hence, $\nabla g_G(\mathbf{v}^*) = 0$ implies $\mathbf{v}^* = \operatorname{argmax} g_G(\mathbf{v})$. By definition, $\nabla f(\mathbf{v}^*) = 0$ yielding $\nabla f_G(\mathbf{v}^*) = -G^2 \mathbf{v}^* + G^2 \mathbf{v}^* = 0$.

Now the results follow from Propositions 5 and 3 applied with $f(\mathbf{v}) = g_G(\mathbf{v}) = f_G(\mathbf{v}) - \langle \mathbf{A}, \mathbf{v} \rangle$, $g(\mathbf{v}) = f_G(\mathbf{v})$, and $\mathbf{A} = G^2 \mathbf{v}^*$.

Define $\mathbb{F}_G = -\nabla^2 f(\mathbf{v}^*) + G^2$, $\mathbf{S}_G = G^2 \mathbf{v}^*$, and

$$\mathbf{m}_G = \mathbb{F}_G^{-1} \{ \mathbf{S}_G + \nabla \mathcal{T}(\mathbb{F}_G^{-1} \mathbf{S}_G) \}$$

with $\mathcal{T}(\mathbf{u}) = \frac{1}{6} \langle \nabla^3 f(\mathbf{v}^*), \mathbf{u}^{\otimes 3} \rangle$.

$(\mathcal{T}_4^*)$ $f(\mathbf{v})$ is strongly concave, $\mathbb{D}^2 \leq \nabla^2 f(\mathbf{v})$, and

$$\sup_{\mathbf{u}: \|\mathbb{D}\mathbf{u}\| \leq r} \sup_{\mathbf{z} \in \mathbb{R}^p} \frac{|\langle \nabla^4 f(\mathbf{v} + \mathbf{u}), \mathbf{z}^{\otimes 4} \rangle|}{\|\mathbb{D}\mathbf{z}\|^4} \leq \tau_4.$$

Typically $\tau_3 \asymp n^{-1/2}$ and $\tau_4 \asymp n^{-1}$.

Proposition

Let f be concave and $\mathbf{v}^* = \operatorname{argmax}_{\mathbf{v}} f(\mathbf{v})$. With $\mathbb{F}_G = -\nabla^2 f(\mathbf{v}^*) + G^2$. Let $f(\mathbf{v})$ follow $(\mathcal{T}_3^*)$ and $(\mathcal{T}_4^*)$ with some $\mathbb{D}^2$, τ_3 , τ_4 , and $\mathbf{r}$ satisfying

$$\mathbb{D}^2 \leq \mathbb{F}_G, \quad \mathbf{r} = \frac{3}{2} \mathbf{b}_G, \quad \tau_3 \mathbf{b}_G < \frac{4}{9}, \quad \tau_4 \mathbf{b}_G^2 < \frac{1}{3}.$$

with $\mathbf{b}_G = \|\mathbb{D} \mathbb{F}_G^{-1} G^2 \mathbf{v}^*\|$. Define

$$\mathbf{m}_G = \mathbb{F}_G^{-1} \{G^2 \mathbf{v}^* + \nabla \mathcal{T}(\mathbb{F}_G^{-1} G^2 \mathbf{v}^*)\}$$

with $\mathcal{T}(\mathbf{u}) = \frac{1}{6} \langle \nabla^3 f(\mathbf{v}^*), \mathbf{u}^{\otimes 3} \rangle$ and $\nabla \mathcal{T} = \frac{1}{2} \langle \nabla^3 f(\mathbf{v}^*), \mathbf{u}^{\otimes 2} \rangle$. Then

$$\|\mathbb{D}^{-1} \mathbb{F}_G(\mathbf{v}^* - \mathbf{v}_G^* - \mathbf{m}_G)\| \leq \frac{\tau_4 + 2\tau_3^2}{2} \mathbf{b}_G^3.$$

- Statistical inference for nonlinear regression. DNN training.
- Gaussian variational inference
- Bayesian optimization



Banach, S. (1938).

Über homogene Polynome in (L^2) .
Studia Mathematica, 7(1):36–44.



Katsevich, A. and Rigollet, P. (2023).

On the approximation accuracy of gaussian variational inference.



Weierstraß-Institut für
Angewandte Analysis und Stochastik



Linearly perturbed optimization: theory and applications

Vladimir Spokoiny ,
WIAS, HU Berlin

8. Oktober 2024

1 Statistical inference

- Linear and SLS models
- Nonlinear regression. Theoretical study

2 Structural modeling: Examples

- Matrix completion
- Gaussian mixture
- Deep Neural Networks

3 Gaussian Variational Inference

4 Optimization vs sampling

Let $L(\mathbf{v})$ be a random function, $\mathbf{v} \in \mathcal{Y} \subseteq \mathbb{R}^p$, $p < \infty$.

Given a quadratic penalty $\|G\mathbf{v}\|^2/2$, define

$$L_G(\mathbf{v}) = L(\mathbf{v}) - \|G\mathbf{v}\|^2/2.$$

Consider

$$\tilde{\mathbf{v}}_G = \operatorname{argmax}_{\mathbf{v}} L_G(\mathbf{v}) = \operatorname{argmax}_{\mathbf{v}} \left\{ L(\mathbf{v}) - \frac{1}{2} \|G\mathbf{v}\|^2 \right\};$$

$$\mathbf{v}_G^* = \operatorname{argmax}_{\mathbf{v}} \mathbb{E} L_G(\mathbf{v}) = \operatorname{argmax}_{\mathbf{v}} \left\{ \mathbb{E} L(\mathbf{v}) - \frac{1}{2} \|G\mathbf{v}\|^2 \right\};$$

$$\mathbf{v}^* = \operatorname{argmax}_{\mathbf{v}} \mathbb{E} L(\mathbf{v});$$

Aim: describe the estimation loss $\tilde{\mathbf{v}}_G - \mathbf{v}^*$ and the prediction loss (excess) $L_G(\tilde{\mathbf{v}}_G) - L_G(\mathbf{v}^*)$.

- A linear model $\mathbf{Y} = \mathbf{\Psi}^\top \mathbf{v} + \boldsymbol{\varepsilon}$
- a quadratic penalty $\text{pen}_G(\mathbf{v}) = \|\mathbf{G}\mathbf{v}\|^2/2$.
- Penalized MLE: with $\mathbb{F}_G \stackrel{\text{def}}{=} \mathbf{\Psi}\mathbf{\Psi}^\top + \mathbf{G}^2$

$$\tilde{\mathbf{v}}_G = \mathbb{F}_G^{-1} \mathbf{\Psi} \mathbf{Y},$$

$$2L_G(\tilde{\mathbf{v}}_G) - 2L_G(\mathbf{v}_G^*) = \|\mathbb{F}_G^{-1/2} \mathbf{\Psi} \boldsymbol{\varepsilon}\|^2.$$

$$\text{where } \mathbf{v}_G^* = \mathbb{F}_G^{-1} \mathbf{\Psi} \mathbb{E} \mathbf{Y}.$$

Loss, bias-variance decomposition:

$$\tilde{\mathbf{v}}_G - \mathbf{v}^* = \mathbb{F}_G^{-1} \mathbf{\Psi} \boldsymbol{\varepsilon} + \mathbb{F}_G^{-1} \mathbf{G}^2 \mathbf{v}^*,$$

$$\mathbb{E} \left\| Q(\tilde{\mathbf{v}}_G - \mathbf{v}^*) \right\|^2 = \text{tr} \{ Q \mathbb{F}_G^{-1} \text{Var}(\mathbf{\Psi} \boldsymbol{\varepsilon}) \mathbb{F}_G^{-1} Q^\top \} + \| Q \mathbb{F}_G^{-1} \mathbf{G}^2 \mathbf{v}^* \|^2.$$

Stochastic component $\zeta(\boldsymbol{v}) \stackrel{\text{def}}{=} L(\boldsymbol{v}) - \mathbb{E}L(\boldsymbol{v})$ is **linear** in $\boldsymbol{v}$:

$$\nabla \zeta \stackrel{\text{def}}{=} \nabla L(\boldsymbol{v});$$

The function $f(\boldsymbol{v}) = \mathbb{E}L(\boldsymbol{v})$ is **smooth** and **concave** in $\boldsymbol{v}$.

Consider

$$\tilde{\boldsymbol{v}}_G = \underset{\boldsymbol{v}}{\operatorname{argmax}} L_G(\boldsymbol{v}) = \underset{\boldsymbol{v}}{\operatorname{argmax}} \left\{ L(\boldsymbol{v}) - \frac{1}{2} \|G\boldsymbol{v}\|^2 \right\};$$

$$\boldsymbol{v}_G^* = \underset{\boldsymbol{v}}{\operatorname{argmax}} \mathbb{E} L_G(\boldsymbol{v}) = \underset{\boldsymbol{v}}{\operatorname{argmax}} \left\{ \mathbb{E} L(\boldsymbol{v}) - \frac{1}{2} \|G\boldsymbol{v}\|^2 \right\};$$

$$\boldsymbol{v}^* = \underset{\boldsymbol{v}}{\operatorname{argmax}} \mathbb{E} L(\boldsymbol{v});$$

$(\mathcal{C}_G)$ The function $\mathbb{E}L_G(\mathbf{v})$ is concave on $\mathcal{V}$ which is open and convex set in $\mathbb{R}^p$.

(ζ) The stochastic component $\zeta(\mathbf{v}) = L(\mathbf{v}) - \mathbb{E}L(\mathbf{v})$ is linear in $\mathbf{v}$, $\nabla\zeta \equiv \nabla\zeta(\mathbf{v}) \in \mathbb{R}^p$.

$f(\mathbf{v}) = \mathbb{E} L_G(\mathbf{v})$ is smooth: for $k = 3$ (and may be $k = 4$)

$(\mathcal{T}_3^*)$ $f(\mathbf{v})$ is strongly concave, $\mathbb{D}^2 \leq \nabla^2 f(\mathbf{v})$, and

$$\sup_{\mathbf{u}: \|\mathbb{D}\mathbf{u}\| \leq r} \sup_{\mathbf{z} \in \mathbb{R}^p} \frac{|\langle \nabla^3 f(\mathbf{v} + \mathbf{u}), \mathbf{z}^{\otimes k} \rangle|}{\|\mathbb{D}\mathbf{z}\|^3} \leq \tau_3.$$

Banach's characterization [Banach, 1938] yields for $k \geq 2$

$$|\langle \nabla^k f(\mathbf{v} + \mathbf{u}), \mathbf{z}_1 \otimes \dots \otimes \mathbf{z}_k \rangle| \leq \tau_k \|\mathbb{D}\mathbf{z}_1\| \dots \|\mathbb{D}\mathbf{z}_k\|.$$

If $f(\mathbf{v}) = \mathbb{E} L_G(\mathbf{v})$ scales with n , then the same holds for $\nabla^k f(\mathbf{v})$ and

$$\tau_3 \asymp n^{-1/2}, \quad \tau_4 \asymp n^{-1}.$$

By (ζ) , it holds for $\zeta(\mathbf{v}) = L(\mathbf{v}) - \mathbb{E}L(\mathbf{v})$

$$\nabla\zeta(\mathbf{v}) \equiv \nabla\zeta.$$

$(\nabla\zeta)$ *There exists $V^2 \geq \text{Var}(\nabla\zeta)$ s.t. $\boldsymbol{\xi} \stackrel{\text{def}}{=} V^{-1}\nabla\zeta$ satisfies for any considered $\mathbf{x} > 0$ and $B \in \mathfrak{M}_p$*

$$\mathbb{P}(\|B^{1/2}\boldsymbol{\xi}\| \geq z(B, \mathbf{x})) \leq 3e^{-\mathbf{x}},$$

$$z^2(B, \mathbf{x}) \stackrel{\text{def}}{=} \text{tr } B + 2\sqrt{\mathbf{x} \text{ tr } B^2} + 2\mathbf{x}\|B\|.$$

Alternative formulation: on $\Omega(\mathbf{x})$ with $\mathbb{P}(\Omega(\mathbf{x})) \geq 1 - 3e^{-\mathbf{x}}$

$$\|B^{1/2}\boldsymbol{\xi}\| \geq z(B, \mathbf{x}).$$

With the metric tensor D from $(\mathcal{T}_3^*)$, define

$$\mathbf{r}_D = z(B_D, \mathbf{x}), \quad B_D \stackrel{\text{def}}{=} \text{Var}(D\mathbb{F}_G^{-1}\nabla\zeta), \quad \mathbb{F}_G = \mathbb{F}_G(\mathbf{v}_G^*).$$

Theorem (Fisher and Wilks expansions)

Assume $(\mathcal{C}_G)$, (ζ) , $(\nabla\zeta)$, and $(\mathcal{T}_3^*)$ with D , $\mathbf{r}$, and τ_3 s.t.

$$D^2 \leq \mathbb{F}_G, \quad \mathbf{r} \geq \frac{3}{2} \mathbf{r}_D, \quad \tau_3 \mathbf{r}_D < \frac{4}{9},$$

Then on $\Omega(\mathbf{x})$

$$\|D^{-1}\mathbb{F}_G(\tilde{\mathbf{v}}_G - \mathbf{v}_G^* - \mathbb{F}_G^{-1}\nabla\zeta)\| \leq \frac{3\tau_3}{4} \|D\mathbb{F}_G^{-1}\nabla\zeta\|^2,$$

$$|2L_G(\tilde{\mathbf{v}}_G) - 2L_G(\mathbf{v}_G^*) - \|\mathbb{F}_G^{-1/2}\nabla\zeta\|^2| \leq \tau_3 \|D\mathbb{F}_G^{-1}\nabla\zeta\|^3.$$

Compare

$$\mathbf{v}_G^* = \operatorname{argmax} \left\{ \mathbb{E} L(\mathbf{v}) - \frac{1}{2} \|G\mathbf{v}\|^2 \right\}, \quad \mathbf{v}^* = \operatorname{argmax} \mathbb{E} L(\mathbf{v}).$$

Proposition

Let

$$\mathbf{b}_G \stackrel{\text{def}}{=} \|D\mathbb{F}_G^{-1}G^2\mathbf{v}^*\|.$$

Assume $(\mathcal{T}_3^*)$ with $\mathbf{r} = (3/2)\mathbf{b}_G$ and let $\tau_3 \mathbf{b}_G \leq 1/2$. Then

$$\|D^{-1}\mathbb{F}_G(\mathbf{v}_G^* - \mathbf{v}^* + \mathbb{F}_G^{-1}G^2\mathbf{v}^*)\| \leq \frac{3\tau_3}{4} \mathbf{b}_G^2.$$

Theorem

For any linear Q

$$\begin{aligned} & \|Q(\tilde{\mathbf{v}}_G - \mathbf{v}^* - \mathbb{F}_G^{-1} \nabla \zeta + \mathbb{F}_G^{-1} G^2 \mathbf{v}^*)\| \\ & \leq \|Q \mathbb{F}_G^{-1} D\| \frac{3\tau_3}{4} (\|D \mathbb{F}_G^{-1} \nabla \zeta\|^2 + \mathbf{b}_D^2) \end{aligned}$$

Fix $Q: \mathbb{R}^p \rightarrow \mathbb{R}^q$ and define

$$\mathbf{p}_D \stackrel{\text{def}}{=} \text{tr Var}(D\mathbf{F}_G^{-1}\nabla\zeta), \quad \mathbf{b}_D = \|D\mathbf{F}_G^{-1}G^2\mathbf{v}^*\|,$$

$$\mathbf{p}_Q \stackrel{\text{def}}{=} \text{tr Var}(Q\mathbf{F}_G^{-1}\nabla\zeta), \quad \mathbf{b}_Q = \|Q\mathbf{F}_G^{-1}G^2\mathbf{v}^*\|,$$

$$\mathcal{R}_Q \stackrel{\text{def}}{=} \mathbb{E}\{\|Q\mathbf{F}_G^{-1}(\nabla\zeta - G^2\mathbf{v}^*)\|^2 \mathbb{I}_{\Omega(\mathbf{x})}\} \leq \mathbf{p}_Q + \mathbf{b}_Q^2.$$

Theorem

$$\mathbb{E}\{\|Q(\tilde{\mathbf{v}}_G - \mathbf{v}^*)\| \mathbb{I}_{\Omega(\mathbf{x})}\} \leq \mathcal{R}_Q^{1/2} + \|Q\mathbf{F}_G^{-1}D\| \frac{3\tau_3}{4} (\mathbf{p}_D + \mathbf{b}_D^2),$$

$$(1 - \alpha_Q)^2 \mathcal{R}_Q \leq \mathbb{E}\{\|Q(\tilde{\mathbf{v}}_G - \mathbf{v}^*)\|^2 \mathbb{I}_{\Omega(\mathbf{x})}\} \leq (1 + \alpha_Q)^2 \mathcal{R}_Q$$

$$\text{provided } \alpha_Q \stackrel{\text{def}}{=} \frac{\|Q\mathbf{F}_G^{-1}D\| (3/4)\tau_3 (\mathbf{p}_D + \mathbf{b}_D^2)}{\sqrt{\mathcal{R}_Q}} < 1.$$

With $n = \lambda_{\min}(D^2)$, $Q = D = n^{1/2} \mathbb{I}_p$, and $\mathcal{R}_G = p_G + b_G^2$

$$\mathbb{E} \{ \|n^{1/2} (\tilde{\mathbf{v}}_G - \mathbf{v}^*)\|^2 \mathbb{I}_{\Omega(\mathbf{x})} \} = \mathcal{R}_G (1 \pm \tau_3 \sqrt{\mathcal{R}_G}).$$

A sharp bound under $\tau_3 \sqrt{p_G} \ll 1$ and $\tau_3 b_G \ll 1$.

Critical dimension: with $\tau_3 \asymp n^{-1/2}$

$$p_G \ll n.$$

Let observations $Y_1, \dots, Y_n$ follow the nonlinear regression model

$$Y_i = m(\mathbf{X}_i, \boldsymbol{\theta}) + \varepsilon_i$$

with independent zero mean errors ε_i .

Target parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ for p large/infinite.

Example in mind: $\boldsymbol{\theta}$ codes the architecture of a DNN.

Aim: estimation/inference on $\boldsymbol{\theta}$.

Least squares estimation (Gauss, Legendre):

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \| \mathbf{Y} - m(\mathbf{X}_i, \boldsymbol{\theta}) \|^2.$$

Problems: $L(\boldsymbol{\theta})$ is **not concave**, the gradient $\nabla \zeta(\boldsymbol{\theta}) = \nabla m(\boldsymbol{\theta}) \varepsilon$ of the stochastic component depends on $\boldsymbol{\theta}$, both **SLS** assumptions **fail**.

Calming = (pre)smoothing + relaxation + regularization.

(Pre)smoothing (or dual representation/kernelization/observables):

$$\mathbf{Z} = \Phi \mathbf{Y}, \quad \Phi: \mathbb{R}^n \rightarrow \mathbb{R}^q.$$

Further, define $\mathbf{M}(\theta) \stackrel{\text{def}}{=} \Phi \mathbf{m}(\theta)$ and represent

$$\mathbf{Y} = \mathbf{m}(\theta) + \varepsilon \quad \rightarrow \quad \Phi \mathbf{Y} \approx \eta + \Phi \varepsilon \quad \text{and} \quad \eta \approx \mathbf{M}(\theta).$$

Then $\|\mathbf{Y} - \mathbf{m}(\theta)\|^2$ transforms to

$$\|\Phi \mathbf{Y} - \eta\|^2 + \lambda \|\Phi \mathbf{m}(\theta) - \eta\|^2 = \|\mathbf{Z} - \eta\|^2 + \lambda \|\mathbf{M}(\theta) - \eta\|^2$$

with a Lagrange multiplier λ . Leads to

$$2\mathcal{L}(\theta, \eta) = -\|\mathbf{Z} - \eta\|^2 - \lambda \|\mathbf{M}(\theta) - \eta\|^2,$$

$$2\mathcal{L}_{\mathcal{G}}(\theta, \eta) = -\|\mathbf{Z} - \eta\|^2 - \lambda \|\mathbf{M}(\theta) - \eta\|^2 - \|G\theta\|^2 - \|\Gamma\eta\|^2$$

Consider (with $\lambda = 1$)

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) &= -\frac{1}{2}\|\mathcal{S}\mathbf{Y} - \boldsymbol{\eta}\|^2 - \frac{1}{2}\|\mathcal{S}\mathbf{m}(\boldsymbol{\theta}) - \boldsymbol{\eta}\|^2 \\ &= -\frac{1}{2}\|\mathbf{Z} - \boldsymbol{\eta}\|^2 - \frac{1}{2}\|\mathbf{M}(\boldsymbol{\theta}) - \boldsymbol{\eta}\|^2\end{aligned}$$

$\tilde{\mathbf{v}}_G$ is given by

$$\begin{aligned}\mathcal{L}_G(\mathbf{v}) &= \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) = -\frac{1}{2}\|\mathbf{Z} - \boldsymbol{\eta}\|^2 - \frac{1}{2}\|\mathbf{M}(\boldsymbol{\theta}) - \boldsymbol{\eta}\|^2 - \frac{1}{2}\|G\boldsymbol{\theta}\|^2, \\ \tilde{\mathbf{v}}_G &= \operatorname{argmax}_{\mathbf{v} \in \mathcal{V}} \mathcal{L}_G(\mathbf{v}).\end{aligned}$$

Profile MLE:
$$\tilde{\boldsymbol{\theta}}_G = \operatorname{argmax}_{\boldsymbol{\theta}} \max_{\boldsymbol{\eta}} \mathcal{L}_G(\mathbf{v}).$$

With $\mathbf{m}^* = \mathbb{E}\mathbf{Y}$ and $\mathbf{M}^* = \mathcal{S}\mathbf{m}^*$

$$\mathbf{v}^* = \operatorname{argmin}_{\mathbf{v}=(\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathcal{Y}} \left\{ \|\mathbf{M}^* - \boldsymbol{\eta}\|^2 + \|\mathbf{M}(\boldsymbol{\theta}) - \boldsymbol{\eta}\|^2 \right\},$$

$$\mathbf{v}_G^* = \operatorname{argmin}_{\mathbf{v}=(\boldsymbol{\theta}, \boldsymbol{\eta}) \in \mathcal{Y}} \left\{ \|\mathbf{M}^* - \boldsymbol{\eta}\|^2 + \|\mathbf{M}(\boldsymbol{\theta}) - \boldsymbol{\eta}\|^2 + \|G\boldsymbol{\theta}\|^2 \right\}.$$

The $\boldsymbol{\theta}$ -component $\boldsymbol{\theta}^*$ of $\mathbf{v}^*$ (resp. $\boldsymbol{\theta}_G^*$ of $\mathbf{v}_G^*$) solves the original problem in which the smoothed response $\mathbf{Z} = \mathcal{S}\mathbf{Y}$ is replaced by the auxiliary parameter $\boldsymbol{\eta}^*$ (resp. $\boldsymbol{\eta}_G^*$):

$$\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta} \in \Theta} \|\mathbf{M}(\boldsymbol{\theta}) - \boldsymbol{\eta}_G^*\|^2,$$

$$\boldsymbol{\theta}_G^* = \operatorname{argmin}_{\boldsymbol{\theta} \in \Theta} \left\{ \|\mathbf{M}(\boldsymbol{\theta}) - \boldsymbol{\eta}_G^*\|^2 + \|G\boldsymbol{\theta}\|^2 \right\}.$$

Let

$$D^2(\boldsymbol{\theta}) = \frac{1}{2} \nabla M(\boldsymbol{\theta}) \nabla M(\boldsymbol{\theta})^\top = \frac{1}{2} \sum_{j=1}^q \nabla M_j(\boldsymbol{\theta}) \nabla M_j(\boldsymbol{\theta})^\top \in \mathfrak{M}_p.$$

For an initial guess $\boldsymbol{\theta}_0$, define $D_0 = D(\boldsymbol{\theta}_0)$ and

$$\Theta^\circ = \{\boldsymbol{\theta} : \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\| \leq r_0\}$$

$(\boldsymbol{\theta}^*)$ It holds $\boldsymbol{\theta}^* \in \Theta^\circ$ and $\boldsymbol{\theta}_G^* \in \Theta^\circ$.

Conditions of this kind are often applied in nonlinear optimization for studying, e.g. Gauss-Newton iterations; see e.g. [Gratton et al., 2007].

With

$$D^2(\boldsymbol{\theta}) = \frac{1}{2} \nabla M(\boldsymbol{\theta}) \nabla M(\boldsymbol{\theta})^\top, \quad D_0 = D(\boldsymbol{\theta}_0),$$

assume

(∇M) For some $\omega^+ \leq 1/3$ and any $\boldsymbol{\theta} \in \Theta^\circ$, it holds

$$(1 - \omega^+) D_0^2 \leq D^2(\boldsymbol{\theta}) \leq (1 + \omega^+) D_0^2.$$

$(\nabla^k M)$ For $k \in \{2, 3, 4\}$ and small $\varkappa \geq 0$, uniformly over $\boldsymbol{\theta} \in \Theta^\circ$ and $\mathbf{u} \in \mathbb{R}^p$

$$\sum_{j=1}^q \langle \nabla^k M_j(\boldsymbol{\theta}), \mathbf{u}^{\otimes k} \rangle^2 \leq \varkappa^{2k-2} \|D_0 \mathbf{u}\|^{2k}.$$

For $\zeta(\mathbf{v}^*) = \mathcal{L}(\mathbf{v}) - \mathbb{E} \mathcal{L}(\mathbf{v})$, it holds

$$\nabla \zeta = \begin{pmatrix} 0 \\ \nabla_{\eta} \zeta \end{pmatrix} = \begin{pmatrix} 0 \\ \mathcal{S} \boldsymbol{\varepsilon} \end{pmatrix},$$

Bounding $\nabla \zeta$ can be easily reduced to a similar question for $\mathcal{S} \boldsymbol{\varepsilon}$.

($\mathcal{S} \boldsymbol{\varepsilon}$) *The vector $\mathcal{S} \boldsymbol{\varepsilon}$ satisfies for all considered $x > 0$*

$$\mathbb{P}(\|\mathcal{S} \boldsymbol{\varepsilon}\| > z(\mathbb{W}^2, x)) \leq 3e^{-x},$$

where

$$\mathbb{W}^2 \stackrel{\text{def}}{=} \text{Var}(\mathcal{S} \boldsymbol{\varepsilon}) = \mathcal{S} \text{Var}(\boldsymbol{\varepsilon}) \mathcal{S}^{\top},$$

$$z(\mathbb{W}^2, x) \stackrel{\text{def}}{=} \sqrt{\text{tr } \mathbb{W}^2} + \sqrt{2x \|\mathbb{W}^2\|}.$$

[Spokoiny, 2024b], [Spokoiny, 2024a].

With

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\eta}) = -\frac{1}{2}\|\mathbf{Z} - \boldsymbol{\eta}\|^2 - \frac{1}{2}\|\mathbf{M}(\boldsymbol{\theta}) - \boldsymbol{\eta}\|^2$$

it holds

$$\mathcal{F}_G(\mathbf{v}) \stackrel{\text{def}}{=} -\nabla^2 \mathcal{L}_G(\mathbf{v}) = \begin{pmatrix} \mathbb{F}_G(\mathbf{v}) & -\nabla \mathbf{M}(\boldsymbol{\theta}) \\ -\nabla \mathbf{M}(\boldsymbol{\theta})^\top & 2\mathbb{I}_q \end{pmatrix}$$

with the upper left diagonal block

$$\mathbb{F}_G(\mathbf{v}) \stackrel{\text{def}}{=} \nabla \mathbf{M}(\boldsymbol{\theta}) \nabla \mathbf{M}(\boldsymbol{\theta})^\top + \sum_{j=1}^q \{M_j(\boldsymbol{\theta}) - \eta_j\} \nabla^2 M_j(\boldsymbol{\theta}) + G^2.$$

Define $\mathcal{F} \stackrel{\text{def}}{=} \mathcal{F}(\mathbf{v}_G^*)$.

With

$$D^2(\boldsymbol{\theta}) = \frac{1}{2} \nabla M(\boldsymbol{\theta}) \nabla M(\boldsymbol{\theta})^\top$$

and $D^2 = D^2(\boldsymbol{\theta}_G^*)$, define

$$\mathcal{D}^2 = \text{block}\{D^2, \mathbb{I}_q\}.$$

Lemma

It holds

$$\mathcal{F}_G^{-1} \leq 2 \begin{pmatrix} (D^2 + 2G^2)^{-1} & 0 \\ 0 & \mathbb{I}_q \end{pmatrix}.$$

With $\mathbf{M}_G = (G^2 \boldsymbol{\theta}^*, 0)$, define

$$\mathbf{b}_D \stackrel{\text{def}}{=} \|\mathcal{D} \mathcal{F}_G^{-1} \mathbf{M}_G\| \leq 2 \|D \mathbf{F}_G^{-1} G^2 \boldsymbol{\theta}^*\|.$$

Theorem

Let $\mathbf{r}_D \stackrel{\text{def}}{=} 2z(\mathbb{V}^2, \mathbf{x})$ and

$$\mathbf{r}_0 \preceq \frac{1}{4}, \quad \mathbf{r}_0 \geq \frac{3}{2} (\mathbf{r}_D \vee \mathbf{b}_D), \quad \tau_3 (\mathbf{r}_D \vee \mathbf{b}_D) < \frac{2}{9}.$$

It holds on $\Omega(\mathbf{x})$ with for any linear mapping Q on $\boldsymbol{\theta}$

$$\begin{aligned} & \|Q\{\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^* - (\mathcal{F}_G^{-1} \nabla \zeta)_{\boldsymbol{\theta}} + (\mathcal{F}_G^{-1} \mathbf{M}_G)_{\boldsymbol{\theta}}\}\| \\ & \leq \|Q D^{-1}\| \frac{3\tau_3}{2} (\|2\mathcal{S}\boldsymbol{\varepsilon}\|^2 + \mathbf{b}_D^2). \end{aligned}$$

Also, define

$$\begin{aligned}
 p_Q &\stackrel{\text{def}}{=} \text{tr Var}\{Q(\mathcal{F}_G^{-1}\nabla\zeta)\boldsymbol{\theta}\}, \\
 \mathcal{R}_Q &\stackrel{\text{def}}{=} \mathbb{E}\{\|Q(\mathcal{F}_G^{-1}\nabla\zeta)\boldsymbol{\theta} - Q(\mathcal{F}_G^{-1}\mathbf{M}_G)\boldsymbol{\theta}\|^2 \mathbb{1}_{\Omega(\mathbf{x})}\} \\
 &\leq p_Q + \|Q(\mathcal{F}_G^{-1}\mathbf{M}_G)\boldsymbol{\theta}\|^2.
 \end{aligned}$$

Theorem

With $\bar{p}_D = \mathbb{E}\|\mathcal{D}\mathcal{F}_G^{-1}\nabla\zeta\|^2 \leq \mathbb{E}\|2\mathcal{S}\epsilon\|^2$, it holds

$$\mathbb{E}\{\|Q(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*)\| \mathbb{1}_{\Omega(\mathbf{x})}\} \leq \sqrt{\mathcal{R}_Q} + \|Q D^{-1}\| \frac{3\tau_3}{2} (\bar{p}_D + \mathbf{b}_D^2).$$

Define the full effective dimension

$$\bar{p}_D = \mathbb{E} \|2\mathcal{S}\epsilon\|^2 \leq 4\sigma^2 q$$

The *effective sample size* n is defined via the constant $\varkappa$ from $(\nabla^k M)$. We use

$$\tau_3 \asymp \varkappa \asymp n^{-1/2}.$$

The results require

$$\bar{p}_G \ll n.$$

1 Statistical inference

- Linear and SLS models
- Nonlinear regression. Theoretical study

2 Structural modeling: Examples

- Matrix completion
- Gaussian mixture
- Deep Neural Networks

3 Gaussian Variational Inference

4 Optimization vs sampling

Suppose that a matrix $\mathbf{Y} = (Y_{ij}) \in \mathbb{R}^{p \times q}$ is partly observed with noise:

$$Y_{ij} = X_{ij} + \varepsilon_{ij}, \quad (i, j) \in \mathcal{G},$$

where $\mathcal{G}$ describes the “design”. The goal is to recover the matrix $\mathbf{X} = (X_{ij})$ under a “low-rank” condition. The latter yields the representation

$$\mathbf{X} = \mathbf{U} \mathbf{\Lambda} \mathbf{V}^\top = \sum_m \lambda_m \mathbf{u}_m \mathbf{v}_m^\top, \quad (1)$$

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_r)$, $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_r) \in \mathbb{R}^{p \times r}$, $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_r) \in \mathbb{R}^{q \times r}$, and the vectors $\mathbf{u}_m$ are orthonormal in $\mathbb{R}^p$ while $\mathbf{v}_m$ are orthonormal in $\mathbb{R}^q$. In the case when all the eigenvalues λ_j are different and ordered by absolute values $|\lambda_1| > \dots > |\lambda_r|$, representation (1) is unique.

Let now $(\mathcal{T}_k)$ be a collection of “templates” in $\mathbb{R}^{p \times q}$, $k = 1, \dots, K$. A typical example is of the form

$$\mathcal{T} = \text{diag}(\delta_i) \mathbf{1}_{p \times q} \text{diag}(\delta'_j),$$

where $\mathbf{1}_{p \times q}$ is the matrix of ones in $\mathbb{R}^{p \times q}$ and $(\delta_1, \dots, \delta_p)$, $(\delta'_1, \dots, \delta'_q)$ are obtained as independent Bernoulli r.v.s. Informally, we include in the template $\mathcal{T}$ each row i with probability $\alpha_{1,i}$ and each column j with probability $\alpha_{2,j}$. Define

$$z(\mathcal{T}) \stackrel{\text{def}}{=} \langle \mathbf{X}, \mathcal{T} \rangle = \sum_{(i,j) \in \mathcal{G}} \mathcal{T}_{ij} X_{ij}$$

$$Z(\mathcal{T}) \stackrel{\text{def}}{=} \langle \mathbf{Y}, \mathcal{T} \rangle = \sum_{(i,j) \in \mathcal{G}} \mathcal{T}_{ij} Y_{ij}.$$

Also introduces “observables” Z_k and the image parameters z_k

$$Z_k = \langle \mathbf{Y}, \mathcal{T}_k \rangle, \quad z_k = \langle \mathbf{X}, \mathcal{T}_k \rangle.$$

The whole set of parameters include orthonormal vectors $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_r)$ in $\mathbb{R}^p$ and $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_r)$ in $\mathbb{R}^q$, the vector of eigenvalues $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_r)^\top$, and the image vector $\mathbf{z} = (z_k)$ leading to the log-likelihood

$$\begin{aligned}\mathcal{L}(\mathbf{U}, \mathbf{V}, \boldsymbol{\lambda}, \mathbf{z}) &= -\frac{1}{2} \|\mathbf{Z} - \mathbf{z}\|^2 - \frac{1}{2} \sum_{k=1}^K |z_k - \langle \mathcal{T}_k, \mathbf{U} \boldsymbol{\Lambda} \mathbf{V}^\top \rangle|^2 \\ &= -\frac{1}{2} \sum_{k=1}^K |\langle \mathbf{Y}, \mathcal{T}_k \rangle - z_k|^2 - \frac{1}{2} \sum_{k=1}^K \left| z_k - \sum_{m=1}^r \lambda_m \mathbf{u}_m^\top \mathcal{T}_k \mathbf{v}_m \right|^2.\end{aligned}$$

(Near) orthonormality of the $\mathbf{u}_m$'s and $\mathbf{v}_m$'s can be enforced by the penalty

$$\mu \left(\|\mathbf{U}^\top \mathbf{U} - \mathbb{I}_r\|_{\text{Fr}}^2 + \|\mathbf{V}^\top \mathbf{V} - \mathbb{I}_r\|_{\text{Fr}}^2 \right).$$

Identifiability of the model is supported by a penalty $\sum_m g_m^2 \lambda_m^2$ on the eigenvalues $\lambda_1, \dots, \lambda_r$ with $g_1^2 < \dots < g_r^2$. Finally, to distinguish between $\mathbf{u}_m$ and $-\mathbf{u}_m$, add the penalty $\|\mathbf{U} - \mathbf{E}\|_{\text{Fr}}^2 = \sum_m \|\mathbf{u}_m - \mathbf{e}_m\|^2$ for given orthonormal vectors $\mathbf{e}_m$ and similarly for $\mathbf{v}_m$.

In total

$$\begin{aligned}
 \mathcal{L}(\mathbf{U}, \mathbf{V}, \boldsymbol{\lambda}, \mathbf{z}) = & -\frac{1}{2} \|\mathbf{Z} - \mathbf{z}\|^2 - \frac{1}{2} \sum_{k=1}^K |z_k - \langle \mathcal{T}_k, \mathbf{U} \boldsymbol{\Lambda} \mathbf{V}^\top \rangle|^2 \\
 & - \frac{\mu_g}{2} \sum_{m=1}^r g_m^2 \lambda_m^2 \\
 & - \frac{\mu_o}{2} (\|\mathbf{U}^\top \mathbf{U} - \mathbb{I}_r\|_{\text{Fr}}^2 + \|\mathbf{V}^\top \mathbf{V} - \mathbb{I}_r\|_{\text{Fr}}^2) \\
 & - \frac{\mu_e}{2} (\|\mathbf{U} - \mathbf{E}\|_{\text{Fr}}^2 + \|\mathbf{V} - \mathbf{E}'\|_{\text{Fr}}^2).
 \end{aligned} \tag{2}$$

The whole procedure includes the following steps:

- Fix a collection of templates $\mathcal{T}_k$ and compute $Z_k = \langle \mathcal{T}_k, \mathbf{Y} \rangle$;
- Fix the matrices $\mathbf{E} = (\mathbf{e}_1, \dots, \mathbf{e}_r) \in \mathbb{R}^{p \times r}$ with $\mathbf{E}^\top \mathbf{E} = \mathbb{I}_r$ and similarly $\mathbf{E}' = (\mathbf{e}'_1, \dots, \mathbf{e}'_r) \in \mathbb{R}^{q \times r}$ with $\mathbf{E}'^\top \mathbf{E}' = \mathbb{I}_r$
- Solve the maximization problem $\tilde{\mathbf{v}} = \operatorname{argmax}_{\mathbf{v}} \mathcal{L}(\mathbf{v})$ for $\mathcal{L}(\mathbf{v})$ from (2) by alternating optimization.
- Build $\tilde{\mathbf{X}}$ using the solution $\tilde{\mathbf{v}}$ e.g.

$$\tilde{\mathbf{X}} = \tilde{\mathbf{U}} \operatorname{diag}(\tilde{\boldsymbol{\lambda}}) \tilde{\mathbf{V}}^\top = \sum_{m=1}^r \tilde{\lambda}_m \tilde{\mathbf{u}}_m \tilde{\mathbf{v}}_m^\top.$$

- If necessary, redesign $\mathbf{E}$ and $\mathbf{E}'$ and repeat.

With $\phi(\mathbf{x}) = \mathcal{C} \exp(-\|\mathbf{x}\|^2/2)$, consider

$$f(\mathbf{x}) = \int \phi\left(\frac{\mathbf{x} - \mathbf{m}}{\sigma}\right) d\mu(\mathbf{m}, \sigma). \quad (3)$$

Usually, the mixing measure μ is discrete with well separated atoms $\{(\mathbf{m}_k, \sigma_k), k \in \mathcal{K}\}$:

$$\mu = \sum_{k \in \mathcal{K}} \mu_k \mathbb{I}_{\mathbf{m}_k, \sigma_k}.$$

Then $\mu = \mu_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta} = \{(\mu_k, \mathbf{m}_k, \sigma_k), k \in \mathcal{K}\}$, where $\sum_k \mu_k = 1$.

$$f(\mathbf{x}) = f(\mathbf{x}, \boldsymbol{\theta}) = \sum_{k \in \mathcal{K}} \mu_k \phi\left(\frac{\mathbf{x} - \mathbf{m}_k}{\sigma_k}\right).$$

Later we consider $\mu = \mu_{\boldsymbol{\theta}} = \sum_j \mu_k \delta_{\mathbf{m}_k, \sigma_k}$.

Let X_i be i.i.d. from f . Consider the problem of recovering the mixing measure μ from the data. Given a family of test functions $(\psi_j(\mathbf{x}))$, define the observables

$$Z_j \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n \psi_j(\mathbf{X}_i).$$

One example is given by a collection $\psi_j(\mathbf{x}) = \psi(\|\mathbf{x} - \mathbf{x}_j\|^2/s_j^2)$ for a kernel ψ , a fixed set of points $\mathbf{x}_j$ and scalings s_j , $j \leq q$. Denote

$$\psi_j(\mathbf{m}, \sigma) \stackrel{\text{def}}{=} \int \psi_j(\mathbf{x}) \phi\left(\frac{\mathbf{x} - \mathbf{m}}{\sigma}\right) d\mathbf{x}$$

Under (3), it holds

$$\begin{aligned} \mathbb{E} Z_j &= \int \psi_j(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \iint \psi_j(\mathbf{x}) \phi\left(\frac{\mathbf{x} - \mathbf{m}}{\sigma}\right) d\mu(\mathbf{m}, \sigma) d\mathbf{x} \\ &= \int \psi_j(\mathbf{m}, \sigma) d\mu(\mathbf{m}, \sigma) = \mu(\psi_j). \end{aligned}$$

The calming device suggests to introduce the image parameter z and the extended log-likelihood $\mathcal{L}(\boldsymbol{v}) = \mathcal{L}(\boldsymbol{\theta}, z)$ with

$$\mathcal{L}(\boldsymbol{\theta}, z) \stackrel{\text{def}}{=} -\frac{1}{2}\|\boldsymbol{Z} - z\|^2 - \frac{1}{2}\|z - \mu_{\boldsymbol{\theta}}(\boldsymbol{\Psi})\|^2,$$

where $\mu_{\boldsymbol{\theta}}(\boldsymbol{\Psi})$ is a vector in $\mathbb{R}^q$ with the entries

$$\mu_{\boldsymbol{\theta}}(\psi_j) = \sum_{k \in \mathcal{K}} \mu_k \psi_j(\boldsymbol{m}_k, \sigma_k).$$

To overcome the issue of identifiability problem, introduce a penalty

$$\text{pen}_{\varkappa}(\boldsymbol{\mu}) \stackrel{\text{def}}{=} \sum_{k \in \mathcal{K}} \varkappa_k^2 \mu_k^2,$$

where $\varkappa_k^2$ strictly increase with k . Such a penalty ensures identifiability if the true weights μ_k^* of each component $\phi_{\mathbf{m}_k, \sigma_k}$ are significantly different. Unfortunately, the problem is not completely resolved if there are different components with nearly the same weights μ_k^* . One possibility to make it fixed is by using an additional penalty $\text{pen}_G(\boldsymbol{\theta})$ based on the distance of each mean $\mathbf{m}_k$ from the origin (or any other fixed point $\mathbf{m}_0$):

$$\text{pen}_G(\mathbf{m}) \stackrel{\text{def}}{=} \sum_{k \in \mathcal{K}} \|G\mathbf{m}_k\|^2,$$

where the matrix G identifies the distance from each mean $\mathbf{m}_k$ to the origin. In particular, one can use $G^2 = \text{diag}(g_j^2)$ with g_j^2 strictly increasing.

Altogether leads to the following approach: for $\mathbf{v} = (\boldsymbol{\theta}, \mathbf{z}) = \{(\mu_k, \mathbf{m}_k, \sigma_k), (z_j)\}$ and $\text{pen}(\mathbf{v}) = \text{pen}_{\mathcal{K}}(\boldsymbol{\theta}) + \text{pen}_G(\mathbf{m})$

$$\begin{aligned}\tilde{\mathbf{v}}_G &= \underset{\mathbf{v}}{\operatorname{argmax}} \left\{ \mathcal{L}(\mathbf{v}) - \frac{1}{2} \text{pen}(\mathbf{v}) \right\} \\ &= \underset{\mathbf{v}}{\operatorname{argmax}} \left\{ -\frac{1}{2} \|\mathbf{Z} - \mathbf{z}\|^2 - \frac{1}{2} \|\mathbf{z} - \mu_{\boldsymbol{\theta}}(\boldsymbol{\Psi})\|^2 \right. \\ &\quad \left. - \frac{1}{2} \sum_{k \in \mathcal{K}} \|G\mathbf{m}_k\|^2 - \frac{1}{2} \sum_{k \in \mathcal{K}} \kappa_k^2 \mu_k^2 \right\}.\end{aligned}$$

The structural penalty $\|\mathbf{z} - \mu_{\boldsymbol{\theta}}(\boldsymbol{\Psi})\|^2$ creates some difficulties for the analysis, however, it is deterministic and smooth in the scope of arguments.

Let, for an input vector $\mathbf{x} = (x_m) \in \mathbb{R}^d$, the hidden layer transformation is given by

$$\mathbf{x}^{(1)} = \sigma(\mathbf{a} + W\mathbf{x}),$$

where $\mathbf{a} \in \mathbb{R}^p$, $W : \mathbb{R}^d \rightarrow \mathbb{R}^p$, and σ is a coordinate-wise activating function, e.g.

$$\sigma(t) = \lambda^{-1} \log(1 + e^{\lambda t}).$$

The transformed vectors $\mathbf{x}^{(1)}$ enter in the logistic regression model for binary labels Y_i

$$\mathbb{P}(Y = 1 \mid \mathbf{x}^{(1)}) = \text{softmax}(\mathbf{x}^{(1)}), \quad \mathbb{P}(Y = 0 \mid \mathbf{x}^{(1)}) = 1 - \text{softmax}(\mathbf{x}^{(1)}).$$

The structure of this neuronal network is described by the structural parameter $\mathbf{v} = (\mathbf{a}, W)$.

Now consider the statistical problem of inference about this parameter given independent data $(\mathbf{X}_i, Y_i)$. The corresponding log-likelihood involves the fidelity term $L(\mathbf{Y}, \mathbf{x}^{(1)}) = \sum_i \ell(Y_i, \eta_i)$ with $\ell(y, \eta) = y\eta - \log(1 + e^\eta)$, $\eta_i = \text{softmax}(\mathbf{x}_i^{(1)})$ and the structural terms $\|\mathbf{X}^{(1)} - \sigma(\mathbf{a} + W\mathbf{X})\|^2$. We also add some penalty on $\mathbf{a} = (a_j)$ and $W = (w_{mj})$:

$$\text{pen}(\mathbf{a}) = \frac{1}{2} \|\mathcal{T}\mathbf{a}\|^2 = \frac{1}{2} \sum_j a_j^2 \mathcal{T}_j^2,$$

$$\text{pen}(W) = \frac{1}{2} \langle \mathcal{G}, W \rangle^2 = \frac{1}{2} \sum_{m,j} w_{mj}^2 \mathcal{G}_{mj}^2,$$

with $\mathcal{T}_j$ and $\mathcal{G}_{mj}$ polynomially growing in j . This enables us to identify the most informative nodes in the hidden layer and control the overall complexity of the network.

This results in maximization of the penalized log-likelihood

$$\mathcal{L}(\mathbf{a}, W, \mathbf{X}^{(1)}) = L(\mathbf{Y}, \text{softmax}(\mathbf{X}^{(1)})) - \frac{\mu}{2} \|\mathbf{X}^{(1)} - \sigma(\mathbf{a} + W\mathbf{X})\|^2 \\ - \frac{1}{2} \|\mathcal{T}\mathbf{a}\|^2 - \frac{1}{2} \langle \mathcal{G}, W \rangle^2$$

with a Lagrange multiplier μ . The structural relation $\mathbf{X}^{(1)} \equiv \sigma(\mathbf{a} + W\mathbf{X})$ is relaxed and replaced by the structural penalty $\frac{\mu}{2} \|\mathbf{X}^{(1)} - \sigma(\mathbf{a} + W\mathbf{X})\|^2$. Introducing the auxiliary variable $\mathbf{X}^{(1)}$ is not mandatory, one can use $\mathbf{X}^{(1)} \equiv \sigma(\mathbf{a} + W\mathbf{X})$. However, it can be useful, e.g. for an additional penalization.

One example of choosing the penalty on $\mathbf{a}$ and W is given by $\mathcal{T}_j^2 = c_a j^{2\beta}$, and $\mathcal{G}_{mj}^2 = \mathcal{G}_j^2 = c_w j^{2\beta}$ for e.g. $\beta = 2$ and some constants c_a, c_w . Any prior information about the input features $\mathbf{X}$ can be incorporated in the penalty coefficients $\mathcal{A}_m$ leading to a structure $\mathcal{G}_{mj}^2 = G_m^2 + \mathcal{G}_j^2$, e.g. $\mathcal{G}_{mj}^2 = c_x m^{2\beta} + c_w j^{2\beta}$.

This construction extends to a K -layer network using recurrence

$$\mathbf{X}^{(k)} = \sigma(\mathbf{a}^{(k)} + W^{(k)} \mathbf{X}^{(k-1)})$$

for $k = 1, \dots, K$ and $\mathbf{X}^{(0)} = \mathbf{X}$. This leads to the log-likelihood

$$\begin{aligned} \mathcal{L}_{\mathcal{G}}(\mathbf{X}^{(1)}, \mathbf{a}^{(1)}, W^{(1)}, \dots, \mathbf{X}^{(K)}, \mathbf{a}^{(K)}, W^{(K)}) &= L(\mathbf{Y}, \mathbf{X}^{(K)}) \\ &- \frac{1}{2} \sum_{k=1}^K \left(\|\mathbf{X}^{(k)} - \sigma(\mathbf{a}^{(k)} + W^{(k)} \mathbf{X}^{(k-1)})\|^2 + \|\mathcal{T}^{(k)} \mathbf{a}^{(k)}\|^2 + \langle W^{(k)}, \mathcal{G}^{(k)} \rangle^2 \right) \end{aligned}$$

1 Statistical inference

- Linear and SLS models
- Nonlinear regression. Theoretical study

2 Structural modeling: Examples

- Matrix completion
- Gaussian mixture
- Deep Neural Networks

3 Gaussian Variational Inference

4 Optimization vs sampling

Let $\mathbb{P}_f \sim \exp f(\mathbf{x})$. Denote by $\mathbb{N}_{\mathbf{x}, \mathbb{Z}}$ the Gaussian measure with the mean $\mathbf{x}$ and covariance $\mathbb{Z}^{-1}$, i.e. $\mathbb{N}_{\mathbf{x}, \mathbb{Z}} \stackrel{\text{def}}{=} \mathcal{N}(\mathbf{x}, \mathbb{Z}^{-1})$.

$$\text{Gauss VI: } (\mathbf{x}_{\text{VI}}, \mathbb{Z}_{\text{VI}}) = \underset{\mathbf{x}, \mathbb{Z}}{\operatorname{arginf}} \mathcal{K}(\mathbb{N}_{\mathbf{x}, \mathbb{Z}} \parallel \mathbb{P}_f).$$

Natural candidates:

1. **Laplace:** $\mathbf{x}_{\text{VI}} \approx \operatorname{argmax} f(\mathbf{x})$, $\mathbb{Z}_{\text{VI}} \approx -\nabla^2 f(\mathbf{x}^*)$;
2. **Moments:** $\mathbf{x}_{\text{VI}} \approx \mathbb{E}_f \mathbf{X}$, $\mathbb{Z}_{\text{VI}}^{-1} \approx \operatorname{Var}_f(\mathbf{X})$.

[Katsevich and Rigollet, 2023] argued for (2).

- [David M. Blei and McAuliffe, 2017] Variational Inference: A review for statisticians
- [Zhang and Gao, 2020] Convergence rates of variational posterior distributions
- [Wang and Blei, 2019] Frequentist consistency of variational Bayes
- [Han and Yang, 2019] Statistical inference in mean-field variational Bayes
- [Challis and Barber, 2013] Gaussian Kullback-Leibler approximate inference
- [Alquier and Ridgway, 2020] Concentration of tempered posteriors and of their variational approximations
- [Lambert et al., 2023] Variational inference via Wasserstein gradient flows

The VI approach assumes minimizing of the KL-divergence $\mathcal{K}(\mathbb{P}_{x,\mathbb{Z}} \parallel \mathbb{P}_f)$ over all feasible $x, \mathbb{Z}$. Here we rewrite this problem in terms of local parameters a and S .

Lemma

For any x and any $\mathbb{Z}$, it holds

$$\mathcal{K}(\mathbb{P}_{x,\mathbb{Z}} \parallel \mathbb{P}_f) = C + \frac{1}{2} \log \det(\mathbb{Z}^{-1}) - \frac{p}{2} - \mathbb{E} f(x + \gamma_{\mathbb{Z}}).$$

with C depending on f and p only.

With $C_f \stackrel{\text{def}}{=} \log \int e^{f(\bar{x}+u)} d\mathbf{u}$ and $C_p = (2\pi)^{-p/2}$, for any $\mathbf{u} \in \mathbb{R}^p$

$$\frac{d\mathbb{P}_f}{d\mathbf{u}}(\mathbf{x} + \mathbf{u}) = e^{-C_f} e^{f(\mathbf{x}+\mathbf{u})},$$

$$\frac{d\mathbb{P}_{\mathbf{x},\mathbb{Z}}}{d\mathbf{u}}(\mathbf{x} + \mathbf{u}) = C_p \det(\mathbb{Z}^{1/2}) e^{-\|\mathbb{Z}^{1/2}\mathbf{u}\|^2/2}.$$

This yields with $\gamma_{\mathbb{Z}} \sim \mathcal{N}(0, \mathbb{Z}^{-1})$ and $\gamma \sim \mathcal{N}(0, \mathbb{I}_p)$

$$\begin{aligned} \mathbb{E}_{\mathbf{x},\mathbb{Z}} \log \frac{d\mathbb{P}_{\mathbf{x},\mathbb{Z}}}{d\mathbb{P}_f} \\ = C_f + \log C_p - \mathbb{E} f(\mathbf{x} + \gamma_{\mathbb{Z}}) - \frac{1}{2} \mathbb{E} \|\gamma\|^2 - \frac{1}{2} \log \det(\mathbb{Z}^{-1}), \end{aligned}$$

and the result follows in view of $\mathbb{E} \|\gamma\|^2 = p$.

With $\mathbb{F} = -\nabla^2 f(\bar{\mathbf{x}})$, represent $\mathbb{Z}$ in the form

$$\mathbb{Z}^{-1/2} = \mathbb{F}^{-1/2} + S \quad \text{or} \quad \mathbb{F}^{1/4} \mathbb{Z}^{-1/2} \mathbb{F}^{1/4} = \mathbb{I}_p + \mathbb{F}^{1/4} S \mathbb{F}^{1/4}.$$

A vicinity of $\mathbb{F}$ using Kullback-Leibler divergence $\mathcal{K}(\mathbf{N}_{\bar{\mathbf{x}},\mathbb{F}} \parallel \mathbf{N}_{\bar{\mathbf{x}},\mathbb{Z}})$.

Lemma

Let $\mathbb{Z}^{-1/2} = \mathbb{F}^{-1/2} + S$ and $U = \mathbb{F}^{1/4} S \mathbb{F}^{1/4}$ fulfill $\|U\| \leq \nu < 1$. Then

$$\begin{aligned} \mathcal{K}(\mathbf{N}_{\bar{\mathbf{x}},\mathbb{F}} \parallel \mathbf{N}_{\bar{\mathbf{x}},\mathbb{Z}}) &= -\log \det(\mathbb{I}_p + \mathbb{F}^{1/4} S \mathbb{F}^{1/4}) + \frac{1}{2} \operatorname{tr} \{ \mathbb{F}(\mathbb{F}^{-1/2} + S)^2 - \mathbb{I}_p \} \\ &= -\log \det(\mathbb{I}_p + U) + \operatorname{tr} U + \frac{1}{2} \operatorname{tr}(\mathbb{F} S^2) \geq \frac{1}{2} \operatorname{tr}(\mathbb{F} S^2). \end{aligned} \quad (4)$$

For two Gaussian distributions $\mathcal{N}_{\bar{x}, \mathbb{F}}, \mathcal{N}_{\bar{x}, \mathbb{Z}}$ with the same mean $\bar{x}$

$$\begin{aligned}\mathcal{K}(\mathcal{N}_{\bar{x}, \mathbb{F}} \parallel \mathcal{N}_{\bar{x}, \mathbb{Z}}) &= \frac{1}{2} \left\{ -\log \det(\mathbb{F} \mathbb{Z}^{-1}) + \text{tr}(\mathbb{F} \mathbb{Z}^{-1} - \mathbb{I}_p) \right\} \\ &= -\log \det \{ \mathbb{F}^{1/2} (\mathbb{F}^{-1/2} + S) \} + \frac{1}{2} \text{tr} \{ \mathbb{F} (\mathbb{F}^{-1/2} + S)^2 - \mathbb{I}_p \} \\ &= -\log \det(\mathbb{I}_p + U) + \frac{1}{2} \text{tr}(\mathbb{F} S^2 + 2\mathbb{F}^{1/2} S)\end{aligned}$$

and (4) follows by $x - \log(1+x) \geq 0$ for any $x > -1$.

Consider symmetric matrices $S \in \mathfrak{M}_p$ such that for some $\nu < 1$

$$\|\mathbb{F}^{1/4} S \mathbb{F}^{1/4}\| \leq \nu. \quad (5)$$

Lemma

With $\gamma \sim \mathcal{N}(0, \mathbb{I}_p)$, $\mathbf{a} \in \mathbb{R}^p$, and $S \in \mathfrak{M}_p$ satisfying (5), define

$$H(\mathbf{a}, S) \stackrel{\text{def}}{=} -\log \det(\mathbb{F}^{-1/2} + S) - \mathbb{E} f(\bar{\mathbf{x}} + \mathbf{a} + (\mathbb{F}^{-1/2} + S)\gamma),$$

$$(\hat{\mathbf{a}}, \hat{S}) \stackrel{\text{def}}{=} \underset{(\mathbf{a}, S)}{\operatorname{argmin}} H(\mathbf{a}, S).$$

Then the VI problem leads to minimization of the function $H(\mathbf{a}, S)$:

$$(\hat{\mathbf{x}}, \hat{\mathbb{Z}}) \stackrel{\text{def}}{=} \underset{(\mathbf{x}, \mathbb{Z})}{\operatorname{argmin}} \mathcal{K}(\mathbb{P}_{\mathbf{x}, \mathbb{Z}} \parallel \mathbb{P}_f) = (\bar{\mathbf{x}} + \hat{\mathbf{a}}, (\mathbb{F}^{-1/2} + \hat{S})^{-2}).$$

For $X \sim \mathbb{P}_f \propto e^{f(x)}$, consider

$$\bar{x} = \mathbb{E}_f X, \quad \Sigma = \text{Var}(X), \quad \mathbb{F} = -\nabla^2 f(\bar{x}).$$

Consider

$$H(\mathbf{a}, S) \stackrel{\text{def}}{=} -\log \det(\mathbb{F}^{-1/2} + S) - \mathbb{E} f(\bar{x} + \mathbf{a} + (\mathbb{F}^{-1/2} + S)\gamma),$$
$$(\hat{\mathbf{a}}, \hat{S}) \stackrel{\text{def}}{=} \underset{(\mathbf{a}, S)}{\text{argmin}} H(\mathbf{a}, S).$$

A guess $(\mathbf{a}, S) = (0, 0)$. How far from the solution $(\hat{\mathbf{a}}, \hat{S})$?

Technical issue: **anisotropic** smoothness in $\mathbf{a}$ and S directions.

Fix $\mathbb{Z}^{-1/2} = \mathbb{F}^{-1/2} + S$ and optimize w.r.t. $\mathbf{a}$.

For $\mathbb{Z}^{-1/2} = \mathbb{F}^{-1/2} + S$ fixed, consider $H(\mathbf{a}) = H(\mathbf{a}, S)$

$$\hat{\mathbf{a}} \stackrel{\text{def}}{=} \underset{\mathbf{a}}{\operatorname{argmin}} H(\mathbf{a}) = \underset{\mathbf{a}}{\operatorname{argmax}} \mathbb{E} f(\bar{\mathbf{x}} + \mathbf{a} + \mathbb{Z}^{-1/2} \boldsymbol{\gamma}).$$

Main step: compute $\mathbf{A} = \nabla H(0)$ and $\mathcal{F} = -\nabla^2 H(0)$.

A guess: $\mathcal{F} \approx \mathbb{F} = -\nabla^2 f(\bar{\mathbf{x}})$, $\mathbf{A} \approx 0$ up to fourth order.

Fix $\mathbf{a}$ and consider

$$h(t) = -\mathbb{E} f(\bar{\mathbf{x}} + t\mathbf{a} + \mathbb{Z}^{-1/2}\boldsymbol{\gamma}).$$

Lemma

The function $h(t) = H(t\mathbf{a})$ is strongly convex and satisfies

$$h''(t) = -\langle \mathbb{E} \nabla^2 f(\bar{\mathbf{x}} + t\mathbf{a} + \mathbb{Z}^{-1/2}\boldsymbol{\gamma}), \mathbf{a}^{\otimes 2} \rangle.$$

Concavity of $f(\cdot)$ implies convexity of h .

Lemma

It holds with $\mathbb{F} = -\nabla^2 f(\bar{\mathbf{x}})$

$$h''(0) = -\mathbb{E} \langle \nabla^2 f(\bar{\mathbf{x}} + \gamma_{\mathbb{Z}}), \mathbf{a}^{\otimes 2} \rangle,$$

and with $p = \text{tr}(\mathbb{D} \mathbb{F}^{-1} \mathbb{D})$ and $\alpha = \|\mathbb{D} \mathbb{F}^{-1} \mathbb{D}\|$

$$|h''(0) - \mathbf{a}^\top \mathbb{F} \mathbf{a}| \leq \frac{\tau_4(p + 2\alpha)}{2} \|\mathbb{D} \mathbf{a}\|^2. \quad (6)$$

It holds

$$-\langle \nabla^2 f(\bar{\mathbf{x}}), \mathbf{a}^{\otimes 2} \rangle = \mathbf{a}^\top \mathbb{F} \mathbf{a}.$$

For any $\mathbf{u} \in \mathbb{R}^p$,

$$\begin{aligned} & | -\langle \nabla^2 f(\bar{\mathbf{x}} + \gamma_{\mathbb{Z}}), \mathbf{u}^{\otimes 2} \rangle + \langle \nabla^2 f(\bar{\mathbf{x}}), \mathbf{u}^{\otimes 2} \rangle + \langle \nabla^3 f(\bar{\mathbf{x}}), \gamma_{\mathbb{Z}} \otimes \mathbf{u}^{\otimes 2} \rangle | \\ & \leq \frac{1}{2} \tau_4 \|\mathbb{D} \gamma_{\mathbb{Z}}\|^2 \|\mathbb{D} \mathbf{u}\|^2. \end{aligned}$$

With $p = \text{tr}(\mathbb{D}^2 \mathbb{F}^{-1})$

$$\mathbb{E} \|\mathbb{D} \gamma_{\mathbb{Z}}\|^2 = p.$$

Further, $\mathbb{E} \langle \nabla^3 f(\bar{\mathbf{x}}), \gamma_{\mathbb{Z}} \otimes \mathbf{a}^{\otimes 2} \rangle = 0$ and (6) follows.

Define for any direction $\mathbf{a}$

$$h(t) = -\mathbb{E} f(\bar{\mathbf{x}} + t\mathbf{a} + \mathbb{Z}^{-1/2} \boldsymbol{\gamma}).$$

Lemma

It holds with $p = \text{tr}(\mathbb{D} \mathbb{F}^{-1} \mathbb{D})$, $\alpha = \|\mathbb{D} \mathbb{F}^{-1} \mathbb{D}\|$,

$$|h'(0)| \leq \frac{\tau_4 (p + \alpha)^{3/2}}{6} \|\mathbb{D} \mathbf{a}\| + \frac{\diamond_{4,1}}{1 - \diamond} \|\mathbb{D} \mathbf{a}\|.$$

With $\gamma_{\mathbb{Z}} = \mathbb{Z}^{-1/2}\gamma$, Taylor expansion of $\nabla f(\bar{\mathbf{x}} + \gamma_{\mathbb{Z}})$ yields for any $\mathbf{u} \in \mathbb{R}^p$

$$\begin{aligned} & \left| \langle \nabla f(\bar{\mathbf{x}} + \gamma_{\mathbb{Z}}), \mathbf{u} \rangle - \langle \nabla f(\bar{\mathbf{x}}), \mathbf{u} \rangle - \langle \nabla^2 f(\bar{\mathbf{x}}), \gamma_{\mathbb{Z}} \otimes \mathbf{u} \rangle \right. \\ & \quad \left. - \frac{1}{2} \langle \nabla^3 f(\bar{\mathbf{x}}), \gamma_{\mathbb{F}} \otimes \gamma_{\mathbb{Z}} \otimes \mathbf{u} \rangle \right| \leq \frac{1}{6} \tau_4 \|\mathbb{D}\gamma_{\mathbb{Z}}\|^3 \|\mathbb{D}\mathbf{u}\|. \end{aligned} \quad (7)$$

Also by Laplace approximation

$$\left| \langle \nabla f(\bar{\mathbf{x}}), \mathbf{a} \rangle - \frac{1}{2} \mathbb{E} \langle \nabla^3 f(\bar{\mathbf{x}}), \gamma_{\mathbb{F}} \otimes \gamma_{\mathbb{F}} \otimes \mathbf{a} \rangle \right| \leq \frac{\diamond_{4,1}}{1 - \diamond} \|\mathbb{D}\mathbf{a}\|.$$

Now we apply (7) with $\mathbf{u} = \mathbf{a}$ and $\mathbb{E} \|\mathbb{D}\gamma_{\mathbb{F}}\|^3 \leq (p + \alpha)^{3/2}$. The use of $\mathbb{E} \langle \nabla^2 f(\bar{\mathbf{x}}), \gamma_{\mathbb{F}} \otimes \mathbf{a} \rangle = 0$ yields

$$\left| \mathbb{E} \langle \nabla f(\bar{\mathbf{x}} + \mathbb{Z}^{-1/2}\gamma), \mathbf{a} \rangle \right| \leq \frac{\tau_4 (p + \alpha)^{3/2}}{6} \|\mathbb{D}\mathbf{a}\| + \frac{\diamond_{4,1}}{1 - \diamond} \|\mathbb{D}\mathbf{a}\|.$$

Theorem (3-bound)

$$\|\mathbb{F}^{1/2}\hat{\mathbf{a}} - \mathbb{F}^{-1/2}\mathbf{A}\| \leq \tau_3 \|\mathbb{F}^{-1/2}\mathbf{A}\|^3$$

Theorem (4-bound)

$$\|\mathbb{F}^{1/2}\hat{\mathbf{a}} - \mathbb{F}^{-1/2}\mathbf{A} - \mathbb{F}^{-1/2}\nabla\mathcal{T}(\mathbb{F}^{-1}\mathbf{A})\| \leq \mathcal{C}(\tau_3^2 + \tau_4) \|\mathbb{F}^{-1/2}\mathbf{A}\|^3.$$



Alquier, P. and Ridgway, J. (2020).
Concentration of tempered posteriors and of their variational approximations.
The Annals of Statistics, 48(3):1475 – 1497.



Banach, S. (1938).
Über homogene Polynome in (L^2) .
Studia Mathematica, 7(1):36–44.



Challis, E. and Barber, D. (2013).
Gaussian kullback-leibler approximate inference.
Journal of Machine Learning Research, 14(68):2239–2286.



David M. Blei, A. K. and McAuliffe, J. D. (2017).
Variational inference: A review for statisticians.
Journal of the American Statistical Association, 112(518):859–877.



Gratton, S., Lawless, A. S., and Nichols, N. K. (2007).
Approximate gauss-newton methods for nonlinear least squares problems.
SIAM J. Optim., 18:106–132.



Han, W. and Yang, Y. (2019).
Statistical inference in mean-field variational bayes.
<https://arxiv.org/abs/1911.01525>.



Katsevich, A. and Rigollet, P. (2023).
On the approximation accuracy of gaussian variational inference.



Lambert, M., Chewi, S., Bach, F., Bonnabel, S., and Rigollet, P. (2023).

Variational inference via wasserstein gradient flows.

<https://arxiv.org/abs/2205.15902>.



Spokoiny, V. (2024a).

Deviation bounds for the norm of a random vector under exponential moment conditions with applications.

<https://arxiv.org/abs/2309.02302v1>.



Spokoiny, V. (2024b).

Sharp deviation bounds and concentration phenomenon for the squared norm of a sub-gaussian vector.

<https://arxiv.org/abs/2305.07885>.



Wang, Y. and Blei, D. M. (2019).

Frequentist consistency of variational bayes.

Journal of the American Statistical Association, 114(527):1147–1161.



Zhang, F. and Gao, C. (2020).

Convergence rates of variational posterior distributions.

The Annals of Statistics, 48(4):2180 – 2207.



MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

Unlocking Machine Learning Potentials with Innovative SGD Algorithms

Martin Takáč



INNOPOLIS
UNIVERSITY

ASCOMP 2024 Autumn School
2024/07/12

BIG THANKS TO ...



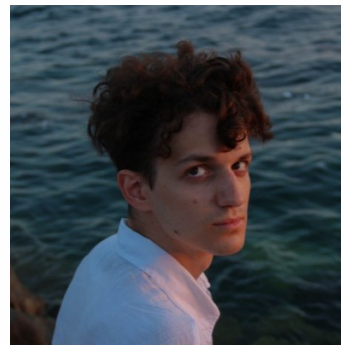
Lam Nguyen



Nicolas Loizou



Jie Liu



Petr Ostroukhov



Katya Scheinberg



Zheng Shi



Konstantin Mishchenko



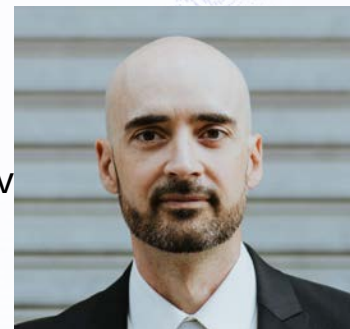
Peter Richtarik



Alexander Beznosikov



Robert M. Gower



Francesco Orabona

Agenda

- What is machine learning (short intro)
- How we train machine learning – SGD
- Approaches to reduce variance
- Towards adaptive SGD
 - Polyak step-size, MoMo
 - D-Adaptation

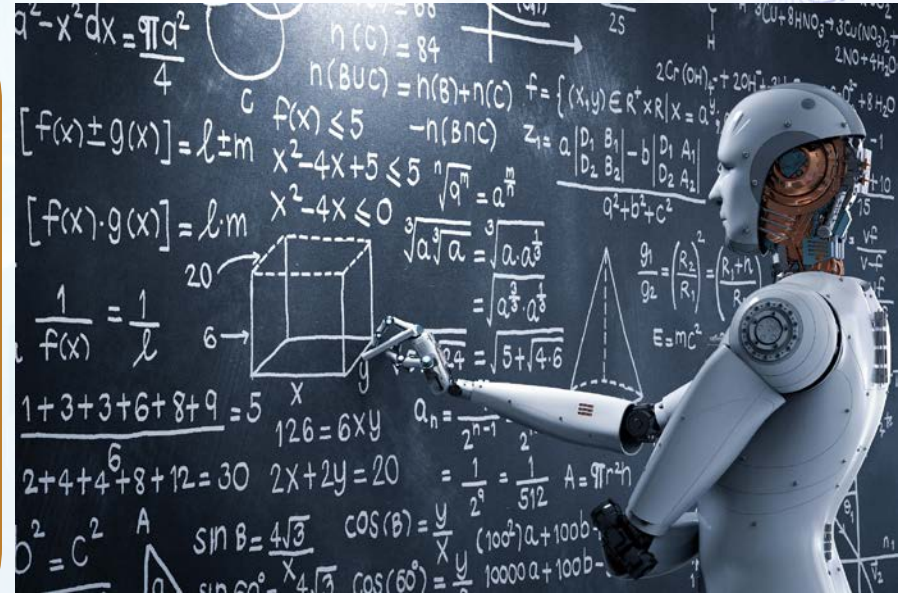


Quick Intro to ML

Problem formulation, notation, algorithms, ...

- Problem:** Assume you are software engineer and need to write a program that decide in on given image is a human.

Question: How would you approach this problem?



Goal: Predict label for given input, how?

Choose a family of prediction functions $\phi(a, x)$ parametrized by $x \in R^d$

Example: linear predictor $\phi(a, x) = a^T x$

Task: Find a good parameter x !

$(a_i, y_i) \sim \text{Distribution}$

$$\xrightarrow{\quad} \phi(a_i, x) \approx y_i$$

Choose a **loss function** to measure the success/failure

$$f(a, b)$$

$$f(a, b) = \|a - b\|^2$$

prediction

true label

Training procedure

 $\min_{x \in \mathbb{R}^d}$

$$E[f(\phi(a_i; x), y_i)]$$



$$(a_i, y_i) \sim \text{Distribution}$$

Sample n points i.i.d.

$$\{(a_i, y_i)\}_{i=1}^n \sim \text{Distribution}$$

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^n f(\phi(a_i; x), y_i) = \frac{1}{n} \sum_{i=1}^n f_i(x)$$

The problem formulation – Empirical risk minimization

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$$

- Large number of samples n

- Large dimension d

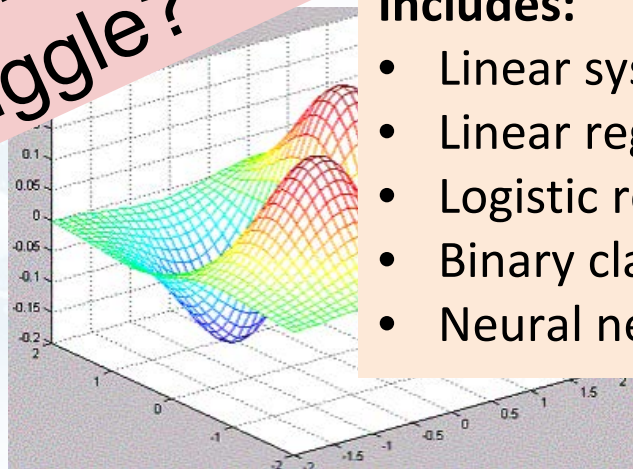
- $f_i(x)$

So many parts to tune....
Does **optimization** have to
be part of the struggle?



Includes:

- Linear systems
- Linear regression
- Logistic regression
- Binary classification
- Neural networks

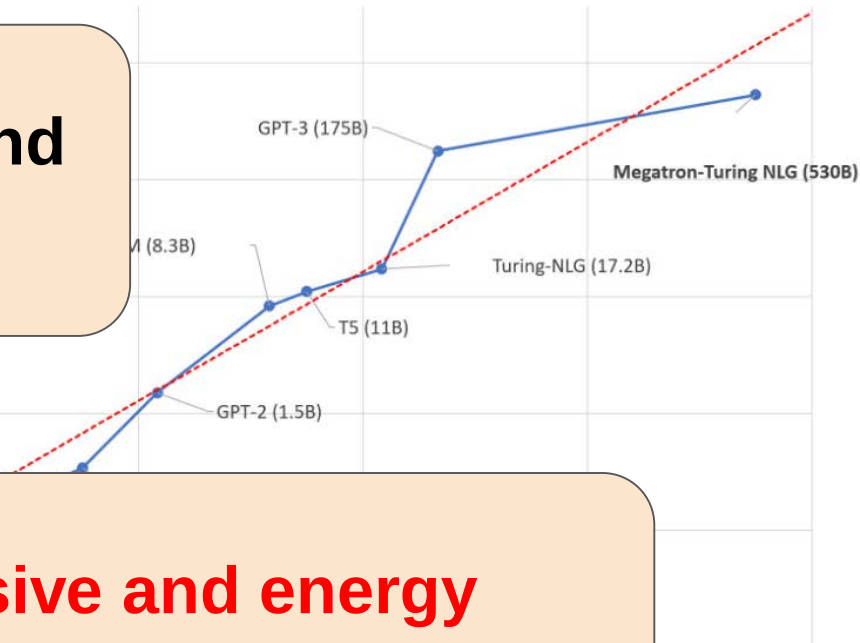


The “Size” Challenge in ML!

Models are getting larger and training takes months!

Bloom	Hugging Face and BigScience	176 billion
GPT-3	OpenAI	175 billion
ESMFold	Meta AI	
LaMDA	Google	
WuDao 2.0	Beijing Academy of Artificial Intelligence	
Gato	DeepMind	

1



Expensive and energy intensive!

Stochastic Gradient Descent (SGD) – (SG?)

The State-Of-The-Art (SOTA) for Training DNN

Stochastic Gradient Descent

n is big

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$$



H. Robbins and S. Monro

A Stochastic Approximation Method

Annals of Mathematical Statistics 22, pp. 400–407, 1951

Sutton Monro

was professor at **Lehigh University**

Stochastic Gradient Descent

stepsize

$$x_{k+1} = x_k - \eta_k \nabla f_i(x_k)$$

$$E[\nabla f_i(x_k)] = \nabla f(x_k)$$

i = chosen uniformly at random

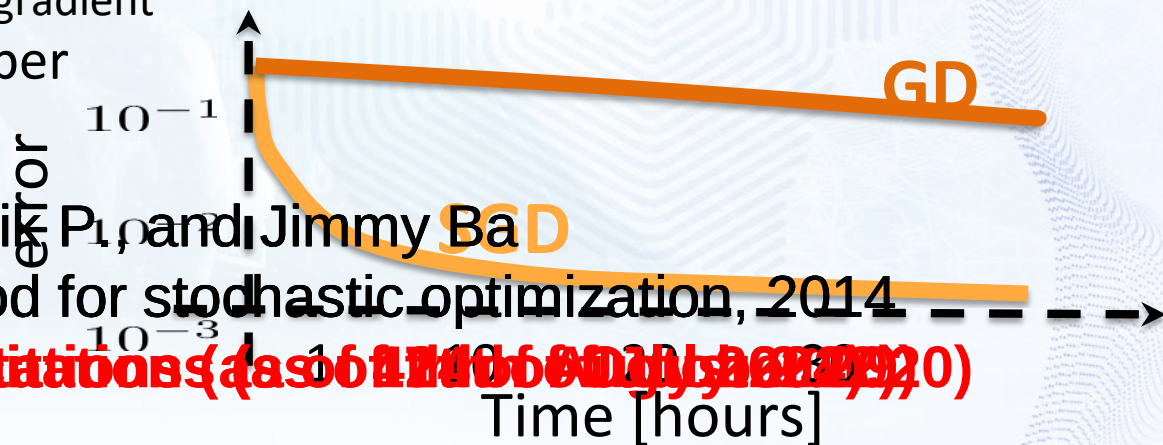
Unbiased estimate of the gradient
Each iteration n times cheaper
than 1 iteration of GD



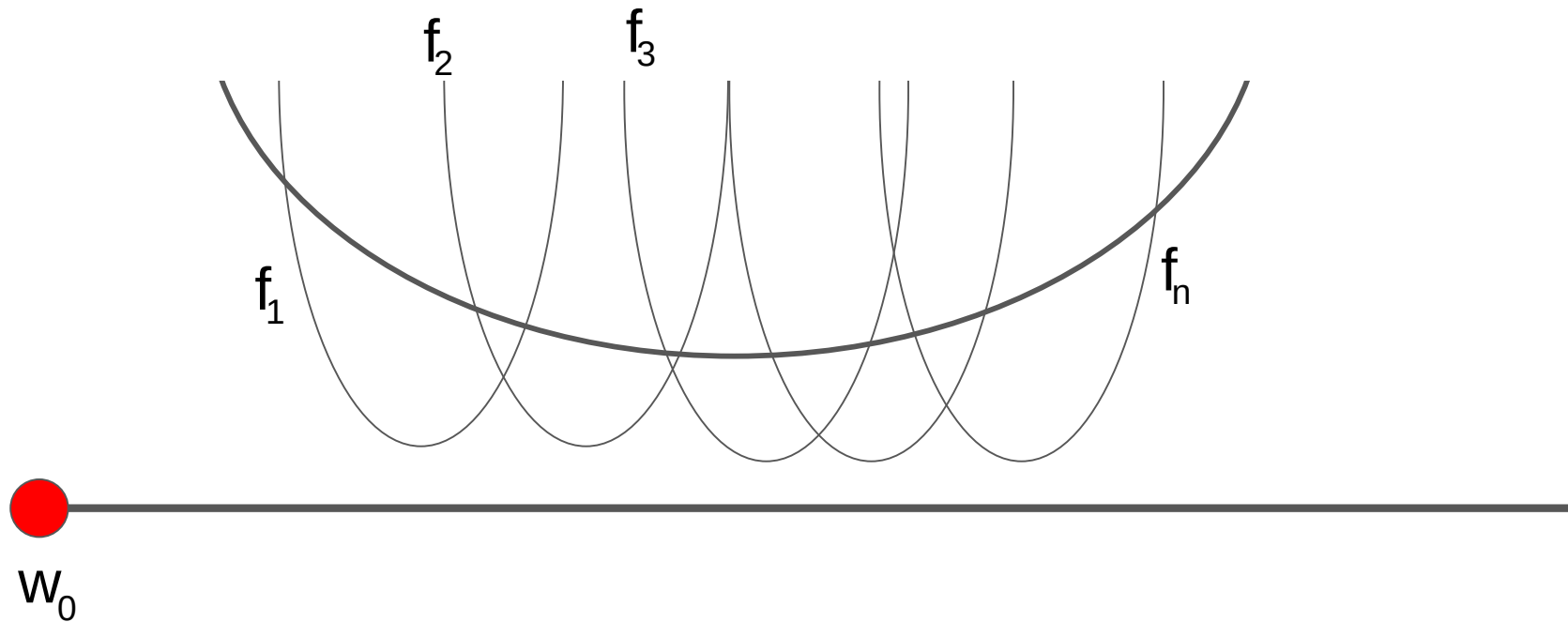
Kingma, Diederik P., and Jimmy Ba

Adam: A method for stochastic optimization, 2014

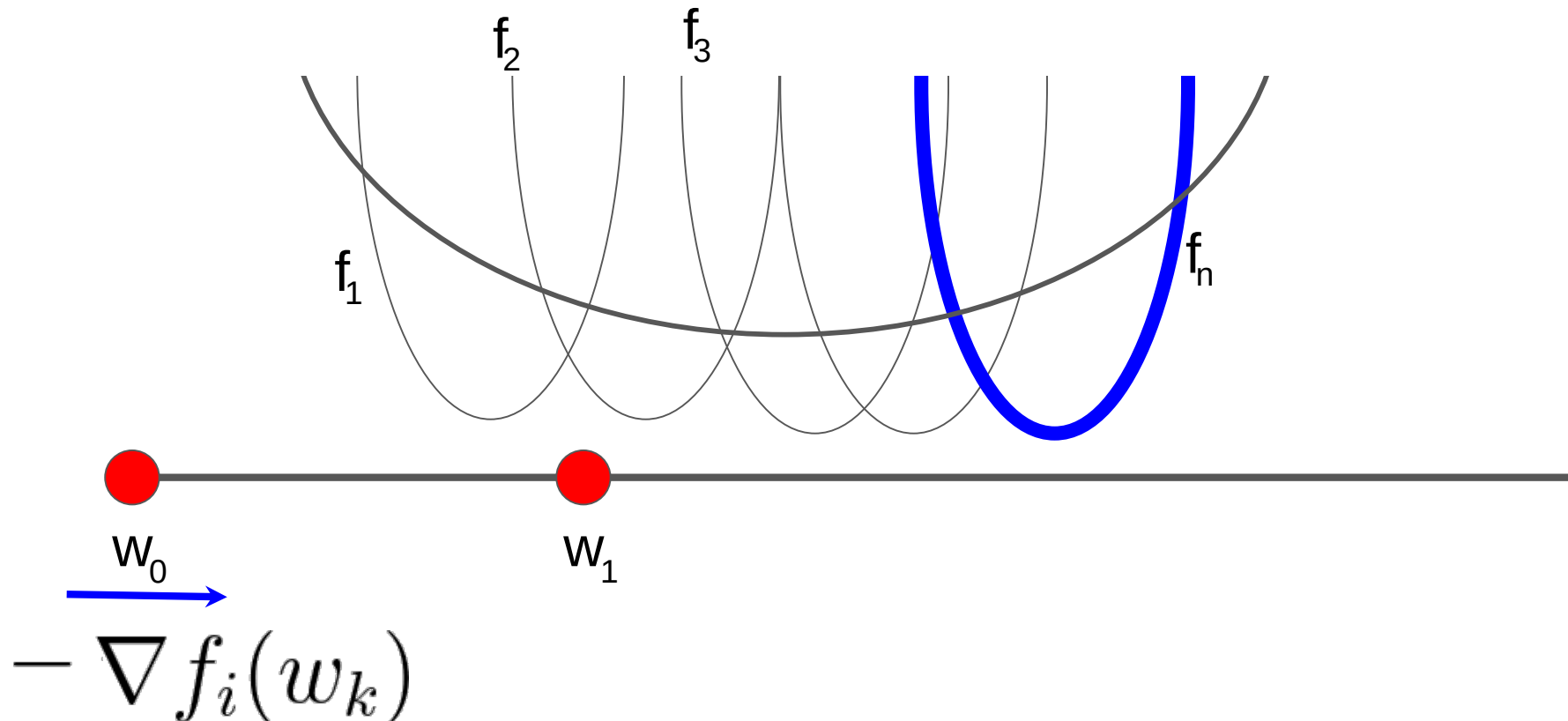
Over 25,000 citations (as of 11/06/2020)



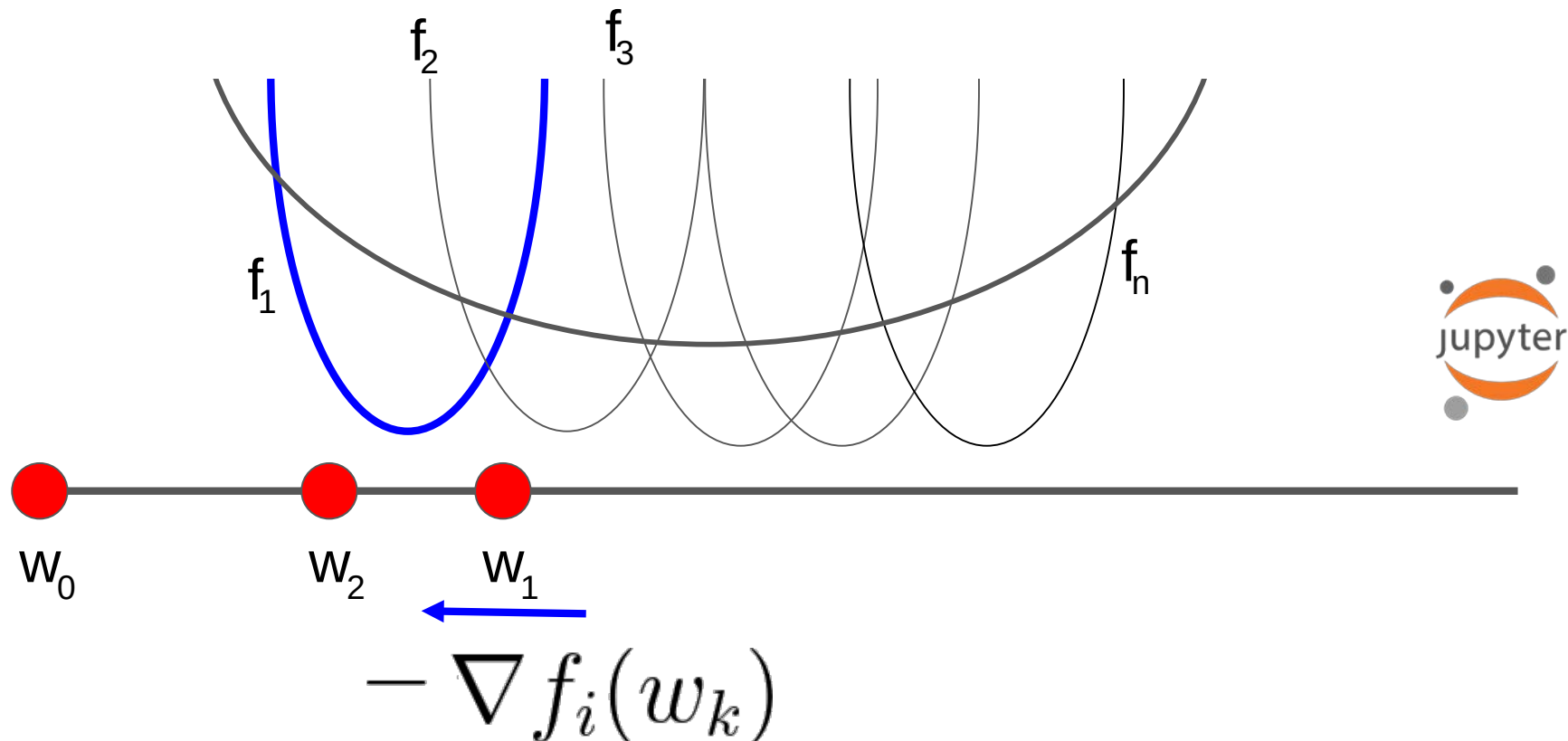
SGD - How it works?



SGD - How it works?



SGD - How it works?



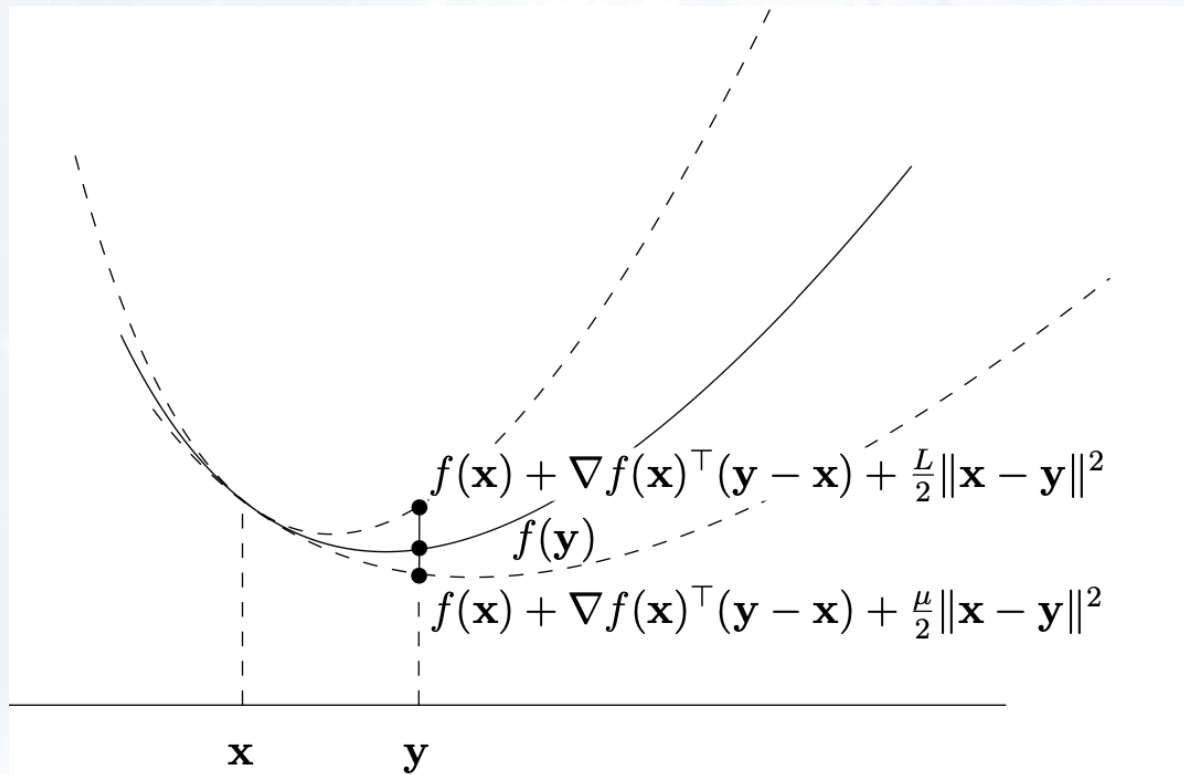
Analysis of SGD



Assumptions

- functions $f(x)$ are convex, L -smooth and μ – strongly convex
- stochastic gradient $g_k = \nabla f_i(x_k)$ has bounded second moment

$$E[\|g_k\|^2] \leq G^2$$



Convergence guarantees - Derivation

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$$

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), -\eta \nabla f(x_k) \rangle + \frac{L}{2} \|\eta \nabla f(x_k)\|^2$$

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), -\eta g_k \rangle + \frac{L\eta^2}{2} \|g_k\|^2$$

$$E[f(x_{k+1})|x_k] - f^* \leq f(x_k) - f^* - \eta \langle \nabla f(x_k), E[g_k|x_k] \rangle + \frac{L\eta^2}{2} E[\|g_k\|^2|x_k]$$

$$E[f(x_{k+1})|x_k] - f^* \leq f(x_k) - f^* - \eta \|\nabla f(x_k)\|^2 + \frac{L\eta^2}{2} G^2$$

$$E[\|g_k\|^2] \leq G^2$$

$$E[f(x_{k+1})|x_k] - f^* \leq f(x_k) - f^* - \eta 2\mu (f(x_k) - f^*) + \frac{L\eta^2}{2} G^2$$

$$\|\nabla f(x)\|^2 \geq 2\mu (f(x) - f^*)$$

$$E[f(x_{k+1})|x_k] - f^* \leq (1 - 2\mu\eta) (f(x_k) - f^*) + \frac{L\eta}{2} G^2$$

Convergence guarantees - Derivation

$$E[f(x_{k+1})] - f^* \leq (1 - 2\mu\eta) (E[f(x_k)] - f^*) + \frac{L\eta^2}{2} G^2$$

$$r_{k+1} \leq (1 - 2\mu\eta) r_k + \frac{L\eta^2}{2} G^2$$

$$r_k = E[f(x_k)] - f^*$$

Q: What can we show about r_k ?

IDEA: Find the fix point!

$$\tilde{r} = (1 - 2\mu\eta)\tilde{r} + \frac{L\eta^2}{2} G^2 \quad \Longrightarrow \quad \tilde{r} = \frac{L\eta}{4\mu} G^2$$

Putting together

$$r_{k+1} - \tilde{r} \leq (1 - 2\mu\eta) r_k - \tilde{r} + \frac{L\eta^2}{2} G^2$$

$$r_{k+1} - \tilde{r} \leq (1 - 2\mu\eta) r_k - \tilde{r} + 2\eta\mu \frac{L\eta}{4\mu} G^2 = (1 - 2\mu\eta) r_k - \tilde{r} + 2\eta\mu \tilde{r}$$

$$r_{k+1} - \tilde{r} \leq (1 - 2\mu\eta) (r_k - \tilde{r}) \quad \Longrightarrow \quad r_k \leq \frac{L\eta}{4\mu} G^2 + (1 - 2\mu\eta)^k (r_0 - \tilde{r})$$

Decreasing variance

$Var[g_k]$ is causing the algorithm to get confused near optimum solution!

Q: How can we make $Var[g_k]$ smaller?

$$g_k^{\{i\}} = \nabla f_i(x_k)$$

Mini-batch SGD

$$S \subset \{1, 2, \dots, n\} \quad g_k^S = \frac{1}{|S|} \sum_{i \in S} \nabla f_i(x_k)$$

$$Var[g_k^S] \leq \frac{1}{|S|^2} \sum_{i \in S} Var[\nabla f_i(x_k)] = \frac{1}{|S|} Var[g_k^{\{i\}}]$$

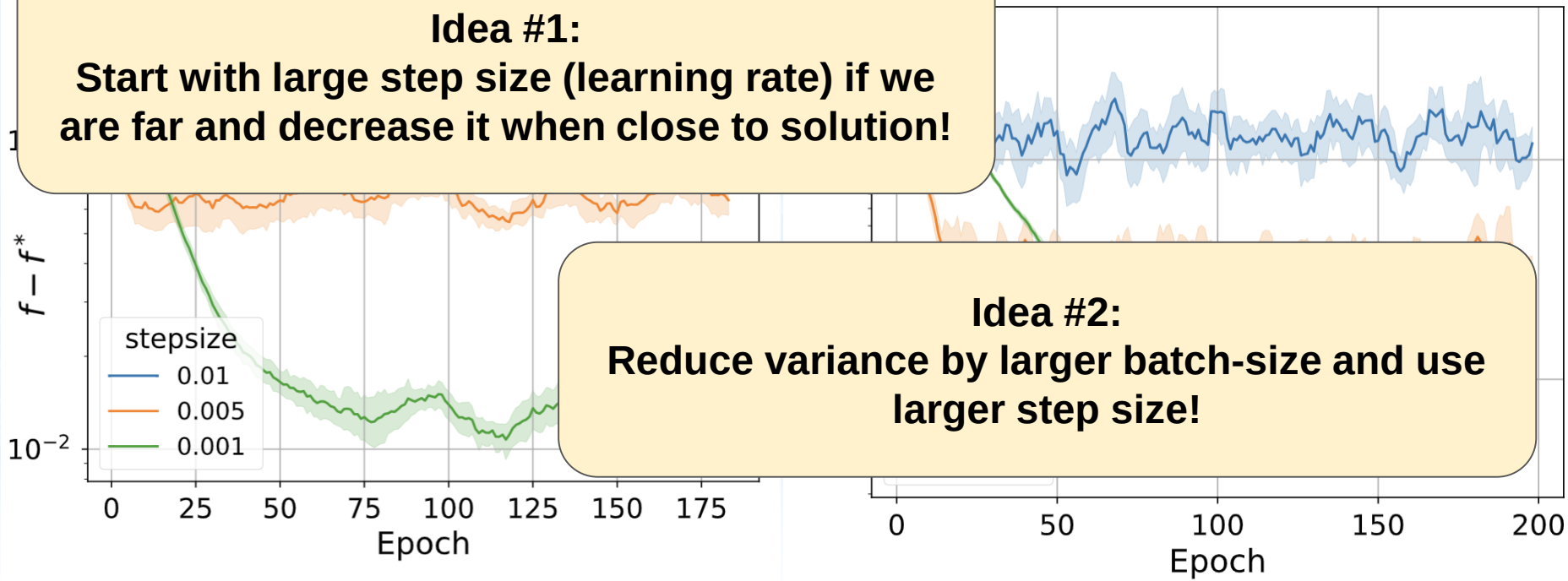
Q: Should we pick large mini-batch from beginning?

How fast is SGD?

$$E[f(x_k) - f^*] \leq (1 - 2\mu\eta)^k (f(x_0) - f^*) + \frac{L\eta}{4\mu} G^2$$

Idea #1:

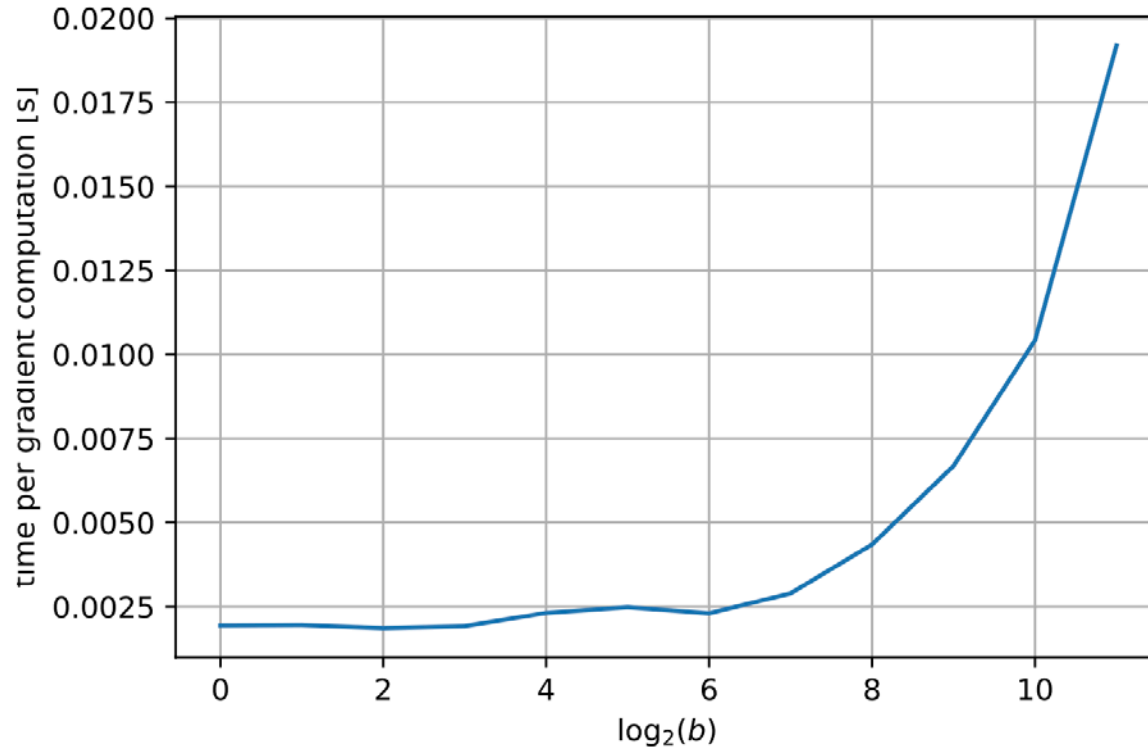
Start with large step size (learning rate) if we are far and decrease it when close to solution!



Idea #2:

Reduce variance by larger batch-size and use larger step size!

GPU – larger batch-size does not mean slower





Smoothing

Smoothing

- choosing only one function (or a mini-batch) can be very noisy
- could we reuse past gradients (with decaying weight) - as a proxy for full gradient? **What do you think?**

SGD

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla f_i(\mathbf{x}_k)$$

SGD + momentum

$$\mathbf{m}_{k+1} = \beta \mathbf{m}_k + (1 - \beta) \nabla f_i(\mathbf{x}_k)$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \mathbf{m}_{k+1}$$

- usual choice of $\beta \in (0.95, 0.99)$
- let $\mathbf{m}_0 = \mathbf{0}$, then $\mathbf{m}_k$ is weighted sum of the past gradients!

**Is it weighted average
or just weighted sum**



Smoothing

$$m_{k+1} = \beta m_k + (1 - \beta) \nabla f_i(x_k)$$

Is it weighted average
or just weighted sum?



$$m_0 = 0$$

$$m_1 = (1 - \beta) \nabla f_i(x_0)$$

$$m_2 = \beta m_1 + (1 - \beta) \nabla f_i(x_1) = \beta(1 - \beta) \nabla f_i(x_0) + (1 - \beta) \nabla f_i(x_1)$$

$$m_k = (1 - \beta) \sum_{l=1}^k \beta^{k-l} \nabla f_i(x_l)$$

SGD + momentum

$$\hat{m}_{k+1} = \frac{m_{k+1}}{1 - \beta^{k+1}}$$

$$x_{k+1} = x_k - \eta \hat{m}_{k+1}$$

Feature Scaling and SGD

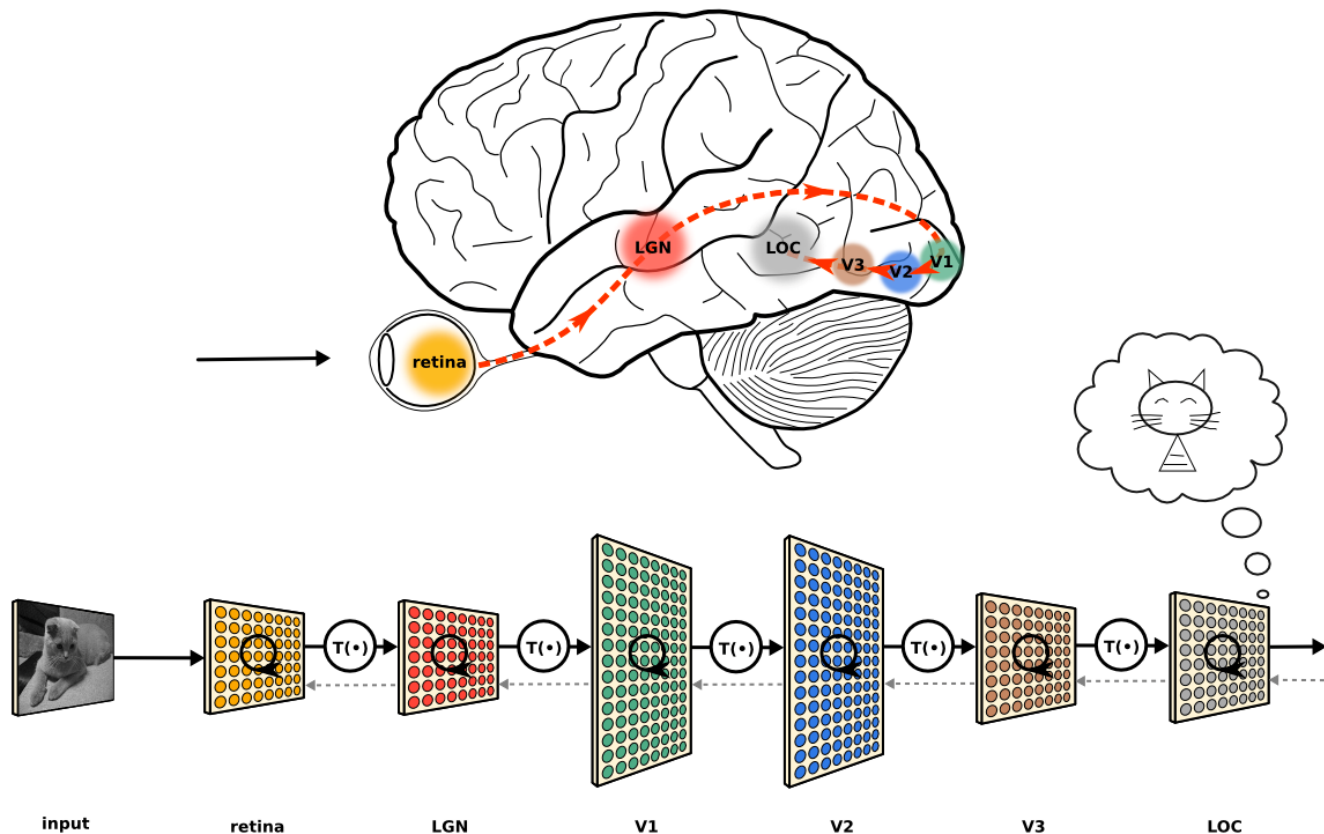
notation change

$$x_k \rightarrow w_k$$

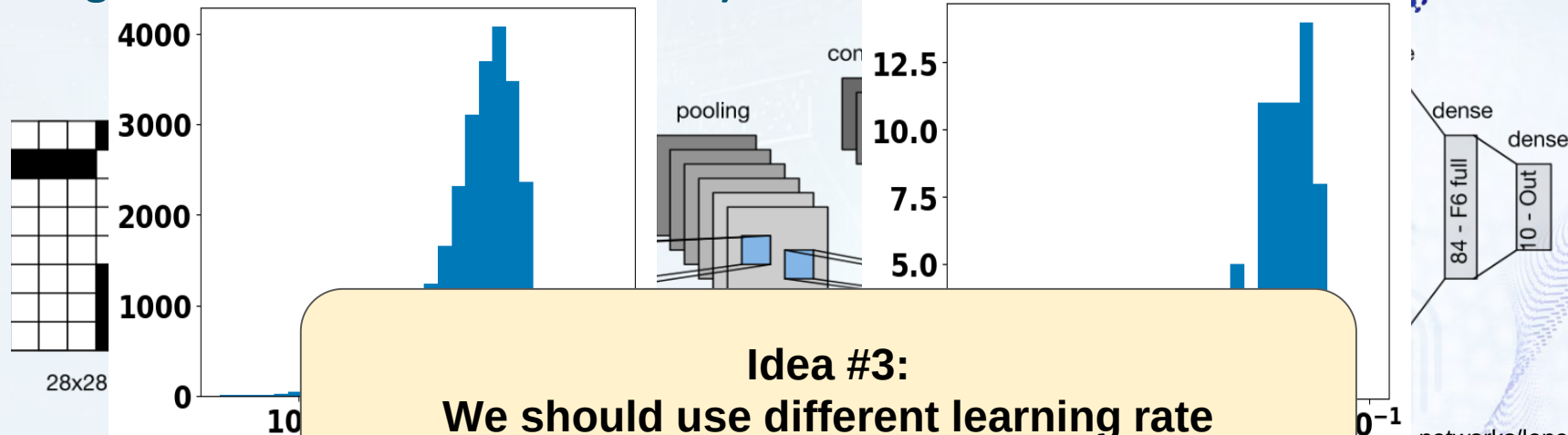
$$f(x) \rightarrow F(w)$$

Training Deep Neural Networks (DNNs)

- Hierarchy of neurons connected
- Multiple types of “layers”
- Can have large number of layers in some architectures!
- Billions or parameters to tune

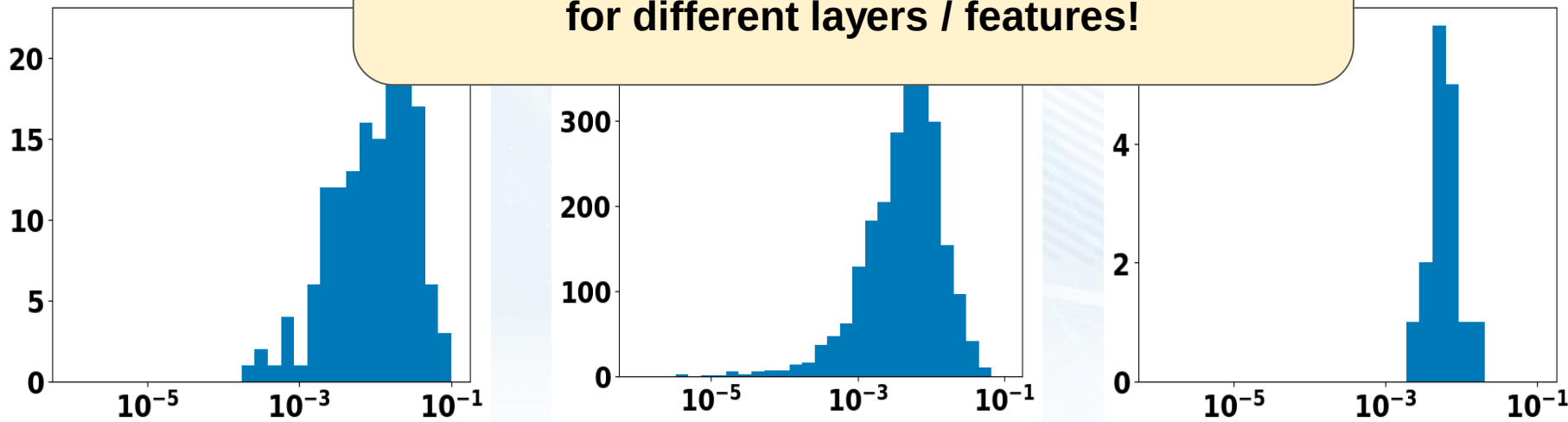


Training Deep Neural Networks (DNNs)

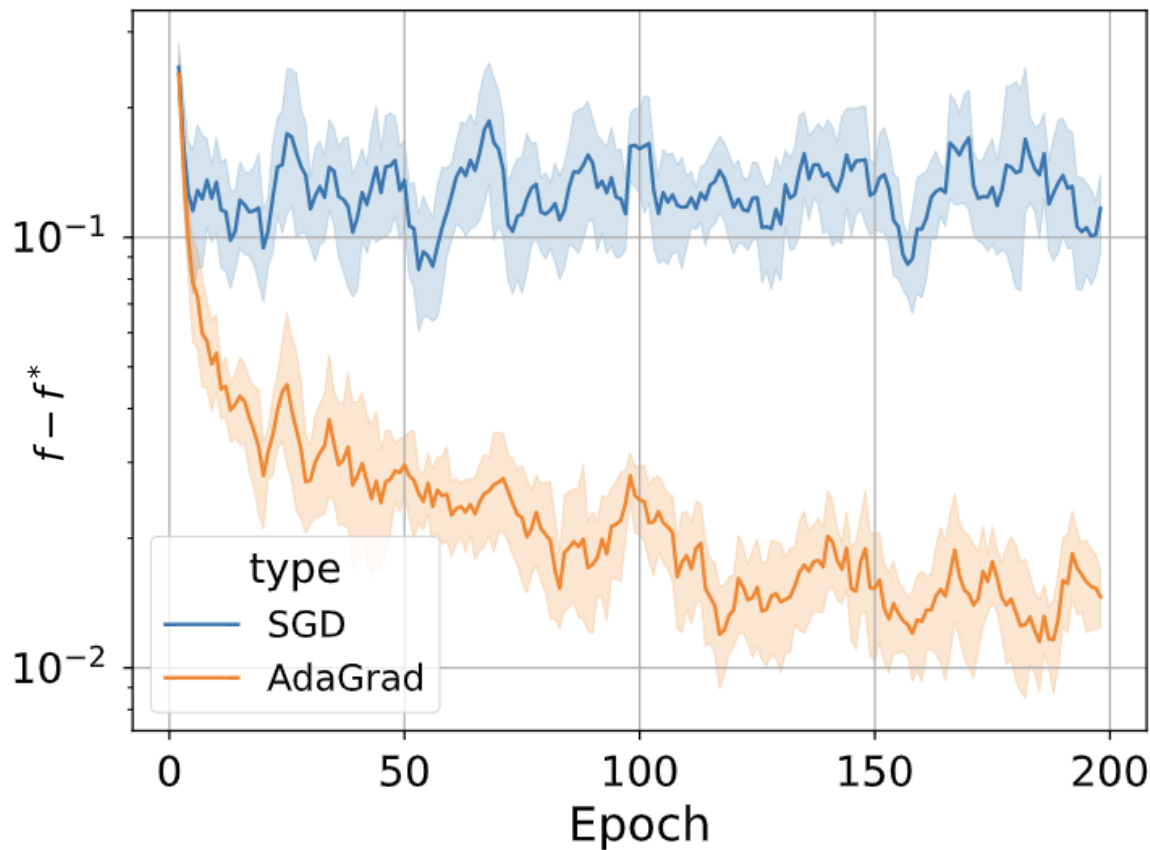


Idea #3:
We should use different learning rate
for different layers / features!

-networks/lenet.html



AdaGrad



John Duchi Elad Hazan Yoram Singer: Adaptive subgradient methods for online learning and stochastic optimization, JMLR 201

AdaGrad

$$G_k = \text{diag} \left(\sum_{l=0}^k g_l \odot g_l \right)$$

$$w_{k+1}^j = w_k^j - \eta \frac{1}{\sqrt{\epsilon + G_{k,jj}}} [g_k]^j$$

Adadelta

- extension of AdaGrad seeks to reduce its **aggressive, monotonically decreasing learning rate**
- instead of accumulating **all past squared gradients**, Adadelta restricts the window of accumulated past gradients to some fixed size T

How to implement it efficiently?

$$G_k = \text{diag} \left(\sum_{l=k-T}^k g_l \odot g_l \right)$$

$$G_k = \beta G_{k-1} + (1 - \beta) g_k \odot g_k$$



- Adam "combines" **momentum** with **Adadelata**

- Adam is storing an exponentially decaying average of past squared gradients

$$v_0 = 0 \quad v_{k+1} = \beta_2 v_k + (1 - \beta_2) g_k \odot g_k$$

- Adam also keeps an exponentially decaying average of past gradients (similar to momentum)

$$m_0 = 0 \quad m_{k+1} = \beta_1 m_k + (1 - \beta_1) g_k$$

Can we just use them?



Algorithm 1: *Adam*, our proposed algorithm for stochastic optimization. See section 2 for details, and for a slightly more efficient (but less clear) order of computation. g_t^2 indicates the elementwise square $g_t \odot g_t$. Good default settings for the tested machine learning problems are $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. All operations on vectors are element-wise. With β_1^t and β_2^t we denote β_1 and β_2 to the power t .

Require: α : Stepsize

Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates

Require: $f(\theta)$: Stochastic objective function with parameters θ

Require: θ_0 : Initial parameter vector

$m_0 \leftarrow 0$ (Initialize 1st moment vector)

$v_0 \leftarrow 0$ (Initialize 2nd moment vector)

$t \leftarrow 0$ (Initialize timestep)

while θ_t not converged **do**

$t \leftarrow t + 1$

$g_t \leftarrow \nabla_{\theta} f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep t)

$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update biased first moment estimate)

$v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)

$\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ (Compute bias-corrected first moment estimate)

$\hat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ (Compute bias-corrected second raw moment estimate)

$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$ (Update parameters)

end while

return θ_t (Resulting parameters)

Practical considerations

- despite having a constant step-size in the algorithm (AdaGrad, Adadelata, Adam) practitioners use “learning rate schedules”

Pytorch schedulers

`lr_scheduler.LambdaLR`

`lr_scheduler.MultiplicativeLR`

`lr_scheduler.StepLR`

`lr_scheduler.MultiStepLR`

`lr_scheduler.ConstantLR`

`lr_scheduler.LinearLR`

`lr_scheduler.ExponentialLR`

`lr_scheduler.PolynomialLR`

`lr_scheduler.CosineAnnealingLR`

`lr_scheduler.ChainedScheduler`

`lr_scheduler.SequentialLR`

`lr_scheduler.ReduceLROnPlateau`

`lr_scheduler.CyclicLR`

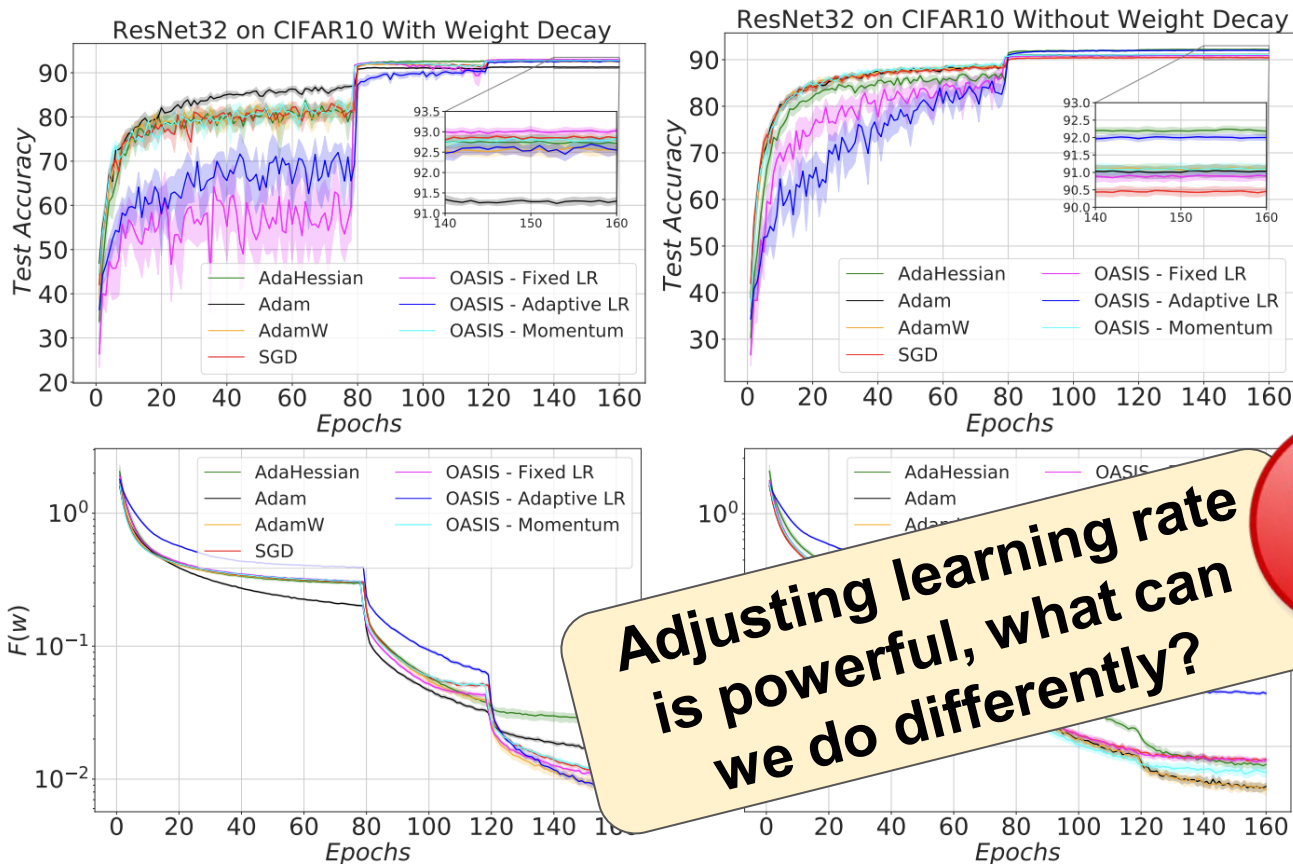
`lr_scheduler.OneCycleLR`

`lr_scheduler.CosineAnnealingWarmRestarts`

$$\eta_k = \frac{\alpha}{\beta + \gamma \cdot k}$$

$$\eta_k = \eta_0 \cdot \gamma^k$$

ResNet32 on CIFAR10



**Adjusting learning rate
is powerful, what can
we do differently?**



Hutchinson's method



Majid Jahani, Sergey Rusakov, Zheng Shi, Peter Richtárik, Michael Mahoney, Martin Takac
OASIS: Doubly Adaptive Scaled Algorithm for Machine Learning Using Second-Order Information ICLR'22

Diagonal Hessian Preconditioning / Scaling

- standard approach in optimization
- simplest - diagonal matrix $P = \text{diag}(\nabla^2 F(w))$

$$w_{t+1} = w_t - \eta_t P^{-1} \nabla f_i(w_t)$$

expensive to compute

- ADADELTA, ADAGRAD, Adam - use past gradients to approximate P

$$P = \sqrt{\sum_i g_i \odot g_i}$$

$$s_t = \beta s_{t-1} + (1 - \beta) g_t \odot g_t$$

$$P = \sqrt{\frac{s_t}{(1 - \beta)^{t+1}}}$$

Diagonal Hessian Preconditioning / Scaling

- standard approach in optimization
- simplest - diagonal matrix $P = \text{diag}(\nabla^2 F(w))$

Hutchinson's method to estimate diagonal of Hessian

$$\text{diag}(\nabla^2 F(w)) = \mathbb{E}_z[z \odot \nabla^2 F(w) z]$$

$z_i = \begin{cases} +1, & \text{with prob. } 0.5 \\ -1, & \text{with prob. } 0.5 \end{cases}$
 IID, randomly drawn from Rademacher distribution

After taking Expectation

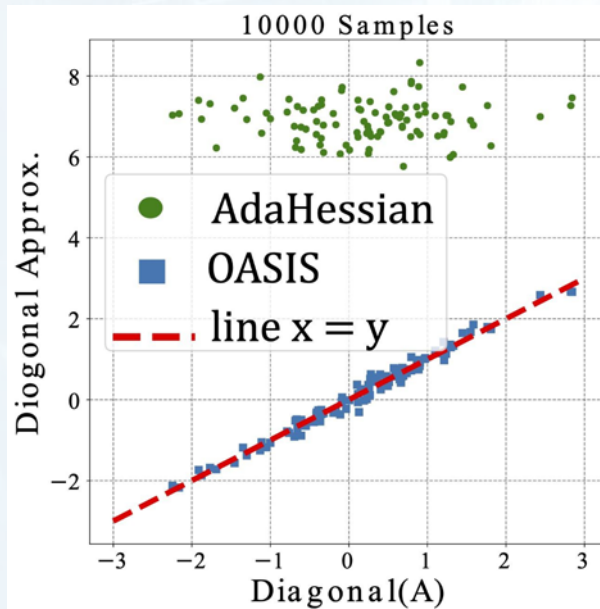
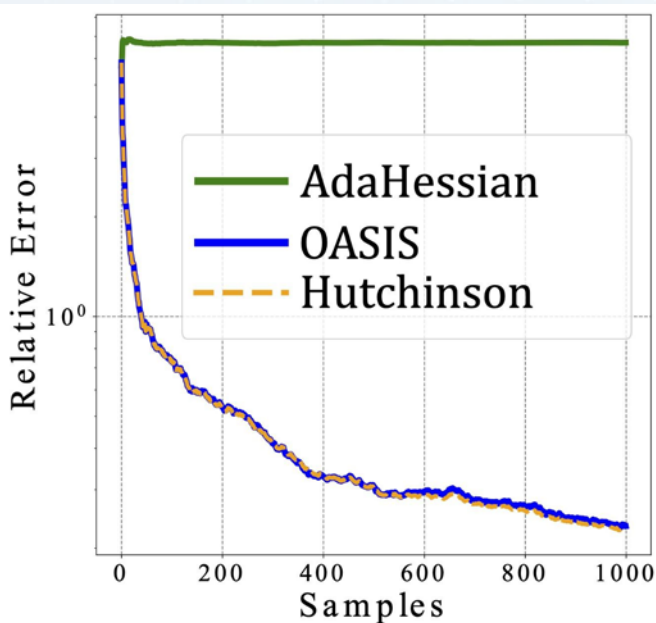
$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \odot \begin{pmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} h_{11} \mathbf{1} + h_{12} \mathbf{0} \\ h_{21} \mathbf{0} + h_{22} \mathbf{1} \end{pmatrix}$$

OASIS vs ADAHESSIAN

$$D_k = \beta_2 D_{k-1} + (1 - \beta_2) \underbrace{\text{diag}(z_k \odot \nabla^2 F(w_k) z_k)}_{v_k}$$

$$(\hat{D}_k)_{i,i} = \max\{|D_k|_{i,i}, \alpha\}$$

$$D_t = \sqrt{\frac{(1 - \beta) \sum_{i=1}^t D_i D_i}{1 - \beta^t}}$$



ADAHESSIAN: An Adaptive Second Order Optimizer for Machine Learning Zhewei Yao, Amir Gholami, Sheng Shen, Mustafa Mustafa, Kurt Keutzer, Michael W. Mahoney, 2020

Adaptive Learning Rate

- Fixed step-size is too conservative
- Oasis extends work of **Konstantin Mishchenko** and **Yura Malitsky**
[Adaptive gradient descent without descent, ICLM 2020]

OASIS Algorithm

Input: $w_0, \eta_0, D_0, \theta_0 = +\infty$

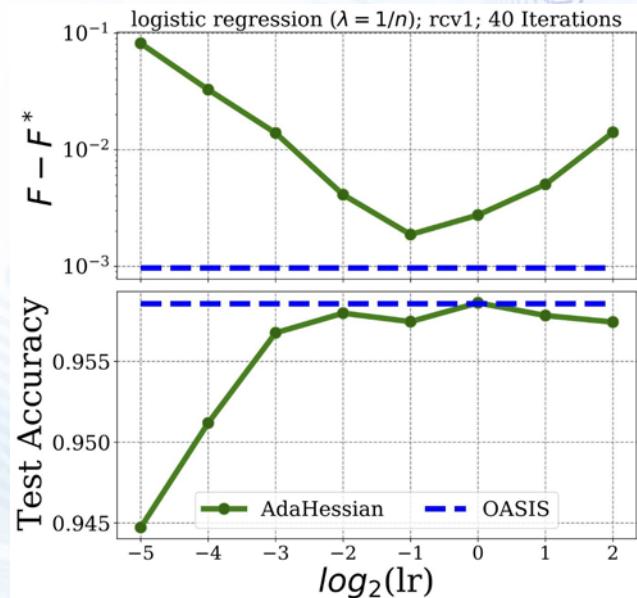
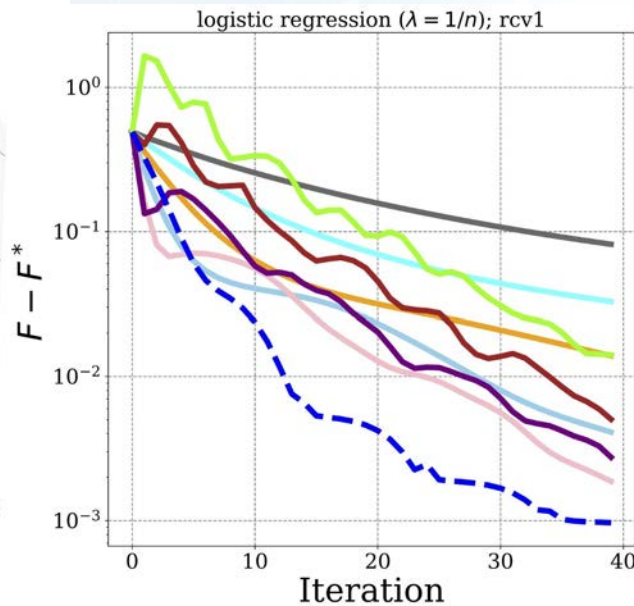
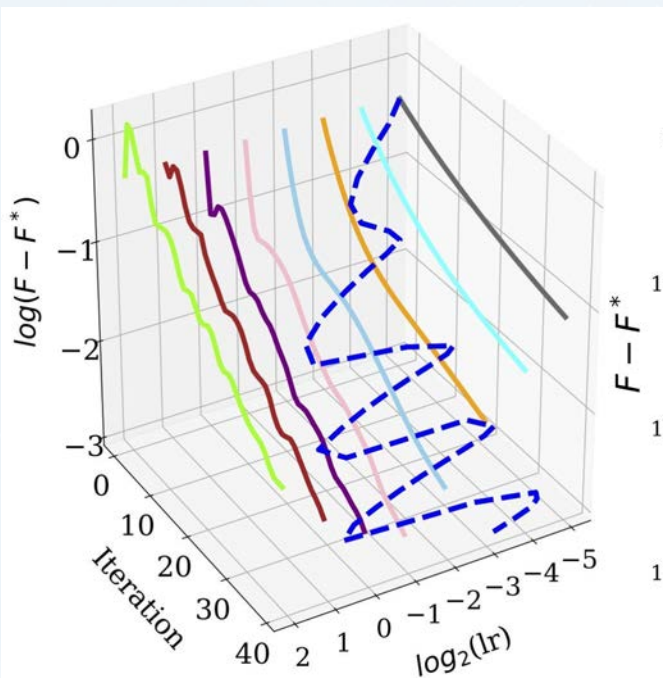
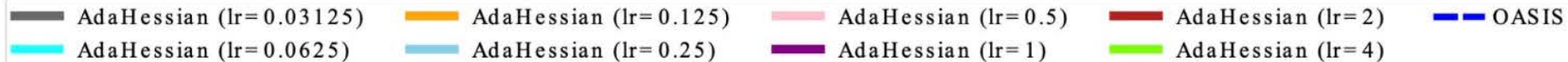
iterate $k = 1, 2, \dots$

sample v_k , update $D_k, \hat{D}_k$

$$\eta_k := \min\left\{\sqrt{1 + \theta_{k-1}\eta_{k-1}}, \frac{\|w_k - w_{k-1}\|_{\hat{D}_k}}{2\|\nabla F(w_k) - \nabla F(w_{k-1})\|_{\hat{D}_k}^*}\right\}$$

$$w_{k+1} = w_k - \eta_k \hat{D}_k^{-1} \nabla F(w_k) \quad \theta_k = \frac{\eta_k}{\eta_{k-1}}$$

Adaptive Learning Rate



Theorem

Let $F(x)$ is twice differentiable, with Lipschitz continuous gradient with parameter L .

Then

- $|(v_k)_i| \leq \Gamma \leq \sqrt{d}L, \quad \forall i \in \{1, 2, \dots, d\}$
 $\exists \delta \leq 2(1 - \beta)\Gamma$ such that $\|D_{k+1} - D_k\|_\infty \leq \delta, \quad \forall k$

recall: $D_k = \beta D_{k-1} + (1 - \beta) \underbrace{\text{diag}(z_k \odot \nabla^2 F(w_k) z_k)}_{v_k}$

Convergence Guarantees

Theorem

Let $F(x)$ is twice differentiable, convex with Lipschitz continuous gradient with parameter L . Then

$$F(\hat{w}_k) - F^* \leq \frac{LC}{k} + 2L(1 - \beta)\Gamma \frac{Q_k}{k}$$

$$C = \frac{2\|w_1 - w^*\|_{\hat{D}_0}^2 + \|w_1 - w_0\|_{\hat{D}_0}^2}{2} + 2\eta_1\theta_1(F(w_0) - F(w^*)),$$

$$Q_k = \sum_{i=1}^k \left(\frac{2\eta_i\theta_i + \alpha}{2\alpha} \|w_{i-1} - w_i\|^2 + \frac{L^2\eta_i\theta_i + \alpha}{\alpha} \|w_i - w^*\|^2 \right).$$

Convergence Guarantees

Theorem

Let $F(x)$ is twice differentiable, μ -strongly convex with Lipschitz continuous gradient with parameter L . If

$\beta \geq \max\left\{1 - \frac{\alpha^4 \mu^4}{4L^2 \Gamma^2 (\alpha^2 \mu^2 + L \Gamma^2)}, 1 - \frac{\alpha^3 \mu^3}{4L \Gamma (2\alpha^2 \mu^2 + L^3 \Gamma^2)}\right\}$, then

$$\Psi^{k+1} \leq \left(1 - \frac{\alpha^2}{2\Gamma^2 \kappa^2}\right) \Psi^k,$$

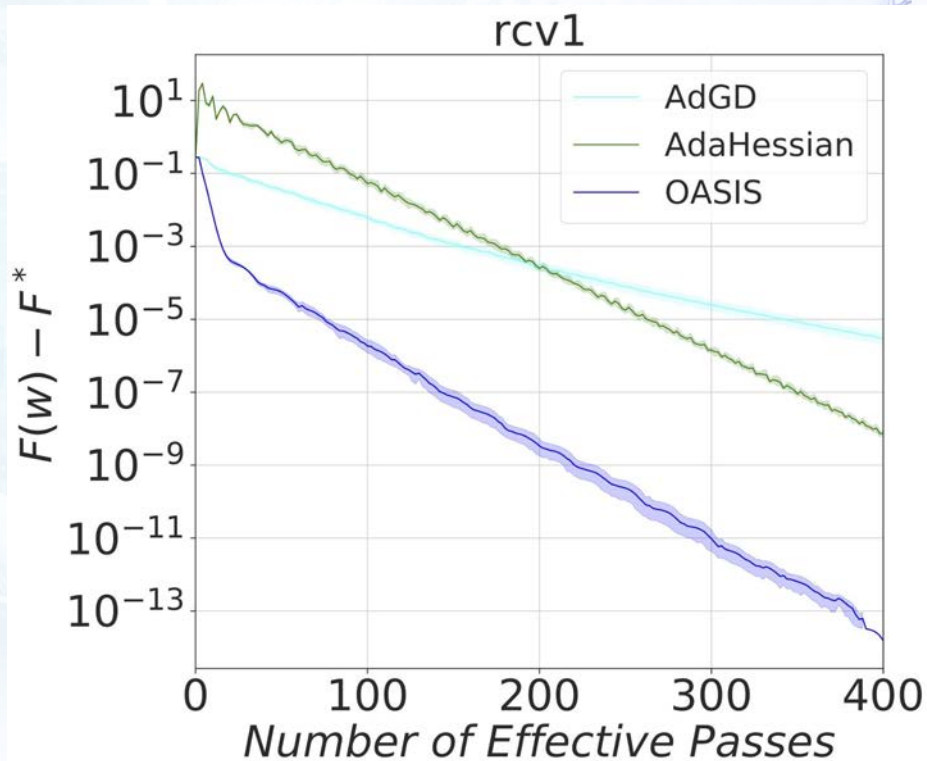
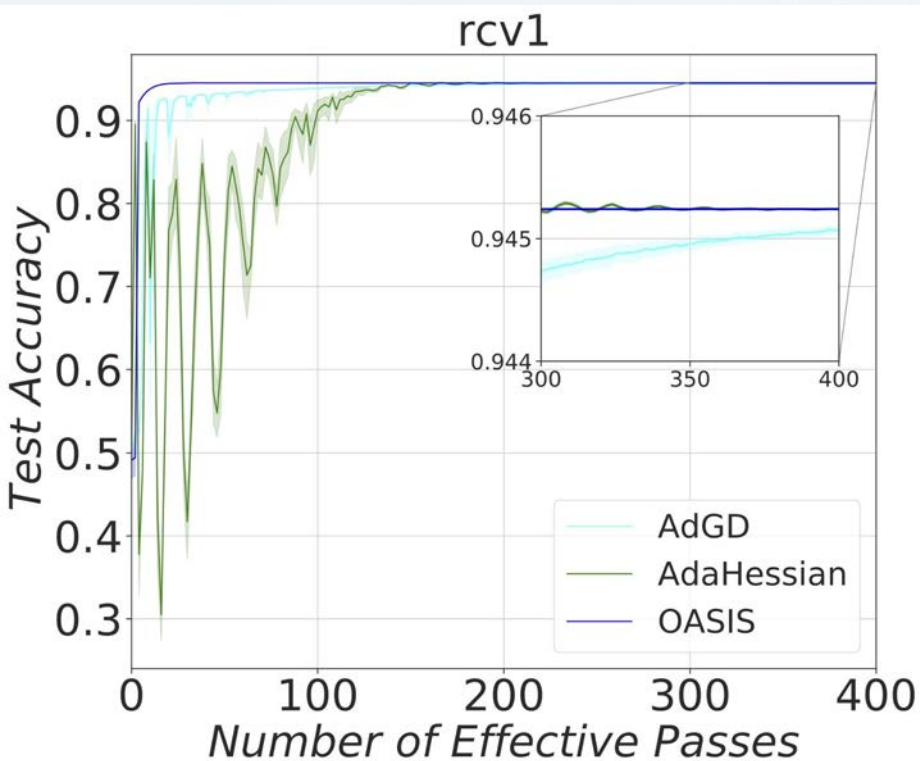
where

$$\Psi^{k+1} = \|w_{k+1} - w^*\|_{\hat{D}_k}^2 + \frac{1}{2} \|w_{k+1} - w_k\|_{\hat{D}_k}^2 + 2\eta_k(1 + \theta_k)(F(w_k) - F(w^*)).$$

In paper we show also convergence for non-convex and stochastic version of OASIS.

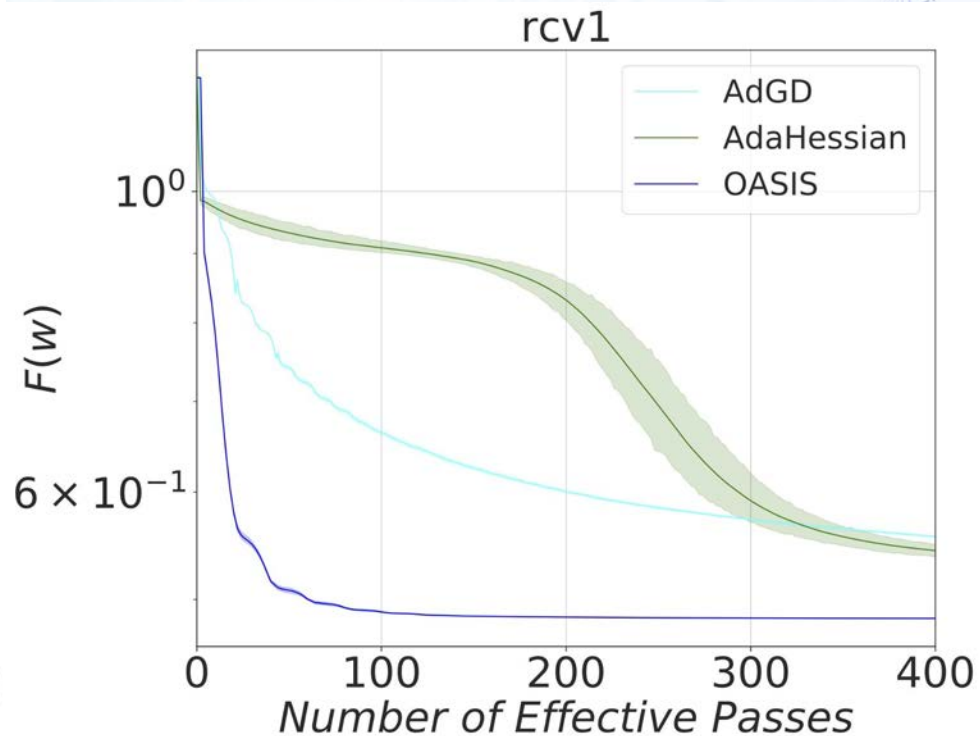
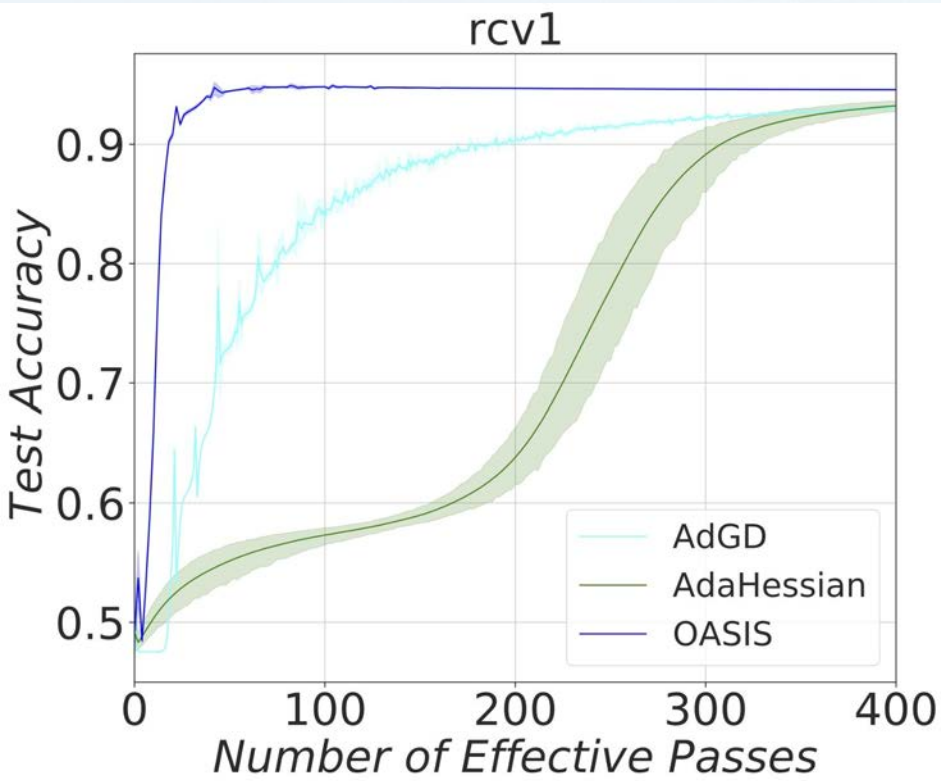
Logistic Regression - Batch Setting

$$F(w) = \frac{1}{n} \sum_{i=1}^n \log(1 + e^{-y_i x_i^T w}) + \frac{\lambda}{2} \|w\|^2$$



Nonlinear regression - Batch Setting

$$F(w) = \frac{1}{n} \sum_{i=1}^n (y_i - 1/(1 + e^{-x_i^T w}))^2$$



SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives

Aaron Defazio
Ambiat *

Australian National University, Canberra

Francis Bach
INRIA - Sierra Project-Team

École Normale Supérieure, Paris, France

Simon Lacoste-Julien
INRIA - Sierra Project-Team
École Normale Supérieure, Paris, France

In this work, SAGA, SDC algorithms, and their variants, like SDC, are shown to be more effective than SAGA.

Accelerating Stochastic Gradient Descent using Predictive Variance Reduction

Rie Johnson
RJ Research Consulting
Tarrytown NY, USA

Tong Zhang
Baidu Inc., Beijing, China
Rutgers University, New Jersey, USA

Stochastic gradient descent convergence analysis, we introduce a new algorithm which we call Stochastic Gradient Descent with Predictive Variance Reduction (SGD-PVR). More generally, we show that the convergence rate of SGD-PVR is faster than that of standard SGD.

SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient

Lam M. Nguyen¹ Jie Liu¹ Katya Scheinberg^{1,2} Martin Takáč¹

Abstract

In this paper, we propose a Stochastic Recursive Gradient algorithm (SARAH), as well as its practical variant SARAH+, as a novel approach to the finite-sum minimization problems. Different from the vanilla SGD and other modern stochastic methods such as SVRG, S2GD, SAGA and SAGA, SARAH admits a simple recursive framework for updating stochastic gradient estimates; when comparing to SAGA/SAGA, SARAH does not require a storage of past gradients. The linear convergence rate of SARAH is proven under strong convexity assumption. We also prove a linear convergence rate (in the strongly convex case) for an inner loop of SARAH, the property that SVRG does not possess. Numerical experiments demonstrate the efficiency of our algo-

Problems of this type arise frequently in supervised learning applications (Hastie et al., 2009). Given a training set $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$, the least squares regression model, for example, is written as (1) with $f_i(w) \triangleq (x_i^T w - y_i)^2 + \frac{\lambda}{2} \|w\|^2$, where $\|\cdot\|$ denotes the ℓ_2 -norm. The ℓ_2 -regularized logistic regression for binary classification is written with $f_i(w) \triangleq \log(1 + \exp(-y_i x_i^T w)) + \frac{\lambda}{2} \|w\|^2$ ($y_i \in \{-1, 1\}$).

In recent years, many advanced optimization methods have been developed for problem (1). While the objective function is smooth and convex, the traditional optimization methods, such as gradient descent (GD) or Newton method are often impractical for this problem, when n – the number of training samples and hence the number of f_i 's – is very large. In particular, GD updates iterates as follows

$$w_{t+1} = w_t - \eta_t \nabla P(w_t), \quad t = 0, 1, 2, \dots$$



MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

Reducing Variance of SGD

SAG/SAGA

- *Minimizing Finite Sums with the Stochastic Average Gradient*, Mark Schmidt, Nicolas Le Roux, Francis Bach 2013
- *SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives*, Aaron Defazio, Francis Bach, Simon Lacoste-Julien, 2014

SAG / SAGA - IDEA

Research Question: Can we modify SGD to achieve convergence without “tricks”

Requirements: choose and evaluate gradient for **ONLY** one stochastic function $f_i(x)$

$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$	$f_6(x)$	$f_7(x)$	$f_8(x)$
0	0	0	0	0	0	0	0

stored
gradients
 y_i^0

when
computed

- x_0
- pick stochastic function f_5
- compute gradient and store in table

SAG / SAGA - IDEA

Research Question: Can we modify SGD to achieve convergence without “tricks”

Requirements: choose and evaluate gradient for **ONLY** one stochastic function $f_i(x)$

$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$	$f_6(x)$	$f_7(x)$	$f_8(x)$
0	0	0	0	$\nabla f_5(x_0)$	0	0	0
				x_0			

stored
gradients
 y_i^0

when
computed

- x_0
- pick stochastic function f_5
- compute gradient and store in table
- update $x_1 = x_0 - \eta \frac{1}{8} \sum_i y_i^0$

SAG / SAGA - IDEA

Requirements: choose and evaluate gradient for **ONLY** one stochastic function $f_i(x)$

$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$	$f_6(x)$	$f_7(x)$	$f_8(x)$
0	0	0	0	$\nabla f_5(x_0)$	0	$\nabla f_7(x_1)$	0
				x_0		x_1	

stored
gradients
 y_i^1

when
computed

- x_1
- pick stochastic function f_7
- compute gradient and store in table

SAG / SAGA - IDEA

Requirements: choose and evaluate gradient for **ONLY** one stochastic function $f_i(x)$

$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$	$f_5(x)$	$f_6(x)$	$f_7(x)$	$f_8(x)$
0	0	0	0	$\nabla f_5(x_0)$	0	$\nabla f_7(x_1)$	0
				x_0		x_1	

stored
gradients
 y_i^1

when
computed

- x_1
- pick stochastic function f_7
- compute gradient and store in table
- update $x_2 = x_1 - \eta \frac{1}{8} \sum_i y_i^1$

Question: should it converge to optimal solution (with fixed step-size)?

SAG / SAGA

(SAG)

$$x_{k+1} = x_k - \eta \left(\frac{\nabla f_i(x_k) - y_i^{k-1}}{n} + \frac{1}{n} \sum_i y_i^{k-1} \right)$$

$$y_i^k = \begin{cases} \nabla f_i(x_k) & \text{if function } i \text{ was selected at } k\text{-th iteration} \\ y_i^{k-1} & \text{otherwise} \end{cases}$$

if function i was selected
at k -th iteration
otherwise

(SAGA)

$$x_{k+1} = x_k - \eta \left(\nabla f_i(x_k) - y_i^{k-1} + \frac{1}{n} \sum_i y_i^{k-1} \right)$$

Theorem:

$$E[\|x_k - x^*\|^2] \leq \left(1 - \frac{\mu}{2(\mu n + L)} \right)^k \cdot \left(\|x_0 - x^*\|^2 + \frac{n}{n\mu + L} (f(x_0) - f^*) \right)$$

Pros:

- **Linear** convergence rate!

Contra:

- Extra storage!
Need to store “ n ” gradients.

SVRG: Stochastic Variance Reduced Gradient

Accelerating Stochastic Gradient Descent using Predictive Variance Reduction, Rie Johnson, Tong Zhang, 2013
A Proximal Stochastic Gradient Method with Progressive Variance Reduction, Lin Xiao, Tong Zhang, 2014

- Modify stochastic gradient (decrease variance overtime)

1. choose x_0
2. set $\tilde{x} = x_0$
3. for $k = 0, 1, 2, \dots$
4. choose $i \in \{1, 2, \dots, n\}$
5. $x_{k+1} = x_k - \eta(\nabla f_i(x_k) - \nabla f_i(\tilde{x}) + \nabla f(\tilde{x}))$

v_k

Unbiased stochastic gradient: $E[v_k|x_k] = \nabla f(x_k)$

Second moment can be bounded (by suboptimality):

$$E[\|v_k\|^2] \leq 4L(f(x_k) - f^* + f(\tilde{x}) - f^*)$$

Convergence of SVRG

- Choose $\eta < \frac{1}{2L}$ and $m \in N$

such that

$$\alpha = \frac{1}{\mu\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta} < 1$$

- Choose $\tilde{x}^+ \in \{x_0, x_1, \dots, x_{m-1}\}$ randomly, then

$$E[f(\tilde{x}^+) - f^*] \leq \alpha E[f(\tilde{x}) - f^*]$$

Note: For fixed step-size η we do **not converge** to optimal solution!

$$\tilde{x}^{(0)} \rightarrow \tilde{x}^{(1)} \rightarrow \tilde{x}^{(2)} \rightarrow \dots \rightarrow \tilde{x}^{(s)}$$

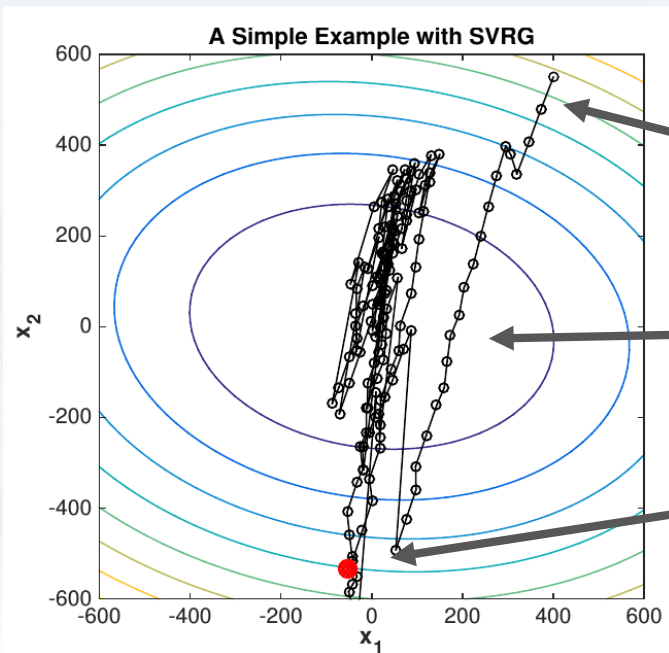


Restarting

$$E[f(\tilde{x}^{(s)}) - f^*] \leq \alpha^s E[f(\tilde{x}^{(0)}) - f^*]$$

SVRG Algorithm

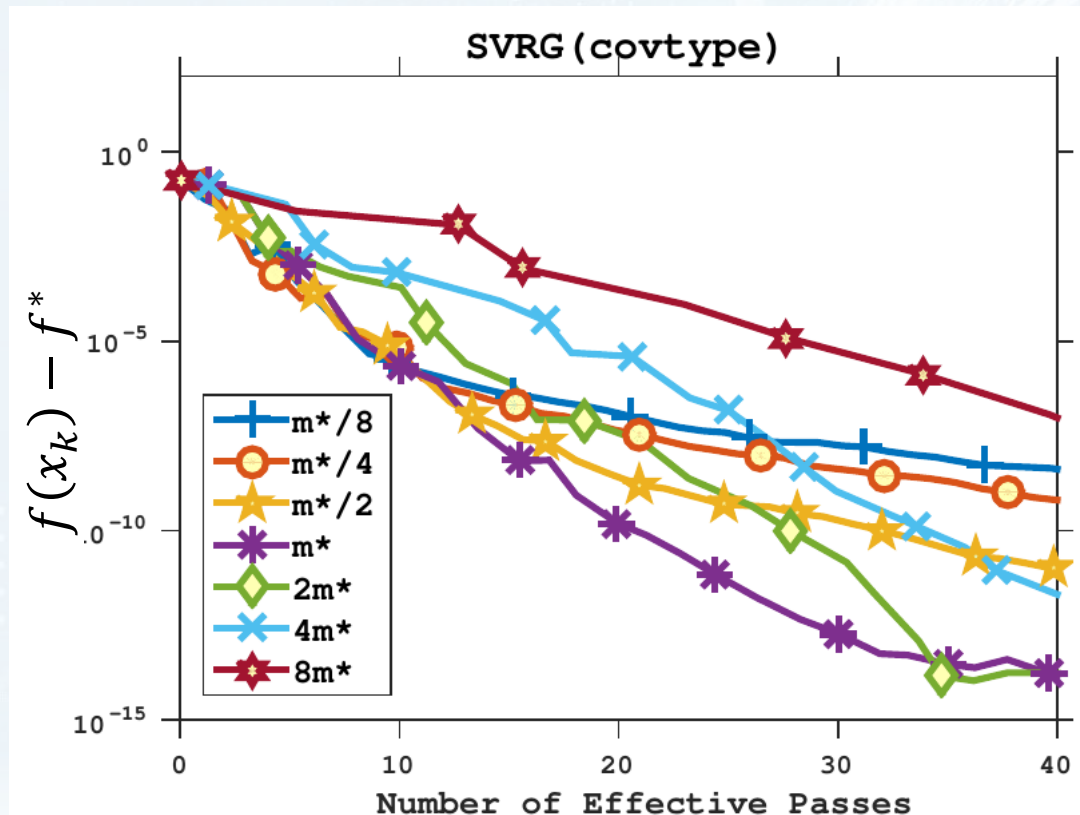
- How to choose “m” in algorithm?
(PS: theory too pessimistic)



“m” too small!
good “m”

“m” too large!

Sensitivity of SVRG on “m”



Research Challenges

- SAG/SAGA - large extra storage!

Can we eliminate it and keep linear convergence?

- SVRG - Performance sensitive on “m”

Can we restart based on runtime criteria?

SARAH

SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient

Lam Nguyen, Jie Liu, Katya Scheinberg, **MIT**, ICML

The New Stochastic Gradient

- We want

$$\mathbf{v}_k \approx \nabla \mathbf{f}(\mathbf{x}_k)$$

- We also want to use only one function $f_i(x)$ to define the stochastic gradient

$$\mathbf{v}_{k+1} = \nabla f_i(x_k) - \nabla f_i(x_{k-1}) + \mathbf{v}_k$$

- A little bit similar to momentum $\mathbf{v}_k = 0.95 \mathbf{v}_{k-1} + 0.05 \nabla \mathbf{f}_i(\mathbf{x}_k)$

The Big Picture

- It does restarting (like SVRG)
- Is “similar” to SAG/SAGA, but **DOESN'T need extra storage**

1. choose x_0 , compute $v_0 = \nabla f(x_0)$
2. set $x_1 = x_0 - \gamma v_0$
3. for $k = 1, 2, \dots, m$
4. choose $i \in \{1, 2, \dots, n\}$
5. update $v_k = \nabla f_i(x_k) - \nabla f_i(x_{k-1}) + v_{k-1}$
6. set $x_{k+1} = x_k - \eta v_k$

No extra storage is needed!

SARAH is Conditionally Biased

- Recall: $v_k = \nabla f_i(x_k) - \nabla f_i(x_{k-1}) + v_{k-1}$

We have

$$E[v_k | \mathcal{F}] = \nabla f(x_k) - \nabla f(x_{k-1}) + v_{k-1} \neq \nabla f(x_k)$$

conditioned on $\{x_0, x_1, \dots, x_{k-1}\}$

However, it holds that $E[v_k] = E[\nabla f(x_k)]$

Theorem:

$$E[\|v_k\|^2] = \begin{cases} \left(1 - \left(\frac{2}{\eta L} - 1\right) \mu^2 \eta^2\right)^t & \leftarrow f(x) \text{ is strongly convex} \\ \left(1 - \frac{2 \mu L \eta}{\mu + L}\right)^t & \leftarrow \forall i: f_i(x) \text{ is strongly convex} \end{cases} \cdot E[\|\nabla F(x_0)\|^2]$$

SARAH is converging (somewhere)!

- Choose $\eta < \frac{2}{L}$, and m such that

$$\alpha = \frac{1}{\mu\eta(m+1)} + \frac{\eta L}{2 - \eta L} < 1$$

- let

$$\tilde{x}^+ \in \{x_0, x_1, \dots, x_{m-1}\}$$

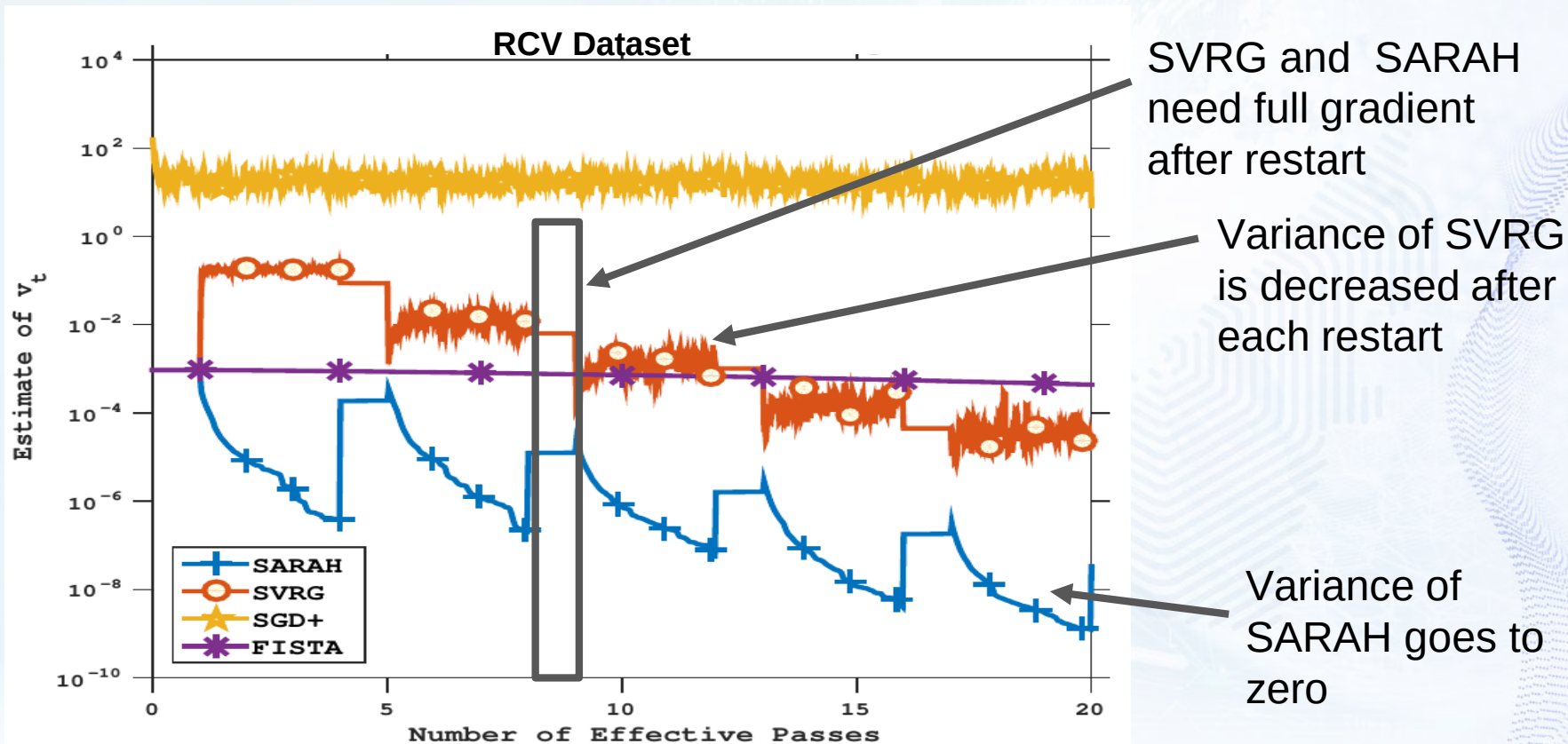
- Then

$$E[\|\nabla f(\tilde{x}^+)\|^2] \leq \alpha \cdot E[\|\nabla f(x_0)\|^2]$$

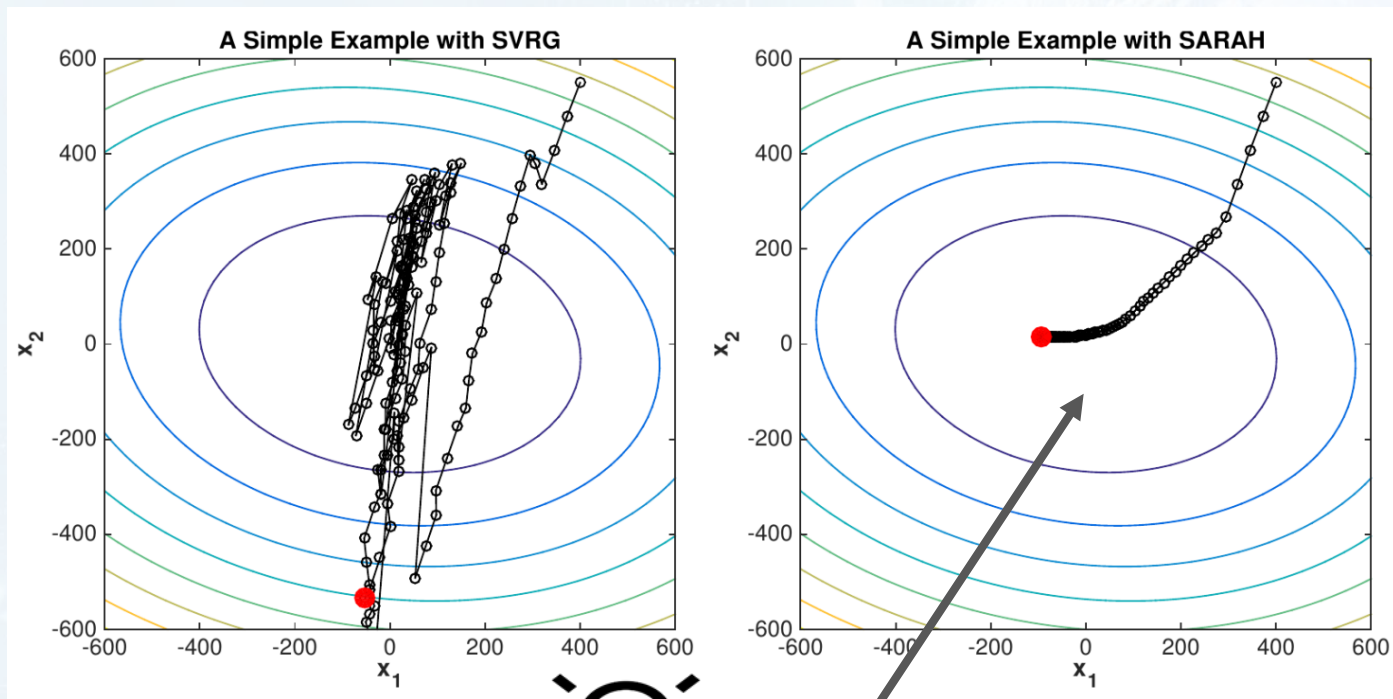
Ok, this is similar to SVRG (a little bit better), but still

why it is cool?

SARAH Demonstration



SARAH Demonstration



Early termination



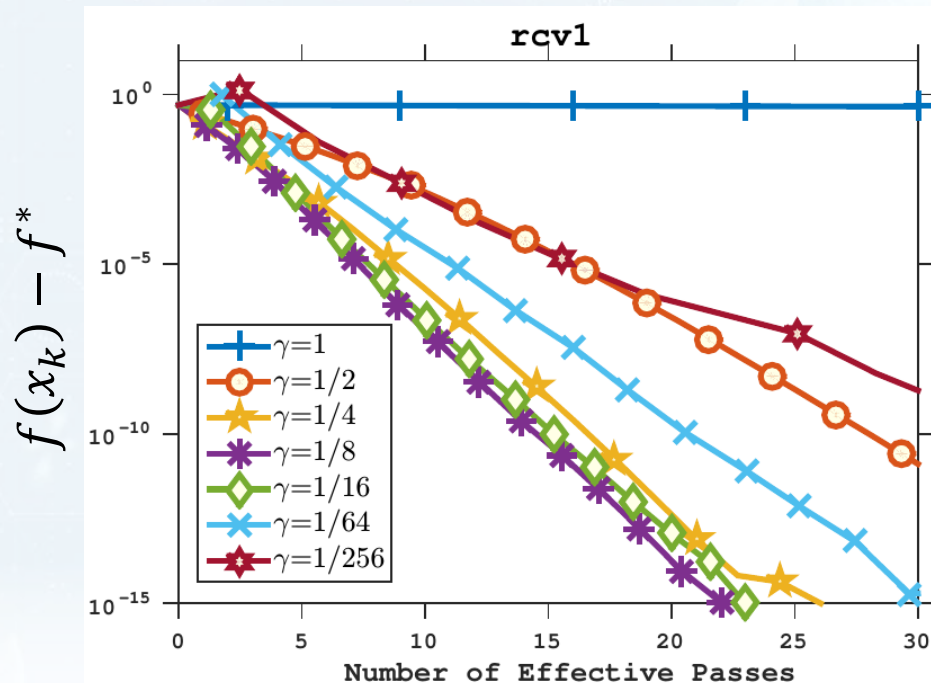
SARAH+

Practical Variant

Fact #1: Size of update is shrinking

It doesn't make sense to do many tiny steps!

Heuristic: **Restart** algorithm when $\|v_k\|^2 < \gamma \|v_0\|^2$

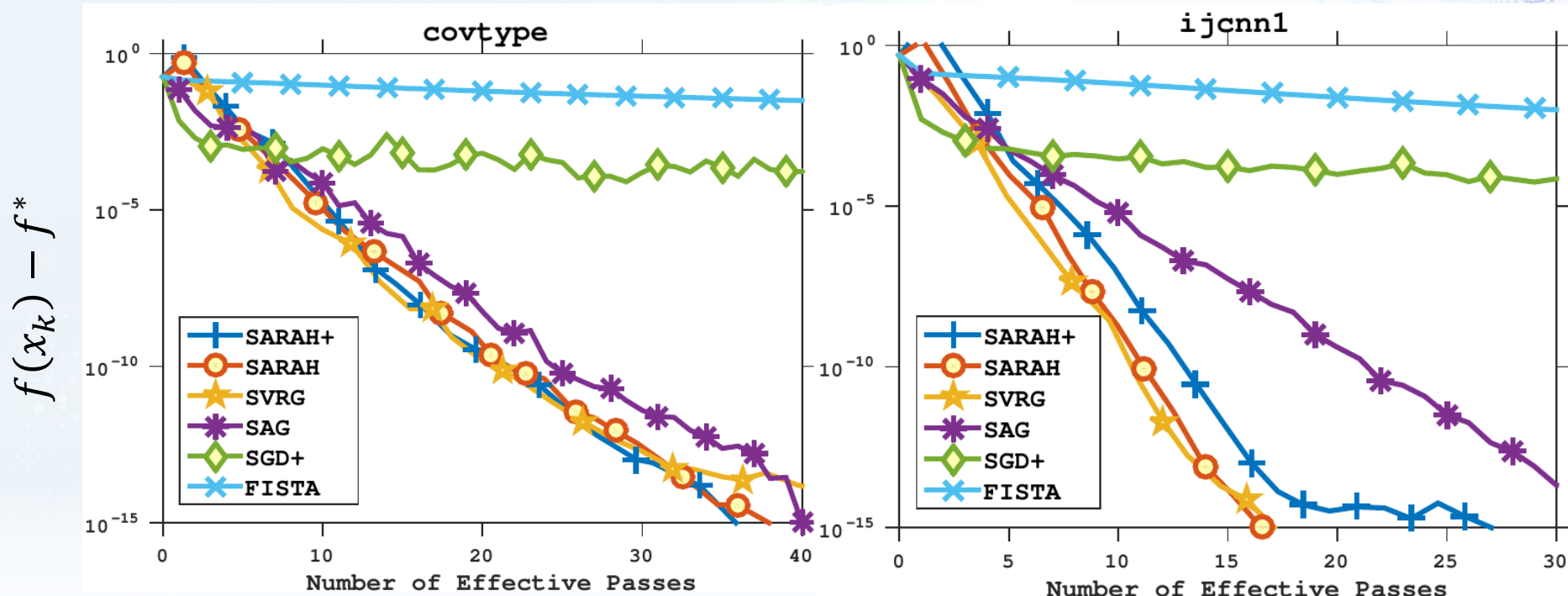


$$\gamma \approx \frac{1}{10}$$

good performance
across many datasets

One has to tune parameters to get a good performance!
Not for SARAH+!

Dataset	SARAH (m^*, η^*)	SVRG (m^*, η^*)	SAG (η^*)	SGD+ (η^*)	FISTA (η^*)
<i>covtype</i>	(2n, 0.9/L)	(n, 0.8/L)	0.3/L	0.06/L	50/L
<i>ijcnn1</i>	(0.5n, 0.8/L)	(n, 0.5/L)	0.7/L	0.1/L	90/L
<i>news20</i>	(0.5n, 0.9/L)	(n, 0.5/L)	0.1/L	0.2/L	30/L
<i>rcv1</i>	(0.7n, 0.7/L)	(0.5n, 0.9/L)	0.1/L	0.1/L	120/L

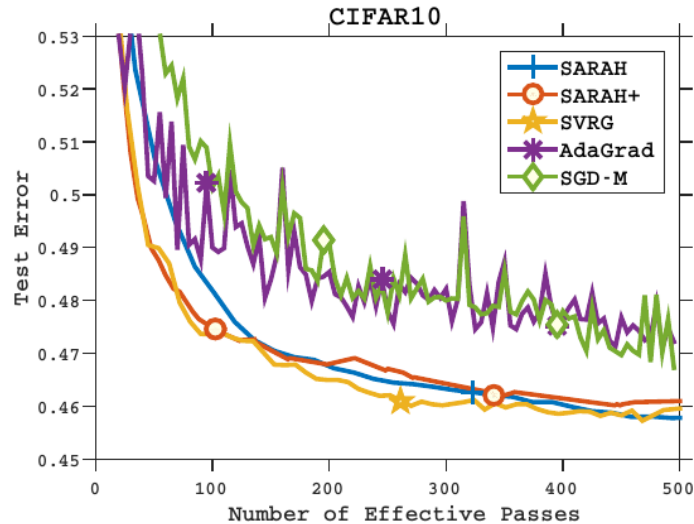
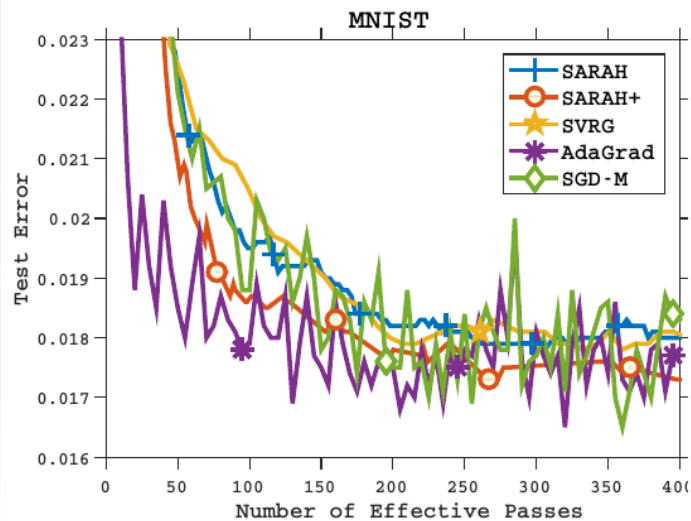
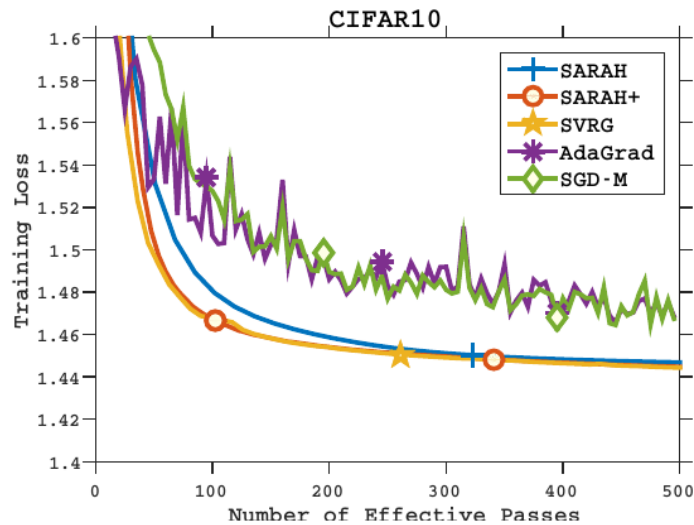
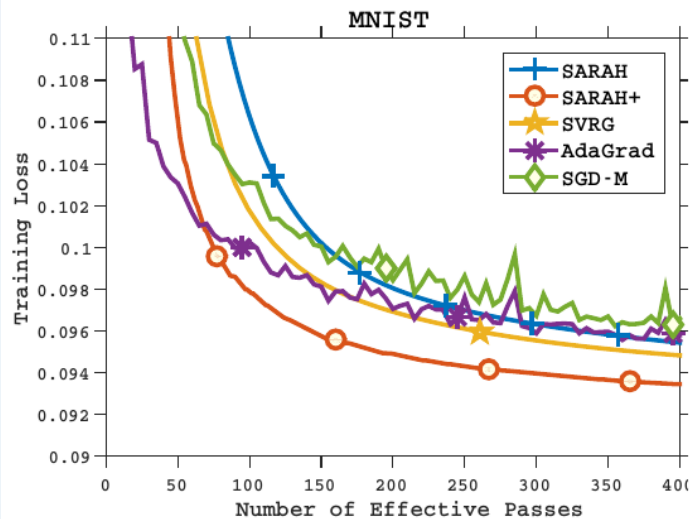


Method	Complexity	Fixed Learning Rate	Low Storage Cost
GD	$\mathcal{O}(n\kappa \log(1/\epsilon))$	✓	✓
SGD	$\mathcal{O}(1/\epsilon)$	✗	✓
SVRG	$\mathcal{O}((n + \kappa) \log(1/\epsilon))$	✓	✓
SAG/SAGA	$\mathcal{O}((n + \kappa) \log(1/\epsilon))$	✓	✗
SARAH	$\mathcal{O}((n + \kappa) \log(1/\epsilon))$	✓	✓

Practical variant available

More cases already covered:

- Smooth, convex objective function (sublinear rate)
- Smooth, non-convex objective function (sublinear rate)
- Smooth, gradient dominated function (linear rate)



The step-size issue

ALL algorithms discussed so far – we need to **PICK step-size** η

Q: why? How to make it better?

- assume that $f(x)$ is convex and L -smooth, i.e. $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ **then**

IF $\eta \leq \frac{1}{L}$ **THEN** Gradient Descent

converges to optimal solution!

What if we do not know L



Backtracking Line-Search

pick x_0 and initial estimate $L_0 \in R_{++}$

iterate $k = 0, 1, 2, 3, \dots$

repeat

choose $\eta = \frac{1}{L_k}$ define $\bar{x} = x_t - \eta \nabla f(x_t)$

IF $f(\bar{x}) \leq f(x_t) - \frac{\eta}{2} \|\nabla f(x_t)\|^2$

$x_{t+1} = \bar{x}$; **break**

ELSE

$L_k = L_k \cdot 2$

$L_k = L_k / 2$

Can we do the same for SGD





Polyak Step-Size

Поляк (Polyak) Step-Size



Борис Поляк

$$\|w_{k+1} - w^*\|^2 = \|w_k - \eta \nabla F(w_k) - w^*\|^2$$

$$= \|w_k - w^*\|^2 + 2\langle w_k - w^*, -\eta \nabla F(w_k) \rangle + \eta^2 \|\nabla F(w_k)\|^2$$

$$F^* - F(w_k) \geq -\langle \nabla F(w_k), w_k - w^* \rangle$$

$$\leq \|w_k - w^*\|^2 - 2\eta(F(w_k) - F^*) + \eta^2 \|\nabla F(w_k)\|^2$$

$\min_{\eta}$

$$\eta = \frac{F(w_k) - F^*}{\|\nabla F(w_k)\|^2}$$

can be replaced
by lowerbound

Stochastic Polyak Step-size (SPS)

$$w_{k+1} = w_k - \eta_k \nabla f_i(w_k)$$

$$\eta_k = \frac{f_i(w_k) - f_i^*}{\|\nabla f_i(w_k)\|^2}$$

SPSmax: $\eta_k = \min \left\{ \gamma_b; \frac{f_i(w_k) - \bar{f}_i^*}{c \|\nabla f_i(w_k)\|^2} \right\}$



Stochastic Polyak Step-size (SPS)

$$\text{SPS}_{\max}: \quad \eta_k = \min \left\{ \gamma_b; \frac{f_i(w_k) - \bar{f}_i^*}{c \|\nabla f_i(w_k)\|^2} \right\}$$

in many ML problems, we can choose $\bar{f}_i^* = 0$

Assumption

$$\sigma^2 = E[f_i(w^*) - \bar{f}_i^*] = F^* - E[\bar{f}_i^*] < +\infty$$

Theorem

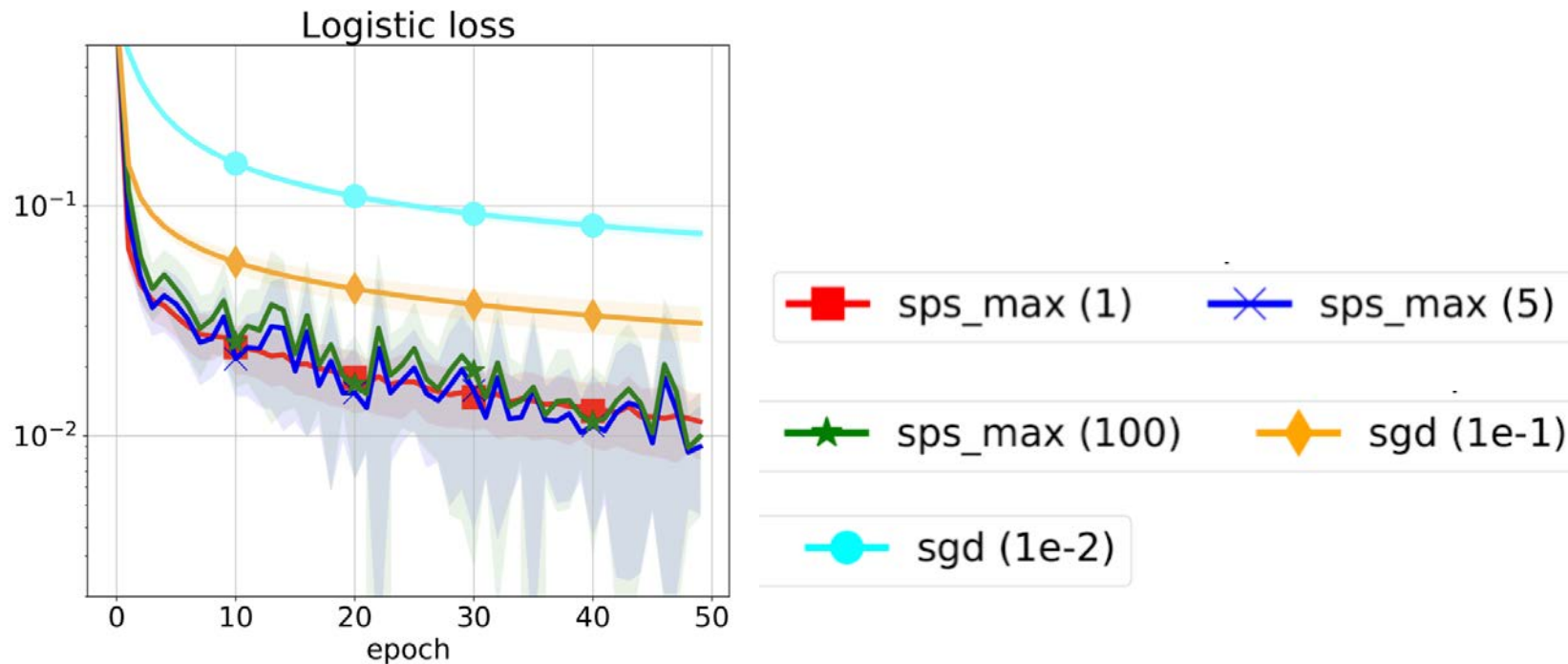
if f_i is L_i smooth, $F(w)$ is μ strongly convex and $c \geq \frac{1}{2}$

then $E[\|w_k - w_*\|^2] \leq (1 - \mu\alpha)^k \|w_0 - w^*\|^2 + \frac{2\gamma_b\sigma^2}{\mu\alpha}$

where $\alpha = \min \left\{ \gamma_b, \frac{1}{2cL_{\max}} \right\}; \quad L_{\max} = \max_i L_i$



Stochastic Polyak Step-size (SPS)



N Loizou, S Vaswani, I Laradji, S Lacoste-Julien: *Stochastic Polyak step-size for SGD: An adaptive learning rate for fast convergence, AISTATS 2020*

Stochastic Polyak Step-size + Interpolation

$$F(w) = \frac{1}{n} \sum_{i=1}^n f_i(w), \quad f_i(w) \geq 0 \quad \forall i \in \{1, 2, \dots, n\}$$

interpolation assumption

Let's assume that: $\exists w^* \in R^d : f_i(w^*) = 0$

Question #1
Why we should use
Euclidean norm?

$$w_{k+1} = \arg \min_w \|w - w_k\|^2$$

such that $f_i(w_k) + \langle \nabla f_i(w_k), w - w_k \rangle = 0$

Question #2
Why linearize?

$$-\frac{f_i(w_k)}{\|\nabla f_i(w_k)\|^2} \nabla f_i(w_k)$$

fazio, Michael Rabbat: *Stochastic Polyak Stepsize with a*

Polyak Step-Size With Scaling

- Stochastic Polyak - the same step size for all features! Can we improve it?

$$w_{k+1} = \arg \min \|w - w_k\|^2_{diag(v)}$$

such that $f_i(w_k) + \langle \nabla f_i(w_k), w - w_k \rangle = 0$

change norm

$$w_{k+1} = w_k - \frac{f_i(w_k)}{(\|\nabla f_i(w_k)\|_v^*)^2} diag(v)^{-1} \nabla f_i(w_k)$$

Adding scaling
to Polyak step size

Polyak Step-Size With Scaling and Curvature

- How can we include curvature information?

$$w_{k+1} = \arg \min \|w - w_k\|_{\text{diag}(v)}^2$$

such that $f_i(w_k) + \langle \nabla f_i(w_k), w - w_k \rangle = 0$



Could replace by Hessian approximation with rank1

$$w_{k+1} = \arg \min \|w - w_k\|_{\text{diag}(v)}^2$$

such that $f_i(w_k) + \langle \nabla f_i(w_k), w - w_k \rangle + \frac{1}{2}(w - w_k)^T \nabla^2 f_i(w - w_k) = 0$

HARD to solve

$$f_i(w) = \phi_i(x_i^\top w - y_i)$$

Lemma 3.1. (SP2) Assume $f_i(w)$ is the loss of a generalized linear model (10) and is non-negative. Let $f_i = f_i(w^t)$ for short. Let $a_i := \phi'_i(x_i^\top w - y_i)$ and $h_i := \phi''_i(x_i^\top w - y_i)$. If

$$a_i^2 - 2h_i f_i \geq 0 \quad (12)$$

then the optimal solution of (7) is as follows

$$w^{t+1} = w^t - \frac{a_i}{h_i} \left(1 - \frac{\sqrt{a_i^2 - 2h_i f_i}}{|a_i|} \right) \frac{x_i}{\|x_i\|^2}. \quad (13)$$

Alternatively if (12) does not hold, since $f_i \geq 0$ we have necessarily that $h_i > 0$, and consequently a Newton step will give the minima of the local quadratic, that is

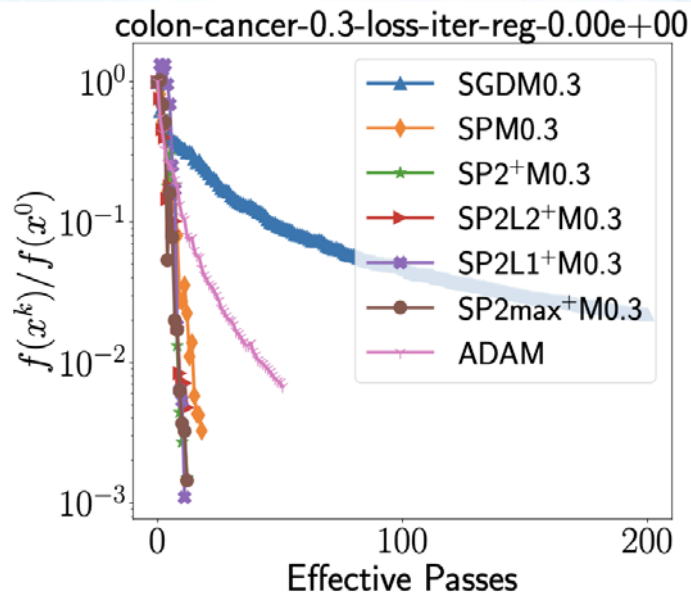
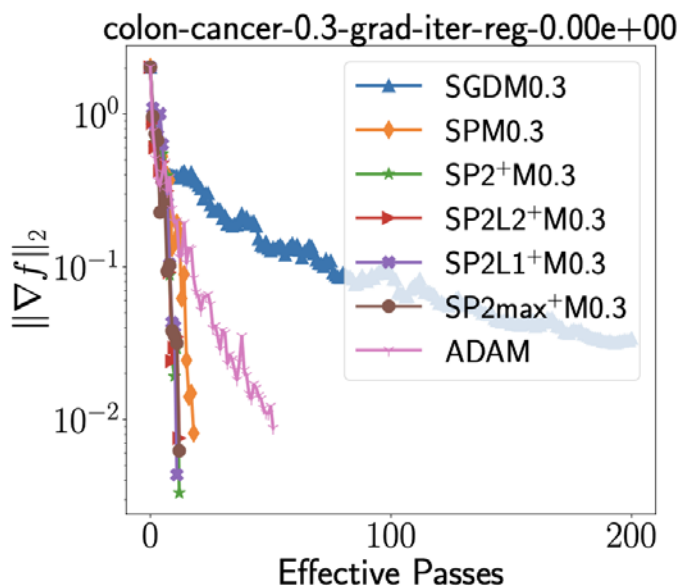
$$w^{t+1} = w^t - \frac{a_i}{h_i} \frac{x_i}{\|x_i\|^2}. \quad (14)$$



Polyak Step-Size With Scaling and Curvature

Lemma 3.3. (SP2⁺) Let $g_t \equiv \nabla f_i(w^t)$ and $\mathbf{H}_t \equiv \nabla^2 f_i(w^t)$

$$w^{t+1} = w^t - \frac{f_i(w^t)}{\|g_t\|^2} g_t - \frac{1}{2} \frac{f_i(w^t)^2}{\|g_t\|^4} \frac{\langle \mathbf{H}_t g_t, g_t \rangle}{\|v^{t+1}\|^2} v^{t+1}, \text{ where } v^{t+1} = \left(\mathbf{I} - \mathbf{H}_t \frac{f_i(w^t)}{\|g_t\|^2} \right) g_t, \quad (20)$$





+ momentum to Polyak step-size (MoMo)

Adding Momentum, how ?

$$w_{k+1} = \arg \min_w m_k(w) + \frac{1}{2\eta_k} \|w - w_k\|^2$$

SGD $m_k(w) = f_i(w_k) + \langle f_i(w_k), w - w_k \rangle$

$$m_k(w) = \frac{1}{\rho_k} \sum_{j=1}^k \rho_j^k f^j(w) \quad \text{with } \rho_k = \sum_{j=1}^k \rho_j^k$$

HARD to solve

$$m_k(w) = \max \left\{ \frac{1}{\rho_k} \sum_{j=1}^k \rho_j^k (f_j(w_j) + \langle \nabla f_j(w_j), w - w_j \rangle), 0 \right\}$$



Adding Momentum, how ?

$$w_{k+1} = \arg \min_w m_k(w) + \frac{1}{2\eta_k} \|w - w_k\|^2$$

$$m_k(w) = \max \left\{ \frac{1}{\rho_k} \sum_{j=1}^k \rho_j^k (f_j(w_j) + \langle \nabla f_j(w_j), w - w_j \rangle), 0 \right\}$$

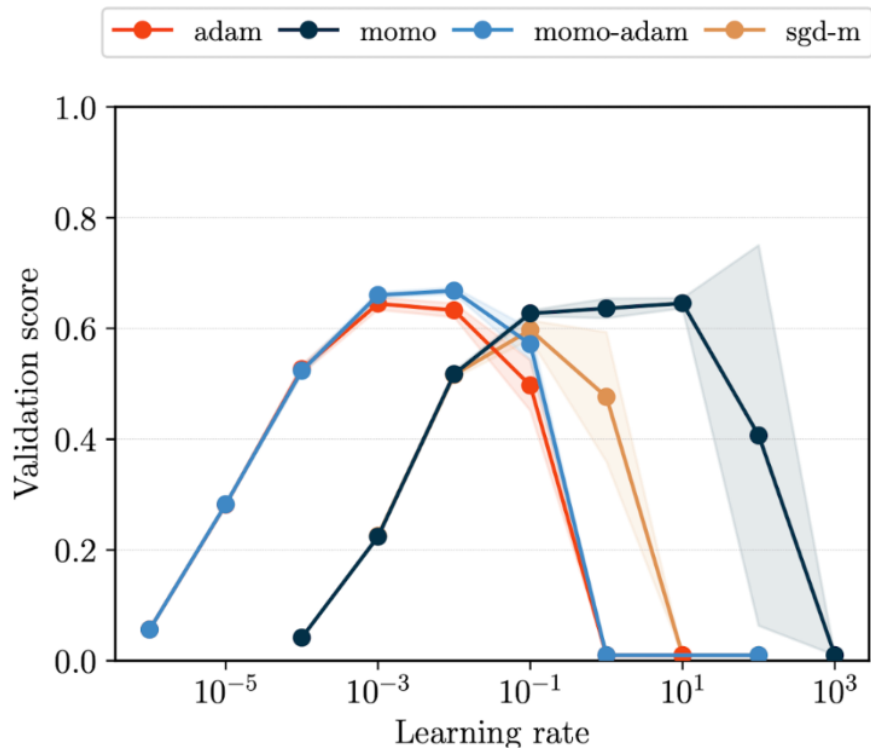
Closed form solution:

$$w_{k+1} = w_k - \min \left\{ \frac{\eta_k}{\rho_k}, \frac{(\bar{f}_k + \langle d_k, w_k \rangle - \gamma_k)_+}{\|d_k\|^2} \right\} d_k$$

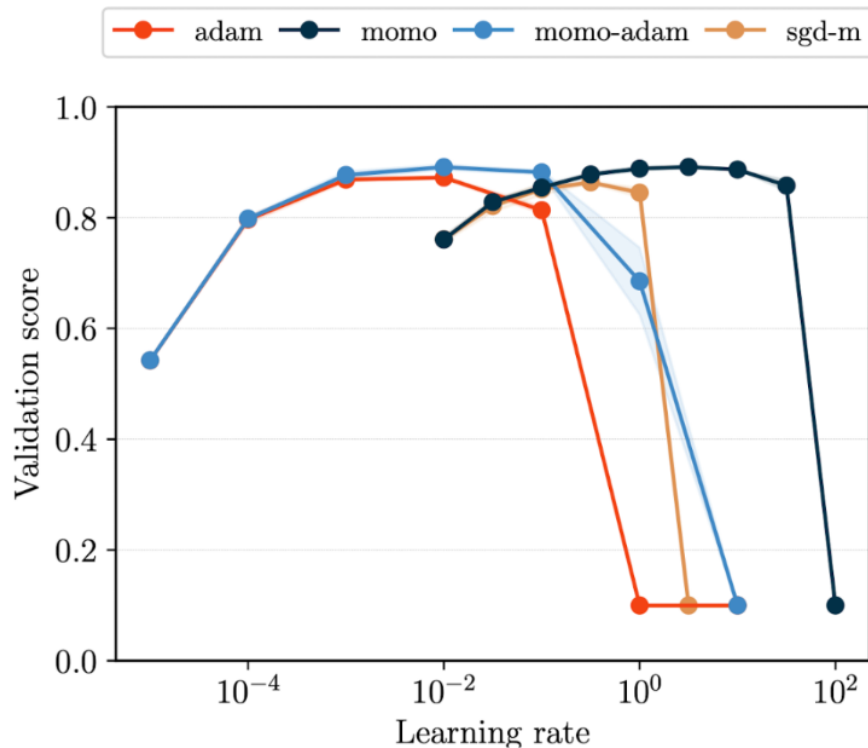
$$d_k = \sum_{j=1}^k \rho_j^k \nabla f_j(w_j) \quad \bar{f}_k = \sum_{j=1}^k \rho_j^k f_j(w_j) \quad \gamma_k = \sum_{j=1}^k \rho_j^k \langle \nabla f_j(w_j), w_j \rangle$$



How it works?



(a) ResNet110 for CIFAR100



(b) ResNet20 for CIFAR10



MoMo

arXiv 2305.07583

Pytorch implementation of MoMo methods. Adaptive learning rates for SGD with momentum (SGD-M) and Adam.

Authors: Fabian Schaipp, Ruben Ohana, Michael Eickenberg, Aaron Defazio, Robert M. Gower

Installation

You can install the package with

```
pip install momo-opt
```



Fabian Schaipp, Ruben Ohana, Michael Eickenberg, Aaron Defazio, Robert M. Gower
MoMo: Momentum Models for Adaptive Learning Rates



AI-SARAH

Adaptive and Implicit SARAH

Research Challenge

- SARAH produces smoother trajectories than SGD
 - we still need to pick a **learning rate**
- can we develop an adaptive strategy for SARAH?**

$$\mathbf{E}[\|v_t\|^2] \leq \begin{cases} (1 - (\frac{2}{\eta L} - 1)\mu^2\eta^2)^t \\ (1 - \frac{2\mu L\eta}{\mu + L})^t \end{cases} \cdot \mathbf{E}[\|\nabla F(w_0)\|^2]$$



pick η that will minimize $\|v_t\|$

An Example

Logistic regression

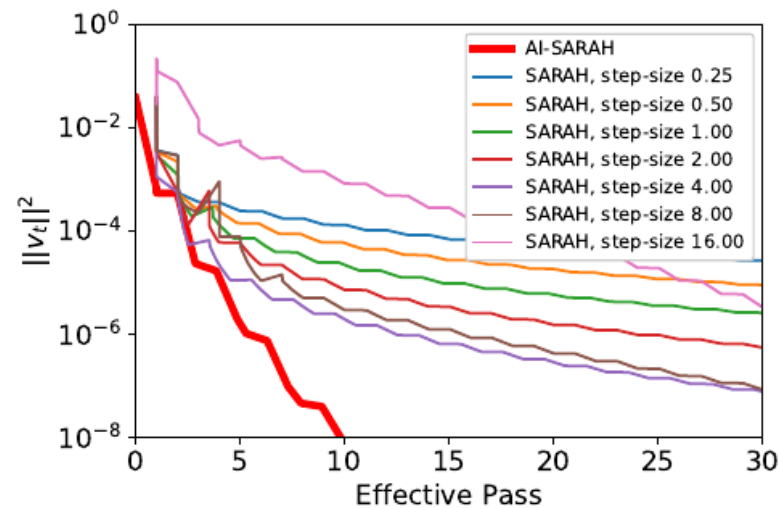
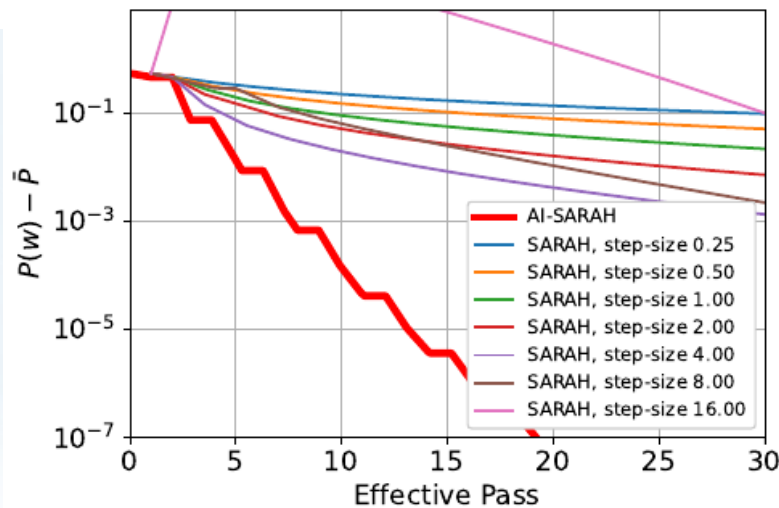
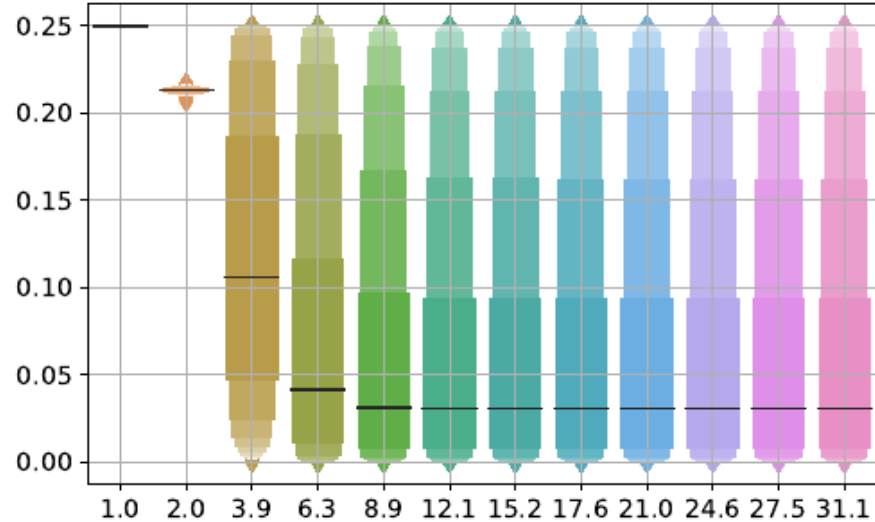
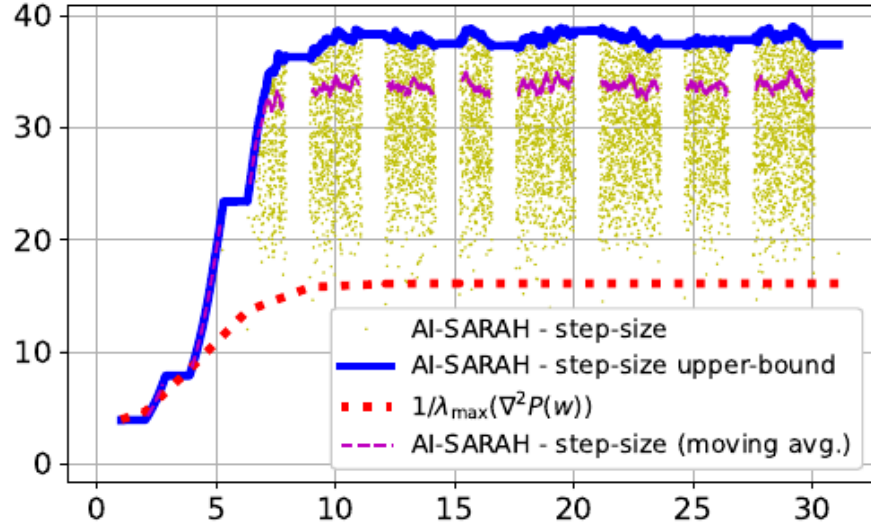
$$P(w) = \frac{1}{n} \sum_{i=1}^n \left(-y_i x_i^T w + \frac{\lambda}{2} \|w\|^2 \right)$$

“... tuning step-size is challenging.... it should be **adaptive to the local curvature**...”

Hessian

$$\nabla^2 P(w) = \frac{1}{n} \sum_{i=1}^n \underbrace{\frac{e^{-y_i x_i^T w}}{(1 + e^{-y_i x_i^T w})^2}}_{s_i(w)} x_i x_i^T + \lambda I$$

$s_i(w) \leq \frac{1}{4}$



Algorithm 2 AI-SARAH - Practical Variant

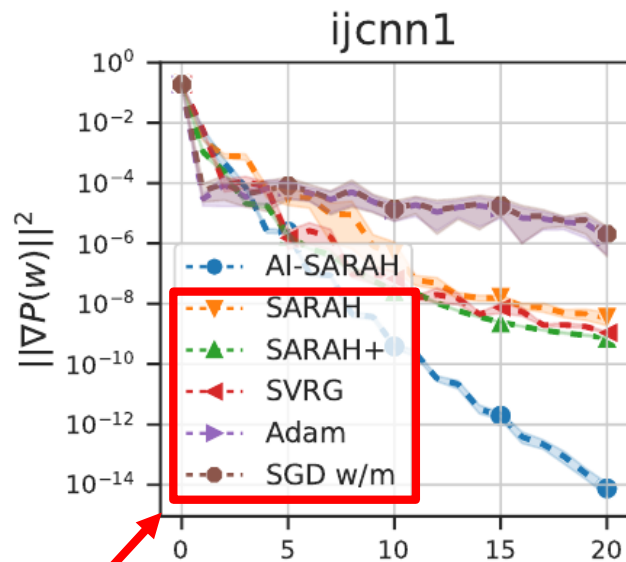
```

1: Parameter:  $0 < \gamma < 1$ 
2: Initialize:  $\tilde{w}_0$ 
3: Set:  $\alpha_{max} = \infty$ 
4: for  $k = 1, 2, \dots$  do
5:    $w_0 = \tilde{w}_{k-1}$ 
6:    $v_0 = \nabla P(w_0)$ 
7:    $t = 1$ 
8:   while  $\|v_t\|^2 \geq \gamma \|v_0\|^2$  do
9:     Select random mini-batch  $S_t$  from  $[n]$  uniformly
       with  $|S_t| = b$ 
10:     $\tilde{\alpha}_{t-1} \approx \arg \min_{\alpha > 0} \xi_t(\alpha)$ 
11:     $\alpha_{t-1} = \min\{\tilde{\alpha}_{t-1}, \alpha_{max}\}$ 
12:    Update  $\alpha_{max}$ 
13:     $w_t = w_{t-1} - \alpha_{t-1} v_{t-1}$ 
14:     $v_t = \nabla f_{S_t}(w_t) - \nabla f_{S_t}(w_{t-1}) + v_{t-1}$ 
15:     $t = t + 1$ 
16:  end while
17:  Set  $\tilde{w}_k = w_t$ .
18: end for
  
```

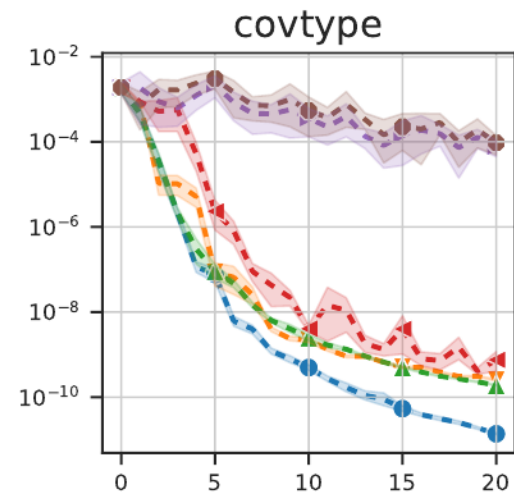
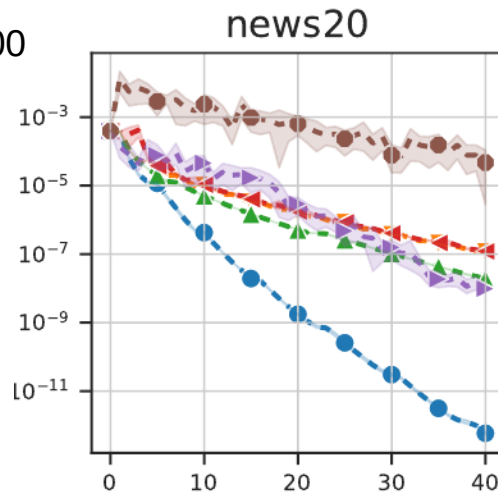
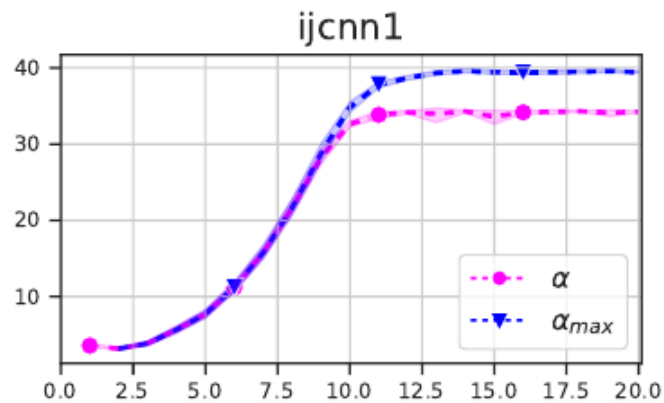
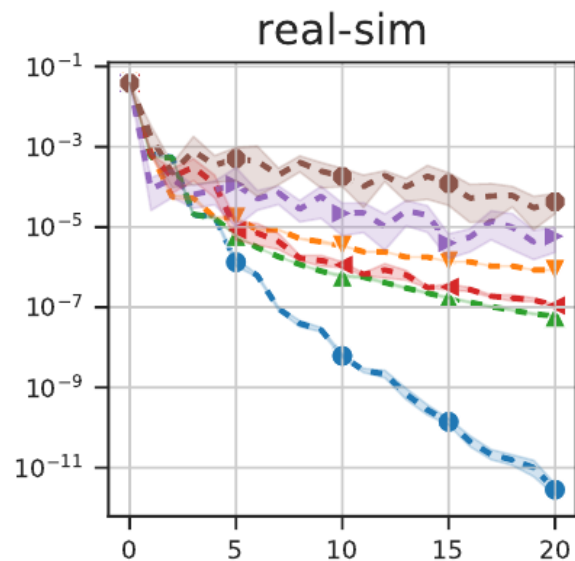
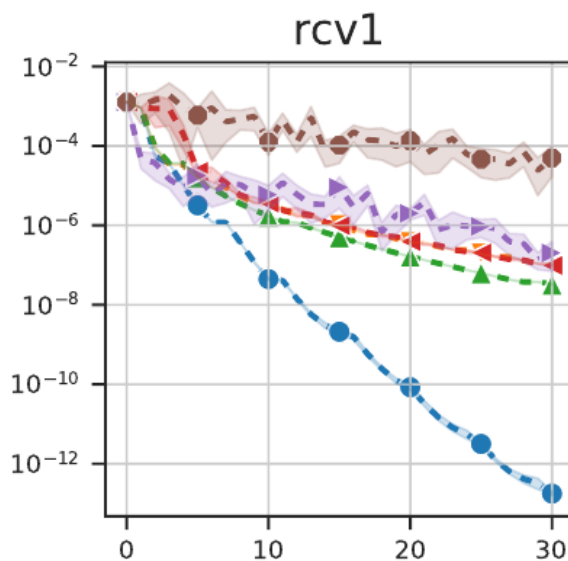
$$\xi_t(\alpha) := \|\nabla f_S(w_{t-1} - \alpha v_{t-1}) - \nabla f_S(w_{t-1}) + v_{t-1}\|^2$$

one Newton step is sufficient, easy to implement in pytorch

local curvature of minibatch S may be different from curvature on the whole dataset. We use exponential smoothing of α_{t-1} 's as upper-bound on step-size



extensive hyper-parameter tuning ~5,000



Mohamed bin Zayed
University of Artificial Intelligence
Masdar City
Abu Dhabi
United Arab Emirates





MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

Unlocking Machine Learning Potentials with Innovative SGD Algorithms

Martin Takáč



INNOPOLIS
UNIVERSITY

ASCOMP 2024 Autumn School
2024/07/12

Training procedure

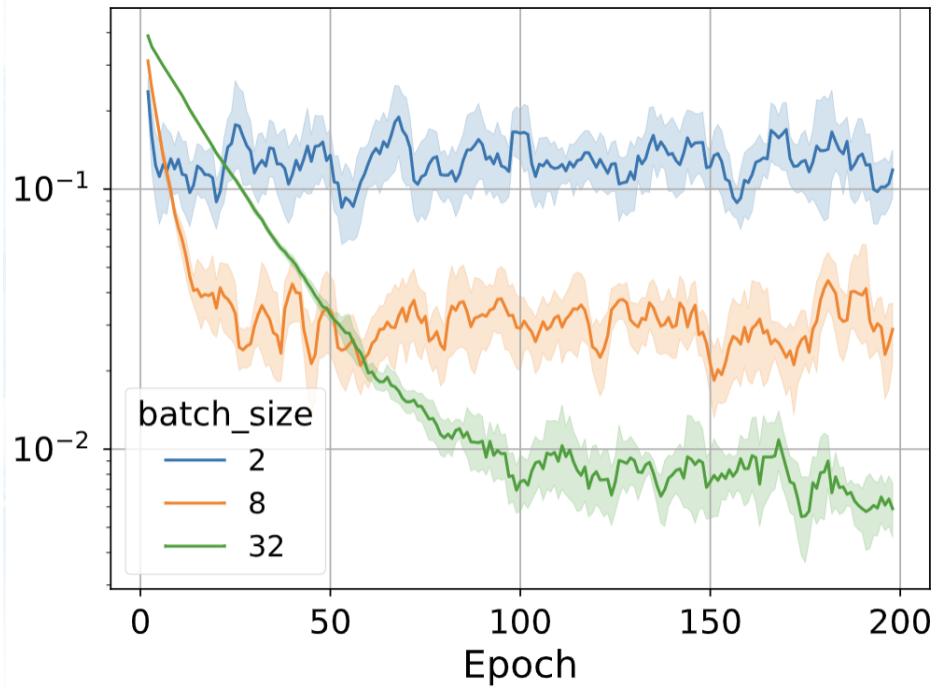
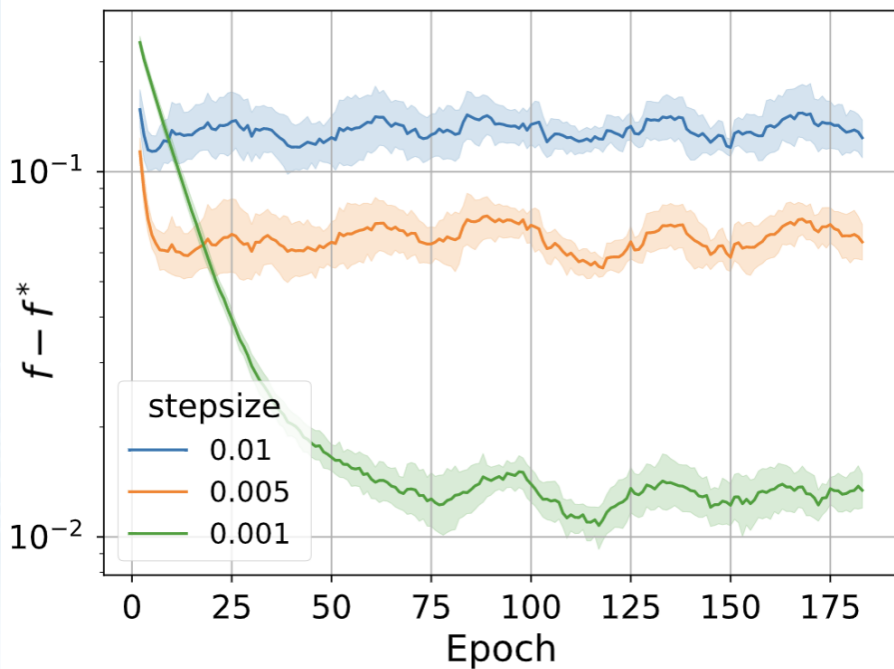
$$\min_{x \in \mathbb{R}^d} E[f(\phi(a_i; x), y_i)]$$



$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^n f(\phi(a_i; x), y_i) = \frac{1}{n} \sum_{i=1}^n f_i(x)$$

How fast is SGD?

$$E[f(x_k) - f^*] \leq (1 - 2\mu\eta)^k (f(x_0) - f^*) + \frac{L\eta}{4\mu} G^2$$

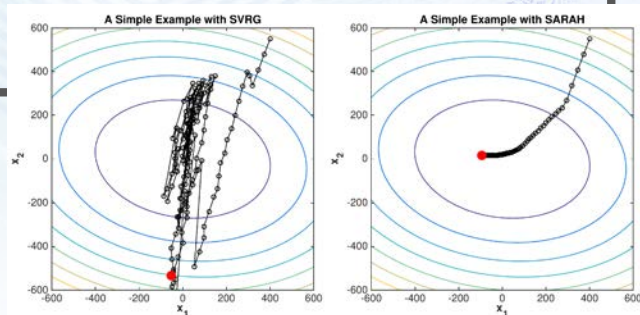


The Big Picture

- It does restarting (like SVRG)
- Is “similar” to SAG/SAGA, but **DOESN'T need extra storage**

1. choose x_0 , compute $v_0 = \nabla f(x_0)$
2. set $x_1 = x_0 - \gamma v_0$
3. for $k = 1, 2, \dots, m$
4. choose $i \in \{1, 2, \dots, n\}$
5. update $v_k = \nabla f_i(x_k) - \nabla f_i(x_{k-1}) + v_{k-1}$
6. set $x_{k+1} = x_k - \eta v_k$

Can we use it in practice?





Polyak Step-Size

Поляк (Polyak) Step-Size



Борис Поляк

$$\|w_{k+1} - w^*\|^2 = \|w_k - \eta \nabla F(w_k) - w^*\|^2$$

$$= \|w_k - w^*\|^2 + 2\langle w_k - w^*, -\eta \nabla F(w_k) \rangle + \eta^2 \|\nabla F(w_k)\|^2$$

$$F^* - F(w_k) \geq -\langle \nabla F(w_k), w_k - w^* \rangle$$

$$\leq \|w_k - w^*\|^2 - 2\eta(F(w_k) - F^*) + \eta^2 \|\nabla F(w_k)\|^2$$

$\min_{\eta}$

$$\eta = \frac{F(w_k) - F^*}{\|\nabla F(w_k)\|^2}$$

can be replaced
by lowerbound

Stochastic Polyak Step-size (SPS)

$$w_{k+1} = w_k - \eta_k \nabla f_i(w_k)$$

$$\eta_k = \frac{f_i(w_k) - f_i^*}{\|\nabla f_i(w_k)\|^2}$$

SPSmax: $\eta_k = \min \left\{ \gamma_b; \frac{f_i(w_k) - \bar{f}_i^*}{c \|\nabla f_i(w_k)\|^2} \right\}$



Stochastic Polyak Step-size (SPS)

$$\text{SPS}_{\max}: \quad \eta_k = \min \left\{ \gamma_b; \frac{f_i(w_k) - \bar{f}_i^*}{c \|\nabla f_i(w_k)\|^2} \right\}$$

in many ML problems, we can choose $\bar{f}_i^* = 0$

Assumption

$$\sigma^2 = E[f_i(w^*) - \bar{f}_i^*] = F^* - E[\bar{f}_i^*] < +\infty$$

Theorem

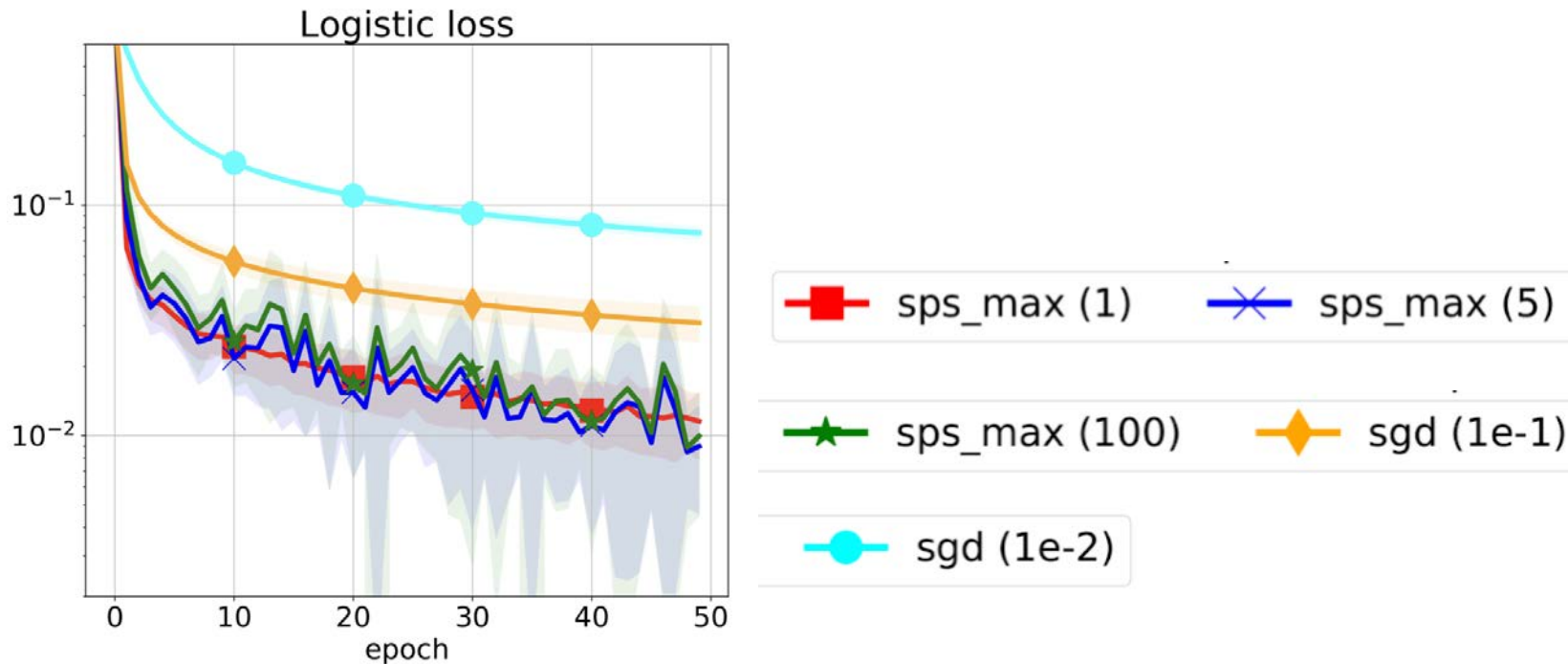
if f_i is L_i smooth, $F(w)$ is μ strongly convex and $c \geq \frac{1}{2}$

then $E[\|w_k - w_*\|^2] \leq (1 - \mu\alpha)^k \|w_0 - w^*\|^2 + \frac{2\gamma_b\sigma^2}{\mu\alpha}$

where $\alpha = \min \left\{ \gamma_b, \frac{1}{2cL_{\max}} \right\}; \quad L_{\max} = \max_i L_i$



Stochastic Polyak Step-size (SPS)



N Loizou, S Vaswani, I Laradji, S Lacoste-Julien: *Stochastic Polyak step-size for SGD: An adaptive learning rate for fast convergence, AISTATS 2020*

Stochastic Polyak Step-size + Interpolation

$$F(w) = \frac{1}{n} \sum_{i=1}^n f_i(w), \quad f_i(w) \geq 0 \quad \forall i \in \{1, 2, \dots, n\}$$

interpolation assumption

Let's assume that: $\exists w^* \in R^d : f_i(w^*) = 0$

Question #1
Why we should use
Euclidean norm?

$$w_{k+1} = \arg \min_w \|w - w_k\|^2$$

such that $f_i(w_k) + \langle \nabla f_i(w_k), w - w_k \rangle = 0$

Question #2
Why linearize?

$$-\frac{f_i(w_k)}{\|\nabla f_i(w_k)\|^2} \nabla f_i(w_k)$$

fazio, Michael Rabbat: *Stochastic Polyak Stepsize with a*

Polyak Step-Size With Scaling

- Stochastic Polyak - the same step size for all features! Can we improve it?

$$w_{k+1} = \arg \min \|w - w_k\|^2$$

change norm

diag(v)

such that $f_i(w_k) + \langle \nabla f_i(w_k), w - w_k \rangle = 0$

$$w_{k+1} = w_k - \frac{f_i(w_k)}{(\|\nabla f_i(w_k)\|_v^*)^2} \text{diag}(v)^{-1} \nabla f_i(w_k)$$

Adding scaling
to Polyak step size

Polyak Step-Size With Scaling and Curvature

- How can we include curvature information?

$$w_{k+1} = \arg \min \|w - w_k\|_{\text{diag}(v)}^2$$

such that $f_i(w_k) + \langle \nabla f_i(w_k), w - w_k \rangle = 0$



Could replace by Hessian approximation with rank1

$$w_{k+1} = \arg \min \|w - w_k\|_{\text{diag}(v)}^2$$

such that $f_i(w_k) + \langle \nabla f_i(w_k), w - w_k \rangle + \frac{1}{2}(w - w_k)^T \nabla^2 f_i(w - w_k) = 0$

HARD to solve

$$f_i(w) = \phi_i(x_i^\top w - y_i)$$

Lemma 3.1. (SP2) Assume $f_i(w)$ is the loss of a generalized linear model (10) and is non-negative. Let $f_i = f_i(w^t)$ for short. Let $a_i := \phi'_i(x_i^\top w - y_i)$ and $h_i := \phi''_i(x_i^\top w - y_i)$. If

$$a_i^2 - 2h_i f_i \geq 0 \quad (12)$$

then the optimal solution of (7) is as follows

$$w^{t+1} = w^t - \frac{a_i}{h_i} \left(1 - \frac{\sqrt{a_i^2 - 2h_i f_i}}{|a_i|} \right) \frac{x_i}{\|x_i\|^2}. \quad (13)$$

Alternatively if (12) does not hold, since $f_i \geq 0$ we have necessarily that $h_i > 0$, and consequently a Newton step will give the minima of the local quadratic, that is

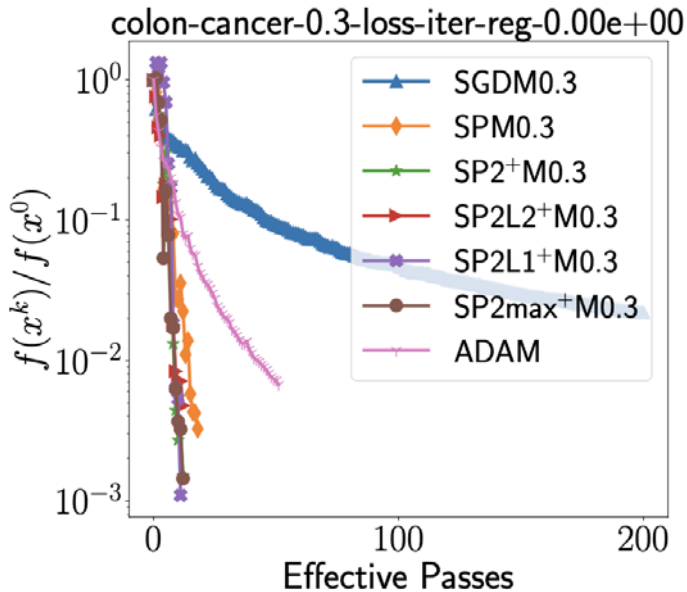
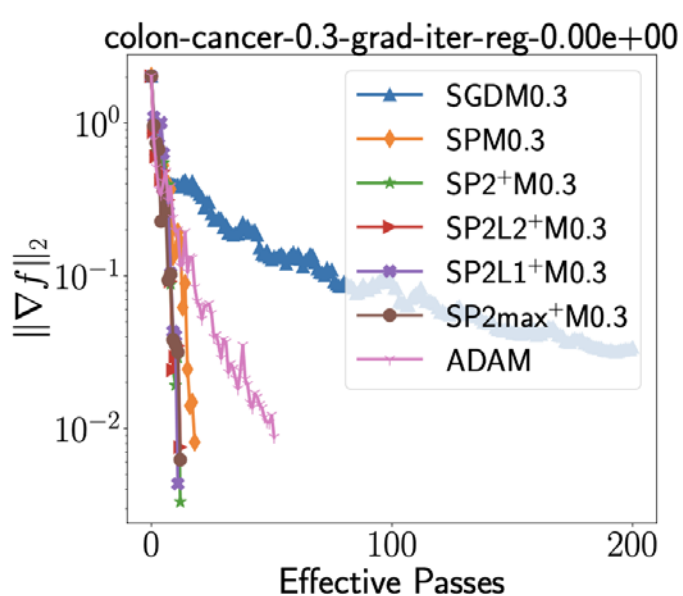
$$w^{t+1} = w^t - \frac{a_i}{h_i} \frac{x_i}{\|x_i\|^2}. \quad (14)$$



Polyak Step-Size With Scaling and Curvature

Lemma 3.3. (SP2⁺) Let $g_t \equiv \nabla f_i(w^t)$ and $\mathbf{H}_t \equiv \nabla^2 f_i(w^t)$

$$w^{t+1} = w^t - \frac{f_i(w^t)}{\|g_t\|^2} g_t - \frac{1}{2} \frac{f_i(w^t)^2}{\|g_t\|^4} \frac{\langle \mathbf{H}_t g_t, g_t \rangle}{\|v^{t+1}\|^2} v^{t+1}, \text{ where } v^{t+1} = \left(\mathbf{I} - \mathbf{H}_t \frac{f_i(w^t)}{\|g_t\|^2} \right) g_t, \quad (20)$$



$$\min_{s \in \mathbb{R}, w \in \mathbb{R}^d} \frac{1}{2} s^2 \text{ subject to } f_i(w) \leq s, \text{ for } i = 1, \dots, n$$
$$\min_{s \geq 0, w \in \mathbb{R}^d} s \text{ s.t. } f_i(w) \leq s, \text{ for } i = 1, \dots, n.$$





+ momentum to Polyak step-size (MoMo)

Adding Momentum, how ?

$$w_{k+1} = \arg \min_w m_k(w) + \frac{1}{2\eta_k} \|w - w_k\|^2$$

SGD $m_k(w) = f_i(w_k) + \langle f_i(w_k), w - w_k \rangle$

$$m_k(w) = \frac{1}{\rho_k} \sum_{j=1}^k \rho_j^k f^j(w) \quad \text{with } \rho_k = \sum_{j=1}^k \rho_j^k$$

HARD to solve

$$m_k(w) = \max \left\{ \frac{1}{\rho_k} \sum_{j=1}^k \rho_j^k (f_j(w_j) + \langle \nabla f_j(w_j), w - w_j \rangle), 0 \right\}$$



Adding Momentum, how ?

$$w_{k+1} = \arg \min_w m_k(w) + \frac{1}{2\eta_k} \|w - w_k\|^2$$

$$m_k(w) = \max \left\{ \frac{1}{\rho_k} \sum_{j=1}^k \rho_j^k (f_j(w_j) + \langle \nabla f_j(w_j), w - w_j \rangle), 0 \right\}$$

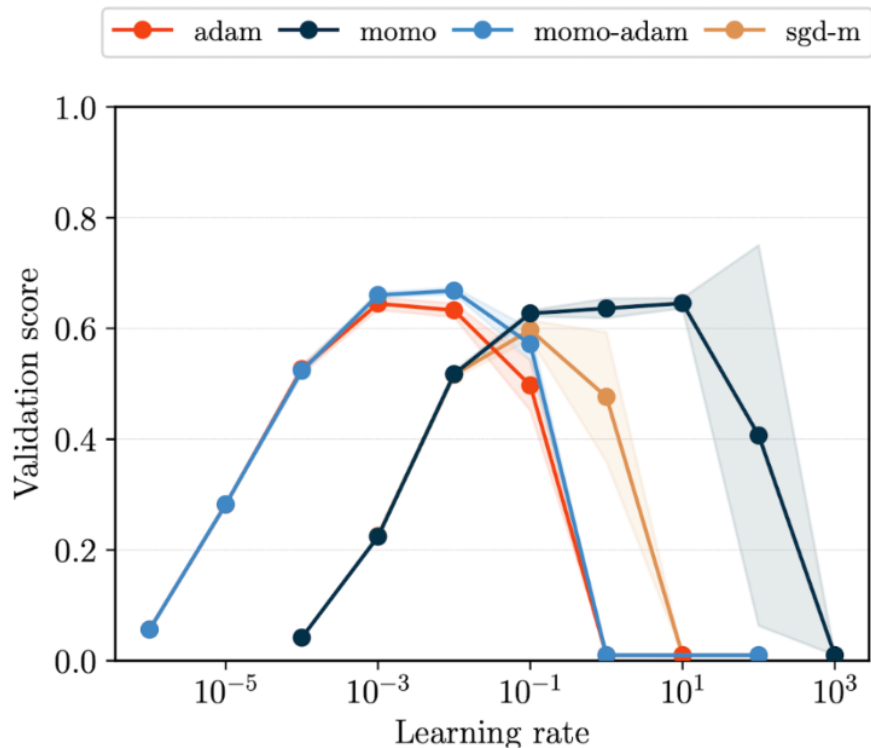
Closed form solution:

$$w_{k+1} = w_k - \min \left\{ \frac{\eta_k}{\rho_k}, \frac{(\bar{f}_k + \langle d_k, w_k \rangle - \gamma_k)_+}{\|d_k\|^2} \right\} d_k$$

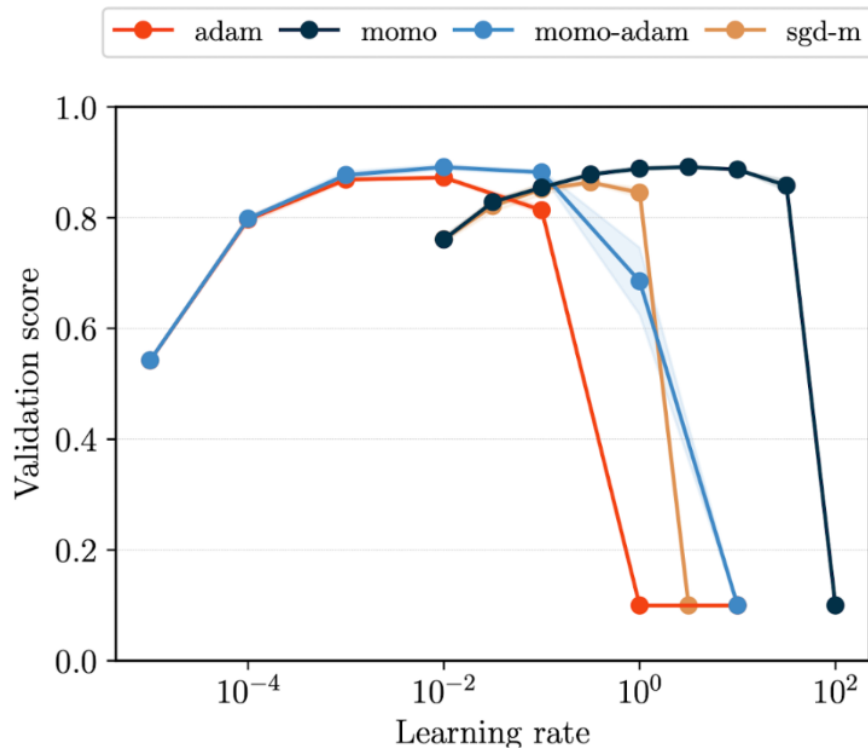
$$d_k = \sum_{j=1}^k \rho_j^k \nabla f_j(w_j) \quad \bar{f}_k = \sum_{j=1}^k \rho_j^k f_j(w_j) \quad \gamma_k = \sum_{j=1}^k \rho_j^k \langle \nabla f_j(w_j), w_j \rangle$$



How it works?



(a) ResNet110 for CIFAR100



(b) ResNet20 for CIFAR10



MoMo

arXiv 2305.07583

Pytorch implementation of MoMo methods. Adaptive learning rates for SGD with momentum (SGD-M) and Adam.

Authors: Fabian Schaipp, Ruben Ohana, Michael Eickenberg, Aaron Defazio, Robert M. Gower

Installation

You can install the package with

```
pip install momo-opt
```



Fabian Schaipp, Ruben Ohana, Michael Eickenberg, Aaron Defazio, Robert M. Gower
MoMo: Momentum Models for Adaptive Learning Rates



AI-SARAH

Adaptive and Implicit SARAH

Research Challenge

- SARAH produces smoother trajectories than SGD
 - we still need to pick a **learning rate**
- can we develop an adaptive strategy for SARAH?**

$$\mathbf{E}[\|v_t\|^2] \leq \begin{cases} (1 - (\frac{2}{\eta L} - 1)\mu^2\eta^2)^t \\ (1 - \frac{2\mu L\eta}{\mu + L})^t \end{cases} \cdot \mathbf{E}[\|\nabla F(w_0)\|^2]$$



pick η that will minimize $\|v_t\|$

An Example

Logistic regression

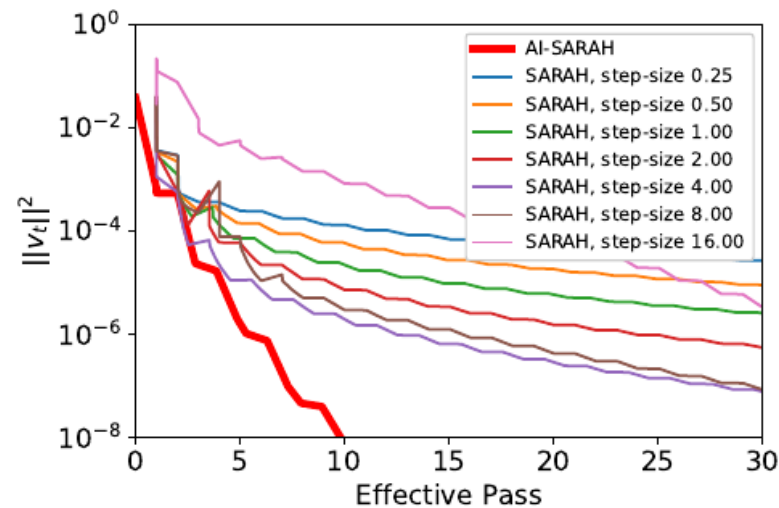
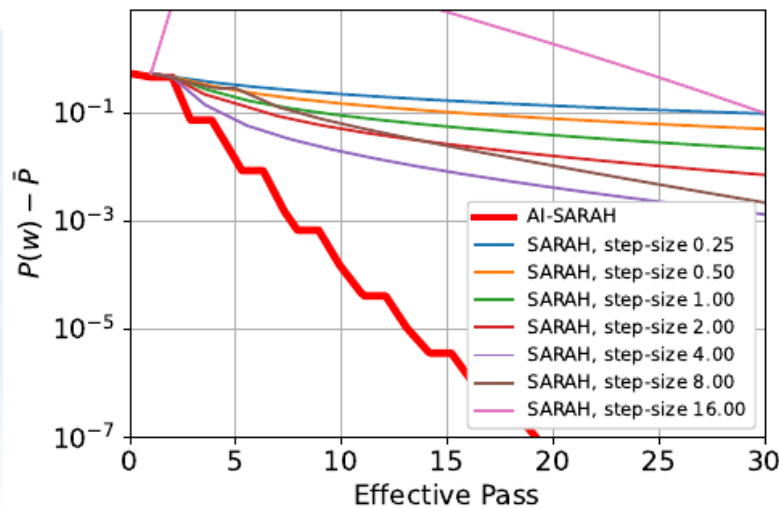
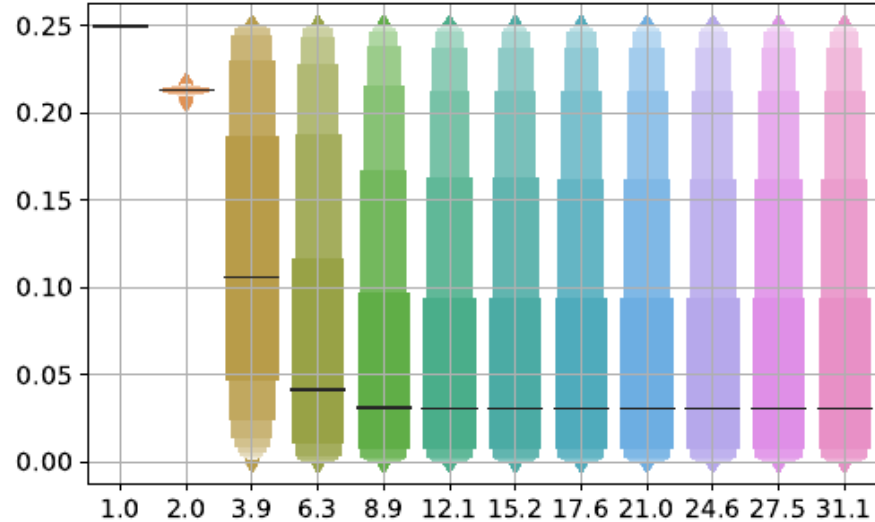
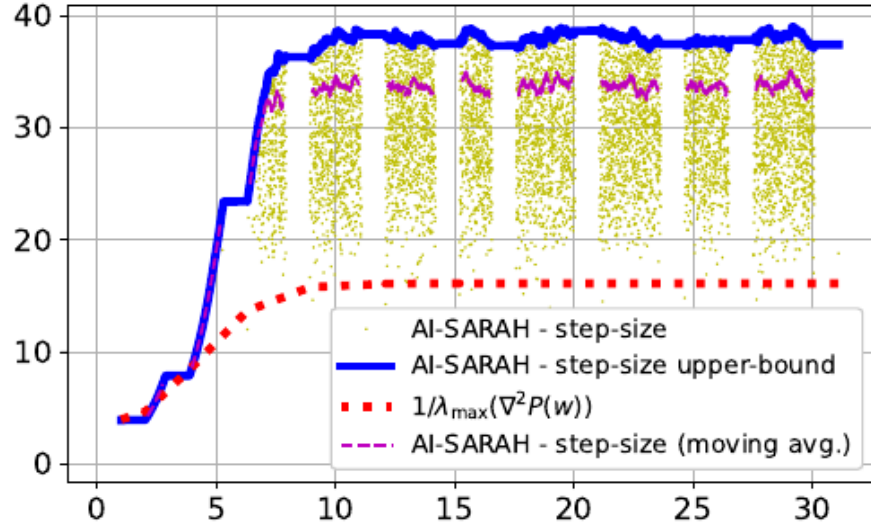
$$P(w) = \frac{1}{n} \sum_{i=1}^n \left(-y_i x_i^T w + \log(1 + e^{x_i^T w}) \right) + \frac{\lambda}{2} \|w\|^2$$

“... tuning step-size is challenging.... it should be **adaptive to the local curvature**...”

Hessian

$$\nabla^2 P(w) = \frac{1}{n} \sum_{i=1}^n \underbrace{\frac{e^{-y_i x_i^T w}}{(1 + e^{-y_i x_i^T w})^2}}_{s_i(w)} x_i x_i^T + \lambda I$$

$s_i(w) \leq \frac{1}{4}$



Algorithm 2 AI-SARAH - Practical Variant

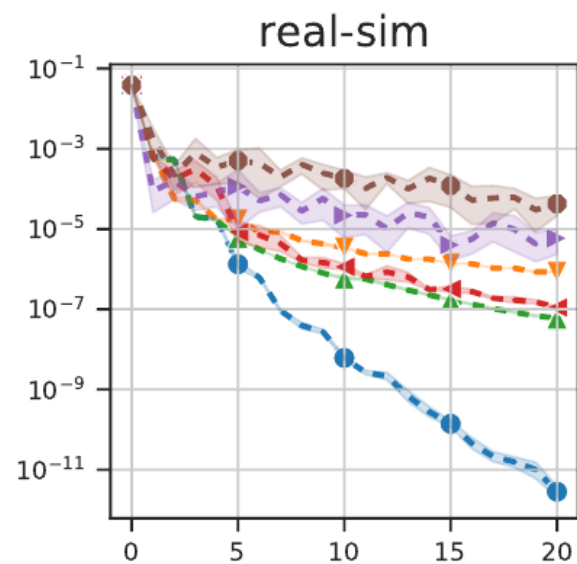
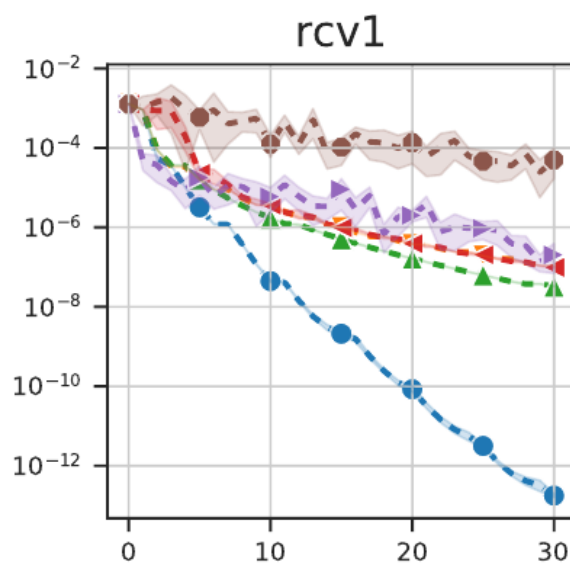
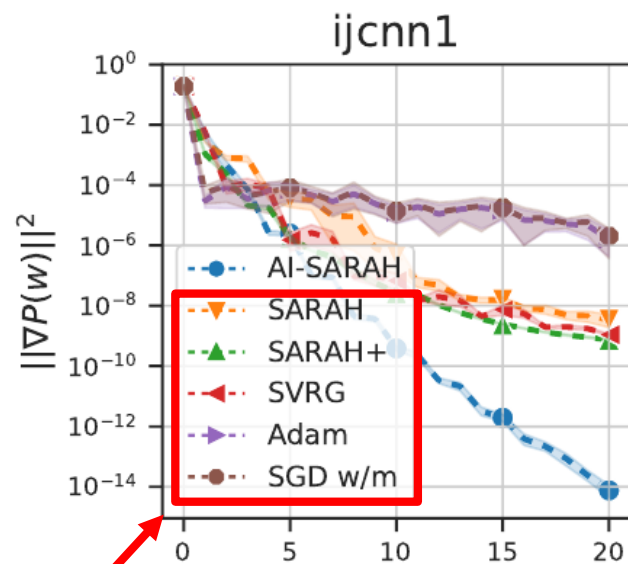
```

1: Parameter:  $0 < \gamma < 1$ 
2: Initialize:  $\tilde{w}_0$ 
3: Set:  $\alpha_{max} = \infty$ 
4: for  $k = 1, 2, \dots$  do
5:    $w_0 = \tilde{w}_{k-1}$ 
6:    $v_0 = \nabla P(w_0)$ 
7:    $t = 1$ 
8:   while  $\|v_t\|^2 \geq \gamma \|v_0\|^2$  do
9:     Select random mini-batch  $S_t$  from  $[n]$  uniformly
       with  $|S_t| = b$ 
10:     $\tilde{\alpha}_{t-1} \approx \arg \min_{\alpha > 0} \xi_t(\alpha)$ 
11:     $\alpha_{t-1} = \min\{\tilde{\alpha}_{t-1}, \alpha_{max}\}$ 
12:    Update  $\alpha_{max}$ 
13:     $w_t = w_{t-1} - \alpha_{t-1} v_{t-1}$ 
14:     $v_t = \nabla f_{S_t}(w_t) - \nabla f_{S_t}(w_{t-1}) + v_{t-1}$ 
15:     $t = t + 1$ 
16:  end while
17:  Set  $\tilde{w}_k = w_t$ .
18: end for
  
```

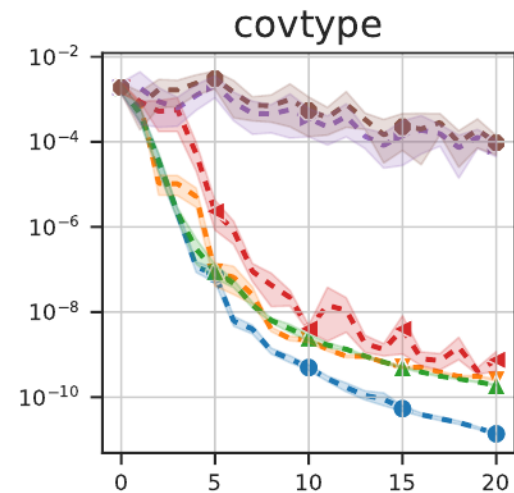
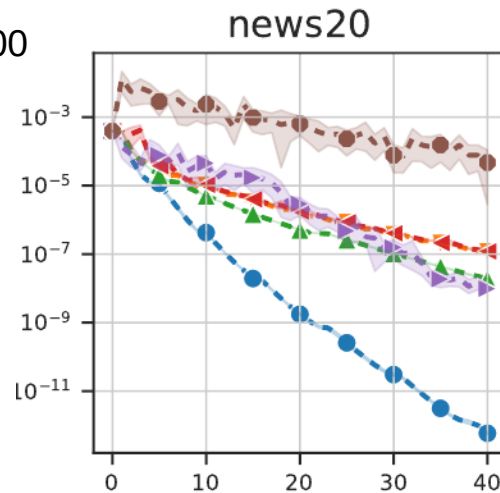
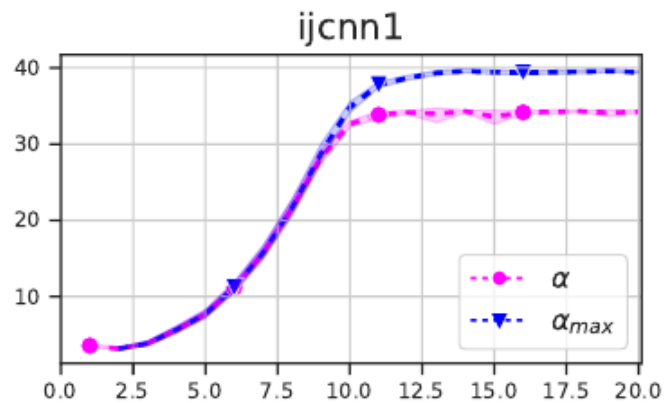
$$\xi_t(\alpha) := \|\nabla f_S(w_{t-1} - \alpha v_{t-1}) - \nabla f_S(w_{t-1}) + v_{t-1}\|^2$$

one Newton step is sufficient, easy to implement in pytorch

local curvature of minibatch S may be different from curvature on the whole dataset. We use exponential smoothing of α_{t-1} 's as upper-bound on step-size



extensive hyper-parameter tuning ~5,000



D-Adaptation

<https://github.com/facebookresearch/dadaptation>



Fork 16



Star 455



Used by 148



downloads 1M

downloads/month 103k

<https://github.com/konstmish/prodigy>



Fork 10



Star 194



96 Repositories 2 Packages

downloads 357k

downloads/month 63k

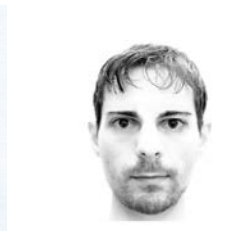


Konstantin Mishchenko

blems, including la
h is available¹.



age problems.



Aaron Defazio

What is a good step-size for Adam?

Adam

```
CLASS torch.optim.Adam(params, lr=0.001, betas=(0.9, 0.999), eps=1e-08, weight_decay=0, amsgrad=False, *,  
    foreach=None, maximize=False, capturable=False, differentiable=False, fused=None) [SOURCE]
```

If you check some libraries, or

```
lr = lr_base
```

A lot of hyper-parameters to tune

```
** 0.5
```

... the same for computer vision

```
sched.G_lr_rate_dict = {128: 0.0015, 256: 0.002, 512: 0.003, 1024: 0.003}
```

<https://github.com/NVlabs/stylegan/blob/master/train.py>

... and in GANs

```
G_lr=5e-5
```

```
D_lr=2e-4
```

<https://github.com/ajbrock/BigGAN-PyTorch>

Sub-gradient Descent

Algorithm: Sub-gradient Descent

input: $x_0, \eta_k > 0$

for: $k = 0, 1, \dots,$

$$g_k \in \partial f(x_k)$$

$$x_{k+1} = x_k - \eta_k g_k$$

Sub-gradient Descent

Algorithm: Sub-gradient Descent

input: $x_0, \eta_k > 0$

for: $k = 0, 1, \dots,$

$$g_k \in \partial f(x_k)$$

$$x_{k+1} = x_k - \eta_k g_k$$

Sub-gradient Descent

Algorithm: Sub-gradient Descent

input: $x_0, \boldsymbol{\eta}_k > 0$

for: $k = 0, 1, \dots,$

$$g_k \in \partial f(x_k)$$

$$x_{k+1} = x_k - \boldsymbol{\eta}_k g_k$$

What is the optimal step-size?

optimal step-size: $\boldsymbol{\eta}_k = \frac{D}{\sqrt{k} G}$

$$D = \|x_0 - x^*\|$$

Normalized Gradient by Наум Зуселевич Шор

Algorithm

input: $x_0, \eta_k > 0$

for: $k = 0, 1, \dots,$

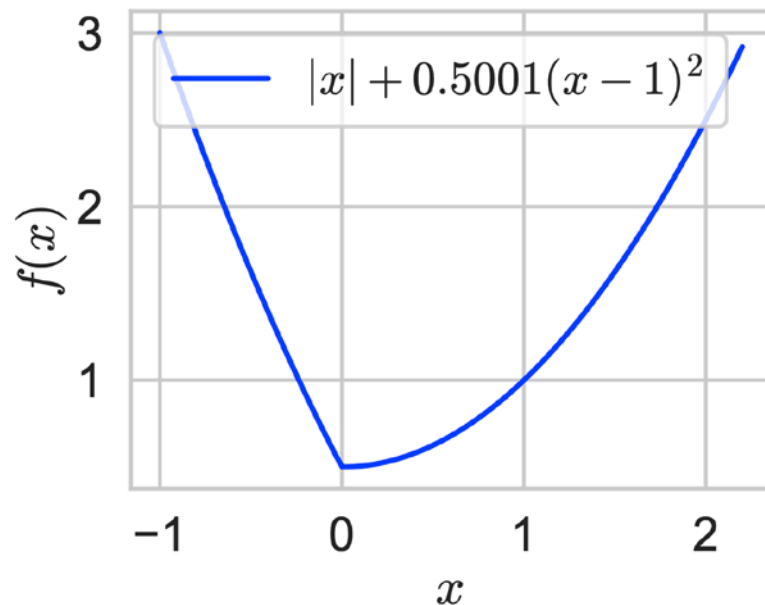
$$g_k \in \partial f(x_k)$$

$$x_{k+1} = x_k - \eta_k \frac{g_k}{\|g_k\|}$$

If $g_k \in \partial f(x_k)$ then $\|g_k\|$ can be meaningless

Why we need to normalize?

$$f(x) = |x| + 0.5001(x - 1)^2$$



Normalized Gradient by Наум Зуселевич Шор

Algorithm

input: $x_0, \eta_k > 0$

for: $k = 0, 1, \dots,$

$$g_k \in \partial f(x_k)$$

$$x_{k+1} = x_k - \eta_k \frac{g_k}{\|g_k\|}$$

BUT....

decreasing and need **D**!

Step-size does not depend on G!

Theorem

For nonsmooth $f(x)$ with $\|g_k\| \leq G$ with $\eta_k = \frac{D}{\sqrt{k}}$ we have

$$\min_{k \leq n} [f(x_k) - f^*] = O\left(\frac{GD}{\sqrt{n}}\right)$$

Adagrad-Norm

Normalized GD

$$\eta_k = \frac{D}{\sqrt{k}}$$



$$x_{k+1} = x_k - \eta_k \frac{g_k}{\|g_k\|}$$

$$x_{k+1} = x_k - \frac{D}{\sqrt{k}} \frac{g_k}{\|g_k\|}$$

AdaGrad norm

$$x_{k+1} = x_k - D \frac{g_k}{\sqrt{\sum_{t=0}^k \|g_t\|^2}}$$

Less Regret via Online Conditioning

Matthew Streeter
Google, Inc.
mstreeter@google.com

H. Brendan McMahan
Google, Inc.
mcmahan@google.com

February 25, 2010

A bit more adaptive, but still need D

DoG (Distance over Gradients)

Recall: Sub-gradient method

$$x_{k+1} = x_k - \eta_k g_k$$

$$\eta_k = \frac{D}{\sqrt{k} G}$$

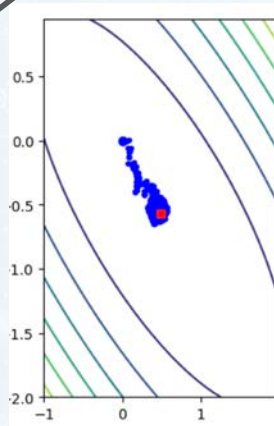
estimate using Adagrad

$$G \sqrt{k} \approx \sqrt{\sum_{t=0}^k \|g_t\|^2}$$

Theorem: if $\|g_k\| \leq G$, then

$$\min_{k \leq t} f(x_k) - f^* = O\left(\frac{DG \log \frac{D}{d_0}}{\sqrt{t}}\right)$$

$$D = \|x_0 - x^*\|$$



DoG is SGD's Best Friend:

A Parameter-Free Dynamic Step Size Schedule

Maor Ivgi
maor.ivgi@cs.tau.ac.il

Oliver Hinder
ohinder@pitt.edu

Yair Carmon
ycarmon@tauex.tau.ac.il

Abstract

We propose a tuning-free dynamic SGD step size formula, which we call Distance over Gradients (DoG). The DoG step sizes depend on simple empirical quantities (distance from the initial point and norms of gradients) and have no “learning rate” parameter. Theoretically, we show that a slight variation of the DoG formula enjoys strong parameter-free convergence guarantees for stochastic convex optimization assuming only *locally bounded* stochastic gradients.¹ Empirically, we consider a broad range of vision and language transfer learning tasks, and show that DoG's performance is close to that of SGD with tuned learning rate. We also propose a per-layer variant of DoG that generally outperforms tuned SGD, approaching the performance of tuned Adam. A PyTorch implementation is available at <https://github.com/formll/dog>.

Idea: approximate D from history of points!

$$d_k = \|x_0 - x_k\| \approx \|x_0 - x^*\|$$

$$\text{DoG: } d_k = \max_{t \leq k} \|x_0 - x_t\| \approx D$$

$$x_{k+1} = x_k - \frac{\max_{t \leq k} \|x_0 - x_t\|}{\sqrt{\sum_{t=0}^k \|g_t\|^2}} g_k$$

DoWG (Distance over Weighted Gradients)

DoG

$$x_{k+1} = x_k - \frac{d_k}{\sqrt{\sum_{t=0}^k \|g_t\|^2}} g_k$$

$$d_k = \max_{t \leq k} \|x_0 - x_t\|$$

DoWG

$$x_{k+1} = x_k - \frac{d_k^2}{\sqrt{\sum_{t=0}^k d_t^2 \|g_t\|^2}} g_k$$

Theorem: if $\|g_k\| \leq G$, then

$$\min_{k \leq t} f(x_k) - f^* = O\left(\frac{DG \sqrt{\log \frac{D}{d_0}}}{\sqrt{t}}\right)$$

DoWG Unleashed: An Efficient Universal Parameter-Free Gradient Descent Method

Ahmed Khaled
Princeton University

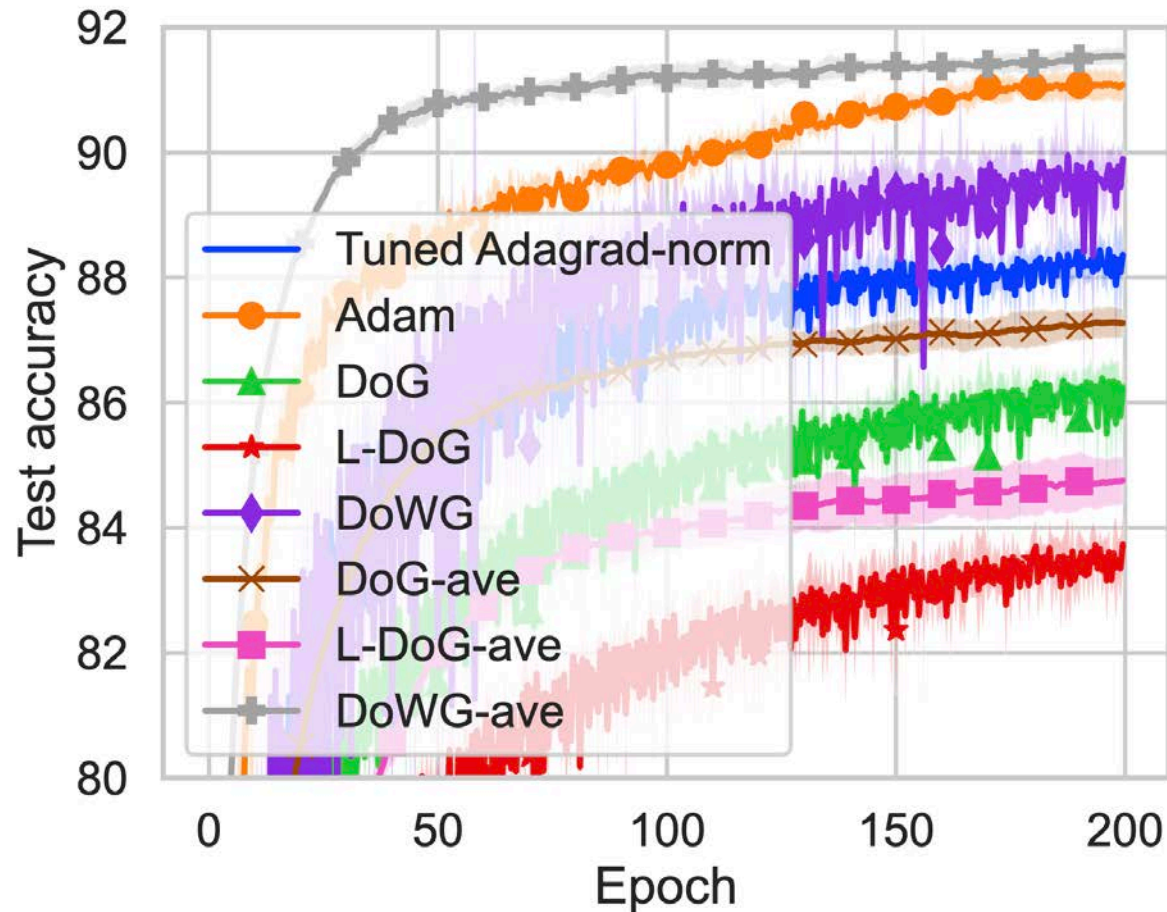
Konstantin Mishchenko
Samsung AI Center

Chi Jin
Princeton University

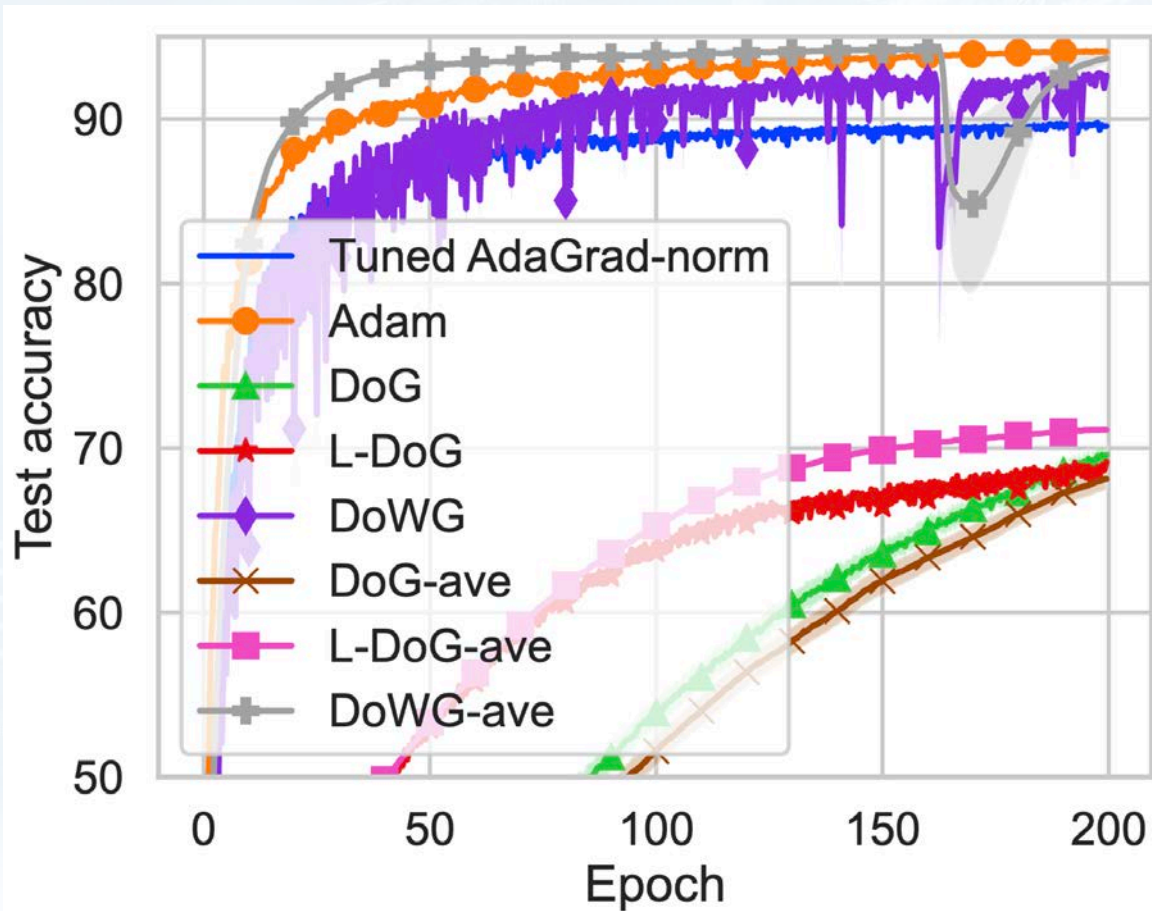
Abstract

This paper proposes a new easy-to-implement *parameter-free* gradient-based optimizer: DoWG (Distance over Weighted Gradients). We prove that DoWG is *efficient*—matching the convergence rate of optimally tuned gradient descent in convex optimization up to a logarithmic factor without tuning any parameters, and *universal*—automatically adapting to both smooth and nonsmooth problems. While popular algorithms following the AdaGrad framework compute a running average of the squared gradients to use for normalization, DoWG maintains a new distance-based weighted version of the running average, which is crucial to achieve the desired properties. To complement our theory, we also show empirically that DoWG trains at the edge of stability, and validate its effectiveness on practical machine learning tasks.

VGG11 on CIFAR10



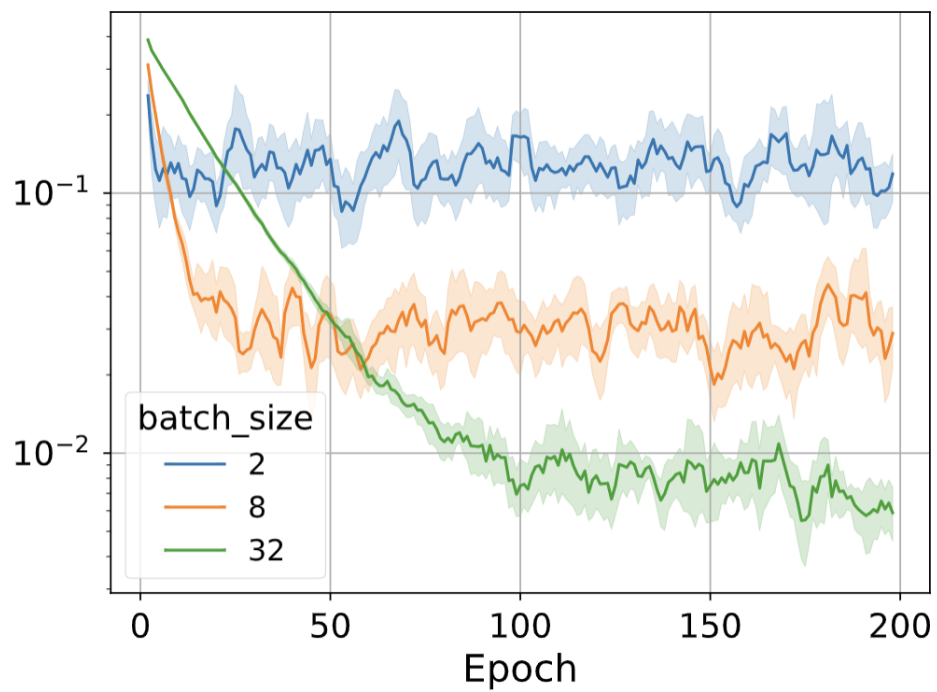
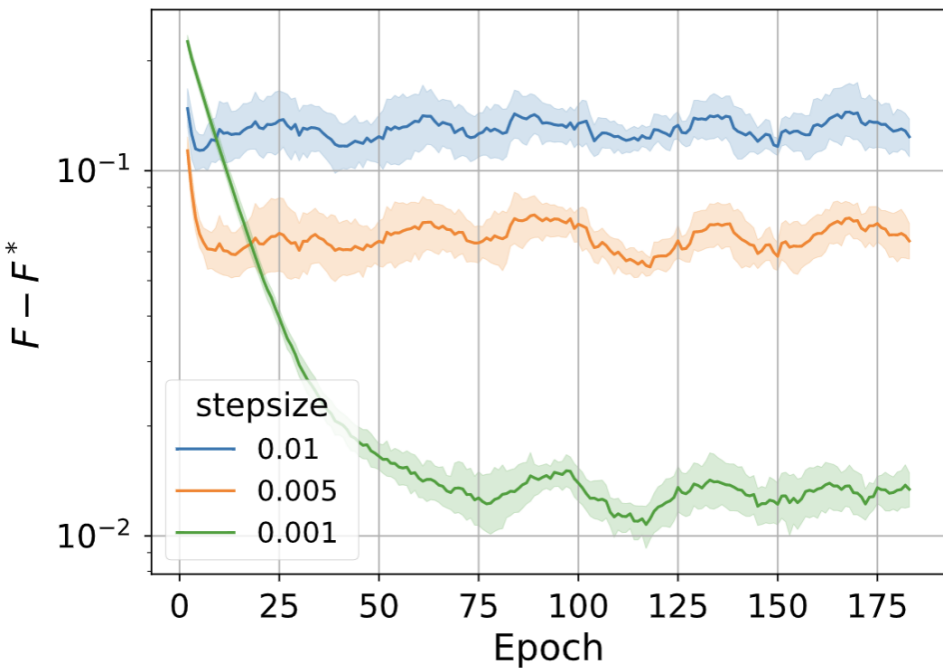
ResNet-50 on CIFAR10



Adaptive Batch-Size

Recall:

$$E[f(x_k) - f^*] \leq (1 - 2\mu\eta)^k (f(x_0) - f^*) + \frac{L\eta}{4\mu} G^2$$



Let's Consider Deterministic Setting

What conditions we need for convergence of approximate gradient descent?



$$\nabla f_S(w) \approx \nabla F(w)$$

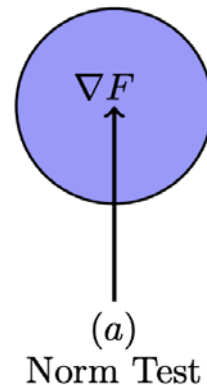
Deterministic
approx. gradient

True gradient

$$\langle \nabla f_S(w), \nabla F(w) \rangle \geq 0$$

Norm test (condition):

$$\|\nabla f_S(w) - \nabla F(w)\| \leq \omega \|\nabla F(w)\|$$



Let's Consider Deterministic Setting

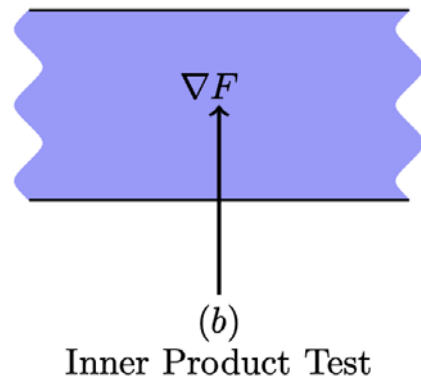
$$\nabla f_S(w) \approx \nabla F(w)$$

$$\langle \nabla f_S(w), \nabla F(w) \rangle \geq 0$$

$$\langle \nabla f_S(w), \nabla F(w) \rangle \approx \langle \nabla F(w), \nabla F(w) \rangle = \|\nabla F(w)\|^2$$

Inner Product test (condition):

$$(\langle \nabla f_S(w), \nabla F(w) \rangle - \|\nabla F(w)\|^2)^2 \leq \theta^2 \|\nabla F(w)\|^4$$



Raghu Bollapragada, Richard Byrd, and Jorge Nocedal Adaptive sampling strategies for stochastic optimization. SIAM Journal on Optimization, 28(4):3312–3343, 2018.

Let's switch to stochastic setting!

$$\nabla f_S(w) := \frac{1}{|S|} \sum_{i \in S} \nabla f_i(w)$$

Condition $\langle \nabla f_S(w), \nabla F(w) \rangle \geq 0$ may not hold, but will be valid “in expectation”

$$E [\langle \nabla f_S(w), \nabla F(w) \rangle] = \|\nabla F(w)\|^2$$

What do we need to hope
for convergence?



Control the variance of LHS

$$E \left[\left(\langle \nabla f_S(w), \nabla F(w) \rangle - \|\nabla F(w)\|^2 \right)^2 \right] \leq \theta^2 \|\nabla F(w)\|^4$$

require the sample size $|S|$ be large enough so that the condition is satisfied

Reformulate the condition

HARD to compute!

$$E \left[\left(\langle \nabla f_S(w), \nabla F(w) \rangle - \|\nabla F(w)\|^2 \right)^2 \right] \leq \theta^2 \|\nabla F(w)\|^4$$

$$E \left[\left(\langle \nabla f_S(w), \nabla F(w) \rangle - \|\nabla F(w)\|^2 \right)^2 \right] \leq \frac{E \left[\left(\langle \nabla f_i(w), \nabla F(w) \rangle - \|\nabla F(w)\|^2 \right)^2 \right]}{|S|} \leq \theta^2 \|\nabla F(w)\|^4$$

exact test 

Easier to compute if we can compute $\nabla F(w)$

Practical (approximate) test

$$\frac{\text{Var}_{i \in S} [\langle \nabla f_i(w), \nabla f_S(w) \rangle]}{|S|} \leq \theta^2 \|\nabla f_S(w)\|^4$$

$$\text{Var}_{i \in S} [\langle \nabla f_i(w), \nabla f_S(w) \rangle] = \frac{1}{|S| - 1} \sum_{i \in S} (\langle \nabla f_i(w), \nabla f_S(w) \rangle - \|\nabla f_S(w)\|^2)^2$$



Raghu Bollapragada, Richard Byrd, and Jorge Nocedal Adaptive sampling strategies for stochastic optimization. SIAM Journal on Optimization, 28(4):3312–3343, 2018.

The Orthogonality Test

$$\langle \nabla f_S(w), \nabla F(w) \rangle \neq 0$$

- Less restricted, needed for proving the convergence!
- We need to make sure that

note that

$$E \left[\left\| \nabla f_S(w) - \frac{\langle \nabla f_S(w), \nabla F(w) \rangle}{\|\nabla F(w)\|^2} \nabla F(w) \right\|^2 \right] = 0$$

need to bound its variance

$$E \left[\left\| \nabla f_S(w) - \frac{\langle \nabla f_S(w), \nabla F(w) \rangle}{\|\nabla F(w)\|^2} \nabla F(w) \right\|^2 \right] \leq \nu^2 \|\nabla F(w)\|^2$$

approximate test

$$\frac{1}{|S| - 1} \frac{\sum_{i \in S} \left\| \nabla f_i(w) - \frac{\langle \nabla f_i(w), \nabla f_S(w) \rangle}{\|\nabla f_S(w)\|^2} \nabla f_S(w) \right\|^2}{|S|} \leq \nu^2 \|\nabla f_S(w)\|^2$$



The Convergence Guarantee

$$\eta \leq \frac{1}{L(1 + \theta^2 + \nu^2)}$$

- Let $F(w)$ be L smooth and μ strongly convex;
- The $|S|$ is picked such that **exact inner and orthogonality tests hold**

then

$$E[F(w_k) - F^*] \leq (1 - \eta\mu)^k (F(w_0) - F^*)$$



Raghu Bollapragada, Richard Byrd, and Jorge Nocedal Adaptive sampling strategies for stochastic optimization. SIAM Journal on Optimization 28(4):3312–3343, 2018.

impossible to evaluate in real applications!!!

Refined AdaGrad analysis by Li & Orabona

$$w_{k+1} = w_k - \eta_k \nabla f_{S_k}(w_k)$$

Batched stochastic gradient:

$$\nabla f_{S_k}(w_k) \equiv \frac{1}{|S_k|} \sum_{i \in S_k} \nabla f_i(w_k)$$

Step size ([Li & Orabona \(2019\)](#)):

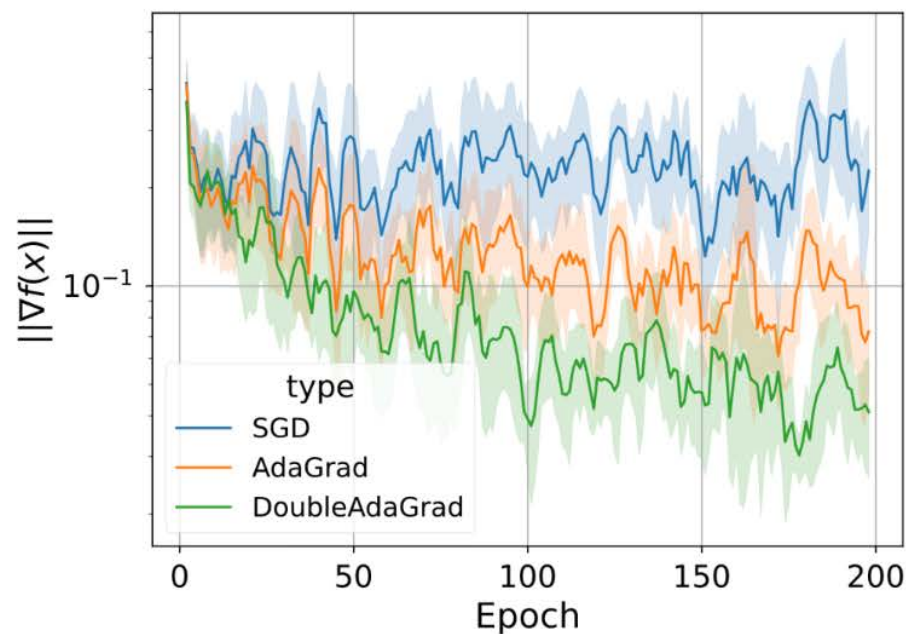
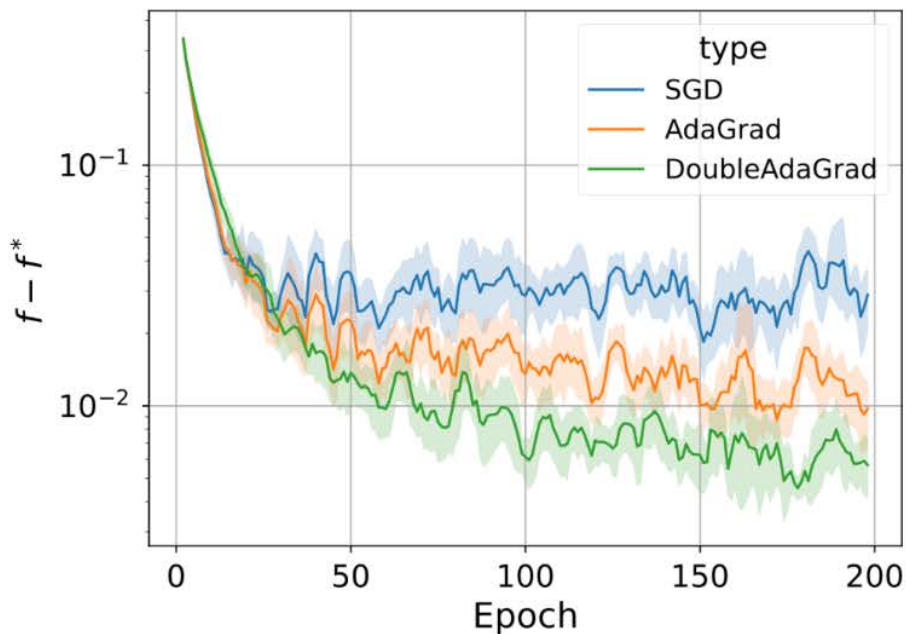
$$\eta_k = \frac{\alpha}{\left(\beta + \sum_{i=1}^{k-1} \|\nabla f_{S_i}(w_i)\|^2\right)^{\frac{1}{2} + \tau}}$$

Convergence: $O(1/\varepsilon)$ or $O(1/\sqrt{\varepsilon})$, depending on the noise variance upper bound σ^2

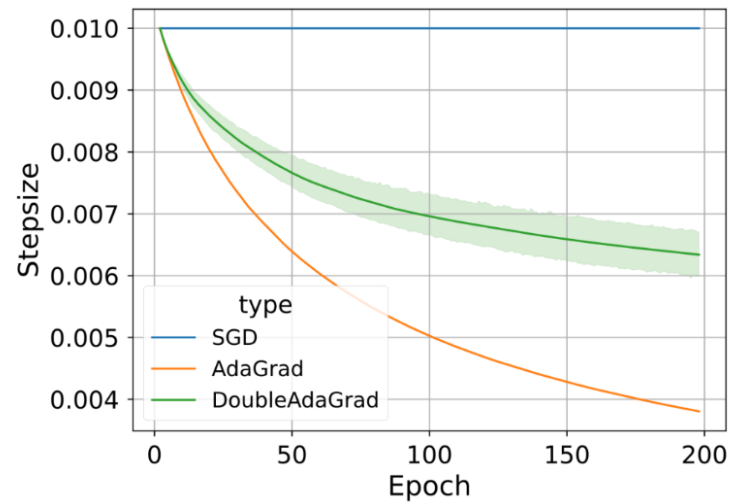
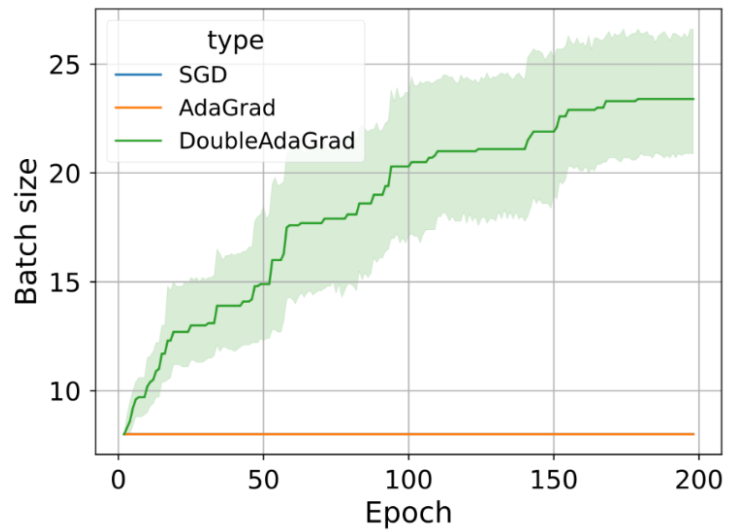
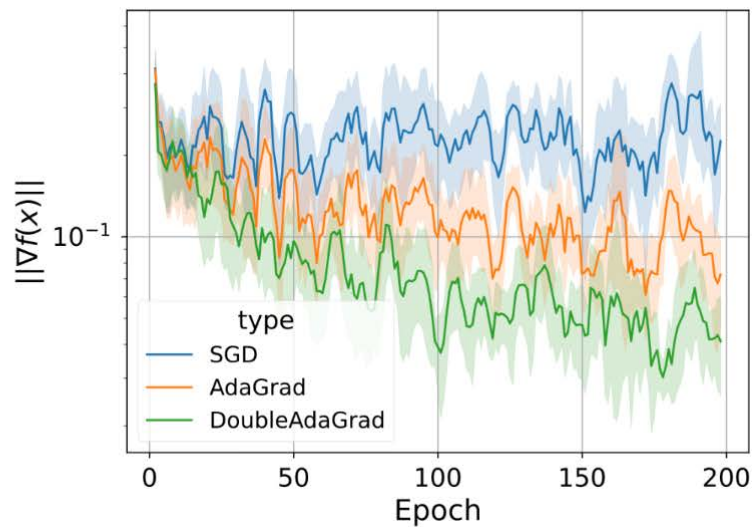
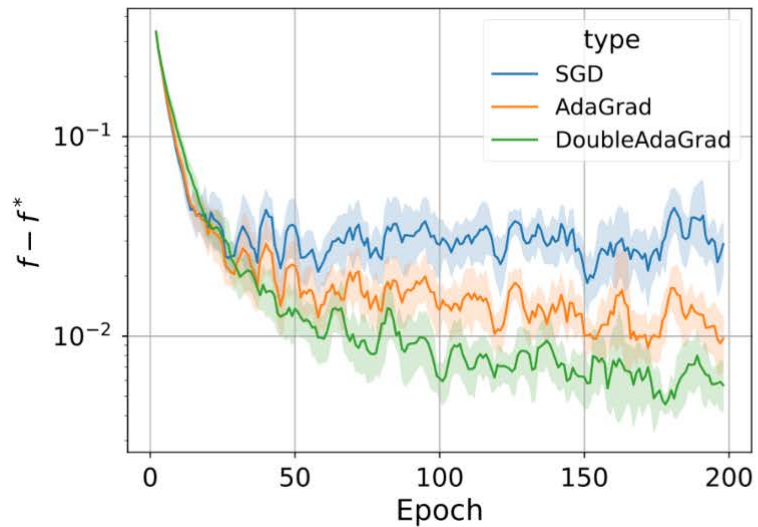


Xiaoyu Li, Francesco Orabona: *On the Convergence of Stochastic Gradient Descent with Adaptive Stepsizes*

Combining Bollapragada et al. (2018) + Li & Orabona (2019)



Petr Ostroukhov, Dmitry Kamzolov, Chulu Xiang, Aigerim Zhumabayeva, Alexander Gasnikov, MT: *DoubleAdaGrad: AdaGrad With Adaptive Batch Size*, 2023



Acknowledgement

Special thanks to

- Martin Jaggi
- Mark Schmidt
- Konstantin Mishchenko
- Samuel Horvath
- Eduard Gorbunov

for sharing their slides that some parts of this lecture are based on

Mohamed bin Zayed
University of Artificial Intelligence
Masdar City
Abu Dhabi
United Arab Emirates

