WASP WINTER CONFERENCE 2022

Poster Catalogue

AI math

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

WASP WINTER CONFERENCE 2022
POSTER CATALOGUE 1/4

AI MATH

Agerberg, Jens
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Data, geometry and homology

Homology-based invariants can be used to characterize the geometry of datasets and thereby gain some understanding of the processes generating those datasets. In this work we investigate how the geometry of a dataset changes when it is subsampled in various ways. In our framework the dataset serves as a reference object; we then consider different points in the ambient space and endow them with a geometry defined in relation to the reference object, for instance by subsampling the dataset proportionally to the distance between its elements and the point under consideration. We illustrate how this process can be used to extract rich geometrical information, allowing for example to classify points coming from different data distributions.

Agerberg, Jens
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Data, geometry and homology

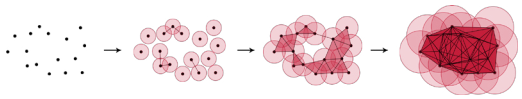
Jens Agerberg, KTH
Math department, Math of data and AI
Joint work with Wojciech Chachólski and Ryan Ramanujam

Abstract

Homology-based invariants can be used to characterize the geometry of datasets and thereby gain some understanding of the processes generating those datasets. In this work (under review) we investigate how the geometry of a dataset changes when it is subsampled in various ways. In our framework the dataset serves as a reference object; we then consider different points in the ambient space and endow them with a geometry defined in relation to the reference object, for instance by subsampling the dataset proportionally to the distance between its elements and the point under consideration. We illustrate how this process can be used to extract rich geometrical information, allowing for example to classify points coming from different data distributions.

Methods

Persistent homology: from point clouds to persistence modules
From a point cloud we can construct a Vietoris-Rips complex, a combinatorial object encoding its geometry, parametrized by $\epsilon \in [0, \infty)$.



By taking homology we get (for each homological degree) a vector space for each ϵ and a linear map for each $\tau \leq \epsilon \in [0, \infty)$. These linear maps are called *persistence modules* and are the output of persistent homology.

Metrics and machine learning: from persistence modules to stable ranks
Persistence modules can be seen as a summary of geometrical aspects of the point cloud. To be useful we need metrics to compare them and ways to develop machine learning algorithms on them.
For this we use the framework of *stable rank*: persistence modules have a discrete invariant called rank, this invariant can be stabilized by considering instead the minimum rank in a growing neighborhood of the module, leading to a type of homology-based invariant in the form of a non-increasing piecewise constant function. Since this function space is a Hilbert space one can consider a kernel based on stable rank [1] for use in machine learning.

From global to local
Homology-based invariants are often used to characterize global aspects of a dataset. In this work, we instead investigate whether they can be useful in describing a single point in the ambient space, by subsampling a dataset (called reference object) according to the distance of its members to the point:

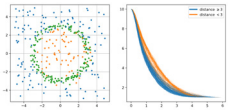
1. Choose a reference object: a finite subset $\mathcal{R} \subset \mathbb{R}^N$ and a point $p \in \mathbb{R}^N$.
2. Attach a probability distribution to $\mathcal{R}$. We are interested here in distributions that attach high probability to points $r \in \mathcal{R}$ which have low distance to p and low probability to more remote points.
3. Sample s points from the reference object according to the probability distribution. Repeat n times and each time compute persistence modules and stable ranks.
4. Average the stable ranks to get a descriptor characterizing the point p.

References

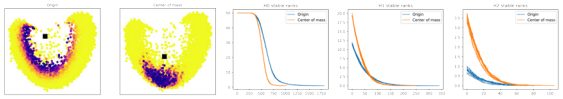
[1] Jens Agerberg, Ryan Ramanujam, Martina Scolamiero, and Wojciech Chachólski. Supervised learning using homology stable rank kernels. *Frontiers in Applied Mathematics and Statistics*, 7:39, 2021.

Selected Results

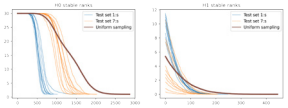
We start with data consisting of random points on the plane. For each point, the reference object (the green points, sampled from a circle) is sampled relative to its distance to the point, Persistent homology and stable ranks are computed. The stable ranks clearly group into orange (for points inside the circle) and blue (outside the circle), indicating that interesting geometric properties can be found.



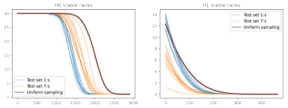
We now use as reference object the MNIST dataset for digit 1. We select two points from the ambient space, $\mathbb{R}^{784}$: the origin and the center of mass of the reference object. Using dimensionality reduction, we can illustrate what it means to sample the reference object relative to those 2 points. Now the stable ranks resulting from the sampling allow to distinguish the 2 points, for all homological degrees.



We now use as reference object the union of MNIST training sets for digit 1 and 7. We select out-of-sample 1:s and 7:s and represent them by their stable rank, obtained by sampling the reference object in the same way as before. In many cases, the geometry of the reference object close to the out-of-sample digits allow to distinguish them. This is further quantified by training an SVM classifier based on the stable rank kernel (in a semi-supervised learning setup: the reference object is used in an unlabeled fashion and the SVM is only trained on 10 samples from each class).



Interestingly, to distinguish a pair of digits, one can also use other digits as reference object. Here 2:s and 3:s are used as reference object to distinguish 1:s and 7:s.



Aronsson, Jimmy
Chalmers

Homogeneous vector bundles and G-equivariant convolutional neural networks

G-equivariant convolutional neural networks (GCNNs) is a geometric deep learning model that uses global symmetry to improve learning. Most GCNNs use convolutional layers to transform data in a translation equivariant manner, like the sliding kernels of CNNs but generalized to other symmetries, e.g. rotation-equivariant transformations of spherical data. We analyze GCNNs and classify those G-equivariant layers that are expressible as convolutional layers. That is, we characterize the expressivity of convolutional layers.

Aronsson, Jimmy
Chalmers

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Equivariant Neural Networks

Jimmy Aronsson
Chalmers University of Technology
Department of Mathematical Sciences

UNIVERSITY OF GOTHENBURG

CHALMERS

Homogeneous vector bundles and G-equivariant convolutional neural networks

G-equivariant convolutional neural networks (GCNNs) is a geometric deep learning model that uses global symmetry to improve learning. Most GCNNs use convolutional layers to transform data in a translation equivariant manner, like the sliding kernels of CNNs but generalized to other symmetries, e.g. rotation-equivariant transformations of spherical data. We analyze GCNNs and classify those G-equivariant layers that are expressible as convolutional layers. That is, we characterize the expressivity of convolutional layers.

Suppose that we are given data defined on a homogeneous space $\mathcal{M}$, e.g. meteorological wind vector fields on the rotation symmetric sphere S^2, or digital images defined on the translation symmetric pixel lattice $\mathbb{Z}^2$. The general idea is to transform such data in *translation equivariant* ways, preserving the global symmetry. A special case is translation *invariant* layers, which produce the same output for all translations of the input. Such layers are useful when classifying images, for example, as they make the same class prediction no matter where objects are located within an image.

Globally symmetric spaces are also called homogeneous:

Definition. Let G be a Lie group. A smooth manifold $\mathcal{M}$ is a *homogeneous space* if there exists a smooth, transitive action

$$G \times \mathcal{M} \to \mathcal{M}, \qquad (g, x) \mapsto g \cdot x.$$

Any homogeneous space $\mathcal{M}$ is diffeomorphic to a quotient space G/K for some closed subgroup $K \leq G$. Examples of homogeneous spaces include the Euclidean spaces $\mathbb{R}^n$, lattices $\mathbb{Z}^n$, spheres $S^n \simeq SO(n+1)/SO(n)$, and many others.

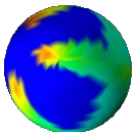
Vector bundles over a homogeneous space $\mathcal{M} = G/K$ often inherit its global symmetry, i.e. there often exists a smooth, transitive action $G \times E \to E$. Such vector bundles are also called *homogeneous* and are uniquely characterized by some K-representation ρ. We write $E = E_\rho$.

Data points are viewed as functions that attach a feature vector/scalar at each point of a homogeneous space. They are formalized as sections of homogeneous vector bundles:

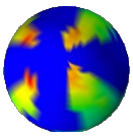
Definition. A *data point* is a square-integrable section $s : \mathcal{M} \to E_\rho$. We denote the space of all data points by $L^2(E_\rho)$.

The global symmetry in G induces a representation $\mathrm{Ind}_K^G \rho$ on $L^2(E_\rho)$ that performs translations of data points:

$$\big(\mathrm{Ind}_K^G \rho(g) s\big)(x) = g \cdot s(g^{-1} x)$$



$s \in L^2(E_\rho)$



$\mathrm{Ind}_K^G \rho(g) s \in L^2(E_\rho)$

Definition. Let E_ρ and E_σ be homogeneous vector bundles over $\mathcal{M}$. A linear transformation $\Phi : L^2(E_\rho) \to L^2(E_\sigma)$ is called a *G-equivariant layer* if it intertwines the induced representations:

$$\mathrm{Ind}_K^G \sigma \circ \Phi = \Phi \circ \mathrm{Ind}_K^G \rho.$$

For example, if Φ produces bounding boxes around objects in an image, then translating the image will also translate the bounding boxes.

Implementations of GCNNs primarily use *convolutional layers*[1]

$$(\Phi s)(g) = \int_G \kappa(g^{-1} g') s(g') \, dg,$$

where κ is a matrix-valued kernel. All convolutional layers are G-equivariant layers but the latter notion is much more general, so implementations that only use convolutional layers could be unnecessarily restrictive. It is thus interesting to analyze the relation between G-equivariant and convolutional layers.

Our main theorem characterizes those G-equivariant layers that are expressible as convolutional layers. It does so for extremely general homogeneous spaces $\mathcal{M} \simeq G/K$, including the Euclidean spaces, grids, spheres, and even Minkowski spacetime (which is homogeneous with respect to the Poincaré group).

Theorem. Consider a homogeneous space $\mathcal{M} = G/K$ with G a unimodular Lie group of type I and $K \leq G$ a compact subgroup. Let E_ρ, E_σ be homogeneous vector bundles over $\mathcal{M}$ and let

$$\Phi : L^2(E_\rho) \to L^2(E_\sigma),$$

be a G-equivariant layer. If Φ maps into a space of bandlimited functions, then Φ is a convolutional layer.

The following corollary is especially useful, since implementations of GCNNs primarily use finite or discrete groups.

Corollary. If G is either discrete abelian or finite, then any G-equivariant layer is a convolutional layer.

Using only convolutional layers is therefore not a restriction, when G is either discrete abelian or finite; Implementations based on these layers are maximally expressive. For instance, convolutional layers

$$[\kappa \star s](x) = \sum_{y \in \mathbb{Z}^2} \kappa(y - x) s(y), \qquad (\mathcal{M} = G = \mathbb{Z}^2)$$

are the only possible equivariant transformations of digital images s.

[1] Convolutional layers actually transform *feature maps* rather than data points, but these objects are equivalent. See [1] for details.

References

[1] Aronsson, J.. "Homogeneous vector bundles and G-equivariant convolutional neural networks." *arXiv preprint arXiv:2105.05400* (2021).

[2] Gerken, J.E., Aronsson, J., et al. "Geometric deep learning and equivariant neural networks." *arXiv preprint arXiv:2105.13926* (2021).

[3] Cohen, T. and Welling, M. "Group equivariant convolutional networks." *International conference on machine learning.* PMLR, 2016.

[4] Kondor, R., and Trivedi, S. "On the generalization of equivariance and convolution in neural networks to the action of compact groups." *International Conference on Machine Learning.* PMLR, 2018.

Bengtsson Bernander, Karl
Uppsala University

Robust learning of geometric equivariances

We extend convolutional neural networks (CNNs) to provide rotation equivariance. We evaluate on the oral cancer dataset to diagnose malignant cancer, using the VGG16 classifier architecture. We also evaluate on the BBB038 dataset of highly varied cell nuclei, this time using the U-net architecture combined with a discriminative loss function for semantic instance segmentation. We expect that incorporating rotation equivariance into CNNs will increase the expressive capacity without increasing the number of parameters, reducing overfitting. Also, since data augmentation can be reduced, misclassification due to interpolation artifacts should decrease. The results indicate that this holds for the classifier network, but more experiments are needed to verify this for the semantic instance segmentation network.

Bengtsson Bernander, Karl
Uppsala University

Robust learning of geometric equivariances

Karl Bengtsson Bernander, Joakim Lindblad, Nataša Sladoje, Robin Strand, Ingela Nyström
Centre for Image Analysis, Department of Information Technology, Uppsala University, Sweden

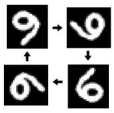
Abstract

We extend convolutional neural networks (CNNs) to provide rotation equivariance. We evaluate on the oral cancer dataset to diagnose malignant cancer, using the VGG16 classifier architecture. We also evaluate on the BBB038 dataset of highly varied cell nuclei, this time using the U-net architecture combined with a discriminative loss function for semantic instance segmentation. We expect that incorporating rotation equivariance into CNNs will increase the expressive capacity without increasing the number of parameters, reducing overfitting. Also, since data augmentation can be reduced, misclassification due to interpolation artifacts should decrease. The results indicate that this holds for the classifier network, but more experiments are needed to verify this for the semantic instance segmentation network.

Classification on the oral cancer dataset

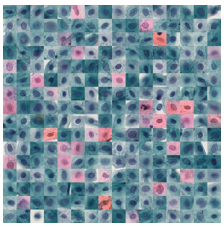
One feature of standard convolutional neural networks (CNNs) is translational invariance: the result of convolving an input with a filter and then shifting the output is identical to shifting the input and then applying the convolution. We are interested in other equivariances, such as rotations and scaling. Recent works on rotation equivariance in CNNs include:

- **Group-equivariant convolutional networks (G-CNNs) [1], using group-convolutions.**
- **General E(2)-Equivariant Steerable CNNs [2], available as a library in Pytorch.**



$$T_g\Phi(f)=\Phi(T_g f)$$

Rotational equivariance: filtering (Φ) an input , then rotating (T_g), gives the same result as filtering on the rotated input.



Microscopy image of cells from the oral cavity

The oral cancer dataset. We modified the VGG16 classifier to use group-equivariant convolutions on the p4 group, consisting of translations and rotations of multiples of 90 degrees.

The baseline CNN version combined with data augmentation of rotations of multiples of 90 degrees yield an accuracy score of around 56 %. The equivariant version , without data augmentation, yields 60 %. The latter architecture is less sensitive to overrfitting.

Semantic Instance Segmentation

We further modify the U-net architecture with a discriminative loss function [3] to be equivariant to rotations of multiples of 90 degrees. The methods yield a DICE score of about 0.7 for both versions of the U-net.

Instance segmentation on the modified BBBC038 dataset. The left image shows the input image, the middle one the results from the baseline U-net architecture, and the right one the results from the U-net architecture modified to be equivariant to rotations of multiples of 90 degrees.

Further directions

We plan to train the instance segmentation network for hundreds of more epochs to verify the hypothesis. That is, that ordinary CNN architectures, combined with data augmentation of multiples of rotations of 90 degrees, can be replaced with networks that are equivariant by design to those same transformations. We also plan to use another clustering method than K-means, preferably one without a predetermined number of clusters.

Both the instance segmentation network and the classifier networks can be tested on other datasets, and with other symmetry groups.

For larger datasets, moving to distributed training over multiple GPUs show promise for speeding up the training phase. Moving to a cloud computational environment could also allow for more flexibility, with the drawback that you lose some control over your own development environment.

References

1. T.S. Cohen, M. Welling. Group Equivariant Convolutional Networks. Proceedings of the International Conference on Machine Learning (ICML), 2016
2. Maurice Weiler and Gabriele Cesa. General E(2)-equivariant´steerable CNNs. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019.
3. Bert De Brabandere, Davy Neven, and Luc Van Gool.Semantic instance segmentation with a discriminative loss´function. In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2017), 2017.

Bozorgpanah, Aso
Doctoral student, Department of Computing Science

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

The Interpretable Protected Machine Learning Model with Privacy

Machine learning (ML) models have the potential to enhance products. It is a type of Artificial Intelligence (AI) that allows software applications to predict outcomes.
Data-driven models built using ML have proven their usefulness. Nevertheless, ML algorithms do not explain their predictions, which is a barrier to ML adoption. To address this issue, the researcher uses eXplainable Artificial Intelligence (XAI). XAI explains why a ML model yields a predicted output for a certain input.
Understanding why a model makes a prediction is important, but it is not enough. So, other principles need to be addressed for ML deployment in the real world. In the current work, privacy is one of the challenges that is discussed.
We studied the effect of data privacy techniques^{[1]} on SHapley Additive exPlanations (SHAP)^{[2]}.
By applying SHAP the output of any ML model can be explained. The output model is interpretable.
Our aim is to understand how data protection affects the measures related to explainability. Hence, we performed a series of experiments comparing the effects of data masking procedures on the explainability of models according to SHAP on the data set.

Bozorgpanah, Aso

Doctoral student, Department of Computing Science

The Interpretable Protected Machine Learning Model with Privacy

Aso Bozorgpanah, Ph.D., Umeå University
Dept. Computing Science, NAUSICA: PrivAcy-AWare traNSparent deCIsions group
Supervisors: Prof. Vicenç Torra (Umu), Associate professor. Lili Jiang (Umu)

Motivation & Research Goals

Machine learning (ML) models have the potential to enhance products. It is a type of Artificial Intelligence (AI) that allows software applications to predict outcomes. Data-driven models built using ML have proven their usefulness. Nevertheless, ML algorithms do not explain their predictions, which is a barrier to ML adoption. To address this issue, the researcher uses eXplainable Artificial Intelligence (XAI). XAI explains why a ML model yields a predicted output for a certain input. Understanding why a model makes a prediction is important, but it is not enough. So, other principles need to be addressed for ML deployment in the real world. In the current work, privacy is one of the challenges that is discussed. We studied the effect of data privacy techniques[1] on SHapley Additive exPlanations (SHAP)[2]. By applying SHAP the output of any ML model can be explained. The output model is interpretable. Our aim is to understand how data protection affects the measures related to explainability. Hence, we performed a series of experiments comparing the effects of data masking procedures on the explainability of models according to SHAP on the data set.

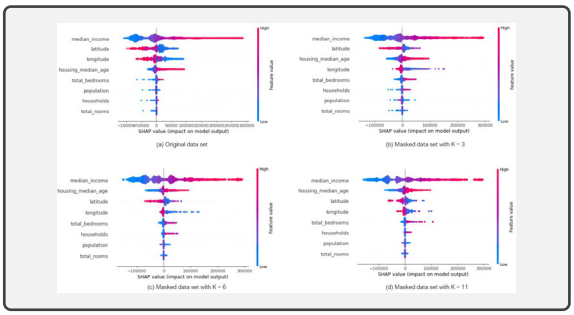
Methods

An implications' analysis of applying data privacy techniques to explainability was performed. It is claimed[3] privacy and explainability are incompatible. While we designed an explainable model along with privacy. In this regard, Maximum Distance to Average Vector (MDAV) was applied for achieving microaggregation. The MDAV is a masking method that provides k-anonymity to protect data[4]. Microaggregation is one of the most efficient approaches in relation to the trade-off risk-utility. It consists of building small clusters with the original data and then replacing each of the data with a cluster center that is representative of the whole cluster. Microaggregation is flexible and permits implementing k-anonymity for any kind of data. We supposed $k = [1, 15]$. Although the range of k is different for various datasets, the k value should be selected in a reasonable range to have high accuracy. After masking the dataset by MDAV. SHapley Additive exPlanations (SHAP) was done on the masked dataset. SHAP[2] is a method to explain individual predictions. It is based on the Shapley Value of game theory. TreeSHAP is an estimation approach of SHAP that was used. TreeSHAP defines the value function in terms of the conditional expectation to estimate effects instead of the marginal expectation.

As the above progress is shown, we present a privacy-preserving explainable ML model. The explainable machine learning algorithms were applied to the protected data to train machine learning models and explain the result of their predictions. They were compared with the one obtained without masking.

Selected Results

A baseline model was trained on the original dataset, then, additional models were trained on the masked datasets. The explainable models were not changed even after protecting data for $k = [1, 12]$. The results showed that explainability for the protected model by MDAV was similar to the one obtained with the original data. Therefore, decisions on the amount of distortion to achieve protection through microaggregation and k-anonymity should be led by the desired trade-off between disclosure risk and model accuracy.
We presented an approach, what kind of data privacy methods are more feasible to explainability after applying SHAP to make an explainable ML model. The ML models trained on masked data were evaluated by their results explainability. We considered feature importance analysis of the final models (based on decision trees) using SHAP. Our approach were applied on 'USA Housing' dataset, and the results were compared between the results for the original and the masked data. The results for $k = [3, 6, 11]$ are shown in (b), (c), and (d) respectively in the below figure. It is clear that the extracted explainability are similar among all four models.



We found that interpretability using SHAP is studied for k-anonymous data. The results showed that qualitative properties of attributes were maintained for masked data. Then, the decision on which level of privacy and the amount of distortion was appropriate needs to focus on the risk-utility trade-off. For instance, a user needs to take into account both the value of k and the utility of the masked data set.

The explainable ML models can be considered along with privacy. We found that how explainability can be affected by data privacy methods and masking methods keep utility. Future research should seek to address other XAI requirements within a privacy-preserving framework to assess to what extent these tools apply in privacy-by-design MI.

References

[1] Data privacy: Foundations, new developments and the big data challenge
Torra, Vicenç
Springer, 2017

[2] A unified approach to interpreting model predictions
Lundberg, Scott and Lee, Su-In
arXiv preprint arXiv:1705.07874

[3] Show Us the Data: Privacy, Explainability, and Why the Law Can't Have Both
Grant, Thomas D and Wischik, Damon J
George Washington Law Review Arguendo, 2020

[4] Efficient k-anonymous microaggregation of multivariate numerical data via principal component analysis
Monedero, David Rebollo and Mezher, Ahmad Mohamad and Colomé, Xavier Casanova and Forné, Jordi and Soriano, Miguel
Information Sciences, Elsevier, 2019

Breitholtz, Adam
Chalmers

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Data dependent bounds for domain adaptation

The study of generalization is one of the cornerstones of machine learning theory. Tight generalization bounds are potential tools for guaranteeing adequate performance and the PAC-Bayes framework has proven useful in deriving such bounds when good model priors are known and test cases match training cases in distribution.
However, in real world tasks, where deep neural networks are the models of choice and training and test cases come from different domains, deriving tight and estimable bounds remains an unresolved challenge.
In our work, we combine recent advances in PAC-Bayes domain adaptation with data-dependent priors to give estimable and informative bounds for problems where classical bounds are vacuous. We apply this method to a domain adaptation image classification task and find that it produces tighter bounds. We study which terms dominate the bounds and identify possible directions for further improvement.

Breitholtz, Adam
Chalmers

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Data Dependent Priors for Domain Adaptation Bounds

Adam Breitholtz, PhD, Chalmers University of Technology
Dept. Computer Science and Engineering, Data Science and AI division
Supervisors: Ass.Prof. Fredrik D. Johansson (CTH) and Prof. Devdatt Dubhashi (CTH)

Motivation & Research Goals

UNIVERSITY OF TECHNOLOGY

The study of generalization is a cornerstone of machine learning theory. Our understanding of how generalization functions is crucial to confidently engineer and deploy models in high stakes, real world domains, such as healthcare. Tight generalization bounds are potential tools for guaranteeing adequate performance and the PAC-Bayes framework has proven useful in deriving such bounds when good model priors are known and test cases match training cases in distribution.

However, in real world tasks, where deep neural networks are the models of choice and training and test cases come from different domains, deriving tight and estimable bounds remains an unresolved challenge. Recent work has shown that using data dependent priors is a promising way to achieve tighter bounds for deep neural networks in stationary domains. In this work, we combine recent advances in PAC-Bayes domain adaptation with data-dependent priors to give estimable and informative bounds for problems where classical bounds are vacuous. We apply this method to a domain adaptation image classification task and find that it produces tighter bounds. We study which terms dominate the bounds and identify possible directions for further improvement.

Methods

We apply data dependent priors[1] on two bounds from the literature[2] for a domain adaptation image classification task. We seek to understand how the addition of data dependent priors affects the sample generalization part of the bound. Further, it is of interest to find if any specific part of the bounds dominates and in which range it does so. Moreover, we also want to investigate if the dominating terms change as the training of the model progresses. I.e., we evaluate the bound at several different points during the training of the model, as the KL term is expected to increase increase as the posterior drifts away from the prior.

Theorem 1 (Additive bound)**.** *For any real numbers $\omega, \alpha > 0$ we have with probability at least $1-\delta$ over the random choice of $S \times T_x \sim (\mathcal{S} \times \mathcal{T}_{\mathbf{X}})^m$; for every posterior ρ on $\mathcal{H}$*

$$\mathop{\mathbb{E}}_{h\sim\rho} R_{\mathcal{T}}(h) \leq \mathop{\mathbb{E}}_{h\sim\rho} \omega' \hat{R}_{\mathcal{S}}(h) + \alpha' \frac{1}{2}\hat{Dis}_{\rho}(S, T_x) + \left(\frac{\omega'}{\omega} + \frac{\alpha'}{\alpha}\right)\frac{KL(\rho\|\pi) + \log\frac{3}{\delta}}{m} + \lambda_{\rho} + \frac{1}{2}(\alpha' - 1),$$

where $\hat{Dis}_{\rho}(S, T_x) = |\hat{d}_{T_x} - \hat{d}_{\mathcal{S}_x}|$ is the empirical domain disagreement, $\lambda_{\rho} = |e_{\mathcal{T}}(\rho) - e_{\mathcal{S}}(\rho)|$ and $\omega' = \frac{\omega}{1-e^{-\omega}}$ and $\alpha' = \frac{2\alpha}{1-e^{-2\alpha}}$.

Theorem 2 (Multiplicative bound)**.** *For any real numbers $a, b > 0$ we have with probability at least $1-\delta$ over the choices $S \sim (\mathcal{S})^m$ and $T_x \sim (\mathcal{T}_x)^n$*

$$\mathop{\mathbb{E}}_{h\sim\rho} R_{\mathcal{T}}(h) \leq a'\frac{1}{2}\hat{d}_{T_x} + b'\beta_{\infty}(\mathcal{T}\|\mathcal{S})\hat{e}_{\mathcal{S}} + \eta_{\mathcal{T}\setminus\mathcal{S}} + \left(\frac{a'}{na} + \frac{b'\beta_{\infty}(\mathcal{T}\|\mathcal{S})}{mb}\right)\left(2KL(\rho\|\pi) + \ln\frac{2}{\delta}\right)$$

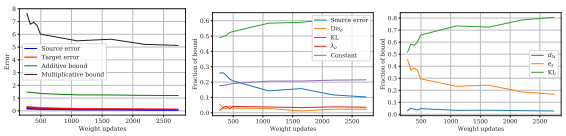
where $a' = \frac{a}{1-e^{-a}}$, $b' = \frac{b}{1-e^{-b}}$,

$$\beta_{\infty}(\mathcal{T}\|\mathcal{S}) = \sup_{(x,y)\sim supp(\mathcal{S})} \frac{\mathcal{T}(x,y)}{\mathcal{S}(x,y)}$$

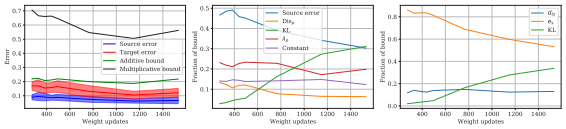
and

$$\eta_{\mathcal{T}\setminus\mathcal{S}} = \Pr_{(x,y)\sim\mathcal{T}}\Big((x,y) \notin supp(\mathcal{S})\Big) \sup_{h\in\mathcal{H}} R_{\mathcal{T}\setminus\mathcal{S}}(h).$$

Selected Results



The two bounds are evaluated on the learning task described earlier. In the center and rightmost figures, corresponding to the additive and multiplicative bound respectively, we show the contribution of different terms in the bounds. The labels refer to the term in the bound including any multiplicative constants.



The same type of figure as the one above, however, here we have used 30% of the source data to inform the prior. We see that the bounds are no longer vacuous and that when the KL divergence is small the unobservable λ_{ρ} term is a significant part of the additive bound. The shaded area around source and target error represents one standard deviation.

References

[1] On the role of data in PAC-Bayes bounds
Dziugaite, G. K.; Hsu, K.; Gharbieh, W.; Arpino, G.; andRoy, D. M.
In Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)

[2] PAC-Bayes and Domain Adaptation
Germain, P.; Habrard, A.; Laviolette, F.; and Morvant, E
Neurocomputing, 2020

Bågmark, Kasper
Chalmers

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Energy-based approach for the nonlinear filtering problem using a deep splitting method

In this work the main goal is to approximate the optimal nonlinear filter of an underlying high dimensional process through deep learning. This work utilise the deep splitting method, developed for the approximation of solutions to (stochastic) partial differential equations. We solve the Zakai equation, which in turn solves the filtering problem, with an energy-based model. Taking the observations as input, a computationally fast filter is obtained. The model is employed on a nonlinear bistable problem and shows promising performance. The bootstrap particle filter is used for comparison.

Bågmark, Kasper
Chalmers

Energy-based approach for the nonlinear filtering problem using a deep splitting method

Kasper Bågmark, PhD student
Department of Mathematical Sciences
Supervisors: Adam Andersson (Chalmers and SAAB), Stig Larsson (Chalmers)

CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG

1. The optimal filtering problem

Consider a system of stochastic differential equations (SDE) (X,Y) given by

$$X_t = X_0 + \int_0^t \mu(X_s)\,\mathrm{d}s + \int_0^t \sigma(X_s)\,\mathrm{d}W_s, \tag{1}$$

$$Y_t = \int_0^t h(X_s)\,\mathrm{d}s + V_t, \tag{2}$$

where X is called the **underlying (unobserved) state process** in $L^2(\Omega;\mathbb{R}^d)$ and Y is the **observation process** in $L^2(\Omega;\mathbb{R}^d)$. W and V are two independent $\mathbb{R}^d$-valued Brownian motions. The optimal filtering problem consists of finding the probability density of the state given the observation, $p\left(X_t|\,(Y_s)_{0\le s\le t}\right)$. This is called the **filtering density**.

2. The Zakai equation

The unnormalized version of the filtering density $p_t := p\left(X_t|\,(Y_s)_{0\le s\le t}\right)$ can be shown to satisfy the stochastic partial differential equation (SPDE) known as the **Zakai equation**. The strong form of the Zakai equation reads

$$p_t = p_0 + \int_0^t \mathcal{A}p\,\mathrm{d}s + \int_0^t p_s h^\top\,\mathrm{d}Y_s, \tag{3}$$

where $\mathcal{A}$ is the second order operator from the Kolmogorov forward equation related to X, and Y is the observed process. By substitution, the second integral contains an Ito integral.

3. Methods

Deep splitting method: In [1] a splitting method for SPDE, including (3), is introduced. The splitting method is based on solving the linear part of the equation analytically via a Feynman–Kac formula, and adding the nonlinearity in a second step. The scheme is formulated as a recursive (in time) nonlinear least squares problem. The recursion reads

$$\begin{aligned} p_t = \arg\min_{u\in C(\mathbb{R}^d,\mathbb{R})} \mathbb{E}\,\big|u(X_{T-t_{n+1}}) - (p_{t_n}(X_{T-t_n}) \\ + f(X_{T-t_n}, p_{t_n}(X_{T-t_n}), (\nabla p_{t_n})(Y_{T-t_n}), \Delta t, \Delta Y))|^2. \end{aligned} \tag{4}$$

Here f is short-notation for the Euler–Maruyama or Milstein schemes for (3). In [1] u is approximated with a deep neural network for every realization of Y. We consider a more general framework where we let the model take the observation sequence as input.

Energy-based approach: In probabilistic model learning, one successful technique in density estimation and maximum likelihood estimation is the use of energy-based methods (EBM) [2]. The idea is to approximate $p(x|y)$ by associating an scalar energy f^θ to each pair of (x,y) where in our setting $x := X_{t_n}$ and $y := Y_{t_1:t_n}$. The model is trained to associate high energies to pairs that are unlikely and low energy to values that are likely. In our setting we use the unnormalized parametric model

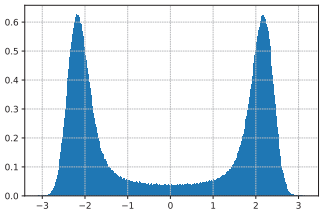
$$\widehat{p}_t(x|y) := e^{-f^\theta(x,y)}, \tag{5}$$

where θ denotes the parameters of our energy-based model.

4. Numerical Results

Our proposed method is to combine the energy-based approach with the deep splitting method. We demonstrate the method on a nonlinear bistable problem.

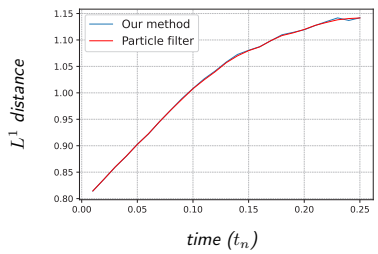
Example. *Consider a process (X,Y) satisfying (1) and (2) with nonlinear drift $\mu(x) = 5x - x^3$, constant diffusion $\sigma(x) = 1$ and linear observation $h(x) = x$ with initial density $p_0 = \mathcal{N}(0,1)$. Below we see the underlying density of the state at time $T = 0.5$.*



We compare the result from our approximation to the bootstrap particle filter by measuring the distance from the true state X_t to the mean of our method and the bootstrap particle filter (PF), respectively. Formally this is the $L^1(\Omega;\mathbb{R}^d)$-norm

$$\mathbb{E}\|X_{t_n} - \mathbb{E}\left[X_{t_n}|\,Y_{t_1:t_n}\right]\| \quad \text{and} \quad \mathbb{E}\|X_{t_n} - \widehat{\mu}_{t_n}\|$$

for $n = 1,\dots,25$, where $\mathbb{E}\left[X_t|\,Y_{t_1:t_n}\right]$ is approximated by the particle filter and $\widehat{\mu}_{t_n}$ is the estimated mean from our method.



References

[1] T. Beck, C., Becker, S., Cheridito, P., Jentzen, A., & Neufeld, A. Deep learning based numerical approximation algorithms for stochastic partial differential equations and high-dimensional nonlinear filtering problems. arXiv preprint arXiv:2012.01194.

[2] Gustafsson, F. K., Danelljan, M., Bhat, G., & Schön, T. B. Energy-based models for deep probabilistic regression. In European Conference on Computer Vision (pp. 325-343). Springer, Cham.

Carlsson, Oscar
Chalmers

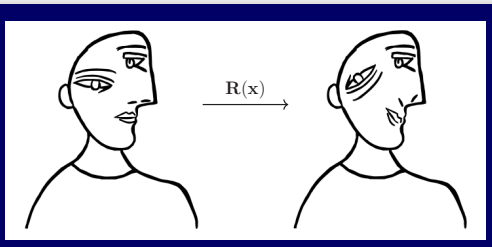
Geometric Deep Learning and Equivariant Neural Networks

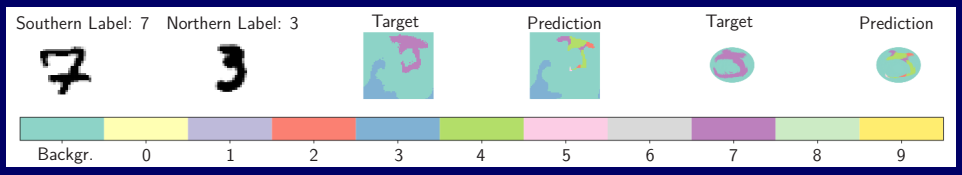
When constructing a convolutional network for image analysis one cannot truly escape the risk that real world data will not respect the the orientation of data the model was trained on. For example, satellite images have, by their nature, no preferred orientation. How can one deal with this problem in an easy way? One solution is to use explicitly equivariant convolutions. This poster discusses some points for why the equivariant convolutions are needed and discusses an implementation made by Cohen et al. in 2016 as well as presents a visualisation of its effect on a network. It also discusses some parts of the mathematical structure as well as current and future work.

Carlsson, Oscar
Chalmers

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Use the **symmetries** in your problems to your advantage.





Geometric Deep Learning and Equivariant Neural Networks

Oscar Carlsson[†⋆] Daniel Persson[†] Jimmy Aronsson[†⋆] Fredrik Ohlsson[•]
Jan Gerken[†] Christoffer Petersson[◇] Hampus Linander[◇]

(**Affiliations** †: Chalmers Univerity of Technology, Department of Mathematical Sciences. ⋆: WASP. •: Umeå University, Department of Mathematics and Mathematical Statistics. ◇: Zenseact.)

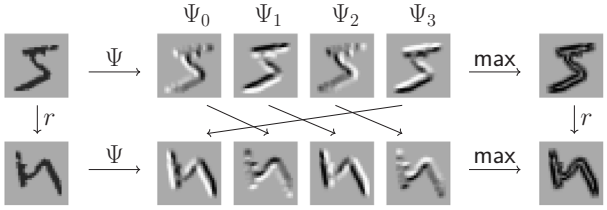
Intro

How does one deal with rotations? Options are:

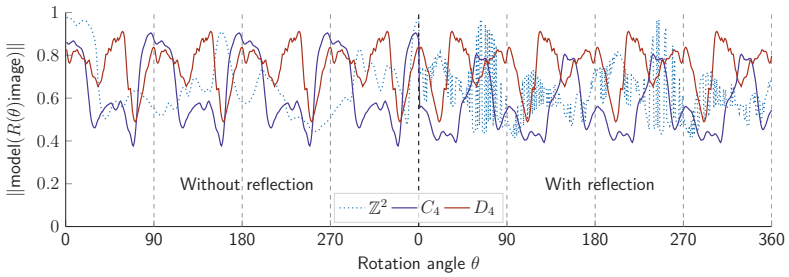
- You don't, you assume all your data will have your prefered orientation. (Dangerous: real life throws curveballs at your models)
- Augement training data so that everything is represented. (Every orientation becomes a **lot** of data to deal with)
- Make sure that all your data has the right orientation. (A lot of work, either manually or making an algorithm to unrotate data)
- Modify your architecture and layers to deal with the rotation automatically. (The easy way.)

(Bullet image source: "rotation" by Adrien Coquet from the Noun Project)

One way: transform kernels [Cohen and Welling 2016]



Example: Magnitude of classification invariance for four fold rotation symmetry applied to single MNIST digit



Some mathematics

A map Φ is equivariant with respect to a transformation T of the data if it doesn't matter if one transforms the data before or after one applies the map Φ:

$$\Phi \circ T = T \circ \Phi. \tag{1}$$

An example is that normal convolutions are equivariant to translation. One can extend this to a larger equivariance if we allow convolution kernels and data to be functions on a group:

$$[\Psi \star f](g) = \int_G \Psi(g^{-1}g')f(g')\,dg'. \tag{2}$$

This is equivariant if the kernel transforms in a special way under the group action:

$$\Psi(hgh') = \rho_2(h)\Psi(g)\rho_1(h'). \tag{3}$$

[Cohen and Welling 2016] discretise this to allow for easy computation, see figure on the left.

This can be generalised to local transformations by taking a viewpoint of local coordinates changes. Equivariant convolutions in this general context was introduced by [Cheng et al. 2019] without much mathematical details. We expand on the details of this and provide some generalisation in our recent article [Gerken et al. 2021].

Current work

We're currently examining the details of how imposed equivariance affects semantic segmentation and how it compares to data augmentation.

References

Cheng, Miranda C. N. et al. (June 6, 2019). *Covariance in Physics and Convolutional Neural Networks*. arXiv: 1906.02481 [hep-th, stat]. URL: http://arxiv.org/abs/1906.02481 (visited on 01/27/2020).

Cohen, Taco S. and Max Welling (June 3, 2016). *Group Equivariant Convolutional Networks*. arXiv: 1602.07576 [cs, stat]. URL: http://arxiv.org/abs/1602.07576 (visited on 11/07/2019).

Gerken, Jan E. et al. (May 28, 2021). *Geometric Deep Learning and Equivariant Neural Networks*. arXiv: 2105.13926 [hep-th]. URL: http://arxiv.org/abs/2105.13926 (visited on 05/31/2021).

← Download [Gerken, Jan E. et al. 2021]

WASP GÖTEBORGS UNIVERSITET

Dadras, Ali
Umeå university

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Solving stochastic/deterministic constrained optimization problems in statistical learning.

Modern learning machines, such as deep neural networks, are often over-parametrized and tuned to perfectly interpolate the training data. Recent works have shown that first-order methods could converge fast in non-convex optimization problems such as overparameterized neural networks, satisfying certain interpolation conditions (e.g., zero training loss). We seek to investigate and understand the convergence of first-order methods in non-convex optimization problems with deterministic or stochastic constraints.

Dadras, Ali
Umeå university

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Solving stochastic/deterministic constrained optimization problems in statistical learning

Ali Dadras, Umeå University
Department of Mathematics and Mathematical Statistics

Abstract

Modern learning machines, such as deep neural networks, are often over-parametrized and tuned to perfectly interpolate the training data. Recent works have shown that first-order methods could converge fast in non-convex optimization problems such as overparameterized neural networks, satisfying certain interpolation conditions (e.g., zero training loss). We seek to investigate and understand the convergence of first-order methods in non-convex optimization problems with deterministic or stochastic constraints.

Problem Statement

Let $\{x_i\}_{i=1}^n$ be a given training set, and let $\{y_i\}_{i=1}^n$ be training labels. We would like to minimize

$$\min_{\theta \in D} f(\theta) = \sum_{i=1}^{n} f_i(\theta;\, x_i, y_i)$$

where θ is the model parameter and D is a set of stochastic or deterministic constraints .

Methods

Considering deterministic constraints, a vast number of studies have been done to solve the above optimization problem. There are different solving strategies, many of them rely on gradient descent and its variants. To improve these gradient-based methods, different strategies are proposed.

- Preconditioning (e.g.,data normalization, layer and batch normalization)
- Mmomentum (e.g., Polyak and Nestrov)
- Variance reduction (e.g., SAG, SDCA, SVRG)
- Adaptive stepsizes (e.g., Adagrad, ADAM)
- Importance sampling

Objectives

- Investigating the potential of first-order methods in optimizing non-convex optimization problems with deterministic or stochastic constraints.
- Investigating the existence of first order methods for solving constrained optimization problems motivated by learning problems.
- Investigating and understanding the convergence of desired first order methods.

References

1. Meng, Si Yi, et al. "Fast and furious convergence: Stochastic second order methods under interpolation." International Conference on Artificial Intelligence and Statistics. PMLR, 2020.
2. Loizou, Nicolas, et al. "Stochastic polyak step-size for SGD: An adaptive learning rate for fast convergence." arXiv preprint arXiv:2002.10542 (2020).
3. Vaswani, Sharan, et al. "Adaptive Gradient Methods Converge Faster with Over-Parameterization (and you can do a line-search)." arXiv preprint arXiv:2006.06835 (2020).

Deligeorgaki, Danai
KTH

Gorenstein discrete decomposable models

Discrete hierarchical models are statistical models that are widely used throughout statistics and data science. An advantage of these models is that there are established methods that can be used to make inference.
The goal of this project is to explore at a deeper level the combinatorial objects arising from discrete decomposable models beyond their graph. Specifically, we aim to answer when enumerative properties, such as the Gorenstein property, hold for the polytope associated to a discrete decomposable model.

Deligeorgaki, Danai
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Gorenstein discrete decomposable models

Danai Deligeorgaki, Department of Mathematics (KTH)
Supervisor: Liam Solus (KTH)

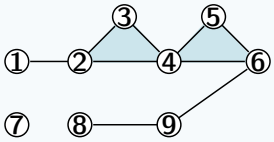
Introduction

We study discrete decomposable models, a family of statistical models that lie in the class of **hierarchical models**. Decomposable models and their corresponding graphs are of wide use throughout **statistics** and data science. For instance, directed acyclic graphs (**DAGs**) can be approximated by decomposable graphs. The complexity of this approximation determines the complexity of probabilistic inference algorithms for DAG models such as **variable elimination**. Therefore, the combinatorics of the graphs defining the decomposable models carry important information in regard to **probabilistic inference**. The goal of this project is to explore at a deeper level the information encoded in combinatorial objects associated to **decomposable models**.

Definition

A **decomposable simplicial complex** Γ is a collection of simplices, i.e. nodes, edges, triangles, tetrahedra, etc., that are glued together (in a certain way). The simplices in Γ are called **faces** and the (non-trivial) inclusion-maximal faces are called **facets**.
For example, the graph on the right denotes a decomposable simplicial complex on 9 nodes. The edge {1,2}, the triangle {2,3,4} and the node {7} are some of its facets.

decomposable simplicial complex



Discrete decomposable models

Let $r_1, ..., r_m \in \mathbb{N}$ be the number outcomes of the discrete variables $X_1, X_2, ..., X_m$, respectively, and let $\mathcal{R} = r_1 \times \cdots \times r_m$ be the set of all possible outcomes. The joint distribution of $X_1, ..., X_m$ lies in the $(\#\mathcal{R} - 1)$-dimensional **probability simplex**

$$\Delta_{\#\mathcal{R}-1} = \{p \in \mathbb{R}^{\#\mathcal{R}} : p_i \geq 0, \text{ for all } i \in \mathcal{R} \text{ and } \sum_{i \in \mathcal{R}} p_i = 1\}.$$

The **decomposable model** associated with a decomposable simplicial complex Γ is

$$M_\Gamma = \{p \in \Delta_{\#\mathcal{R}-1} : p_i = \frac{1}{Z(\theta)} \prod_{F \in \text{facet}(\Gamma)} \theta_{i_F}^{(F)} \text{ for all } i \in \mathcal{R}\},$$

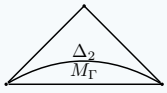
for $\theta_{i_F}^{(F)}$ positive parameters and $Z(\theta)$ normalizing constant.

From the model to the polytope

Apart from the graph Γ, there are other combinatorial objects linked to a decomposable model M_Γ. In fact, M_Γ can be written as the intersection of a toric variety V_{M_Γ} with the probability simplex $\Delta_{\#\mathcal{R}-1}$.

For example, for $\#\mathcal{R} = 3$,

$$M_\Gamma = V_{M_\Gamma} \cap \Delta_2$$



From the toric variety, which is an algebro-geometric object, we can pass to a polytope P_{M_Γ}, a geometric object. It is a property of toric varieties that the geometric properties of V_{M_Γ} are encoded in the **polytope** P_{M_Γ}.

In this project, we are investigating the structure of this polytope to see if it carries useful information in relation to probabilistic inference.

Getting to know the polytope

When investigating a polytope's combinatorics, there are several questions to be explored, such as

- What are the facets of the polytope P_{M_Γ}? Answered in [1].

  The facets are given by $x_{i_F}^F \geq 0$ for $F \in$ facets(Γ) and $i_F \in \mathcal{R}_F$.

- Does P_{M_Γ} admit a regular unimodular triangulation? Answered in [2].

  Yes!

  What combinatorial information does this triangulation carry? Open.

- What are the enumerative properties of P_{M_Γ}? Our results.

To this end, we study the structure of an integer polynomial associated to P_{M_Γ}, called the h^***-polynomial**,

$$h^*(x) = h_0^* + h_1^* x + \cdots + h_{\#\mathcal{R}-1}^* x^{\#\mathcal{R}-1}.$$

This polynomial captures important information about the polytope, including its volume and whether or not the polytope P_{M_Γ}, and hence the model M_Γ, has the Gorenstein property. In fact, the decomposable model M_Γ is **Gorenstein** if and only if $h^*(x)$ is palindromic.
We characterize all Gorenstein discrete decomposable binary models for forests Γ.

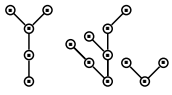
Theorem 1
Let Γ be a forest on m nodes and $X_1, ..., X_m$ be binary variables.
Then M_Γ is Gorenstein if and only if all connected components of Γ have

- exactly one vertex,
- strictly more than one vertex.

Gorenstein examples

Γ_1

Γ_2



Future work

We will continue exploring the combinatorial properties of discrete decomposable models, and their interpretation in terms of statistics. Our current goals are to

1. Interpret the observations in Theorem 1 statistically.
2. Generalize Theorem 1 to characterize all discrete decomposable models. We already have a conjecture in this direction.
3. Analyze the information that the triangulation constructed in [2] carries.

References

[1] Markov bases of binary graph models
M. Develin, S. Sullivant
Annals of Combinatorics 7, 2003

[2] Gröbner bases and polyhedral geometry of reducible and cyclic models
S. Hoşten, S. Sullivant
Journal of Combinatorial Theory, Series A 100.2, 2002

Ekström, Henrik
Lund University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Deducing function from structure

Neuroscientists are working hard to map and understand the intricate connections of neurons in a brain. What will the knowledge of that structure give us? Using which mathematical framework can we in a useful yet practical way describe the (supposed) link between the structure of a network and the tasks it can perform? To find a rigorous answer, we study the impact that different structures and dynamics can have on networks. The aim is to combine pure mathematics and neuroscience, using methods from statistical physics, combinatorics, geometry, percolation and probability theory.

Even endowing a simple structure with simple dynamics can yield surprisingly intricate results. We now study emerging structures in the Hopfield model as well as cellular automata containing inhibitory and excitatory 'neurons'. The latter can be thought of as a generalisation of bootstrap percolation with highly non-monotone behaviour!

Ekström, Henrik
Lund University

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Neuronal networks: connecting structure and function

Searching for mathematics that can help us understand the functioning of the brain.

Cellular Automata

Neuronal networks can be thought of as cellular automata:
Small units with simple local dynamics giving rise to complex global phenomena.

In neural networks, some neurons are inhibitory, having a dampening effect. Combining this feature with CA gives an interesting class of complex dynamical systems with non-monotone behaviour.

Limiting states and the set of initial states leading to them.
↕
'Patterns' and their 'basins of attraction'.

How many patterns are there and how are their basins distributed? This is a surprisingly involved question!

The model we study

- Consider $\mathbb{Z}^2$ where one in four vertices is inhibitory as in the figures.
- Each vertex is connected to its four nearest neighbours.
- Let $A(t=0) \subset \{0,1,\dots,n\}^2 \subset \mathbb{Z}^2$ be a set of 'active' vertices (given n).
- Let each vertex in $\mathbb{Z}^2$ become active at time $(t+1)$ if, at time t,
  $|\{\text{adjacent } \bigcirc\}| > |\{\text{adjacent } \ominus\}|$

Question: How does the size of $A(t) \subset \mathbb{Z}^2$ as t increases vary for different $A(0)$? If every vertex is active with probability p?
Answer: In a *highly* non-trivial way!
We simplify by considering the processes 'B' and 'C' separately (Figure 1), but we still only understand the behaviour in certain regimes. The 'B' case is elaborated here.

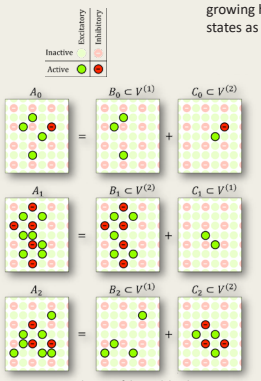


Figure 1. An initialisation of the model with active set A_0 and two evolution steps. Considering the subsets B_0, C_0 separately yields the same total evolution.

Some regimes

Let all vertices within $\{0,1,\dots,n\}^2$ independently be initially active, in $A(0)$, with probability p.

Dense activation
- When p is high, inhibitory vertices will quickly 'shut down' the activation.
- What remains is determined by the boundary between complete and incomplete activation.
- If p is large enough, this leads to the sparse case.

Sparse activation
- A single active vertex spreads activation along a line as in the top of Figure 2.
- This locally repeats with period four.
- All 7 limiting cycles resulting from only two initially active vertices are shown in Figure [right,#]
- For low p there are only pairwise interactions.
- The bottom pattern ('T-junction') is by far the most common interaction
- The dynamics (now monotone!) can be thought of as growing horizontal/vertical lines, and the limiting states as quadrangulations of space.

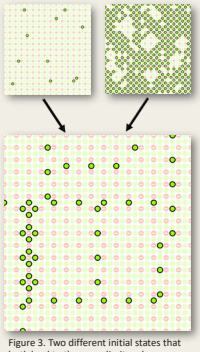


Figure 3. Two different initial states that both lead to the same limit cycle

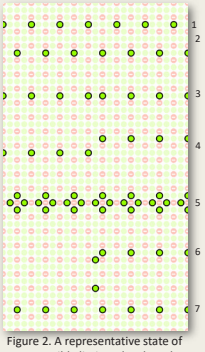


Figure 2. A representative state of evey possible limit cycle when there are two initially active vertices. All cycles have a period of four.

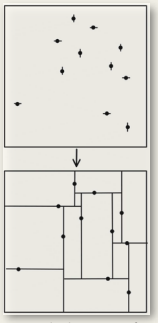


Figure 4. The alternate view of the sparse case, now monotone increasing! Activation in original model is proportional to the total length of lines in guadrangulation.

The Hopfield model for auto-associative memory

Constructing the original Hopfield model

Meta-input: M 'patterns' $\boldsymbol{\xi}^1, \dots, \boldsymbol{\xi}^M$, where $\boldsymbol{\xi}^\mu \in \{-1,1\}^N$
Structure: Originally, a complete graph on N vertices.
'Training': Set the weight $i \leftrightarrow j$ in the graph to $W_{ij} = \frac{1}{N}\sum_{\mu=1}^{M} \xi_i^\mu \xi_j^\mu$.
States: The configuration $\boldsymbol{\sigma}(t) \in \{-1,1\}^N$ of 'spins' at time t, with dynamics

$$\sigma_i(t+1) = \operatorname{sign}\left(\sum_{j=1}^{N} \sigma_i(t) W_{ij}\right).$$

The Hopfield model (ideally) maps input states $\boldsymbol{\sigma}(0)$ to the nearest pattern $\boldsymbol{\xi}^\mu$, which are by construction the minima of the Hopfield Hamiltonian:

$$H_N(\sigma) = -\sum_{i,j=1}^{N} W_{ij}\sigma_i\sigma_j\,.$$

('Spurious' minima typically also appear.)

Dissecting the model

Some (even many) weights W_{ij} can become quite small and contribute little. Randomly removing weights has been shown to preserve pattern retrievability surprisingly well. What if we remove the *smallest* weights first?

- Study the structure of the remaining graph
- What is the connection between the data $\{\xi^\mu\}$ and the structure?
- Are certain graphs more 'efficient' than others, i.e. large M but few non-zero W_{ij}?
- Can the bounds of random dilution be improved with this method?

Classifying models

Impose a structure on the graph: cycle/tree/path with various degrees

- Study the minima of the energy function and retrievable patterns.
- What is the largest set of patterns?
- Are they 'close' by some metric?

The goal: relate the learned patterns with the model structure

1. Given a graph, can we predict what kind of patterns it can learn?
2. Find (unique?) minimal graph capable of learning given patterns.
3. Identify classes of patterns that are suited for classes of graphs.

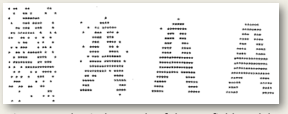


Figure 5. A classical example of the Hopfield model mapping a noisy input state (left) to a fix state corresponding to a learned pattern (right). Each pixel represents the state of $\sigma_i(t)$, the weights are not shown.
(W.Kinzel)

Henrik Ekström | Lund University | Matematisk statistik, Sölvegatan 18, 223 62 Lund | Henrik.Ekstrom@matstat.lu.se

Isaksson, Martin
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Adaptive Expert Models for Federated Learning

Federated Learning is a promising framework for distributed learning when data is private and sensitive, but not optimal when data is heterogeneous and non-IID. We propose a robust approach to personalization in FL that adjusts to heterogeneous data and non-IID distributions using a Mixture Of Experts. We evaluated our method on three datasets representing different non-IID settings, and found that our proposed approach achieve superior performance with two of the datasets, and is robust in the third. Even though we tune our algorithm and hyper-parameters in the IID setting, it still generalizes well in non-IID settings.

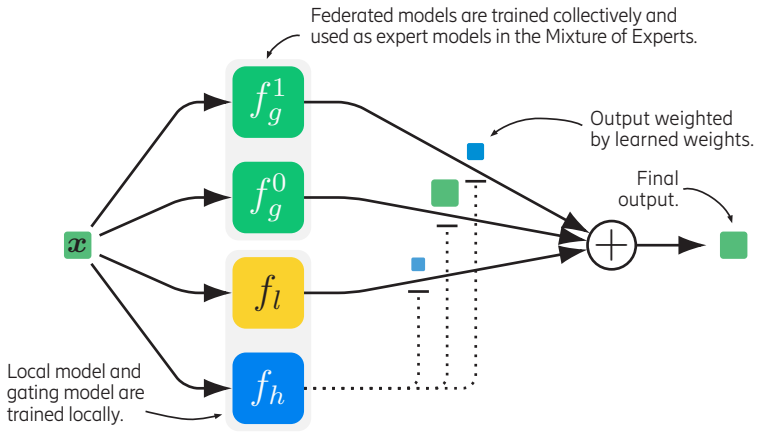
Isaksson, Martin
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

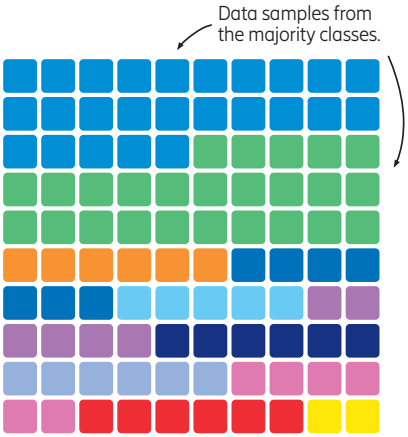
ericsson.com/research

Adaptive Expert Models for Federated Learning

Federated Learning [1] is a promising framework for distributed learning when data is private and sensitive, but not optimal when data is heterogeneous and non-IID. We propose a robust approach to personalization in FL that adjusts to heterogeneous data and non-IID distributions using a Mixture Of Experts. We evaluated our method on three datasets representing different non-IID settings, and found that our proposed approach achieve superior performance with two of the datasets, and is robust in the third. Even though we tune our algorithm and hyper-parameters in the IID setting, it still generalizes well in non-IID settings.



A client has one local expert model and share expert cluster models with other clients. A gating model is used to weight the expert cluster models to produce a personalized inference.



Non-IID class sampling allows us to vary non-IID-ness. With p = 0.5, two classes make up half of the samples.

Our contributions

1. We improve the clustering algorithm from [2] by weighing exploration and exploitation to produce better cluster models;
2. We use said cluster models as expert models, improving [3];
3. An extensive analysis of our approach with respect to different non-IID aspects that also considers the distribution of client performance.

References

[1] B. McMahan, et al., "Communication-efficient learning of deep networks from decentralized data," in AISTATS, 2017
[2] Ghosh, A. et al. "An Efficient Framework for Clustered Federated Learning", NeurIPS, 2020
[3] Listo Zec, E. et al. "Federated learning using a mixture of experts", ArXiv, 2020

Martin Isaksson, Edvin Listo Zec, Rickard Cöster, Daniel Gillblad, Sarunas Girdzijauskas

KTH RI.SE WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM AI SWEDEN CHALMERS UNIVERSITY OF TECHNOLOGY

Jal, Aryaman
KTH

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Polyhedral geometry of Wasserstein distances

Every discrete probability distribution corresponds to a point in the standard simplex. Given a model consisting of probability distributions and sample data, we want to find a candidate in the model that best explains the data. Studying the Wasserstein distance between probability distributions is one route to this. The approach we take is to use polyhedral geometry - in particular bisectors and bisection fans - to better understand the Wasserstein distance.

Jal, Aryaman
KTH

Polyhedral geometry of Wasserstein distances

Aryaman Jal Katharina Jochemko
Department of Mathematics, Royal Institute of Technology (KTH)

Motivation

Every probability distribution on $[n] = \{1, 2, \dots, n\}$ corresponds to a point $\mu \in \Delta_{n-1} = \{x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = 1, x_i \geq 0 \; \forall i = 1, \dots, n\}$.

Given a finite model $\mathcal{M} = \{\mu_1, \dots, \mu_k\} \subseteq \Delta_{n-1}$ of discrete probability distributions and samples $x_1, \dots, x_N \in [n]$, we want to find $\mu_i \in \mathcal{M}$ that best explains the samples. We follow the approach in [1] and use polyhedral geometry to minimise the Wasserstein distance between the empirical distribution and the model.

Wasserstein distance

Given a metric d on $[n]$, the Wasserstein distance is defined by

$$W_d(\mu, \nu) = \operatorname{dist}(\mu, \nu) = \min\{\alpha \in \mathbb{R}_{\geq 0} : \nu - \mu \in \alpha B_d\}$$

where

$$B_d = \operatorname{conv}\left\{\frac{1}{d_{ij}}(e_i - e_j) : 1 \leq i \neq j \leq n\right\}$$

is called the Wasserstein unit ball with respect to W_d. Given $x_1, \dots, x_N$, the Wasserstein distance estimator equals

$$\min_{i=1,\dots,k} W_d\left(\frac{1}{N}\sum_{j=1}^{N} \delta_{x_j}, \mu_i\right)$$

Bisectors

If $N = 2$, the decision boundary is given by the bisector

$$\operatorname{bis}(\mu, \nu) = \{x \in \mathbb{R}^n : \operatorname{dist}(\mu, x) = \operatorname{dist}(\nu, x)\}.$$

Proposition. [2] The bisector $\operatorname{bis}(\mu, \nu)$ is a polyhedral complex.

Remark.

1. Polyhedrality of the bisectors is true for all norms.
2. The bisector can have non-empty interior (see Figure 3). This happens if and only if μ and ν lie on a hyperplane parallel to a facet of B_d.

Euclidean and Wasserstein bisectors

Figure 1 shows the bisector of two points with respect to the Euclidean distance. In Figures 2 and 3, the bisector for W_d for $n = 3$ (projected down to $\mathbf{1}^{\perp}$) and $d_{ij} = 1$, for all $i \neq j$, are depicted.

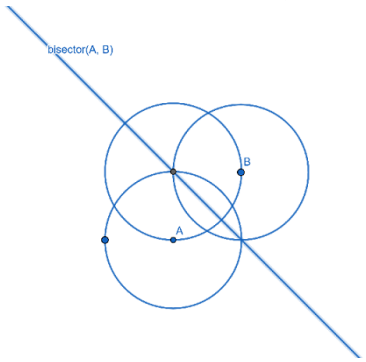


Figure 1: Bisector w.r.t. Euclidean distance

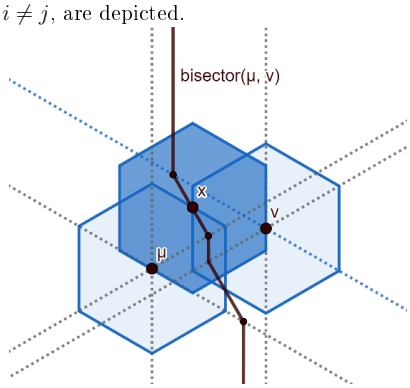


Figure 2: One-dimensional bisector w.r.t. W_d

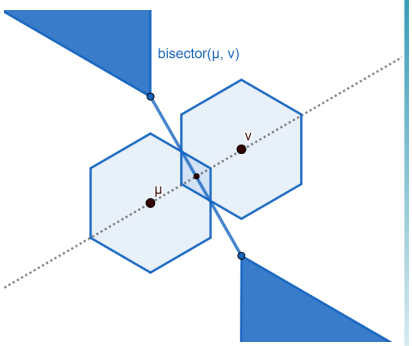


Figure 3: Two-dimensional bisector w.r.t. W_d

Goals

1. Investigating the combinatorics of $\operatorname{bis}(\mu, \nu)$ when moving ν. This leads to the concept of a bisection fan (see [2]).
2. Determining the number of maximal cells in $\operatorname{bis}(\mu, \nu)$ as a measure of complexity of the decision boundary.
3. Determining Voronoi diagrams with respect to Wasserstein distances.

References

[1] Çelik, T.Ö., Jamneshan, A., Montúfar, G., Sturmfels, B. and Venturello, L., 2021. Wasserstein distance to independence models. Journal of Symbolic Computation, 104, pp.855-873.

[2] Criado, F., Joswig, M. and Santos, F., 2019. Tropical bisectors and Voronoi diagrams. arXiv preprint arXiv:1906.10950.

[3] Higashitani, A., Jochemko, K. and Michaek, M., 2019. Arithmetic aspects of symmetric edge polytopes. Mathematika, 65(3), pp.763-784.

Partial Results

For every graph $G = ([n], E)$, define $d_{ij} = 1$ if $ij \in E$ and ∞ otherwise. In this case, B_d is a symmetric edge polytope ([3]). If G is a tree, then W_d equals the 1-norm up to an affine transformation and B_d is affinely isomorphic to the cross-polytope

$$\diamondsuit_{n-1} = \operatorname{conv}\{\pm e_i : i \in [n-1]\}.$$

Proposition (Jal, Jochemko 2021+)**.** The bisection fan of W_d corresponding to the case of d being a graphical metric on a tree is, up to affine transformation, induced by the hyperplane arrangement

$$\mathcal{H} = \bigcup_{i=1}^{n-1}\{x_i = 0\} \bigcup_{I \subseteq [n-1]} \left\{\sum_{i \in I} x_i = \sum_{i \in I^c} x_i\right\}$$

A similar, more involved result was obtained in the case of G equal to the complete graph K_n.

Jansson, Erik
Chalmers

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

ResNets Understood as Sub-Riemannian Landmark Matching

Residual neural networks can be interpreted as time discretizations of optimal control problems. This observation means that it is possible to use sub-Riemannian landmark matching, a method from the field of shape analysis, to study and understand ResNets. For instance, as demonstrated in the poster, the impact of regularization on the smoothness of transformations can be studied from a diffeomorphic point of view. The connection between the ResNets and sub-Riemannian landmark matching demonstrates that it is possible to study and understand neural networks using shape analysis methods.

Jansson, Erik
Chalmers

ResNets understood as Sub-Riemannian Landmark matching

Erik Jansson, Chalmers and University of Gothenburg
Supervisors: Klas Modin (Chalmers and GU)

sub-Riemannian landmark matching

1. Problem: find a diffeomorphism, determined by a vector field, warping *initial landmarks* $x_1, ..., x_n$ on a *manifold* M to targets $c_1, ..., c_n$ on a metric space N.
2. ***Idea: Consider only vector fields parametrized by variables in some Hilbert space*** $\mathcal{U}$.

This is an optimal control problem.

$$\min_{u:[0,1]\to\mathcal{U}} \sum_{i=1}^{m} d_N^2(h(y_i(1), c_i)) + \int_0^1 \ell(F(u))\mathrm{d}t,$$
$$\text{s.t.} \quad \dot{y}_i = F(u)(y_i),\ t \in [0,1],\ y_i(0) = x_i.$$

Note that F maps from controls to vector fields and formally parametrizes the vector fields!

We derive an *equaton of motion* describing the ***evolution of the controls.***

$$A(u)\dot{u} = (D_u^T m \circ L^{-1}) \cdot \left(\mathrm{ad}_v^T m + \mathrm{div}(v)m\right)$$
$$Lv = (1 - \alpha\Delta)^k v = m$$

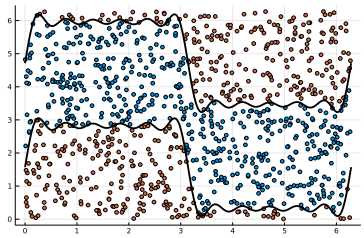
Connection with residual neural networks

Residual neural networks can be considered as *time discretizations* of time-continuous control problems.

In some settings: ResNets discretize a high dimensional sub-Riemannian landmark matching problem!

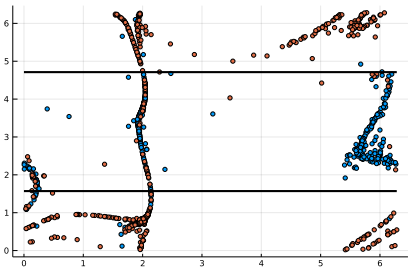
Interpretation 1: If we know the dynamics of the control, increasing the number of layers does not increase the number of parameters we need to optimize.
Interpretation 2: Shape analysis, a mathematically well understood subject, can be used to interpret neural networks, *and vice-versa*.

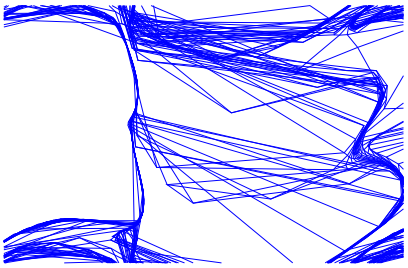
Numerical example

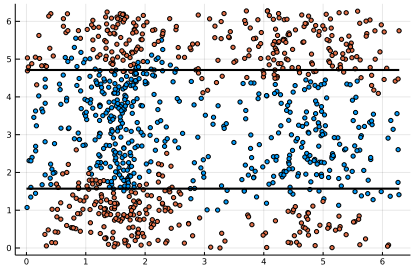


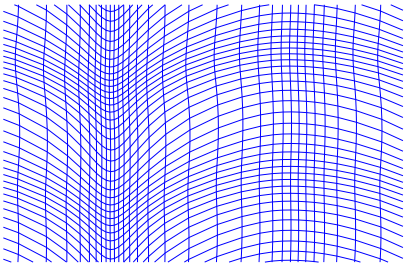
We let $\mathcal{U} = \mathbb{R}^{16}$ and set $F(u) = X^0 + \sum_{i=1}^{m} u_i X^i$ for vector fields on $\mathbb{R}^2/(2\pi\mathbb{Z})^2$ that diagonalizes the Laplace-Beltrami operator. We generate 1000 points and classify those in a variable band of width π as a 1 and those outside as a 0. We want to determine an initial value $u(0)$ so that the points are moved towards a uniform band.

We first try without regularization. Note that the points are moved very erratically. This is not computationally stable.









We increase the regularity sharply. Note that the resulting warp is much smoother. This transformation is more robust.

Lindbäck, Jacob
KTH

A splitting-based algorithm for faster and more accurate optimal transport

We have explored splitting methods for solving large-scale optimal transport (OT) problems, which has resulted in an algorithm that combines speed and accuracy. Built on the celebrated Douglas-Rachford splitting technique, our method tackles the original OT problem directly instead of solving an approximate regularized problem, as many state-of-the-art techniques do. This allows us to provide sparse transport plans and avoid numerical issues of methods that use entropic regularization. Each iteration can be executed efficiently, in parallel, and the proposed method enjoys an iteration complexity $O(1/\epsilon)$ compared to the best-known $O(1/\epsilon 2)$ of the Sinkhorn method. In addition, we establish a linear convergence rate for our formulation of the OT problem.

Lindbäck, Jacob

KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

A splitting-based algorithm for faster and more accurate optimal transport

V. V. Mai, J. Lindbäck, M. Johansson
EECS KTH

KTH VETENSKAP OCH KONST

Our Contributions

We propose a splitting algorithm for optimal transport, that achieves better iteration complexity than the state-of-the-art while preserving memory efficiency. Moreover, it:

- achieves high numerical stability,
- gives sparse solutions,
- is hyperparameter free,
- is readily parallelizable on multi-core CPU or GPU!

Some background

The Discrete Optimal Transport (D-OT) problem follows as:

$$\begin{aligned} \underset{X \geq 0}{\text{minimize}} \quad & \langle C, X \rangle \\ & X\mathbf{1}_n = p, \ , X^\top \mathbf{1}_m = q, \end{aligned}$$

for some fixed cost matrix C. The decision variable X is an $m \times n$ matrix (typically large). That, together with the dense constraints, render standard LP solves, such as simplex or IP-methods, intractable for larger OT problems.

For improved memory efficieny, the Sinkhorn method has been proposed, which is the most popular OT-solver. It finds an approximate solution to the D-OT, by replacing the non-negativity constraint by adding an entropic regularization to the objective. That is, the objective is updated as follows:

$$\langle C, X \rangle \to \langle C, X \rangle - \eta H(X)$$

where $H(X) = -\sum_{ij} X_{ij} \log(X_{ij})$ is the Entropy function, which promotes positivity. This approximate version of the D-OT problem, enables simple dual and primal variable updates, which finds ϵ-accurate solutions in $O(1/\epsilon^2)$ iterations. Moreover, the updates only involve matrix-vector multiplies and element-wise arithmetic operations, which can easily be parallelized. Yet, a consequence of the entropic regularization is that the solution will always be dense, and it requires tuning. Moreover, it induces a trade-off between accuracy and numerical stability.

References

A Fast and Accurate Splitting Method for Optimal Transport: Analysis and Implementation
Vien V. Mai, Jacob Lindbäck, Mikael Johansson, Under revision to ICLR 2021, available at arxiv: https://arxiv.org/abs/2110.11738

The splitting method

It turns out that the celebrated Douglas-Rachford (DR) splitting method gives rise to a fast algorithm in terms of iteration complexity, while keeping the memory footprint low, without having to introduce entropic regularization. This makes it both fast, accurate, and stale! When applied to the D-OT problem, the DR-splitting algorithm reads:

$$X_{k+1} = [Y_k - \rho C]_+, \ Z_{k+1} = P_{\mathcal{X}}(2X_k - Y_k), \ Y_{k+1} = Y_k + Z_{k+1} - X_{k+1}.$$

where ρ is a stepsize, $P_{\mathcal{X}}$ denotes the projection onto the set $\mathcal{X}$, which is given by:

$$\mathcal{X} := \left\{ X \in \mathbf{R}^{m \times n} \,\middle|\, Xe = p \text{ and } X^\top f = q \right\}.$$

This projection admits a closed formula solution, that is linear, and only involved matrix-vector multiplication and rank-one updates, which are hence easy to parallelize. This can further be used to eliminate the Z and Y variable blocks completely, yielding the memory-optimized algorithm:

Algorithm 1 Douglas-Rachford Splitting for Optimal Transport (DROT)

Input: OT(C, p, q), initial point X_0, penalty parameter ρ

$\phi_0 = 0, \varphi_0 = 0$
$a_0 = X_0 e - p, b_0 = X_0^\top f - q, \alpha_0 = f^\top a_0/(m+n)$
for $k = 0, 1, 2, \ldots$ **do**
  $X_{k+1} = \left[X_k + \phi_k e^\top + f\varphi_k^\top - \rho C\right]_+$
  $r_{k+1} = X_{k+1}e - p, s_{k+1} = X_{k+1}^\top f - q, \beta_{k+1} = f^\top r_{k+1}/(m+n)$
  $\phi_{k+1} = (a_k - 2r_{k+1} + (2\beta_{k+1} - \alpha_k)f)/n$
  $\varphi_{k+1} = (b_k - 2s_{k+1} + (2\beta_{k+1} - \alpha_k)e)/m$
  $a_{k+1} = a_k - r_{k+1}, b_{k+1} = b_k - s_{k+1}, \alpha_{k+1} = \alpha_k - \beta_{k+1}$
end for

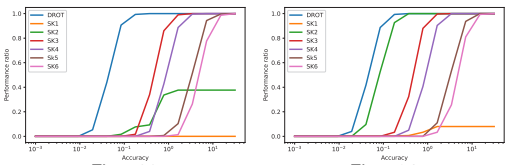
Output: X_K

Theoretical guarantees

We establish a sublinear rate $O(1/\epsilon)$ and a linear rate $O(\log 1/\epsilon)$ for our splitting method, both of which are indeed better than that of Sinkhorn! Although the linear rate is asymptotically stronger than the sublinear rate, we found in numerical experiments that the sublinear rate typically dominates the first iterations.

Selected Results

We generated 300 random image pairs from MNIST and computed how many times our splitting algorithm, as well as Sinkhorn using different neutralizers, manage to find an optimal transportation plan within different target accuracies. Note not our approach is consistently more accurate and robust.



Float 32 Float 64

Maskan, Hoomaan
Umeå University

Accelerated Deterministic Methods in Optimization

Optimization problems play an important role in the process of learning a machine using previously available data. This process, can be time consuming and therefore many researchers have tried to reduce it through various techniques. One method to attack this problem is to reduce the optimization time of the learning process. As a result, accelerated methods in optimization gain a remarkable attention. Among the first order algorithms for smooth convex functions, Nesterov's accelerated gradient descent(NAG) is proven to be the fastest. For decades, various studies tried to enlighten the essence of acceleration through Nesterov updates. Recently, using Ordinary Differential Equations(ODE), it is shown that for a fixed convergence rate, accelerated algorithms may not be unique. This research, proposes a general algorithm which can achieve various convergence rates for different choices of parameters.

Maskan, Hoomaan
Umeå University

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Accelerated Deterministic Methods in Optimization

Hoomaan Maskan, Umeå University
Department of Mathematics and Mathematical Statistics
Main Advisor: Armin Eftekhari

UMEÅ UNIVERSITY

Abstract

Optimization problems play an important role in the process of learning a machine using previously available data. This process, can be time consuming and therefore many researchers have tried to reduce it through various techniques. One method to attack this problem is to reduce the optimization time of the learning process. As a result, accelerated methods in optimization gain a remarkable attention. Among the first order algorithms for **smooth convex functions, Nesterov's accelerated gradient descent(NAG)** is proven to be the fastest. For decades, various studies tried to enlighten the essence of acceleration through Nesterov updates. Recently, using **Ordinary Differential Equations(ODE)**, it is shown that for a fixed convergence rate, accelerated algorithms may not be unique. This research, proposes a general algorithm which can achieve various convergence rates for different choices of parameters.

Motivation and Methods

The learning problem can be formulized as

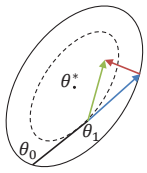
$$\min_{\theta\in\mathbb{R}^{n\times n}} \frac{1}{N}\sum_i \mathcal{L}(\boldsymbol{y_i}, f(\boldsymbol{x_i},\boldsymbol{\theta}))$$

which is known as **Empirical Risk Minimization** (ERM) problem. Depending on the features of f and $\mathcal{L}$, this problem can be non-linear and non-convex. Therefore, if not impossible, it would be so hard to find the global minimizer(s) of this problem. For simplicity, from now on we consider the objective function to be smooth and μ-strongly convex and denote it as $F_{\mathcal{L}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta})$.

NAG updates for $\min_{\theta\in\mathbb{R}^{n\times n}} F_{\mathcal{L}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta})$ are[]

$$\begin{cases}\boldsymbol{v}_{k+1} = \gamma\boldsymbol{v}_k + h\nabla F_{\mathcal{L}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta_k} - \gamma\boldsymbol{v}_k)\\ \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \boldsymbol{v}_{k+1}\end{cases},$$

which gets the best of the momentum term $\boldsymbol{v}_k$ and calculates the gradient near the future point $\boldsymbol{\theta}_{k+1}$(see Figure).



Shi, et. al. proposed high-resolution ODEs for modeling acceleration methods [1]. Specifically, if $\boldsymbol{\vartheta}$ denotes the continuous trajectory of the NAG, then

$$\begin{cases}\dot{\boldsymbol{\vartheta}} = -\sqrt{h}\nabla F_{\mathcal{L}}(\boldsymbol{\vartheta}) - \sqrt{\mu}(\boldsymbol{\vartheta} - V)\\ \dot{V} = -\sqrt{\mu}(V-\boldsymbol{\vartheta}) - \left(\frac{1}{\sqrt{\mu}}\right)\nabla F_{\mathcal{L}}(\boldsymbol{\vartheta})\end{cases} \quad (1)$$

with $\boldsymbol{\vartheta}(0) = \boldsymbol{\vartheta_0}, \dot{\boldsymbol{\vartheta}}(0) = \frac{-2\sqrt{h}\nabla F_{\mathcal{L}}(\boldsymbol{\vartheta}_0)}{1+\sqrt{\mu h}}$ will converge to the global minimizer with rate

$$F_{\mathcal{L}}(\boldsymbol{\vartheta}) - F_{\mathcal{L}}(\boldsymbol{\vartheta}^*) \le \frac{2||\boldsymbol{\vartheta_0}-\boldsymbol{\vartheta}^*||^2}{h} e^{\frac{-\sqrt{\mu}t}{4}}.$$

Interestingly, if one discretizes the above ODE with semi-implicit Euler scheme, then with small step size NAG is approximately a symplectic method. Also, implicit Euler scheme leads to acceleration, but it is not easy to use in practice [2].

References

1. Shi, B., Du, S.S., Jordan, M.I. et al. Understanding the acceleration phenomenon via high-resolution differential equations. Math. Program. (2021). https://doi.org/10.1007/s10107-021-01681-8
2. Shi, Bin. "Acceleration via Symplectic Discretization of High-Resolution Differential Equations." (2019).
3. Zhang, Peiyuan, et al. "Revisiting the Role of Euler Numerical Integration on Acceleration and Stability in Convex Optimization." International Conference on Artificial Intelligence and Statistics. PMLR, 2021.

Main Theorem

The ODE (1) can be generalized by replacing the coefficients with positive parameters m, n, p, q

$$\begin{cases}\dot{\boldsymbol{\vartheta}} = -m\nabla F_{\mathcal{L}}(\boldsymbol{\vartheta}) - n(\boldsymbol{\vartheta} - V)\\ \dot{V} = -q(V-\boldsymbol{\vartheta}) - p\nabla F_{\mathcal{L}}(\boldsymbol{\vartheta})\end{cases},$$

which can be rephrased as the ODE

$$\ddot{\boldsymbol{\vartheta}} + ((n+q) + m\nabla^2 F_{\mathcal{L}}(\boldsymbol{\vartheta}))\dot{\boldsymbol{\vartheta}} + (mq+np)\nabla F_{\mathcal{L}}(\boldsymbol{\vartheta}) = 0 \quad (2)$$

Preliminary Result: The following theorem shows that for a fixed rate of convergence, one can find many accelerated ODEs.

Theorem 1: Assume $F_{\mathcal{L}}(\boldsymbol{\vartheta})$is L-smooth and μ-strongly convex. Then if $\boldsymbol{\vartheta}(t)$ and $V(t)$ are such that (2) holds, the **Lyapunov function**

$$\varepsilon(t) = F_{\mathcal{L}}(\boldsymbol{\vartheta}(t)) - F_{\mathcal{L}}(\boldsymbol{\vartheta}^*) + A\|V(t) - \boldsymbol{\vartheta}^*\|^2$$

will decrease as

$$\varepsilon(t) \le e^{-\min\left\{n,\frac{q}{4}\right\}t}\varepsilon(0)$$

with $\max\left\{\frac{m}{q}, \frac{n\mu}{q}\right\} \le A \le \min\left\{\frac{n}{2(q+p)}, \frac{4n\mu}{3q}\right\}$ and $m, n, p, q \ge 0$.

We can apply **semi-implicit Euler integrator** to achieve the corresponding algorithm

$$\begin{cases}\boldsymbol{v}_{k+1} - \boldsymbol{v}_k = -p\sqrt{h}\nabla F_{\mathcal{L}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta_{k+1}}) - q\sqrt{h}(\boldsymbol{v}_k - \boldsymbol{\theta_k})\\ \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k = -m\sqrt{h}\nabla F_{\mathcal{L}}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta_k}) - n\sqrt{h}(\boldsymbol{\theta}_k - \boldsymbol{v}_k)\end{cases} \quad (3)$$

The following theorem shows the convergence rate of this algorithm.

Theorem 2: Assume $F_{\mathcal{L}}(\boldsymbol{\vartheta})$is L-smooth and μ-strongly convex. Then if $\boldsymbol{\theta}_k$ and $\boldsymbol{v}_k$ follow the updates (3), the Lyapunov function

$$\varepsilon(k) = F_{\mathcal{L}}(\boldsymbol{\vartheta}_k) - F_{\mathcal{L}}(\boldsymbol{\vartheta}^*) + A'\|\boldsymbol{v}_k - \boldsymbol{\vartheta}^*\|^2$$

will decrease as

$$\varepsilon(k+1) \le (1-\lambda)^k\varepsilon(0)$$

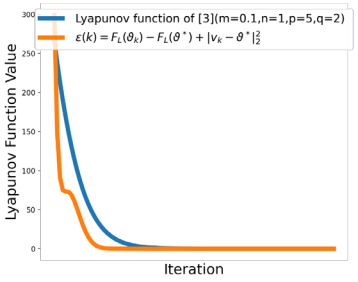
with

$$\frac{n}{2q(1-q\sqrt{h}-p\sqrt{h})} \le A' \le \min\left\{\frac{1}{4L(q\sqrt{h}-q^2h)}, \frac{\mu^2}{4L(q\sqrt{h}-p\sqrt{h\mu^2}-p^2h\mu^2)}\right\},$$

$$p\sqrt{h} \le q\sqrt{h} \le \frac{1}{2}, \frac{1}{2L\sqrt{h}} \le m \le \frac{1}{L\sqrt{h}}, n \le \frac{1-q\sqrt{h}-p\sqrt{h}}{2L(1-q\sqrt{h})\sqrt{h}}, m, n, p, q \ge 0,$$

$$\lambda = \min\{q\sqrt{h} - p\sqrt{h}, \frac{2}{L}(\frac{\mu^2}{4L} + A'p\sqrt{h}\mu^2 + A'p^2h\mu^2 - Aq\sqrt{h})\}.$$

For comparison, the trajectory of Lyapunov function above (3) is compared with the Lyapunov function in [3]. Of course, there are conditions under which the behaviours are different. Deeper analysis is left for future work.



Mellema, René
Umeå University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Normative reasoning for Social AI

Norms are both a regular occurrence in human reasoning, as well as a useful tool for governing the behaviour of agent populations. However, what exactly norms are, and how we can effectively use them in computer science is poorly understood. For example, in CS, norms are often purely used as constraints of behaviour, while they can also have a strong motivating component.
In our research, we try to address this issue by formalizing sociological and psychological theories of norms. This gives us a framework for studying norms and their interactions. Besides this, we also study how norms influence human reasoning, and how agents can use them in their reasoning.

Mellema, René
Umeå University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Normative reasoning for Social AI

René Mellema, Umeå University
Computing Science
Supervisors: Frank Dignum
Juan Carlos Nieves Sanchez

UMEÅ UNIVERSITY

Motivation & Research goals

Norms are both a regular occurrence in human reasoning, as well as a useful tool for governing the behaviour of agent populations. However, what exactly norms are, and how we can effectively use them in computer science is poorly understood. For example, in CS, norms are often purely used as constraints of behaviour, while they can also have a strong motivating component.
In our research, we try to address this issue by formalizing sociological and psychological theories of norms. This gives us a framework for studying norms and their interactions. Besides this, we also study how norms influence human reasoning, and how agents can use them in their reasoning.

Why norms?

In human societies, norms are used to build **accountability** and **cooperation**. This means they play an integral part in all our day-to-day interactions, they can have internal effects, and norm following/breaking behaviour carries meaning as well. This means that they have a strong **motivational** component.
In **social simulation** we want to model human societies to e.g. study them or make predictions about reactions to changes. Since norms play such a pivotal role in human societies, they can have a large effect in these simulations.

Why new formalizations?

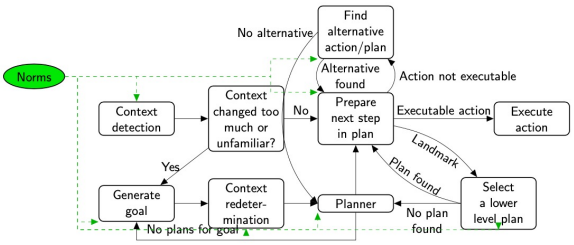
However, in CS norms get used in multi-agent systems to control the behaviour of heterogenous populations of agents. This means the focus is often on norms as **constraints**. This means current formalizations ignore the motivational aspects, as well as **sanctioning** behaviours. Both of these are important for **norm change**, which is currently also not well understood.

The formalizations

We are interested in **representing** norms in our simulations such that agents can **reason** with them. This requires that various aspects of the norm are incorporated in the design:

- Activation/deactivation conditions
- Violation condition
- Sanctions for breaking the norm

Similarly, the agents need to be able to react to other agents norm breaking behaviour, which means the representation also needs to take violations into account. Current research is ongoing on how to best represent violations, and which aspects are necessary to differentiate between them. Using this, a framework for normative reasoning in CTL is being developed.



Agents with normative deliberation

If we want the agents to use norms, the norms do not only need to be represented, but the agents also need to know when to take them into account in their reasoning. This needs to happen in:

- **Goal generation**: Norms can have obligations attached, which can motivate an agent to take actions it otherwise would not have
- **Context detection**: Part of a context is what norms are active, which can influence what actions the agent considers
- **Planning**: Not every plan an agent can follow has to follow all norms, but it should make the decision whether or not to follow a plan/take an action based on the norms it might follow/break

In all of these, whether or not the agent should follow a norm is an important consideration.

Norm change

While norms can stabilize a society, they are not static. Since we are interested in peoples possible responses to new policy, how this influences old norms and shapes new ones is vital. Currently, our work here focusses on how agents can break norms, and why they might do so.

References

1. Mellema R., Jensen M., Dignum F. (2021) Social Rules for Agent Systems.
2. Dignum, F. (2021) Social Simulation for a Crisis
3. Brennan G., Eriksson L., Goodin R., Southwood N. (2013) Explaining norms
4. Vázquez-Salceda, J., Aldewereld, H., Grossi, D., & Dignum, F. (2008). From human regulations to regulated software agents' behavior

Further questions

- Where in the decision making process should norms exert an influence?
- How do we combine normative reasoning with other types of motivation?
- When should an agent break a norm?
  - What factors influence this decision?
- How do identities and norms interact?
- When should an agent sanction another agent for breaking a norm?

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Nilsson, Viktor
KTH

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Interacting Particle Dynamics for Deep Learning

Neural networks (MLPs) and GANs can be interpreted/represented as systems of interacting particles. This may enable using techniques from statistical physics, probability theory and partial differential equations in the understanding of neural networks. Future work includes establishing laws of large deviations (LDP) to help make these connections.
This poster shows two different frameworks in which single hidden layer neural networks, an GANs are treated from this perspective.

Nilsson, Viktor
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Interacting Particle Dynamics for Deep Learning

Viktor Nilsson, Pierre Nyquist - KTH Mathematics Dept.

Introduction

Several frameworks have been proposed that establish a particle dynamic view of neural networks. In two different fashions, one can see the training and inference of a network as the behavior of a many-particle system, consisting of say N particles. Further, such systems with N particles have a 'mean-field' behavior when letting $N \to \infty$, i.e., having the characteristic of a 'smooth' distribution. This lends itself to so called mean-field approximation, where for large N, the system is approximated by the limit behavior instead. Thus, the discrete probability distribution of the N-particle system is replaced by a continuous distribution instead. This distribution and its evolution under the training dynamics can then be described by a PDE, or a so-called gradient flow.
Several questions remain about the convergence to the mean-field limit.
Current literature establishes some convergence results of law of large numbers (LLN) central limit theorem (CLT) type, while not giving any convergence rates. The current goal is to go beyond these results and use the *theory of large deviations* to develop a *large deviations principle* (LDP), which gives convergence rate guarantees based on a *rate function*.

One hidden layer neural network

Consider a one hidden layer neural network $f_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}$. Its prediction can be seen as an average of the N hidden neurons, i.e.,

$$f_{\boldsymbol{\theta}}(x) = \frac{1}{N}\sum_{i=1}^{N} \varphi(x, \boldsymbol{\theta}_i).$$

How is the behavior when $N \to \infty$? With an L2-loss function it turns out that the loss can be written as

$$l(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N) = \sum_{i=1}^{N} F(\boldsymbol{\theta}_i) + \frac{1}{2N}\sum_{i,j=1}^{N} K(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j). \quad (1)$$

Defining the *empirical measure* of the weights, $\mu_t = \sum_{i=1}^{N} \delta_{\boldsymbol{\theta}_{i,t}}$, we have that a standard gradient descent (with infinitesimal timestep) follows the PDE

$$\partial_t \mu_t = \nabla \cdot (\mu_t \nabla V), \quad (2)$$

in the many-particle limit.

GANs

Generative adversarial networks consist of a pair of networks, $G : \mathcal{Z} \to \mathcal{X}$ and $D : \mathcal{X} \to [0, 1]$, that compete in some two-player game, for instance the following zero-sum game.

$$\min_G \max_D \mathbb{E}_{\mathbf{x}}[\log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z}}[\log(1 - D(G(\mathbf{z})))] \quad (3)$$

Existence of pure Nash equilibria are not guaranteed in continuous games. However, the existence of *mixed Nash equilibria* is guaranteed. A mixed Nash equilibrium is a Nash equilibrium for the relaxed game

$$\mathcal{L}(\mu_x, \mu_y) := \int\int l(x, y)\, \mu_x(dx)\mu_y(dy). \quad (4)$$

Thus, we consider *mixed strategies* μ_x, μ_y instead of pure strategies x, y. In practice, this is done by having multiple "particles" $\{x_t^i\}_{i=1}^n, \{y_t^i\}_{i=1}^n$ and letting their empirical measures approximate μ_x, μ_y.

$$\mu_{x,t}^n := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_t^i}, \quad \mu_{y,t}^n := \frac{1}{n}\sum_{i=1}^{n} \delta_{y_t^i} \quad (5)$$

How should we optimize $\delta_{x_t^i}$, $\delta_{y_t^i}$? Gradient descent-ascent (DA) dynamics correspond to

$$\begin{aligned} dX_t^i &= -\frac{1}{n}\sum_{j=1}^{n} \nabla_x l(X_t^i, Y_t^j)dt, \\ dY_t^i &= \frac{1}{n}\sum_{j=1}^{n} \nabla_y l(X_t^j, Y_t^i)dt. \end{aligned} \quad (6)$$

Future work

The mean-field behavior is described in [2] and [1]. Currently, we want to strengthen those results by providing a LDP.
The dynamics of equation (2) and equation (6) can be modified by including a diffusion term, e.g. adding the term $\sqrt{2\beta^{-1}}dW_t^i$ to equation (6). We would further like to see how the size of the inverse temperature β affects the convergence.

References

[1] Carles Domingo-Enrich et al. "A mean-field analysis of two-player zero-sum games". In: *arXiv preprint arXiv:2002.06277* (2020).
Grant Rotskoff et al. "Global convergence of neuron birth-death dynamics". In: *arXiv preprint arXiv:1902.01843* (2019).

Osipov, George
Linköping University

Resolving Inconsistencies in Simple Temporal Problems: A Parameterized Approach

Constraint satisfaction problems (CSPs) have many applications in AI including planning, knowledge representation, and reasoning. Given a set of variables and constraints, the goal in a CSP is to find an assignment that satisfies all constraints. The computational complexity of a CSP depends on the set of allowed constraints. For all sets of constraints over a finite domain, the dichotomy theorem of Bulatov and Zhuk distinguishes between problems that are in P and NP-complete. Almost CSP is an optimization version, where the goal is to find an assignment that violates as few constraints as possible. Applications include handling noise, dealing with faulty measurements, and repairing merge conflicts in databases. With an additional assumption that the number of violated constraints is small, Almost CSP becomes interesting from the point of view of parameterized complexity. Here one needs to distinguish between problems in P, in FPT, W[1]-hard and NP-hard. In our work, we give a full classification for Almost Simple Temporal Problem (STP), an influential reasoning formalism for temporal information.

Osipov, George
Linköping University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Almost Constraint Satisfaction Problems

George Osipov
Linköping University

Toy Problem

Setting:
You teach a class and n students signed up.
You need to assign them into 2 groups.
Some students are friends and asked to be in the same group.
Some students are foes and asked to be in different groups.

Question 1: Can you divide students and respect all preferences?

Mathematical model:
Introduce a variable x_i for every student $i \in \{1, \dots, n\}$.
Let's set $x_i = 1$ ($x_i = 2$) if student i is assigned to group 1 (2).
If i and j are friends, the assignment should satisfy the constraint $x_i = x_j$.
If i and j are foes, the assignment should satisfy the constraint $x_i \neq x_j$.
Find an assignment $f : \{x_1, \dots, x_n\} \to \{1, 2\}$ that satisfies all constraints.

Follow-up:
What if no such assignment exists?
One can try to find an assignment that violates as few constraints as possible (upsets the students as little as possible).

Question 2. Can you divide students and disregard as few preferences as possible?

General Problem: Constraint Satisfaction

Let D be a domain (i.e., a set of values) and $\mathcal{A}$ be a set of relations over D.

Constraint Satisfaction Problem (CSP($\mathcal{A}$))
Instance: A set of variables V and a set of constraints C of the form $R(v_1, \dots, v_r)$, where R is a relation from $\mathcal{A}$ and $v_1, \dots, v_r \in V$.
Question: Is there an assignment $f : V \to D$ that satisfies all constraints in C?

Almost Constraint Satisfaction

Almost Constraint Satisfaction Problem (AlmostCSP($\mathcal{A}$))
Instance: An instance (V, C) of CSP($\mathcal{A}$) and an integer k.
Question: Is there an assignment $f : V \to D$ that satisfies all but k constraints in C?

Toy Problem as a CSP

Variables: $V = \{x_1, \dots, x_n\}$ (one variable per student).
Domain: $\{1, 2\}$ (groups).
Relations: $\{=, \neq\}$.
Call the problem CSP($\mathcal{T}$).

Complexity:
CSP($\mathcal{T}$) is in P. AlmostCSP($\mathcal{T}$) is NP-hard and *fixed-parameter tractable*.

Parameterized Complexity of AlmostCSP

Parameter k – number of violated constraints.
PC of AlmostCSP($\mathcal{A}$) is only interesting if it is NP-hard and CSP($\mathcal{A}$) is in P.
Can be solved in $n^{O(k)}$ time by considering every subset of $n - k$ constraints.
Impractical even for small values of k.
If it can be solved in FPT time, i.e. $f(k) \cdot n^C$ for some function f of k and a constant C independent of k, then the algorithm is efficient for small values of k.

AlmostCSP Classification Project
For which $\mathcal{A}$ does AlmostCSP($\mathcal{A}$) admit an FPT algorithm?

Parameterized Complexity Theory

k-Vertex Cover

Goal: Cover all edges with k vertices.
Solvable in $f(k) \cdot n^C$ time - in FPT.

k-Independent Set

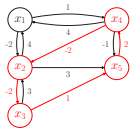
Goal: Find k non-adjacent vertices.
Solvable in $n^{O(k)}$ time - W[1]-hard.

Main Conjecture: FPT $\neq$ W[1].

Simple Temporal Problem (STP)

Introduced by Dechter, Meiri, and Pearl in 1989.
Variables $x_1, x_2, \dots, x_n$ represent points in time.
Constraints: $a \leq x_i - x_j \leq b$, where $a, b \in \mathbb{Q} \cup \{-\infty, \infty\}$.
Denote the problem by CSP($\mathcal{S}$).

$$
\begin{aligned}
-1 &\leq x_4 - x_1 \leq 4 \\
2 &\leq x_2 - x_1 \leq 4 \\
2 &\leq x_3 - x_2 \leq 3 \\
1 &\leq x_5 - x_4 \leq 2 \\
2 &\leq x_2 - x_4 \leq \infty \\
-\infty &\leq x_3 - x_5 \leq 1 \\
-\infty &\leq x_2 - x_5 \leq 3.
\end{aligned}
$$



An instance is satisfiable iff its distance graph has no negative cycle. In P.

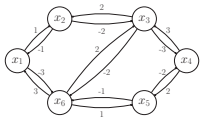
AlmostSTP Classification

Theorem: Let $\mathcal{A} \subseteq \mathcal{S}$. Then AlmostCSP($\mathcal{A}$) is
- in constant time if $\mathcal{A}$ is trivial (all relations satisfied by setting $x_i = x_j$).
- in FPT if $\mathcal{A}$ only contains left/right-sided relations ($x_i - x_j \lesseqgtr c$ for $c \in \mathbb{Q}_{\geq 0}$).
- in FPT if $\mathcal{A}$ only contains equation relations ($x_i - x_j = c$ for $c \in \mathbb{Q}$).
- W[1]-hard otherwise.

Equation STP

An interesting special case. Satisfiable if and only if has no non-zero cycles.

$$
\begin{aligned}
x_1 - x_2 &= 1 \\
x_3 - x_2 &= 2 \\
x_4 - x_3 &= 3 \\
x_4 - x_5 &= 2 \\
x_6 - x_5 &= 1 \\
x_6 - x_1 &= 3 \\
x_3 - x_6 &= 2.
\end{aligned}
$$



Generalize balanced signed graphs. Special case of balanced group-labelled graphs.

LINKÖPING UNIVERSITY

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

All results for Almost STP can be found in "*Resolving Inconsistencies in Simple Temporal Problems: A Parameterized Approach*" by Konrad K. Dabrowski, Peter Jonsson, Sebastian Ordyniak, George Osipov (to appear in AAAI 2022). Access a preliminary version here: https://tinyurl.com/almostp.

Papenmeier, Leonard
Lund University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

High-Dimensional Bayesian Optimization with Gaussian Processes

Standard Bayesian Optimization (BO) is known to perform only well for up to 20-30 input dimensions. Optimizing higher-dimensional functions requires changes to the model or further assumptions on the problem itself. One line of current research focuses on sparse problems, where one assumes that the problem lies in a low-dimensional subspace of the higher-dimensional (ambient) space. Such methods perform BO in a lower-dimensional subspace that ideally captures the true effective subspace. Most algorithms for such problems, however, require an appropriate guess for the effective dimensionality as they rely on fixed embeddings. We present an algorithm that softens this requirement by introducing adaptive embeddings that increase the lower-dimensional subspace over time. Our algorithm outperforms the state-of-the-art on many benchmarks while being more computationally efficient than many contemporary approaches.

Papenmeier, Leonard
Lund University

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

High-dimensional Bayesian Optimization with Adaptive Embeddings

Leonard Papenmeier, PhD Student, Lund University
Department of Computer Science
Supervisor: Dr. Luigi Nardi, Lund University, Coauthor: Matthias Poloczek, Amazon

LUND UNIVERSITY

Motivation & Research Goals

Standard Bayesian Optimization (BO) is known to perform only well for up to 20 input dimensions [2]. Optimizing higher-dimensional functions requires changes to the model or further assumptions on the problem itself. Current research considers sparse problems where the problem lies in a low-dimensional subspace of a higher-dimensional (ambient) space. Such methods perform BO in a lower-dimensional subspace that aims at capturing the true effective subspace. Yet, most algorithms for such problems require an appropriate guess on the effective dimension. We present an algorithm (AdaTheSBO) that softens this requirement by using adaptive embeddings that increase the subspace dimension over time. [*Unpublished, preliminary state.*]

Problem and Algorithm

Minimization of expensive-to-evaluate black-box function $f : \mathbb{R}^D \mapsto \mathbb{R}$: $x^* \in \arg\min_{x \in \mathcal{X}} f(x)$, where $\mathcal{X}$ is D-dimensional ($D \gg 20$). We assume that there exists a low-(d-)dimensional ($d \leq 20$) subspace $\mathcal{Y}$ that can be mapped to by a linear embedding, and f is axis-aligned.

Use HeSBO embedding [3] to train an information-preserving Gaussian Process (GP) in trust region [1] of a subspace of increasing dimension.

Figure 1: Increasing the dimension of the embedding from 2 to 3 (y_2 is *split*). Information can be preserved with the HeSBO embedding when increasing the dimension.

Algorithm 1 Ada-TheSBO Algorithm Outline

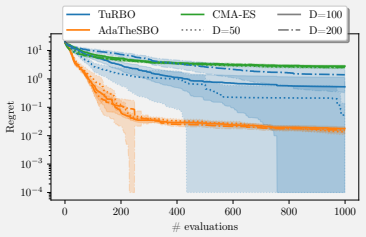
Require: initial latent dimension d
  sample random HeSBO embedding, defining up-projection S^T
  while not converged or budget available **do**
    while trust region sufficiently large **do**
      find candidate $\boldsymbol{x}^{(t)}$ by maximizing Thompson sample
      evaluate $f(S^T \boldsymbol{x}^{(t)})$; update GP; update TR
    end while
    if no progress in inner `while`-loop **then**
      re-start with new embedding and new GP
    else
      split latent dimension(s) with smallest GP length scale
    end if
    $d \leftarrow d + 1$
  end while
  Return Overall best $\boldsymbol{x}$ so far

Contributions

- First algorithm with an embedding of increasing dimension
- Outperforms state-of-the-art on a variety of problems
- Works in arbitrarily high-dimensions as long as d bounded

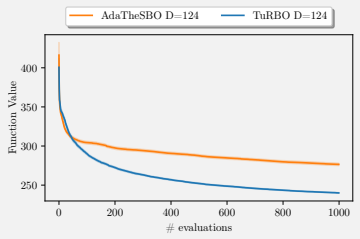
Selected Results

AdaTheSBO (with TuRBO and CMA-ES for comparison) on different input dimensions of the Ackley2 function. Empirically, AdaTheSBO is agnostic with respect to the ambient dimension for a fixed effective dimension.



Strong performance on sparse-axis aligned, competitive on sparse, non-axis aligned problems. Poor performance on truly high-dimensional problems.

AdaTheSBO and TuRBO for comparison on different input dimensions of the Mopta08 function. The benchmark is assumed to be dense.



References

References

[1] David Eriksson et al. "Scalable global optimization via local bayesian optimization". In: Advances in Neural Information Processing Systems 32 (2019), pp. 5496–5507.

[2] Peter I Frazier. "A tutorial on Bayesian optimization". In: arXiv preprint arXiv:1807.02811 (2018).

[3] Amin Nayebi, Alexander Munteanu, and Matthias Poloczek. "A Framework for Bayesian Optimization in Embedded Subspaces". In: Proceedings of the 36th International Conference on Machine Learning. 2019.

Razavikia, Saeed
KTH

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Over the Air Computation For Machine Learning Over Wireless

With the increasing popularity of mobile devices and the development of the internet of things (IoT), accessibility to vast amounts of data has been grown. Further, the global number of connected IoT devices will reach more than 4 billion by 2024. On the flip side, taking advantage of large data sets can aid us in solving many complex problems in Machine Learning (ML). The primary challenges are communication latency, bandwidth consumption, energy limitations, privacy, and security. With limited communication resources, it is challenging to achieve efficient data aggregation over a large volume of IoT devices, as a critical point for exploiting the potential of the distributed ML. Unlike the standard "transmit-then-compute" approach, the over-the-air computation approach integrates communication and computation steps and provides ultra-fast wireless data aggregation in IoT networks.

Razavikia, Saeed
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

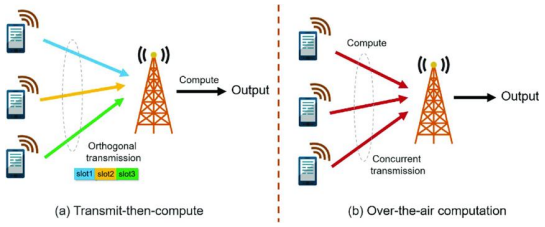
Over the Air Computation For Machine Learning Over Wireless

Saeed Razavikia, KTH Royal Institute of Technology
Electrical Engineering and Computer science
Main Advisor: Carlo Fischione

Motivation

With the increasing popularity of mobile devices and the development of the internet of things (IoT), accessibility to vast amounts of data has been grown. Further, the global number of connected IoT devices will reach more than 4 billion by 2024. On the flip side, taking advantage of large data sets can aid us in solving many complex problems in Machine Learning (ML). The primary challenges are communication latency, bandwidth consumption, energy limitations, privacy, and security. With limited communication resources, it is challenging to achieve efficient data aggregation over a large volume of IoT devices, as a critical point for exploiting the potential of the distributed ML. Unlike the standard "transmit-then-compute" approach, the over-the-air computation approach integrates communication and computation steps and provides ultra-fast wireless data aggregation in IoT networks.

Problem Setup

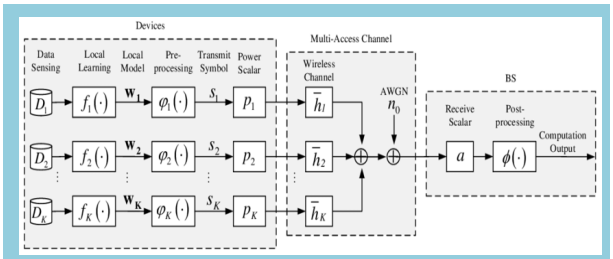


Chen et al. [1]

- Computing a function of input data at edge server by leveraging superposition property of electromagnetic waves
- Wireless communications become wireless computations

Benefits:

1. Integrate the communication and computation
2. Bandwidth is shared among all users in the time, frequency and code domain
3. More spectral efficient than transmit-then-compute
4. Transmission relies on analog communication
5. Fast wireless data aggregation

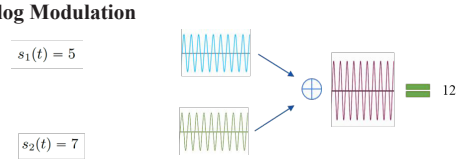


Ni et al. [2]

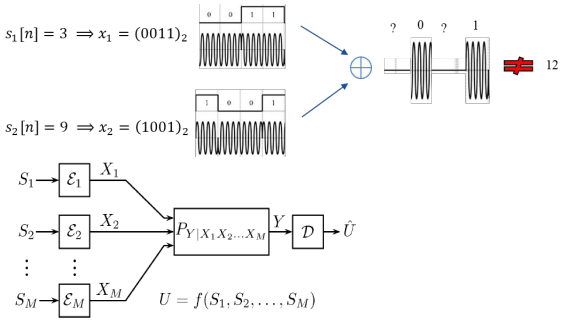
- Nomographic functions can be calculated
  - Mean $r = \frac{1}{K}\sum_{k=1}^{K} w_k$
  - Multiplication $r = \prod_{k=1}^{K} w_k$
  - Maximum $r = \max_k w_k$

Digital Communication

Analog Modulation



Digital Modulation



- Digital Communications can be seen as a probabilistic digital map/function between input code-words and output code-words
- In so doing, we can allow digital communication to perform probabilistic function computations
- The probability map i.e. $P_{(Y|X_1, X_2, \ldots, X_M)}$ can be designed by artificially enforcing the probability correspondence input output (changing the modulation formats)

References

1. X. Chen and Q. Qi, Convergence of Energy, Communication and Computation in B5G Cellular Internet of Things. Berlin, Germany: Springer, 2020
2. N. Wanli, L. Yuanwei, Y. Zhaohui, T. Hui, and S. Xuemin. "Federated learning in multi-RIS aided systems". IEEE Internet of Things Journal, 2021.

Restadh, Petter
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Greedy Causal Discovery is Geometric

Finding a directed acyclic graph (DAG) that best encodes the conditional independence statements observable from data is a central question within causality. Algorithms that greedily transform one candidate DAG into another given a fixed set of moves have been particularly successful, for example the GES, GIES, and MMHC algorithms. In 2010, Studený, Hemmecke and Lindner introduced the characteristic imset polytope, CIM_p , whose vertices correspond to Markov equivalence classes, as a way of transforming causal discovery into a linear optimization problem. We show that the moves of the aforementioned algorithms are included within classes of edges of CIM_p and that restrictions placed on the skeleton of the candidate DAGs correspond to faces of CIM_p . Thus, we observe that GES, GIES, and MMHC all have geometric realizations as greedy edge-walks along CIM_p.

Restadh, Petter
KTH

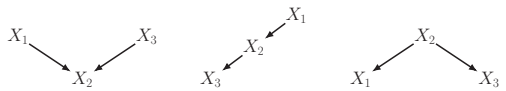
WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Greedy Causal Discovery is Geometric

S. Linusson, P. Restadh and L. Solus
Email: linusson@kth.se, petter@kth.se, and solus@kth.se
KTH Royal Institute of Technology

Graphical Models

Directed Acyclic graphs (DAGs) are commonly used for complex models and causal interference [2]. DAGs encode conditional independence statements in the following way:



$X_1 \perp\!\!\!\perp X_3|\emptyset$, $X_1 \not\perp\!\!\!\perp X_3|X_2$ | $X_1 \perp\!\!\!\perp X_3|X_2$, $X_1 \not\perp\!\!\!\perp X_3|\emptyset$ | $X_1 \perp\!\!\!\perp X_3|X_2$, $X_1 \not\perp\!\!\!\perp X_3|\emptyset$

Goal 1 *Given data D on $X_1, \ldots, X_p$, find the DAG that "best" fits our data.*

Several algorithms have been proposed, for example PC, greedy SP, GES, GIES, and MMHC.

Characteristic Imset Polytope

The **characteristic imset** [4] of a DAG $\mathcal{G}$ is a vector in $\mathbb{R}^{\Upsilon}$ where $\Upsilon := \{S \subseteq [p] : |S| \geq 2\}$. It is defined as

$$c_{\mathcal{G}}(S) = \begin{cases} 1 & \text{if } \exists i \in S \text{ s.t. } S \subseteq \mathrm{pa}_G(i) \cup \{i\} \\ 0 & \text{otherwise.} \end{cases}$$

It was shown [4] that any (additive) decomposable score equivalent function can be written as a linear function in $\mathbb{R}^{\Upsilon}$. Thus we consider the polytopes

$$\mathrm{CIM}_p := \mathrm{conv}(c_{\mathcal{G}} : \mathcal{G} \text{ is a DAG with } p \text{ nodes})$$

and if we have two graphs $G \subseteq H$ we define

$$\mathrm{CIM}_{G,H} := \mathrm{conv}(c_{\mathcal{G}} : \mathcal{G} \text{ a DAG with skeleton } D \text{ such that } G \subseteq D \subseteq H).$$

Example 1 *Let us consider* CIM.

Proposition 1 $\mathrm{CIM}_{G,H}$ *is a face of* CIM_p.

Hence Goal 1 can be formulated as the following.

Goal 2 *Given a (additive) decomposable score equivalent function s_D. Maximize s_D over* CIM_p.

GES, GIES, and MMHC all use the Bayesian Information Criterion (BIC) as a score function.

Edges of CIM_p and CIM_G

If $\mathcal{G}$ is a directed graph with $i \to j \in \mathcal{G}$ we denote by $\mathcal{G}_{i \leftarrow j}$ the directed graph identical to $\mathcal{G}$ except that the edge $i \to j$ is replaced with $i \leftarrow j$.

Proposition 2 *If $\mathcal{G}_{i \leftarrow j}$ is a DAG, then either $\mathcal{G}$ and $\mathcal{G}_{i \leftarrow j}$ are Markov equivalent, or* $\mathrm{conv}\left(c_{\mathcal{G}}, c_{\mathcal{G}_{i \leftarrow j}}\right)$ *is an edge of* CIM_G.

Let $\mathcal{G}$ be a DAG and assume i and j are not adjacent in the skeleton of $\mathcal{G}$. We denote by $\mathcal{G}_{+i \leftarrow j}$ the directed graph identical to $\mathcal{G}$ with the edge $i \leftarrow j \in \mathcal{G}_{+i \leftarrow j}$.

Proposition 3 *If $\mathcal{G}_{+i \leftarrow j}$ is a DAG, then* $\mathrm{conv}\left(c_{\mathcal{G}}, c_{\mathcal{G}_{+i \leftarrow j}}\right)$ *is an edge of* CIM_p.

We also show several more classes of edges.

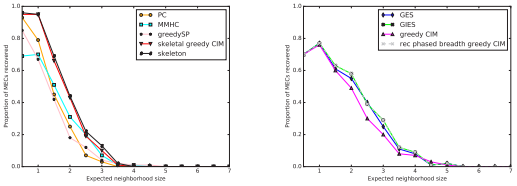
Theorem 4 *The following causal discovery algorithms are greedy edge-walks along a face of* CIM_p*:*

1. *GES,*
2. *GIES with purely observational data,*
3. *MMHC, and*
4. *Greedy SP [1].*

Thus we can define two algorithms **Greedy CIM** and **Skeletal Greedy CIM** as greedy depth-first edge-walks along CIM_p and CIM_G respectively. By the above propositions **Greedy CIM** generalize GES and GIES with observational data. The graph G in **Skeletal Greedy CIM** was determined using conditional independence test similar to PC. A recurrent phased breadth-first version of **Greedy CIM** was as well implemented, as an easier comparison to GES and GIES.

Computational Results

The algorithms were implemented on simulated data using linear structural equation models with Gaussian noise. The code is available at [3].



We see that among algorithms using conditional independence test **Skeletal Greedy CIM** performs better than previous algorithms. In the purely score based methods the breadth first algorithms have a higher recovery rate, but that can change if more edges of CIM_p were found and classified.

[1] F. Mohammadi, C. Uhler, C. Wang, and J. Yu, *Generalized permutohedra from probabilistic graphical models*, SIAM Journal on Discrete Mathematics, 32 (2018), pp. 64–93.

[2] J. Pearl, *Causality : Models, Reasoning, and Inference*, Cambridge University Press, Cambridge, U.K. New York, 2000.

[3] P. Restadh and L. Solus, *causalCIM. GitHub Repository*, 2021.

[4] M. Studený, R. Hemmecke, and S. Lindner, *Characteristic imset: A simple algebraic representative of a bayesian network structure*, Proceedings of the 5th European Workshop on Probabilistic Graphical Models, PGM 2010, (2010), pp. 257–265.

Rydell, Felix
KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Algebraic Vision

The applications of reconstructing 3D models from 2D images include modelling of cities and objects for movies and video games, modelling clouds to predict the wheather, and helping robots and vehicles to orient themselves in new environments. Algebraic vision, which describes the algebraic component, is a prominent connection between methods from algebraic geometry and artificial intelligence. I investigate the geometry of points and lines projected onto the images of a set of cameras, and the stability of different approaches in the algebraic part of the reconstruction. This can help engineers in building new algorithms.

Rydell, Felix
KTH

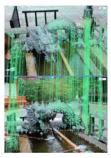
WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

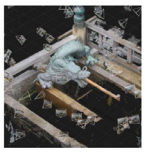
Algebraic Vision

Felix Rydell, KTH
Mathematics for Data and AI

Motivation & Research goals

The applications of reconstructing 3D models from 2D images include modelling of cities and objects for movies and video games, modelling clouds to predict the wheather, and helping robots and vehicles to orient themselves in new environments. *Algebraic vision*, which describes the algebraic component, is a prominent connection between methods from algebraic geometry and artificial intelligence. I investigate the geometry of points and lines projected onto the images of a set of cameras, and the stability of different approaches in the algebraic part of the reconstruction. This can help engineers in building new algorithms.



(a) Input images (b) Image matching (c) Reconstruct cameras and 3D points (d) Output

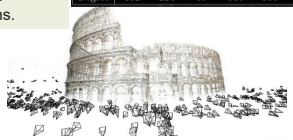
The reconstruction pipeline. Algebraic vision concerns itself with part c).

This table gives alternatives to part c). The first rows stand for number of cameras and the last row the number of solutions of camera positions **[1]**. The fewer number of solutions, the faster the algorithms.

m views	6	6	6	5	5	5	4	4	4	4	4	4
$p^1p^{01}l^1_0$	1021_1	1013_3	1005_5	2011_1	2003_2	2003_2	1030_0	1022_2	1014_4	1006_6	3001_1	2110_0	2102_2
$(p,l,\mathcal{I})$													
Minimal	Y	N	N	Y	Y	Y	Y	Y	N	N	Y	Y	Y
Degree	> 450k*			11306*	26240*	11008*	3040*	4524*			1728*	32*	544*

m views	4	3	3	3	3	3	3	3	3	3	3	3	3
$p^1p^{01}l^1_0$	2102_2	1040_0	1032_2	1024_4	1016_6	1008_8	2021_1	2013_2	2013_3	2005_3	2005_4	2005_5	3010_0
$(p,l,\mathcal{I})$													
Minimal	Y	Y	Y	Y	N	N	Y	Y	Y	Y	Y	Y	Y
Degree	544*	360	552	480			264	432	328	480	240	64	216

m views	3	3	3	3	3	3	3	3	2	2	2	2	2
$p^1p^{01}l^1_0$	3002_1	3002_2	2111_1	2103_1	2103_2	2103_3	3100_0	2201_1	5000_2	4100_3	3200_3	3200_4	2300_2
$(p,l,\mathcal{I})$													
Minimal	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	N	N
Degree	312	224	40	144	144	144	64		20	16	12		

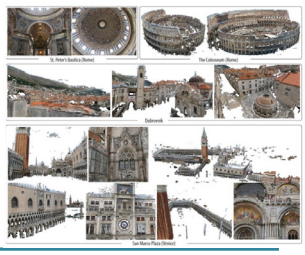


Two examples of reconstructions **[2]**. Once we have found the camera positions using an approach from the table above, we find the point/line cloud in 3D.

A systematic understanding of reconstruction is provided via 1) *projective* and 2) *algebraic* geometry.

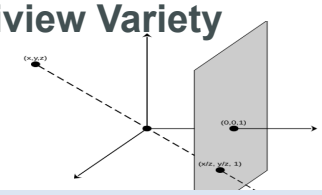
1) *Projective space* $\mathbb{P}^n$ is the set of lines in $\mathbb{R}^{n+1}$ passing through the origin. It is a compactification of $\mathbb{R}^n$, meaning it is equal to $\mathbb{R}^n$ and some additional points at infinity.

2) *Algebraic geometry* is the study of systems of polynomials equations. A set of solutions to such a system is an *algebraic variety*.



The Multiview Variety

Let $\mathtt{C} = \{C_1, \ldots, C_m\}$ be a collection of cameras, meaning a set of m 3×4 matrices that project points and lines in $\mathbb{P}^3$ to their images in the cameras. Note that this map is not linear affinely.



The *Grassmannian* $Gr(k,n)$ is the set of linear k-dimensional subspaces of $\mathbb{P}^n$ **[3]**.

$$\Upsilon_C : \mathrm{Gr}(1,3) \dashrightarrow \mathrm{Gr}(1,2)^m,$$
$$L \mapsto (\ell_1, \ldots, \ell_1) = (C_1 L, \ldots, C_m L).$$

The image of Υ_C describes the geometry of projected lines in the camera planes. Its Zariski closure is the smallest variety containing it, and is called the *multi-view variety*.

The point case is well-understood **[4]**.

A computer can understand an algebraic variety as a finite set of generators of a *defining ideal* in $\mathbb{R}[x_1, \ldots, x_n]$.

There is a one-to-one correspondence sending an ideal $\mathcal{I}$ to the variety $\mathcal{V}(\mathcal{I}) = \{x \in \mathbb{R}^n : f(x) = 0 \ \ \forall f \in \mathcal{I}\}$

Algebraic vision can help practitioners in
- 1: Finding camera specifications,
- 2: Choosing point/line correspondences,
- 3: Improving speed of algorithms,
- 4: Determining robustness under noise.

Results

In joint work with E. Shehu, P. Breiding and A. Torres, we describe the multi-view variety geometrically and its defining ideal.

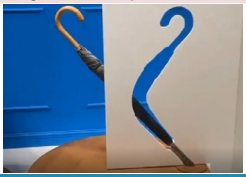
Theorem. The multi-view variety is the set of image lines whose back-projected planes meet in a line.

$$\begin{aligned}
&\text{if } m = 3, \text{ then } \mathrm{EDD} \geq 74;\\
&\text{if } m = 4, \text{ then } \mathrm{EDD} \geq 934;\\
&\text{if } m = 5, \text{ then } \mathrm{EDD} \geq 3651;\\
&\text{if } m = 6, \text{ then } \mathrm{EDD} \geq 9887;\\
&\text{if } m = 7, \text{ then } \mathrm{EDD} \geq 21807;\\
&\text{if } m = 8, \text{ then } \mathrm{EDD} \geq 42073;
\end{aligned}$$

There is numerical evidence that lines are more numerically robust than points. In this direction we consider the 1) *ED degree* **[5]**, the 2) *multi-degree* and numerical experiments.

1) The ED degree tells us how quickly we can ``denoise'' the data. The smaller the better.
2) The multi-degree tells us how ``curvy'' the variety is.

Future directions include explicit computations of ED degrees, theoretical understanding of numerical stability, and investigation of *rolling shutter cameras*.



References

1. Duff, T., Kohn, K., Leykin, A., & Pajdla, T. (2019). *Plmp-point-line minimal problems in complete multi-view visibility.*
2. Agarwal, S., Furukawa, Y., Snavely, N., Curless, B., Seitz, S. M., & Szeliski, R. (2010). *Reconstructing rome*.
3. Gathmann, A. (2002). *Algebraic geometry*. Notes for a class taught at the University of Kaiserslauten *(2002/2003)*
4. Trager, M., Hebert, M., & Ponce, J. (2015). *The joint image handbook*.
5. Draisma, J., Horobeţ, E., Ottaviani, G., Sturmfels, B., & Thomas, R. R. (2016). *The Euclidean distance degree of an algebraic variety*.

Šehić, Kenan
Lund University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

LassoBench:
A High-Dimensional Hyperparameter Optimization Benchmark Suite for Lasso

Even though Weighted Lasso regression has appealing statistical guarantees, it is typically avoided due to its complex hyperparameter space described with thousands of hyperparameters. On the other hand, the latest progress with high-dimensional HPO methods for black-box functions demonstrates that high-dimensional applications can indeed be efficiently optimized. Despite this initial success, the high-dimensional hyperparameter optimization (HPO) approaches are typically applied to synthetic problems with a moderate number of dimensions which limits its impact in scientific and engineering applications. To address this limitation, we propose LassoBench, a new benchmark suite tailored for an important open research topic in the Lasso community that is Weighted Lasso regression. LassoBench consists of benchmarks on both well-controlled synthetic setups (number of samples, SNR, ambient and effective dimensionalities, and multiple fidelities) and real-world datasets, which enable the use of many flavors of HPO algorithms to be studied and extended to the high-dimensional Lasso setting. We evaluate 6 state-of-the-art HPO methods and 3 Lasso baselines, and demonstrate that Bayesian optimization and evolutionary strategies can improve over the methods commonly used for sparse regression while highlighting limitations of these frameworks in very high-dimension and noisy settings. Remarkably, TuRBO and CMA-ES improve the Lasso baselines on 60, 100, 300, and 1000 dimensional synthetic benchmarks, and the real-world benchmark based on the leukemia dataset by 42.3%, 23%, 22.3%, 12.6% and 75%, respectively.

Šehić, Kenan
Lund University

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

LassoBench: A High-Dimensional HPO Benchmark Suite for Lasso

Kenan Šehić, Lund University (Computer Science)

Join work with Alexandre Gramfort[1], Joseph Salmon[2] and Luigi Nardi[3,4]

1-Université Paris-Saclay, Inria, CEA, France 2-IMAG, Université de Montpellier, CNRS,Montpellier, France 3-Lund University, Sweden 4-Stanford University, (USA)

Summary

Even though Weighted Lasso regression has appealing statistical guarantees, it is typically avoided due to its complex hyperparameter space described with thousands of hyperparameters. On the other hand, the latest progress with high-dimensional HPO methods for black-box functions demonstrates that high-dimensional applications can indeed be efficiently optimized. Despite this initial success, the high-dimensional hyperparameter optimization (HPO) approaches are typically applied to synthetic problems with a moderate number of dimensions which limits its impact in scientific and engineering applications. To address this limitation, we propose **LassoBench**, a new benchmark suite tailored for an important open research topic in the Lasso community that is Weighted Lasso regression. LassoBench consists of benchmarks on both well-controlled synthetic setups (number of samples, SNR, ambient and effective dimensionalities, and multiple fidelities) and real-world datasets, which enable the use of many flavors of HPO algorithms to be studied and extended to the high-dimensional Lasso setting. We evaluate 6 state-of-the-art HPO methods and 3 Lasso baselines, and demonstrate that Bayesian optimization and evolutionary strategies can improve over the methods commonly used for sparse regression while highlighting limitations of these frameworks in very high-dimension and noisy settings. Remarkably, **TuRBO** [Eriksson, 2019] and **CMA-ES** [Hansen, 2016] improve the Lasso baselines on 60, 100, 300, and 1000 dimensional synthetic benchmarks, and the real-world benchmark based on the *leukemia* dataset by 42.3%, 23%, 22.3%, 12.6% and 75%, respectively.

LassoBench

We introduce a benchmark suite called **LassoBench** that addresses the limitations of current high-dimensional optimization benchmarks found in the literature while providing an opportunity for AutoML researchers to help advance Lasso research. New insights from the AutoML community will reflect directly on Lasso applications, whose seminal paper has so far been cited more than 40,000 times [Tibshirani, 1996].

LassoBench revolves around the non-convex optimization problem named Weighted Lasso regression, where the objective is to improve a linear model by optimizing the hyperparameters λ_j of the penalty term that promotes the sparsity in regression coefficients β [Bertrand, 2020]. The challenge is that the penalty term is defined typically in a high-dimensional setting (e.g., d=10^6) .

$$\boldsymbol{\beta}^*(\boldsymbol{\lambda}) \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \frac{1}{2n}\|\boldsymbol{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{d} e^{\lambda_j}|\beta_j|$$

LassoBench exposes a number of features, such as both noisy and noise-free benchmarks, well-defined effective dimensionality subspaces, and multiple fidelities, which enable the use of many flavors of Bayesian optimization algorithms to be improved and extended to the high-dimensional setting.

LassoBench includes the baselines commonly used in the Lasso community such as LassoCV [Massias, 2018], AdaptiveLassoCV [Massias, 2018] and Sparse-HO [Bertrand, 2020] for the comparison.

Benchmark Name	# Samples n	Ambient Dimensions d	Effective Dimensions d_e
synt_simple	30	60	3
synt_medium	50	100	5
synt_high	150	300	15
synt_hard	500	1000	50

Table 1 Predefined synthetic benchmarks in LassoBench when the true regression coefficients β_{true} are known.

Benchmark Name	# Samples n	Ambient Dimensions d	Approx. Effective Dimensions $\bar{d}_e$
breast_cancer	683	10	3
diabetes	768	8	5
leukemia	72	7,129	22
dna	2,000	180	43
rcv1	20,242	19,959	75

Table 2 Real-world benchmarks in LassoBench. d_e is derived with Sparse-HO as $\bar{d}_e = \|\bar{\beta}\|_0$.

For a simple 4-line tutorial on how to run LassoBench follow github.com/ksehic/LassoBench

References

1. R. Tibshirani. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267–288. 1996
2. Q. Bertrand, Q. Klopfenstein, M. Blondel, S. Vaiter, A. Gramfort, and J. Salmon. Implicit differentiation of Lasso-type models for hyperparameter optimization. In Proceedings of the 37th International Conference on Machine Learning, pages 119:810–821, 2020
3. D. Eriksson, M. Pearce, J. Gardner, R. D. Turner, and M. Poloczek. Scalable global optimizationvia local Bayesian optimization. In Advances in Neural Information Processing Systems, pages 5496–5507, 2019
4. M. Massias, A. Gramfort, and J. Salmon. Celer: a Fast Solver for the Lasso with Dual Extrapolation. In ICML, volume 80, pages 3315–3324, 2018.
5. Hansen N. The CMA Evolution Strategy: A Tutorial. ArXiv e-prints, arXiv:1604.00772 [cs.LG], 2016.

Results

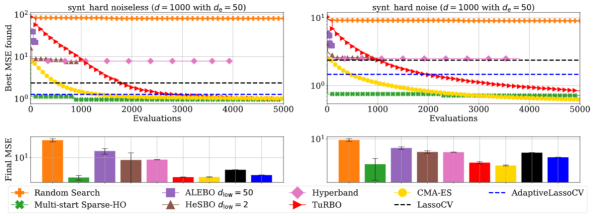


Figure 1 Baselines and HPO algorithms comparisons on synt_hard, left and right are noiseless and noisy, respectively. The bottom subplots show the best found estimations from each method, with confidence interval (for random methods) defined by one standard deviation out of 30 replications.

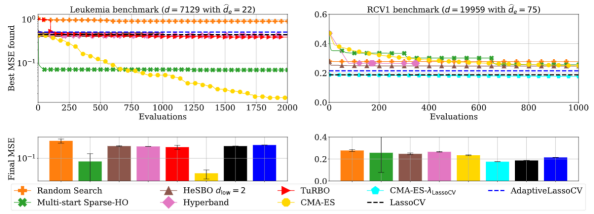


Figure 2 Comparison between the Lasso baselines and the HD-HPO methods for the Leukemia benchmark (left) and the RCV1 benchmark (right). The bottom subplot includes the best found MSE from each method and confidence intervals for random methods defined by one standard deviation out of 30 replications.

Method	Noise	synt_simple (d=60) (N=1000)	synt_medium (d=100) (N=1000)	synt_high (d=300) (N=5000)	synt_hard (d=1000) (N=5000)	Leukemia (d=7,129) (N=2000)	RCV1 (d=19,959) (N=1000)
LassoCV	False	4.73	1.67	2.48	2.37	0.44	0.18
	True	4.58	1.65	2.48	2.38	NA	NA
Adaptive LassoCV	False	2.06	1.52	1.18	1.27	0.51	0.21
	True	7.98	2.48	1.32	1.46	NA	NA
Multi-start Sparse-HO	False	0.697 ± 0.34	1.23	1.11 ± 0.92	**0.96** ± 0.27	0.06 ± 0.1	0.25 ± 0.17
	True	0.59 ± 0.31	0.73 ± 0.49	0.76 ± 0.37	0.71 ± 0.58	NA	NA
Random Search	False	67.22 ± 58.9	60.68 ± 35.5	69.41 ± 21.3	78.45 ± 13.6	0.85 ± 0.21	0.27 ± 8e-3
	True	8.31 ± 6.9	7.93 ± 3.6	8.83 ± 2.0	8.96 ± 1.1	NA	NA
CMA-ES	False	**0.695** ± 0.08	1.07 ± 0.06	0.96 ± 0.03	1.00 ± 0.02	**0.015** ± 7e-3	0.23 ± 3e-3
	True	0.34 ± 0.1	**0.48** ± 0.08	0.64 ± 0.06	**0.62** ± 0.03	NA	NA
CMA-ES λ_{LassoCV}	False	NA	NA	NA	NA	NA	**0.17** ± 2e-4
	True	NA	NA	NA	NA	NA	NA
ALEBO	False	14.59 ± 26.1	18.16 ± 14.1	21.68 ± 18.4	21.84 ± 7.1	NA	NA
	True	4.95 ± 3.5	4.48 ± 2.6	4.75 ± 1.7	3.89 ± 0.5	NA	NA
HeSBO	False	3.20 ± 0.2	1.74 ± 0.2	2.66 ± 0.2	7.57 ± 10.0	0.45 ± 2e-2	0.24 ± 7e-3
	True	3.56 ± 0.6	1.74 ± 0.1	2.82 ± 0.4	2.56 ± 0.3	NA	NA
Hyperband	False	1.52 ± 0.3	4.53 ± 3.2	5.38	7.87	0.43	0.26 ± 2e-3
	True	1.44 ± 0.3	1.94	2.49	2.51	NA	NA
TuRBO	False	0.78 ± 0.7	**0.95** ± 0.1	**0.90** ± 0.03	1.01 ± 0.02	0.39 ± 9e-2	NA
	True	**0.30** ± 0.07	0.55 ± 0.1	**0.59** ± 0.1	0.84 ± 0.09	NA	NA

Table 3 Best-found MSE obtained for all optimizers, the synthetic benchmarks with both conditions (noiseless and noisy) and the real-world benchmarks based on the leukemia dataset and RCV1. We report means and standard deviation across 30 runs of each optimizer with N as the number of evaluations. For each benchmark, bold face indicates the best MSE.

For more details, follow our preprint via arxiv.org/abs/2111.02790

Sharma, Abhijat
Linköping University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Sparsification of Infinite-domain CSPs

Many problems encountered in computer science and mathematics can be viewed as CSPs: for example in spatio-temporal reasoning, computer vision, machine learning, scheduling, and bioinformatics, and this makes the CSP a problem of central importance.
The goal of this research is to study the complexity of constraint satisfaction problems (CSPs) over infinite domain.
There has been a lot of results for sparsification of finite domain CSPs, but not as many in the case of infinite-domain CSPs.
It is an important direction of research since a domain of infinite size can capture many problems encountered in AI and logical reasoning, that cannot be formalised in finite-domain.
We aim to construct faster algorithms and methods that may be useful to analyse the kernelisation of parameterized versions of infinite-domain CSPs.

Sharma, Abhijat
Linköping University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Sparsification of Infinite-domain CSPs

Abhijat Sharma, PhD Student, Linköping University
TCSLab, IDA
Supervisors: Prof. Peter Johnsson (LiU) and Dr. Victor Lagerkvist (LiU)

LINKÖPINGS UNIVERSITET

Motivation & Research Goals

Many problems encountered in computer science and mathematics can be viewed as CSPs: for example in spatio-temporal reasoning, computer vision, machine learning, scheduling, and bioinformatics, and this makes the CSP a problem of central importance. The goal of this research is to study the complexity of constraint satisfaction problems (CSPs) over infinite domain. There has been a lot of results for sparsification of finite domain CSPs, but not as many in the case of infinite-domain CSPs. It is an important direction of research since a domain of infinite size can capture many problems encountered in AI and logical reasoning, that cannot be formalised in finite-domain. We aim to construct faster algorithms and methods that may be useful to analyse the kernelisation of parameterized versions of infinite-domain CSPs.

Background

What is a CSP?

An instance of the constraint satisfaction problem consists of the following as input:

- A set of variables V and a domain D of allowed values for the variables.
- A set of constraints C imposing certain restrictions on the value assignments to the variables.

The solution to a CSP is a value-assignment $f : V \mapsto D$ such that all constraints are satisfied. The most prominent complexity-theoretic questions concerning CSPs is the following: given a set of relations Γ (the constraint language), what is the complexity of CSP(Γ), i.e. the CSP where the constraints contain only the relations from Γ.

Finite Domain Dichotomy: Bulatov([4]) and Zhuk([1]) independently proved the long-standing conjecture: a finite-domain CSP(Γ) is either polynomial-time solvable or NP-complete.

Infinite-domain CSPs are undecidable in general, however there exist many well-understood fragments which admit complexity dichotomies. There seems to be significant variance in the time-complexity of several CSPs, even if most of them are NP-complete. Even though there is vast research on fine-grained complexity of finite-domain and infinite-domain CSPs([2]), it is interesting to study the fine-grained complexity of infinite-domain CSPs restricted to certain kinds of constraints.

Sparsification and Kernelisation

Kernelisation is a pre-processing algorithm that takes an input instance and reduces it to an smaller equivalent instance, called a "kernel". Our goal is to compute efficient kernels for CSPs parameterized by the number of input variables.
In the context of CSPs, we achieve kernelisation by efficiently reducing the number of constraints in terms of the number of variables, while preserving the solution. This is also known as sparsification of CSP instances.

References

[1] A Proof of the CSP Dichotomy Conjecture
Zhuk, Dmitriy
Foundations of Computer Science (FOCS), 2017

[2] A Survey on the Fine-grained Complexity of Constraint Satisfaction Problems Based on Partial Polymorphisms
Couceiro, Miguel and Haddad, Lucien and Lagerkvist, Victor
Journal of Multiple-Valued Logic and Soft Computing, 2020

[3] The Complexity of Equality Constraint Languages
Bodirsky, Manuel and Kára, Jan
Theory of Computing Systems, July 2008

[4] A Dichotomy Theorem for Nonuniform CSPs
Bulatov, Andrei A.
Foundations of Computer Science (FOCS), 2017

[5] Optimal Sparsification for Some Binary CSPs Using Low-degree Polynomials
Jansen, Bart M. P. and Pieterse, Astrid
Mathematical Foundations of Computer Science (MFCS), 2016

Preliminary Results

Equality CSPs

For initial results, we focus on a specific class of CSPs, where the constraint language Γ only consists of equality relations.([3])

A relation $R \subseteq N^k$ is an equality relation of arity k if it can be defined as $R = \{(x_1, x_2, ..., x_k) : \phi(x_1, x_2, ..., x_k)\}$ where ϕ is a first-order formula over the structure $(N; =)$.

When a constraint language Γ contains only relations of arity at-most k, there exists a trivial sparsification of any CSP(Γ) instance to $O(n^k)$ constraints. Our goal is to either achieve sparsification that is better than this bound, or prove that such sparsification does not exist.

Kernel Lower Bounds

One of the most powerful tools used to analyse complexity of finite-domain CSPs is the standard *algebraic approach*. This involves constructing a framework that allows polynomial-time solution-preserving reductions between constraint languages, and their associated CSPs.
We have introduced algebraic methods that extend the above framework for obtaining stronger lower bounds on kernel size. The following result makes use of QFPP-definitions and additionally some novel reduction techniques inspired from the existing framework.

Lower Bound: Let Γ be an equality language such that CSP(Γ) is NP-hard. Then CSP(Γ) admits no kernel of size $O(n^{2-\epsilon})$ where n is the number of variables and $\epsilon > 0$, unless NP $\subseteq$ co-NP/poly.

Sparsification Techniques

We have introduced new sparsification methods based on the *linear-algebraic framework* of viewing constraints as low-degree polynomials([5]). We have applied these ideas to equality constraints and obtained optimal results in certain cases. A natural research direction now is to better understand sparsification of equality constraints, and preferably obtain optimal bounds for all equality languages. This will need the development of even more powerful methods.

Beyond Equality Relations

Apart from equality relations, our aim is to generalise the linear-algebraic techniques to more interesting cases such as temporal constraints over the domain of rational numbers. For example, consider the following well-studied relation, used for gene mapping in bioinformatics.

Betweenness: $B = \{(x, y, z) \in Q^3 \mid x < y < z \text{ or } z < x < y\}$

Our techniques allow us to sparsify both the above relations to a kernel of $O(n^2)$ constraints. This is encouraging as it illustrates that the our methods are applicable to CSPs far beyond equality relations. Our goal is to utilize these sparsification techniques to more complex spatio-temporal constraints used in AI, for instance the Allen's Interval Algebra and RCC-8 Calculus.

Toft, Albin
KTH

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Scalable Causal Inference in Mass Media

This project centers around the theory of causality and its applications to the complex data sets arising from social media platforms, mass media and the financial market. The primary industrial objective of the project is to gain a systematic understanding of the cause-effect network underlying (i) events reported in mass media, (ii) individual interactions in social media and (iii) measurable financial and economic indicators in the globally-coupled markets. As a first step in this direction, we are exploring a different approach to causal inference in time dependent data using Hawkes processes in contrast to the more classical time series approach.

Toft, Albin

KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Scalable Causal Inference in Mass Media

Albin Toft, Ind. PhD, Combient Mix and KTH
Dept. Statistics, Causal Inference in Financial Time Series
Supervisors: Liam Solus (KTH), Raazesh Sainudiin (Combient Mix) and Tatjana Pavlenko (UU)

Motivation & Research Goals

This project centers around the theory of causality and its applications to the complex data sets arising from social media platforms, mass media and the financial market. The primary industrial objective of the project is to gain a systematic understanding of the cause-effect network underlying (i) events reported in mass media, (ii) individual interactions in social media and (iii) measurable financial and economic indicators in the globally-coupled markets. As a first step in this direction, we are exploring a different approach to causal inference in time dependent data using Hawkes processes in contrast to the more classical time series approach.

Granger Causality & Time Series Analysis

[1] When considering causality in time series, Granger causality and its variations are popular approaches. Consider a multivariate time series $(\mathbf{X}_t)_{t\in Z}$, such that the induced joint distribution is faithful with respect to the corresponding full time graph. Then the summary graph has an arrow $X^j \to X^k$ if and only if there exist a $t \in Z$ such that

$$X_t^k \not\perp X_{past(t)}^j | X_{past(t)}^{-j}.$$

Typically, one way of determining whether one time series Granger causes another, is by modelling the multivariate time series as vectorized auto regressive (VAR) processes

$$\mathbf{X}_t = \mathbf{c} + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{X}_{t-i} + \epsilon_t.$$

The task of determining the Granger causal relationships among the time series, boils down to assessing which elements of the matrices $\mathbf{A}_i, i = 1...p$ are non-zero.

Hawkes Processes

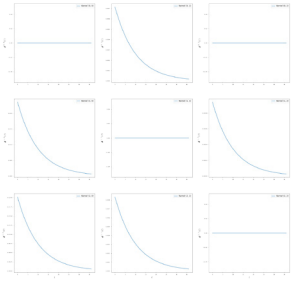
[2] A multi-dimensional Hawkes process is a counting process, where the intensity of each separate counting process at time t can be written as

$$\lambda_i(t) = \mu_i + \sum_{j=1}^{D} \sum_{t_k^j < t} \phi_{ij}(t - t_k^j).$$

The $\phi_{ij}(t)$ functions are called kernel functions, and typically one uses exponential kernels where

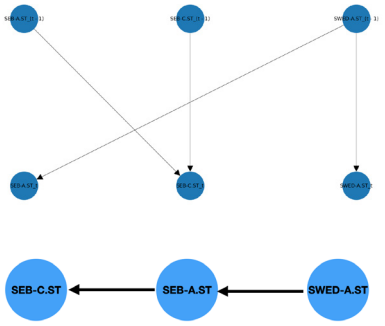
$$\phi_{ij}(t) = \alpha_{ij}\beta_{ij} exp\{-\beta_{ij} t\}.$$

These kernels can be used to describe how events of type j might increase the intensity of events of type i occurring.



Selected Results

VAR-Results



Modelling the log-returns of the stocks SEB-A.ST, SEB-C.ST and SWED-A.ST, as VAR processes the current results show that some Granger causal connections between the time series can be found.

Hawkes-Results

By instead estimating trend reversal points in the time series, it was possible to extract event sequences for the stock prices. The interactions between these event sequences were then modelled using multi dimensional Hawkes processes with exponential kernels.

Planned Work

Most of the research is still to be done in this project, and some ideas are:

- Model expansion: Extend the models to analyse more stocks and over a longer time period.
- Mass media action as intervention: Include news and mass-media data into the analysis.

References

[1] Elements of Causal Inference
J.Peters, D.Janzing, B.Schölkopf
The MIT Press
2017

[2] Learning Granger Causality for Hawkes Processes
H.Xu, M. Farajtabar, H.Zha
2018

Tombari, Francesca
KTH

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Homotopical decompositions of simplicial and Vietoris-Rips complexes

When we decompose a simplicial complex and reassemble it, it might happen that the resulting complex has a different homotopy type from the initial one. However, it is sometimes possible to understand this change by looking at subcomplexes living in the intersection of the two decomposing pieces, the so called obstruction complexes. In this poster it is outlined how the homotopy type of a simplicial complex is related to the one of its decompositions. It is also explained with an example how to use these ideas to find out the homotopy type of given Vietoris-Rips complexes. This is a joint work with Wojciech Chachólski, Martina Scolamiero and Alvin Jin.

Tombari, Francesca

KTH

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Homotopical decompositions of Vietoris-Rips complexes

Wojciech Chachólski, Alvin Jin, Martina Scolamiero, Francesca Tombari

Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden

① Introduction

The use of algebraic topology is rapidly growing in understanding data. The general pipeline of TDA can be summarized by the following:

Dataset - Points with relations between them

⇓

Geometrical structure - Simplicial complex

⇓

Algebraic invariants - Homology

⇓

Features for statistics and machine learning

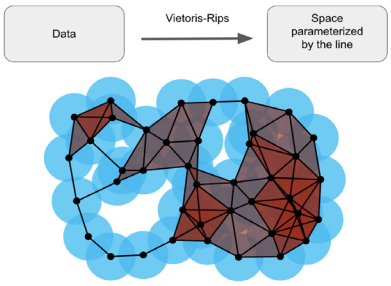


Figure 1: An example of a Vietoris-Rips complex, given some $r > 0$.

The computational challenges motivates the following type of question: given a decomposition of our data set $Z = X \cup Y$, what information can we recover about the Vietoris-Rips complex of Z from the component Vietoris-Rips complexes?

② A general approach

Let K be a simplicial complex, $K_0 = X \cup Y$ the set of its vertices and $A = X \cap Y$. Let $K_X = K \cap X$ and $K_Y = K \cap Y$. One can easily notice that the union of K_X and K_Y does not give the initial simplicial complex K. A natural question that might arise is whether the following inclusion is a homotopy equivalence or not

$$K_X \cup K_Y \hookrightarrow K.$$

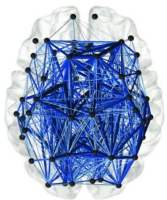


Figure 2: Example of a simplicial complex with high complexity. (Image courtesy of the authors of arXiv:1608.03520)

A special case of this problem occurs when a pseudo-metric space (Z, d) is considered. Fixing $r > 0$ and a covering of Z consisting in two subspaces X and Y, we get the inclusion

$$VR_r(X) \cup VR_r(Y) \hookrightarrow VR_r(Z).$$

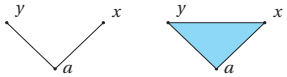


Figure 3: The two figures show a simplicial complex K (on the right) and $K_X \cup K_Y$ (on the left), where $X = \{x, a\}$ and $Y = \{y, a\}$.

③ Main result

We define the obstruction complex:

$$F(\sigma, A) := \{\mu \subset A \mid \mu \cup \sigma \in K\}.$$

Theorem. Let $\mathscr{C}$ be a closed collection of simplicial sets. If, for every σ in $\{\sigma \in K \mid \sigma \cap X \neq \emptyset$ and $\sigma \cap Y \neq \emptyset$ and $\sigma \cap A = \emptyset\}$, the simplicial complex $F(\sigma, A)$ satisfies $\mathscr{C}$, then the homotopy fibers of $K_X \cup K_Y \subset K$ also satisfy $\mathscr{C}$.

Corollary. If, for every σ as above, the simplicial complex $F(\sigma, A)$ is contractible, then $K_X \cup K_Y \subset K$ is a weak equivalence.

We get a long exact sequence in the case when adding one vertex:

$$H_n(F(x, A)) \to H_n(K_A) \to H_n(K) \to H_{n-1}(F(x, A)) \to H_{n-1}(K_A)$$

and another when adding two vertices:

$$H_n(\Sigma F(x, y, A) \to H_n(K_X \cup K_Y) \to H_n(K) \to H_{n-1}(\Sigma F(x, y, A) \to H_{n-1}(K_X \cup K_Y).$$

These sequences give information about the global homology of K with respect to local information.

④ Examples

Consider the metric space $Z = \{x_1, x_2, a_1, a_2, y\}$, with the metric such that every two points of Z has distance 1 except for x_1, a_2 and x_2, a_1 having distance 1.1. Let $X = \{x_1, x_2, a_1, a_2\}$, $Y = \{y, a_1, a_2\}$ be a cover for Z. We can easily see that $VR_1(X) \cup VR_1(Y)$ has the homotopy type of S^1, while $VR_1(Z)$ is contractible. This is due to the fact that $F(\sigma, A)$ is empty, hence non-contractible, when σ is the 2-simplex with vertices x_1, x_2 and y.

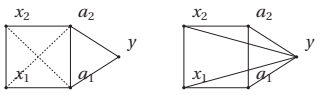


Figure 4: $K_X \cup K_Y$ on the left and K on the right. Notice that all the triangles in this example are filled, because K is a clique complex.

The following picture shows an example of a decomposition of $Z = \{x_1, x_2, y_1, y_2, a_{11}, a_{12}, a_{21}, a_{22}\}$ that has the same homology as the total simplicial complex up to degree 2, but different H_3.

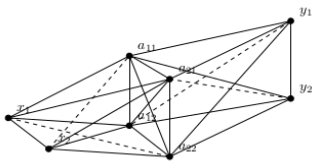


Figure 5: The figure represents a 2-dimensional visualization of the Vietoris-Rips complex $VR_r(X) \cup VR_r(Y)$. $VR_r(X \cup Y)$ is obtained by the above simplicial complex adding the simplex $\{x_1, x_2, y_1, y_2\}$

The metric is given by:

$$\begin{aligned}
&d(a_{11}, a_{21}) = d(a_{11}, a_{12}) = d(a_{21}, a_{22}) = d(a_{12}, a_{22}) = 4,\\
&d(a_{11}, a_{22}) = d(a_{12}, a_{21}) = 6,\\
&d(x_1, a_{11}) = d(y_1, a_{21}) = d(x_2, a_{22}) = d(y_2, a_{12}) = 3,\\
&d(x_1, a_{12}) = d(y_1, a_{11}) = d(x_2, a_{21}) = d(y_2, a_{22}) = 5,\\
&d(x_1, a_{21}) = d(y_1, a_{22}) = d(x_2, a_{12}) = d(y_2, a_{11}) = 7,\\
&d(x_1, a_{22}) = d(y_1, a_{12}) = d(x_2, a_{11}) = d(y_2, a_{21}) = 9,\\
&d(x_1, x_2) = d(y_1, y_2) = 6,\\
&d(x_1, y_1) = d(x_1, y_2) = d(x_2, y_1) = d(x_2, y_2) = 8.
\end{aligned}$$

As we have already noticed, the study of this problem for Vietoris-Rips complexes is actually a consequence of the same problem stated for generic simplicial complexes. Analogously, the conditions that we put on a metric space are just a translation of hypothesis on simplicial complexes.

⑤ References

[1] W. Chacholski, A. Jin, M. Scolamiero, and F. Tombari. Homotopical decompositions of simplicial and Vietoris–Rips complexes. J Appl. and Comput. Topology 5, 215–248 (2021). https://doi.org/10.1007/s41468-021-00066-2

[2] Adamaszek et. al. On Homotopy Types of Vietoris–Rips Complexes of Metric Gluings. Proceedings of the 34th Symposium on Computational Geometry (2018), 3:1-3:15.

Upadhyaya, Manu
Lund University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Performance estimation of iterative algorithms and closed-loop systems

The aim of this research is to combine ideas from the performance estimation problem (PEP) framework in the optimization literature and integral quadratic constraints (IQC) framework from the control theory literature into a novel computer-aided automated Lyapunov analysis framework. Applications that we are considering are 1) the analysis of the worst-case performance of optimization algorithms and iterative algorithms in general and 2) the stability verification of neural network controlled systems and model predictive control (MPC) schemes. Moreover, besides the analysis in 1) and 2), the framework allows for a systematic approach to optimize algorithm or system performance with respect to design parameters.

Upadhyaya, Manu
Lund University

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Performance estimation of iterative algorithms and closed-loop systems

Manu Upadhyaya, Lund University
Department of Automatic Control
Main advisor: Pontus Giselsson

www.control.lth.se

Motivation & research goals

The aim of this research is to combine ideas from the performance estimation problem (PEP) framework in the optimization literature and integral quadratic constraints (IQC) framework from the control theory literature into a novel computer-aided automated Lyapunov analysis framework. Applications that we are considering are 1) the analysis of the worst-case performance of optimization algorithms and iterative algorithms in general and 2) the stability verification of neural network-controlled systems and model predictive control (MPC) schemes. Moreover, besides the analysis in 1) and 2), the framework allows for a systematic approach to optimize algorithm or system performance with respect to design parameters.

Methods

Performance estimation problem (PEP)

The PEP framework, first introduced in [1], provides a systematic method to analyze the worst-case performance of optimization algorithms, and iterative algorithms in general. Roughly speaking, in the optimization algorithm case and in the basic set-up, we have the following components:

- An appropriate class of functions $\mathcal{F}$ with members $f: \mathcal{H} \to \mathbb{R} \cup \{\pm\infty\}$ for some real Hilbert space $\mathcal{H}$.
- Each function $f \in \mathcal{F}$ has a minimizer x^*.
- Some fixed oracle $\mathcal{O}_f(x)$ that provides information about f at x. This could include the function value, and/or the gradient, etc.
- Some initial iterate $x_0 \in \mathcal{H}$.
- A fixed algorithm $\mathcal{A}$ that is allotted N iterations such that it generates a sequence

$$x_1 = \mathcal{A}_1(x_0, \mathcal{O}_f)$$
$$x_2 = \mathcal{A}_2(x_0, x_1, \mathcal{O}_f)$$
$$\vdots$$
$$x_N = \mathcal{A}_N(x_0, x_1, \cdots, x_{N-1}, \mathcal{O}_f).$$

- An appropriate performance metric $\mathcal{P}(x^*, x_0, x_1, \cdots, x_N, \mathcal{O}_f)$. Some simple examples include function value suboptimality $f(x_N) - f(x^*)$, norm of gradient $\|\nabla f(x_N)\|$ and distance to an optimal solution $\|x^* - x_N\|$.

The performance estimation problem (PEP) is then to find the worst-case performance: I.e., maximize

$$\mathcal{P}(x^*, x_0, x_1, \cdots, x_N, \mathcal{O}_f)$$

subject to

$$f \in \mathcal{F}$$
$$x^* \text{ is optimal for } f$$
$$x_1, \cdots, x_N \text{ is generated by } \mathcal{A} \text{ with initial point } x_0$$

with some additional technical assumptions such that the problem becomes well-posed.

Calculating the worst-case performance is in general an infinite-dimensional optimization problem. Luckily, there exist standard techniques in the PEP literature that render the problem tractable by transforming it into a finite-dimensional semidefinite program and do so tightly via so called interpolation conditions. See [2] for additional details. Moreover, the framework allows to select "optimal" design parameters in the algorithm $\mathcal{A}$ by minimizing the worst-case performance of $\mathcal{A}$.

Recently in [3], this framework has been adapted to finding tight contraction factors of fixed-point operators used in splitting schemes to solve monotone inclusion problems.

Integral quadratic constraints (IQC)

The IQC framework, see [4], can be used to analyze the case of a linear system interconnected in feedback to a, possibly uncertain, nonlinear system. In particular, [5] noticed that the IQC framework can be used in the analysis and design of optimization algorithms and [6] highlighted the close connection to the PEP framework. Moreover, in [7], tools from the IQC framework were used to develop a method to certify asymptotic stability of neural network-controlled systems.

Current work

We are currently considering optimization algorithms and splitting schemes that:

- Can be written as a linear system with a nonlinear feedback given by some operator, in accordance with the IQC framework.
- The operators involved have interpolation conditions that only involve quadratic inequalities enabling the use of the PEP framework.
- A quadratic Lyapunov function ansatz to obtain worst-case performance guarantees and extract either linear or sublinear rates of convergence.

Concurrently, within the same framework, we are considering:

- Analyzing the stability of MPC schemes given an iteration count constraint on the optimization algorithm.
- The stability verification and training of neural network-controlled systems.

References

1. Y. Drori and M. Teboulle (2014). Performance of first-order methods for smooth convex minimization: a novel approach. Mathematical Programming 145: 451-482
2. A.B. Taylor, J.M. Hendrickx and F. Glineur (2017). Smooth strongly convex interpolation and exact worst-case performance of first-order methods. Mathematical Programming 161: 307-345
3. E. K. Ryu, A. B. Taylor, C. Bergeling, and P. Giselsson (2020). Operator splitting performance estimation: tight contraction factors and optimal parameter selection, SIAM Journal on Optimization 30:3, 2251-2271
4. A. Rantzer (1997). System analysis via integral quadratic constraints. IEEE Transactions on Automatic Control 42(6): 819-830
5. L. Lessard, B. Recht and A. Packard (2016). Analysis and design of optimization algorithms via integral quadratic constraints (2016). Siam Journal on Optimization 26(1): 57-95
6. A. Taylor, B. Van Scoy, L. Lessard (2018). Lyapunov functions for first-order methods - tight automated convergence guarantees. 35th International Conference on Machine Learning, PMLR 80: 4897-4906
7. H. Yin, P. Seiler and M. Arcak (2021). Stability analysis using quadratic constraints for systems with neural network controllers. IEEE Transactions on Automatic Control

Vallin, Jonatan
Umeå University

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Geometry and Approximation of ReLU Networks

In our upcoming paper we study the geometry and approximation properties of fully-connected ReLU networks. We start by describing the structure of a standard ReLU layer by introducing a convenient partition of the input space. Using this structure, we characterize the geometry of the decision boundary for shallow networks. We end our analysis by deriving approximation results for deep ReLU networks (not presented in this poster).

Vallin, Jonatan
Umeå University

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Geometry and Approximation of ReLU Networks

UMEÅ UNIVERSITY
Jonatan Vallin, Umeå University
Department of Mathematics and Mathematical Statistics

Abstract

In our upcoming paper we study the geometry and approximation properties of fully-connected ReLU networks. More precisely, we consider networks $F\colon \mathbb{R}^d \to \mathbb{R}$ of the form $F(x) = L \circ T_N \circ \cdots \circ T_1(x)$, where each mapping $T_i\colon \mathbb{R}^d \to \mathbb{R}_+^d$ is a ReLU layer and $L\colon \mathbb{R}^d \to \mathbb{R}$ is an affine functional. Since $\mathrm{ReLU}\,(x) = \max(x, 0)$ is a piecewise linear map, the network F defines a piecewise linear function subordinate to a polygonal partition of the input space. We start by describing the structure of a standard ReLU layer by introducing a convenient partition of the input space. Using this structure, we characterize the geometry of the decision boundary $\Gamma = \{x \in \mathbb{R}^d\colon F(x) = 0\}$ for shallow networks. We end our analysis by deriving approximation results for deep ReLU networks (not presented in this poster).

Structure of ReLU Layers

A ReLU layer is a mapping $T\colon \mathbb{R}^d \to \mathbb{R}_+^d$ of the form

$$T(x) = \mathrm{ReLU}(Ax + b)$$

where $A \in \mathbb{R}^{d\times d}$ is a matrix, assumed to have full rank, with rows $a_i \in \mathbb{R}^d$ and $b \in \mathbb{R}^d$ is a vector with elements $b_i \in \mathbb{R}$. To describe the action of T, we introduce a set of dual vectors $\{a_i^*\colon i \in I\}$, where $I = \{1, \dots, d\}$, defined by the equation $a_i^* \cdot a_j = \delta_{ij}$. It follows that these vectors constitute a basis of $\mathbb{R}^d$ which will be convenient when analyzing the structure of T.

To that end, consider a partition $I_+ \cup I_- \cup I_0$ of the index set I, denoted by the three-tuple (I_+, I_-, I_0). Some of the sets may be empty as long as their union is I. For each such partition, we define

$$S_{(I_+,I_-,I_0)} = \left\{x_0 + \textstyle\sum_{i\in I_+} \alpha_i a_i^* - \sum_{i\in I_-} \alpha_i a_i^* : \alpha_i > 0\right\} \subset \mathbb{R}^d$$

where x_0 is the unique solution to $Ax = -b$. By construction, $\dim\left(S_{(I_+,I_-,I_0)}\right) = |I_+ \cup I_-|$ and if $\mathcal{I}$ is the set of all such three-tuples, the family $\mathcal{S} = \{S_{\boldsymbol{I}}\colon \boldsymbol{I} \in \mathcal{I}\}$ is a partition of $\mathbb{R}^d$ with pairwise disjoint sets. In the special case when $A = I_d$ (I_d being the identity matrix) and $b = 0$, the sets reduce to

$$\hat{S}_{(I_+,I_-,I_0)} = \left\{0 + \textstyle\sum_{i\in I_+} \alpha_i e_i - \sum_{i\in I_-} \alpha_i e_i : \alpha_i > 0\right\} \subset \mathbb{R}^d$$

where e_i is the i:th Euclidean basis vector. We call the corresponding partition $\hat{\mathcal{S}}$. Examples of the families $\hat{\mathcal{S}}$ and $\mathcal{S}$ are shown in Figure 1.

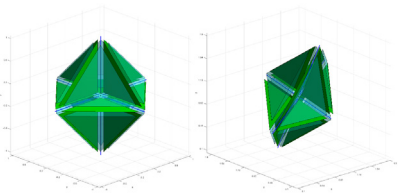


Figure 1. A visualization of the families $\hat{\mathcal{S}}$ (left) and $\mathcal{S}$ (right) when $d = 3$. Both families partition $\mathbb{R}^3$ in eight 3-dimensional sets (the transparent volumes), twelve 2-dimensional sets (the green faces), six 1-dimensional sets (the blue lines) and one 0-dimensional set (the black point in the center). The figure only shows slices of the sets and we have also intentionally added space between the sets for illustrative purposes.

Note, $\mathbb{R}_+^d = T(\mathbb{R}^d)$ can be partitioned as $\partial\hat{S} = \left\{\hat{S}_{(J,\emptyset,I\setminus J)}\colon J \subseteq I\right\}$ and it follows that the image of a set $S_{(I_+,I_-,I_0)} \in \mathcal{S}$ under T is exactly

$$T\left(S_{(I_+,I_-,I_0)}\right) = \hat{S}_{(I_+,\emptyset,I_0\cup I_-)} \in \partial\hat{S}$$

Thus, $T\left(S_{(I_+,I_-,I_0)}\right) = T\left(S_{(J_+,J_-,J_0)}\right)$ whenever $I_+ = J_+$ and T reduces to the affine map $x \mapsto Ax + b$ on the closure $\bar{S}_{(I,\emptyset,\emptyset)}$. Since $\dim\left(T\left(S_{(I_+,I_-,I_0)}\right)\right) \le \dim\left(S_{(I_+,I_-,I_0)}\right)$ it is clear that T has contracting properties. If ω is the preimage under the affine map of a set $\hat{\omega} \subset \hat{S}_{(J,\emptyset,I\setminus J)}$ then

$$T^{-1}(\hat{\omega}) = \left\{x - \textstyle\sum_{i\in I\setminus} \alpha_i a_i^* : \alpha_i \ge 0, x \in \omega \cap S_{(J,\emptyset,I\setminus J)}\right\}$$

Examples of preimages can be seen in Figure 2.

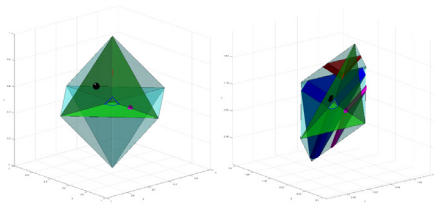


Figure 2. Four different subsets of $\mathbb{R}_+^3$ are shown (left) together with their corresponding preimages (right) under a ReLU layer T. Preimages of subsets intersecting the boundary $\partial\mathbb{R}_+^3$ will be spanned by a subset of the dual basis vectors.

Decision Boundaries

A shallow network has the form $F(x) = L \circ T(x)$ where $L\colon \mathbb{R}^d \to \mathbb{R}$ is an affine functional with a hyperplane $\hat{P}$ as its kernel. The decision boundary of F can be expressed as

$$\Gamma = T^{-1}\left(\hat{P} \cap \mathbb{R}_+^d\right) = \cup_{\hat{S}_{\boldsymbol{I}}\in\partial\hat{S}} T^{-1}\left(\hat{P} \cap \hat{S}_{\boldsymbol{I}}\right)$$

Using the structure of T, we can expand each set in the union as

$$T^{-1}\left(\hat{P} \cap \hat{S}_{(J,\emptyset,I\setminus J)}\right) = \{x - \textstyle\sum_{i\in I\setminus} \alpha_i\, a_i^* : \alpha_i \ge 0, x \in P \cap S_{(J,\emptyset,I\setminus J)}\}$$

where P is the preimage of $\hat{P}$ under the affine map $x \mapsto Ax + b$. Thus, each non-empty intersection $\hat{P} \cap \hat{S}_{(J,\emptyset,I\setminus J)}$ will generate a unique linear piece of Γ spanned by a subset of the dual vectors. Moreover, Γ is completely determined by the preimages $T^{-1}\left(\hat{P} \cap \hat{S}_{(I\setminus\{i\},\emptyset,\{i\})}\right)$, $i \in I$, the intersections of $\hat{P}$ with the $d-1$ dimensional faces in $\partial\hat{S}$. The remaining pieces are essentially linear transitions between these parts. If n is a unit normal to P, then the signs of $n \cdot a_i^*$ indicate how Γ curves since the dual vectors are tangents to the pieces $T^{-1}\left(\hat{P} \cap \hat{S}_{(I\setminus\{i\},\emptyset,\{i\})}\right)$ and n is normal to the central piece $P \cap S_{(I,\emptyset,\emptyset)}$ to which all other pieces are connected. If the hyperplane $\hat{P}$ is in general position, that is, not parallel with any of the standard coordinate axes and $0 \notin \hat{P}$ then there are $t_i \in \mathbb{R} \setminus \{0\}$ such that $t_i e_i \in \hat{P}$ for each $i \in I$. It turns out that $\mathrm{sgn}(t_i) = \mathrm{sgn}(n \cdot a_i^*)$, thus the values $\{t_i\colon i \in I\}$ determine how Γ curves.

We show that the number of linear pieces of Γ is $2^d - 2^m$ where $m = |\{i \in I\colon t_i < 0\colon\}|$. Further, we show that for a shallow ReLU network $F\colon \mathbb{R}^d \to \mathbb{R}$ each possible decision boundary can be obtained by applying an invertible affine map to one of d canonical decision boundaries. Hence, it suffices to understand the properties of these canonical decision boundaries. Figure 3 shows the canonical decision boundaries when $d = 3$.

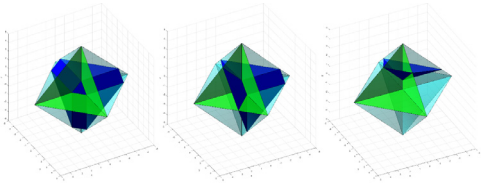


Figure 3. An illustration of the 3 canonical decision boundaries for a shallow network $F\colon \mathbb{R}^3 \to \mathbb{R}$.

Williamson, Måns
Lund university

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

A Stochastic Runge-Kutta Optimization Algorithm

Runge--Kutta--Chebyshev (RKC) methods are used to solve numerical differential equations. They have the advantage of being explicit methods with large stability regions. We propose a stochastic optimization scheme for machine learning problems based on the Runge-Kutta-Chebshev methods.

Williamson, Måns
Lund university

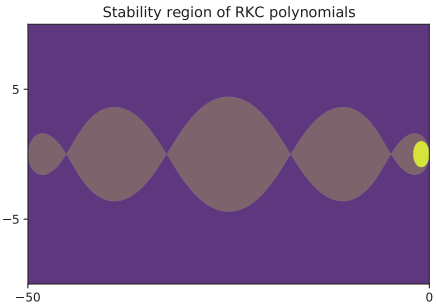
WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

A Stochastic Runge-Kutta Optimization Algorithm

Måns Williamson
Centre for Mathematical Sciences

Runge–Kutta–Chebyshev methods

Runge–Kutta–Chebyshev (RKC) methods are methods used to solve numerical differential equations. They have the advantage of being explicit methods with large stability regions.



In the plot above we see the stability region of an RKC method with 5 stages and that of the explicit Euler scheme (in bright yellow).

Gradient flow & optimization

We can view the gradient descent algorithm as the explicit Euler scheme applied to the gradient flow equation

$$\dot{w} = -\nabla F(w).$$

As we saw above, the explicit Euler scheme (and thus the gradient descent) has a small stability region which puts a severe stepsize restricition on it.

Stochastic Runge–Kutta–Chebyshev descent

We here present a stochastic optimization algorithm that make use of the RKC methods for minimizing a cost function F:

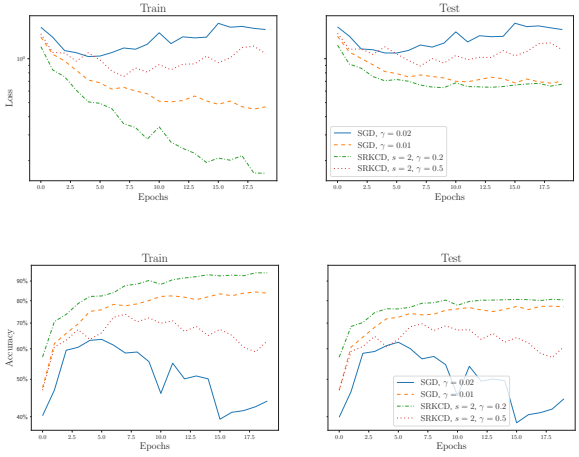
- Choose an initial iterate w_1 and a sequence of jointly independent random variables $\{\xi_k\}$. For $k = 1, 2, \ldots$
- Set $w_{k0} \leftarrow w_k$ and recieve a stochastic approximation $g(\xi_k, \cdot)$ to $\nabla F(\cdot)$.
- Set $w_{k1} = w_{k0} + \tilde{\mu}_1 \alpha_k g(\xi_k, w_{k0})$.
  For $j = 2, \ldots, s$.
  – Set $w_{kj} = \mu_j w_{k,j-1} + \nu_j w_{k,j-2} + \tilde{\mu}_j \alpha_k g(\xi_k, w_{k,j-1})$.
- Set the new iterate as $w_{k+1} \leftarrow w_{ks}$.

Under various standard assumptions (such as strong convexity of the objective function F) we can show that the sequence $\{w_k\}_{k\geq 1}$ generetated by the SRKCD algorithm converges sub-linearly in expectation. Under the assumption that F is twice differentiable we can show that the algorithm converges in expectation to a stationary point in the non-convex case:

$$\lim_{k\to\infty} \mathbb{E}\|\nabla F(w_k)\|^2 = 0.$$

Numerical experiments

Below we see the results from testing the SRKCD-scheme with 2 stages on a VGG-network using the Cifar-10 dataset.



Poster made for WASP winter conference 2022. (Joint work with Tony Stillfjord.)

Zetterqvist, Olof
Chalmers

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Regularised Weights in Statistical Models: A General Strategy for Bias Reduction and Increased Stability in Overparameterised Settings

Two challenging aspects of machine learning are label contamination in training data in supervised classification tasks and bias reduction in classical regularisation settings. Our research focuses on a general strategy to non-interactively deal with both problems by expanding the loss function with newly introduced weights. In the first article, we focus on reducing the impact of contaminated labels in training data by localising incorrect data points and reducing their contribution to the loss function. In the second article, we focus on reducing the added bias introduced by classical regularisation methods, like Lasso and Ridge, in a linear regression setting. By doing this, we can, under certain circumstances, keep key properties from the original regularisation penalty and reduce the bias giving us consistent estimators. This leads to the regularisation methods ”entropy weighted Lasso” (EWL) and ”entropy weighted Ridge” (EWR).

Zetterqvist, Olof
Chalmers

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Regularised Weights in Statistical Models

A General Strategy for Bias Reduction and Increased Stability in Overparameterised Settings

Olof Zetterqvist
Chalmers University of Technology and University of Gothenburg
Olofze@chalmers.se

CHALMERS UNIVERSITY OF TECHNOLOGY

Introduction

Two challenging aspects of machine learning are label contamination in training data in supervised classification tasks and bias reduction in classical regularisation settings. Our research focuses on a general strategy to non-interactively deal with both problems. The material and experiments are distributed as follows:

1. (Article 1) Reduce the impact of contaminated labels in training data by localising incorrect data points and reducing their contribution to the loss function in a deep convolution neural network setting.
2. (Article 2) Reduce the added bias introduced by classical regularisation methods, like Lasso and Ridge, in a linear regression setting. This leads to the regularisation methods "entropy weighted Lasso" (EWL) and "entropy weighted Ridge" (EWR).

Our approach

Our approach is to expand the loss function with more parameters ω that can be considered weights of different terms. In the presence of label noise, the weights can be put on the data loss terms, and for bias reduction, they can be put on the regularisation terms. To find the "optimal" weight setting, we increase the minimisation task to also include the weights ω with an extra regularisation term $\tilde{g}(\omega) = \sum_i (\omega_i \log(\omega_i) - \omega_i + 1)$

Using weights to find contaminated labels

$$\tilde{\theta}, \tilde{\omega} = \underset{\theta,\omega\,:\,\bar{w}_c=\rho_c,\forall c}{\arg\min} \sum_i \omega_i \ell(X_i, Y_i; \theta) + \lambda g(\theta) + \alpha \tilde{g}(\omega)$$

$$\tilde{\theta} = \underset{\theta}{\arg\min} -\alpha \sum_k \rho_k |C_k| \log\left(\sum_{i \in C_k} e^{-\frac{\ell(x_i, y_i; \theta)}{\alpha}} \right) + \lambda g(\theta)$$

Using weights to reduce bias

$$\tilde{\theta}, \tilde{\omega} = \underset{\theta,\omega}{\arg\min} \frac{1}{2} ||Y - X\theta||_2^2 + \lambda \sum_i \omega_i g(\theta_i) + \gamma \tilde{g}(\omega)$$

$$\tilde{\theta} = \underset{\theta}{\arg\min} \frac{1}{2} ||Y - X\theta||_2^2 + \gamma \sum_i (1 - e^{-\frac{\lambda}{\gamma} g(\theta_i)})$$

Results (contaminated labels)

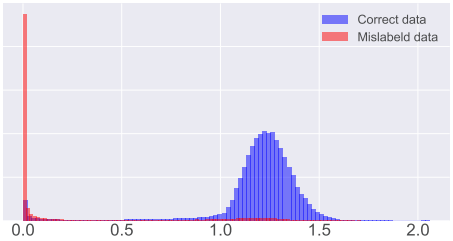


Figure 1: Histograms of the observation weight distributions of both correct and mislabelled training data.

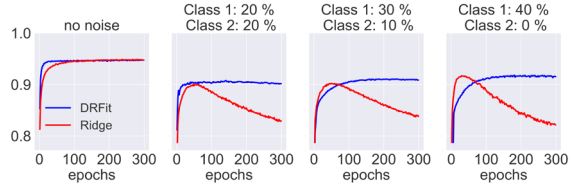
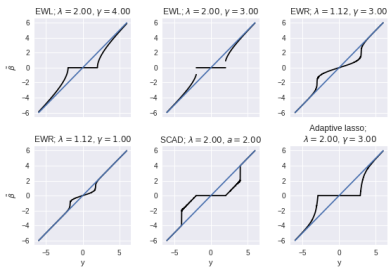


Figure 2: Validation accuracy during training in four different noise settings.

Results (bias reduction)



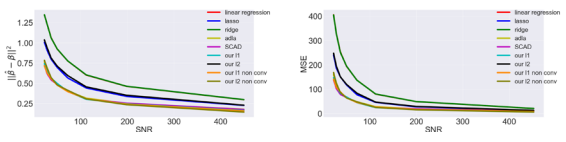


Figure 3: The average L_2 distance between the estimated parameters $\hat{\beta}$ and the true parameters β (**left**) and the mean squared error of predictions on test data (**right**) over 100 runs as functions of the signal to noise ratio (SNR) for nine models on uncorrelated covariates.

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM