Geometric properties of risk functions of neural networks

Vu Dinh
Department of Mathematical Sciences
University of Delaware

Outline

Dynamical systems

Semi-algebraic geometry

$\longrightarrow$

Optimization

Deep learning

$\longrightarrow$

Explainable

deep learning

Feature selection

Landscape of the risk function

$\longleftarrow$

Lojasiewicz's inequality

Outline

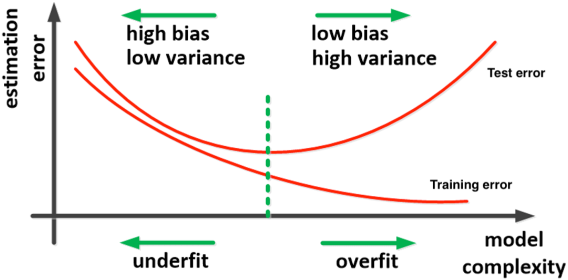
- Motivation and background
    - Feature/variable selection in modern settings
    - Linear feature selection, Lasso and Adaptive Lasso
- Geometric properties of neural networks risk functions
- Consistent feature selection for analytic deep neural networks
- Open problems and future directions

Feature/variable selection in supervised learning settings

- Given sequence of data $\{(X^{(i)}, Y^{(i)}\}_{i=1}^{n}$ where X is composed of p features
$$X = (X_1, X_2, \ldots, X_p)$$
- Question: Which features of X actually influence the outcome Y?

Feature/variable selection

- Create better model. Reduce computational cost.
- Prediction accuracy



- Model explainability

Feature selection: modern machine learning

- Deep learning: double descent phenomenon

- Big Models: PaLM (540B parameters), ViT-G/14 (2B)
- Model explainability has become increasingly more important

Belkin, Mikhail, et al. Proceedings of the National Academy of Sciences 116.32 (2019): 15849-15854.

Looking through the black box

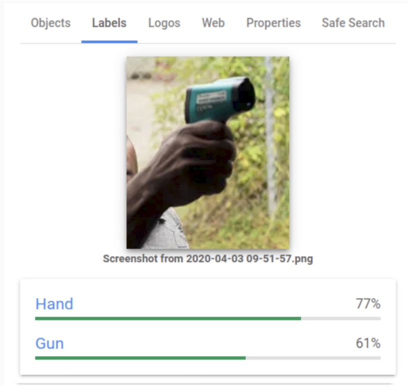
ARTIFICIAL INTELLIGENCE

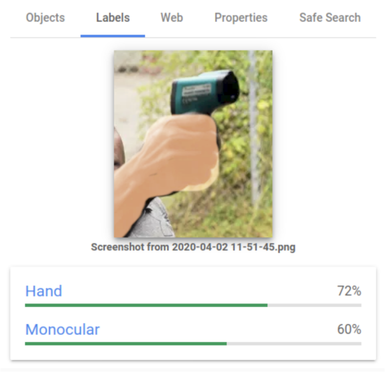
Hundreds of AI tools have been built to catch covid. None of them helped.

- some AIs were found to be picking up on the text font that certain hospitals used to label the scans
- patients scanned while lying down were more seriously ill $\rightarrow$ the AI learned to predict serious covid risk from a person's position
- some dataset contained chest scans of healthy children as negative examples $\rightarrow$ the AIs learned to identify kids

Roberts, Michael, et al. Nature Machine Intelligence 3.3 (2021): 199-217.

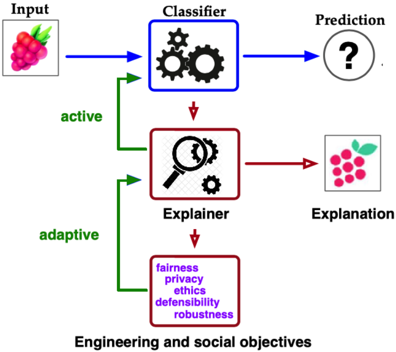
Looking through the black box





Google Vision (2020).

Research: Integrated explainable AI



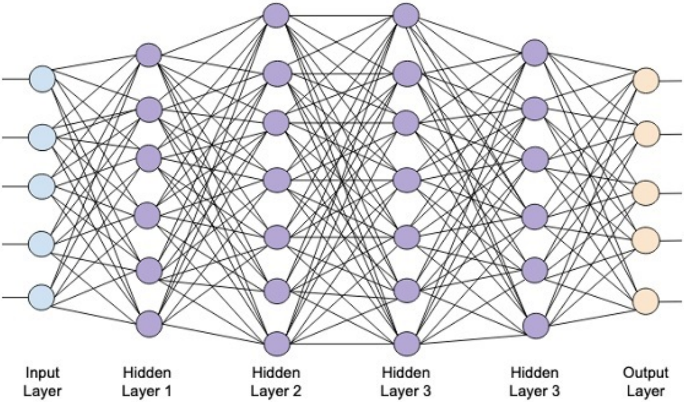
Joint works with Lam Ho (Dalhousie) and Cuong Nguyen (Florida International University)

Some other learning contexts

- Predicting walking activity post-stroke*
  - features: clinical measures of physical function, other clinical and demographic variables
  - question: which test/practice should be pursued for rehabilitation?
- Stem-cell origin of colon cancer
  - features: gene expression data from patients with/without different mutations
  - question: which component of a biological pathway causes the difference in the behavior of normal/cancer cells?

*Miller, A. E., Russell, E., Reisman, D. S., Kim, H. E., & Dinh, V. (2022). A machine learning approach to identifying important features for achieving step thresholds in individuals with chronic stroke. Plos One, 17(6), e0270105.

Question: Can we do feature selection with deep neural networks?



Linear feature selection and Adaptive Lasso

Linear model and Lasso

- Linear model

$$Y = \beta^{(0)} + \beta^{(1)}X^{(1)} + \beta^{(2)}X^{(2)} + \dots \beta^{(p)}X^{(p)} + \epsilon$$

- If $\beta^{(j)} = 0$, then the feature $X^{(j)}$ has no influence on the output
- Lasso

$$\hat{\beta}^{Lasso} = \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta^{(j)}|$$

Lasso's alternative form

- Standard form

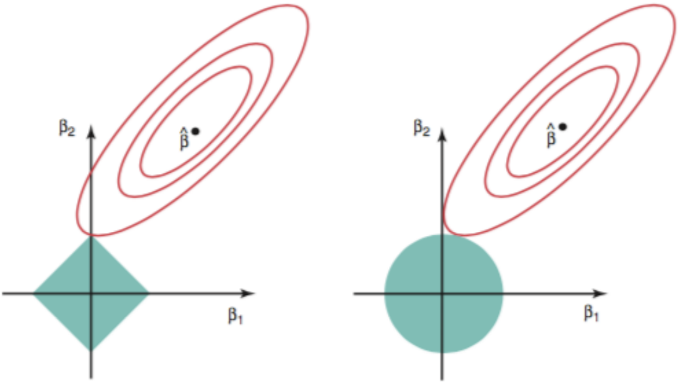
$$\hat{\beta}^{Lasso} = \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta^{(j)}|$$

- Alternative form

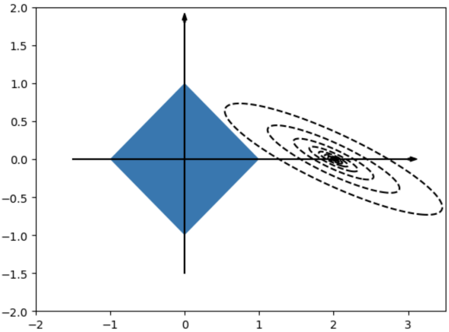
$$\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta^{(j)}| \leq s$$

Lasso vs ridge regression



When Lasso fails ($p = 2$)



- correlated features
- strong non-linearity

Irrepresentable condition

- Necessary condition for Lasso's selection consistency: There exists some sign vector s such that

$$|C_{21}C_{11}^{-1}s| \leq 1$$

where C is the (block) covariance matrix of X (1 and 2 correspond to the group of significant features and insignificant ones, respectively)

Zhao, Peng, and Bin Yu. "On model selection consistency of Lasso." The Journal of Machine Learning Research 7 (2006): 2541-2563.

Irrepresentable condition

- Classical example: Covariance matrix

$$\left(\begin{array}{ccc|c} 1 & -a & -a & b \\ -a & 1 & -a & b \\ -a & -a & 1 & b \\ \hline b & b & b & 1 \end{array}\right)$$

  for appropriate $a, b > 0$

- In high-dimension, pathologies happen when there are strong collinearity with "conflicting" correlational relations among the variables

Adaptive Lasso

- Adaptive Lasso

$$\hat{\beta}^{Adaptive-Lasso} = \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p} \frac{1}{|\hat{\beta}(j)|^{\gamma}} |\beta^{(j)}|$$

  where $\hat{\beta}$ is a base estimator and $\gamma > 0$
- Idea: If $\hat{\beta}$ is consistent with quantifiable convergence rate, then Adaptive Lasso is selection consistent with appropriate regularization

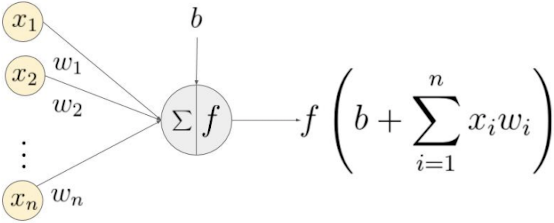
Zou, Hui. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101.476 (2006): 1418-1429.

Consistent feature selection for analytic deep networks

Dinh and Ho. Consistent feature selection for analytic deep neural networks. Advances in Neural Information Processing Systems 33 (2020): 2420-2431.

Ho and Dinh. Searching for minimal optimal neural networks. Statistics & Probability Letters (2022): 109353.

neural network = linear model + non-linear activation



If the weight of an input is zero, then it does nor influence the output node

Mathematical formulation

Given an input x that belongs to be a bounded open set $\mathcal{X} \subset \mathbf{R}^{d_0}$, the output map $f_\alpha(x)$ with $\alpha = (P, p, S, Q, q)$ is defined by

- input layer:

$$h_1(x) = P \cdot x + p$$

- hidden layers:

$$h_j(x) = \phi_{j-1}(S, h_{j-1}(x), h_{j-2}(x), \ldots, h_1(x))$$

- output layer:

$$f_\alpha(x) = h_L(x) = Q \cdot h_{L-1}(x) + q$$

where $\phi_1, \phi_2, \ldots, \phi_{L-2}$ are analytic functions parameterized by the hidden layers' parameter S.

Model-based settings for regression

Assumption

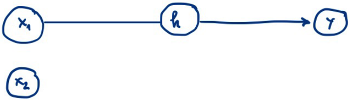
Data $\{(X_i, Y_i)\}_{i=1}^{n}$ *are independent and identically distributed (i.i.d) samples generated from* $P^*_{X,Y}$ *such that*

- *the input density* p_X *is positive and continuous on its domain* $\mathcal{X}$ *and*
- $Y_i = f_{\alpha^*}(X_i) + \epsilon_i$ *where* $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

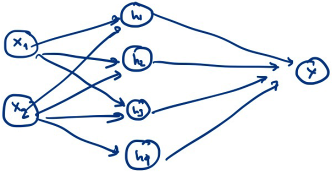
Limit setting: the network is assumed to be fixed, while $n \to \infty$.

Model-based settings for regression

Generating model



Hypothesis space



The set of risk minimizers

- Define

$$R(\alpha) = \mathbb{E}_{(X,Y)\sim P^*_{X,Y}}[(f_\alpha(X) - Y)^2]$$

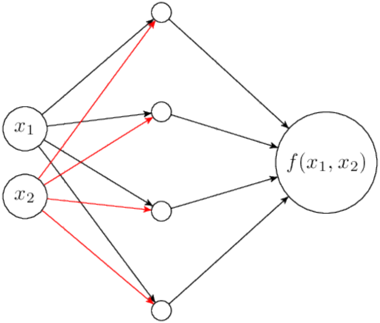
  and

$$\mathcal{H}^* = \{\alpha : R(\alpha) = R(\alpha^*)\}$$

- A "good" estimator will converge to $\mathcal{H}^*$ when $n \to \infty$.
- Under appropriate regularity conditions, we also have

$$\mathcal{H}^* = \{\alpha \in \mathcal{W} : f_\alpha = f_{\alpha^*}\}$$

Group Lasso for neural networks



- parameters correspond to an input node are grouped together
- only parameters of the input layer should be penalized

Group Lasso

- A simple GL estimator for neural networks is thus defined by

$$\hat{\alpha}_n := \underset{\alpha=(u,v,b_1,b_2,S,Q,q)}{\operatorname{argmin}} \quad \frac{1}{n}\sum_{i=1}^{n} \ell(\alpha, X_i, Y_i) + \lambda_n L(\alpha)$$

  where

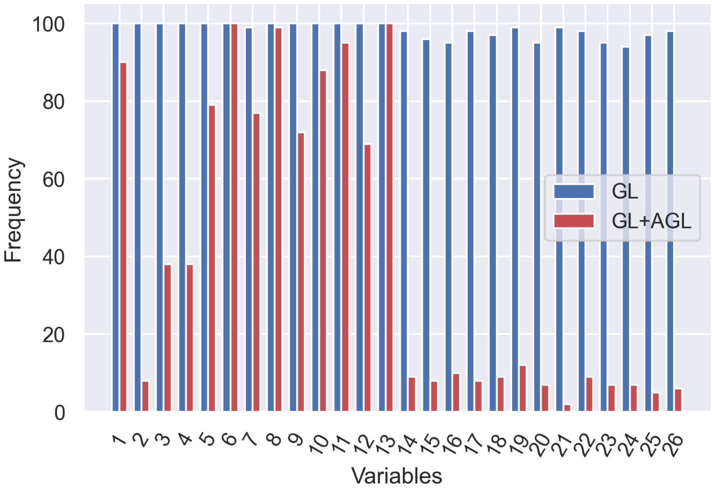
$$L(\alpha) = \sum_{k=1}^{d_0} \|u^{[:,k]}\|$$

  $\ell(\alpha, x, y) = (y - f_\alpha(x))^2$ is the square-loss, $\lambda_n > 0$, $\|\cdot\|$ is the standard Euclidean norm and $u^{[:,k]}$ is the vector of parameters associated with k-th input.
- Group Lasso (and its variants) are very popular in the field.
- No theoretical support in terms of feature selection

Example: Boston housing dataset

- Dataset consists of 506 observations of house prices and 13 predictors
- 13 random Gaussian noise predictors are added
- Question: Can Group Lasso eliminate the noise predictors?

Failure of Group Lasso



Adaptive Group Lasso

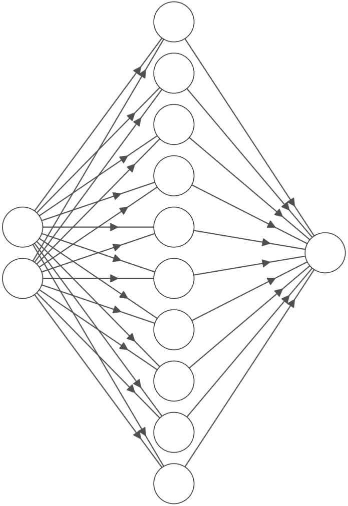
- Adaptive group Lasso (GL+AGL)

$$\tilde{\alpha}_n := \arg\min_{\alpha} \frac{1}{n}\sum_{i=1}^{n}(Y_i - f_\alpha(X_i))^2 + \zeta_n \left(\sum_{k=1}^{d_0} \frac{1}{\|\hat{u}_n^{[:,k]}\|^\gamma}\|u^{[:,k]}\|\right),$$

  where $\hat{u}_n$ denotes the Group Lasso estimate.
- Hope:
  - GL estimation of a significant input stay away from zero
  - GL estimation of an insignificant input converges to zero with a quantifiable convergence rate

Geometric properties of shallow and irreducible neural networks

Shallow networks with *tanh* activation



Unidenitifiability

For *tanh* activation function, the input-output map of a neural network does not change if

- two hidden nodes are swapped
- the weights associated with a hidden node (inward and outward) are multiplied by -1
- a node is cloned and their outward weights are divided by 2

Irreducible network with one hidden layer

A feed-forward model f_{u,v,w,b_1,b_2} is **irreducible** if

(i) $(u^{[i,:]}, v^{[i,:]}) \neq 0$ and $w^{[i]} \neq 0$ for all i.

(ii) For any two different indices i and j

$$(u^{[i,:]}, v^{[i,:]}, b_1^{[i]}) \neq \pm(u^{[j,:]}, v^{[j,:]}, b_1^{[j]}).$$

Theorem

For an "irreducible" single-output feed-forward neural network with one hidden layer and hyperbolic tangent activation function, functionally equivalence are compositions of node interchange and sign flip equivalence.

Kurkova and Kainen. Neural Computation 6.3 (1994): 543-558

The set of risk minimizers

Assumption

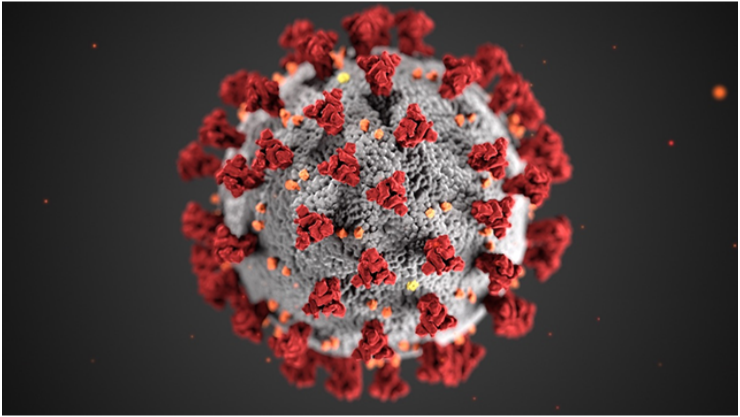
Data $\{(X_i, Y_i)\}_{i=1}^n$ are independent and identically distributed (i.i.d) samples generated from $P^*_{X,Y}$ such that

- *the input density p_X is positive and continuous on its domain $\mathcal{X}$ and*
- $Y_i = f_{\alpha^*}(X_i) + \epsilon_i$ *where* $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

Recall that

$$\mathcal{H}^* = \{\alpha : R(\alpha) = R(\alpha^*)\} = \{\alpha \in \mathcal{W} : f_\alpha = f_{\alpha^*}\}$$

Geometry of $\mathcal{H}^*$



Geometry of the risk function near $\mathcal{H}^*$

Lemma (Information bound)

For any $\alpha \in \mathcal{H}^*$, there exist $c_2(\alpha) > 0$ and a neighborhood $\mathcal{U}_\alpha$ such that

$$R(\beta) - R(\alpha) \geq c_2 \|\beta - \alpha\|^2$$

for all $\beta \in \mathcal{U}_\alpha$.

Uncertainty bound

Lemma

For any $\delta > 0$*, there exist* $c_1(\delta) > 0$ *such that*

$$\left| \frac{1}{n} \sum_{i=1}^{n} (Y_i - f_\alpha(X_i))^2 - R(\alpha) \right| \leq c_1 \frac{\log n}{\sqrt{n}}, \qquad \forall \alpha \in \mathcal{W}.$$

with probability at least $1 - \delta$*.*

Convergence of Group Lasso

Theorem

Assuming that $\lambda_n \to 0$. *For any* $\delta > 0$, *there exist* $C_\delta > 0$, $N_\delta > 0$ *such that for all* $n \geq N_\delta$,

$$\min_{\alpha \in \mathcal{H}^*} \|\hat{\alpha}_n - \alpha\| \leq C_\delta \left(\frac{\log n}{\sqrt{n}} + \lambda_n^2 \right)^{1/2}$$

with probability at least $1 - \delta$.

Feature selection consistency of Adaptive Group Lasso

Theorem

For $\gamma > 0$, $\mu \in (0, \gamma/4)$ *and* $\zeta_n = \Omega(n^{-\gamma/4+\mu})$, *then GL+AGL is consistent for feature selection.*

That is, for any $\delta > 0$, *there exists* N_δ *such that for* $n > N_\delta$,

- *(consistently recovers the significant features)* $\tilde{u}_n^{[:,k]} \neq 0, \ \forall k = 1, \dots, n_s$, *and*
- *(consistently eliminates the insignificant features)* $\tilde{v}_n^{[:,k]} = 0, \ \forall k = 1, \dots, n_z$

with probability at least $1 - \delta$.

Geometric properties of deep neural networks risk functions

Unidenitifiability: deep networks

For *tanh* activation function, the input-output map of a neural network does not change if

- two hidden nodes are swapped
- the weights associated with a hidden node (inward and outward) are multiplied by -1
- a node is cloned and the outward weights are adjusted accordingly

A bigger problem: are they the only sources of unidentifiability of the networks?

Unidenitifiability

A bigger problem: are they the only sources of unidentifiability of the networks?
$\rightarrow$ no definite answer

Existing results (for *tanh* activation and irreducible networks)

- There is no other equivalent graph transformation outside the span of node interchange and sign flip
- Generically, functionally equivalence are compositions of node interchange and sign flip equivalence.

An Mei Chen, Haw-minn Lu, and Robert Hecht-Nielsen. Neural Computation, 5(6):910–927, 1993.
Charles Fefferman and Scott Markel. Advances in Neural Information Processing Systems, pages 335–342, 1994

Unidenitifiability: deep networks

- Existing results cannot be used for statistical consistency
- The model is degenerate when the data-generating network α^* is not irreducible
- $\mathcal{H}^*$ is a high-dimensional algebraic set (that we cannot fully characterize)
- the Hessian of the risk function at an optimum is singular

Characterizing the set of risk minimizers

Lemma

(i) *There exists* $c_0 > 0$ *such that* $\|u_\alpha^{[:,k]}\| \geq c_0$ *for all* $\alpha \in \mathcal{H}^*$ *and* $k = 1, \ldots, n_s$ *(i.e., for significant features).*

(ii) *For* $\alpha \in \mathcal{H}^*$, *the vector* $\phi(\alpha)$, *obtained from* α *be setting its insignificant components to zero, also belongs to* $\mathcal{H}^*$.

Note: the base estimator must be Group Lasso to kill off the insignificant components

Lojasiewicz's inequality

Lemma

There exist $c_2, \nu > 0$ and such that $R(\beta) - R(\alpha^*) \geq c_2 d(\beta, \mathcal{H}^*)^\nu$ for all $\beta \in \mathcal{W}$.

Note:

- generically, $\nu = 2$
- when $\mathcal{H}^*$ is finite, this reduces to the standard Taylor's inequality around a local optimum, with $\nu = 2$ if the Hessian matrix at the optimum is non-singular

Convergence of Group Lasso

Theorem

For any $\delta > 0$*, there exist* $C_\delta, C' > 0$ *and* $N_\delta > 0$ *such that for all* $n \geq N_\delta$,

$$d(\hat{\alpha}_n, \mathcal{H}^*) \leq C_\delta \left(\lambda_n^{\nu/(\nu-1)} + \frac{\log n}{\sqrt{n}} \right)^{1/\nu}$$

and

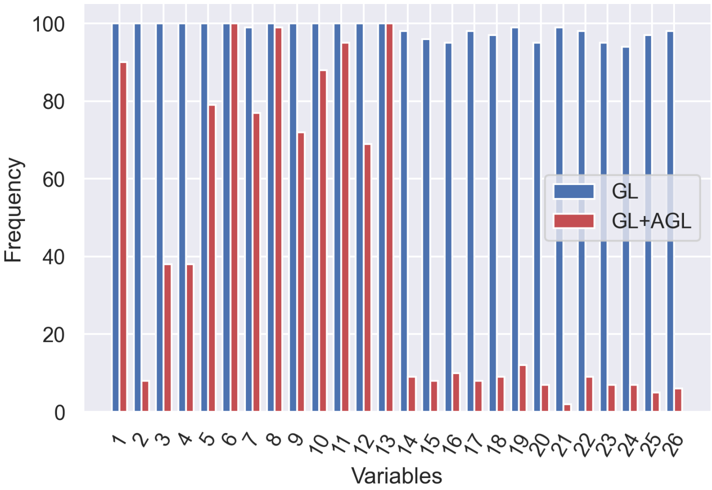
$$\|\hat{v}_n\| \leq 4c_1 \frac{\log n}{\lambda_n \sqrt{n}} + C' \; d(\hat{\alpha}_n, \mathcal{H}^*)$$

Feature selection consistency of GL+AGL

Theorem

Let $\gamma > 0$, $\epsilon > 0$, $\lambda_n \sim n^{-1/4}$, *and* $\zeta_n = \Omega(n^{-\gamma/(4\nu-4)+\epsilon})$, *then the GroupLasso+AdaptiveGroupLasso is feature selection consistent.*

GL vs. GL+AGL

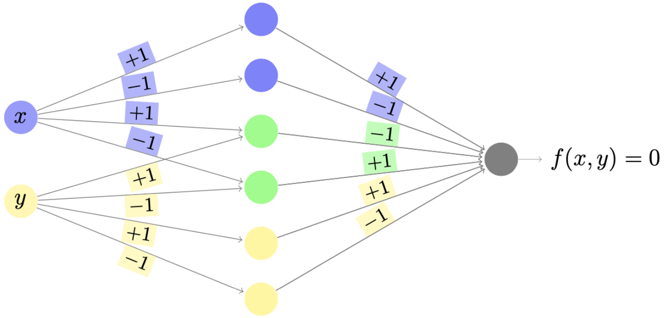


100
80
60
40
20
0
Frequency
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Variables
GL
GL+AGL

Some open questions

Characterizing equivalent networks

- Known: shallow networks with *tanh* activation
- Partially known: deep networks with *tanh* activation
- ReLU network: open



Morse-Bott properties around risk minimizers

- Morse-Bott: Hessian at local minimizers are non-degenerate in the normal direction to the level set
- For matrix factorization using shallow and linear networks with Dropout, the risk function is Morse-Bott around global minimizers
- It has been conjectured that the risk function for ReLU network is also Morse-Bott.

Mianjy, Arora, and Vidal. International Conference on Machine Learning (ICML 2018)

Poggio, T., A. Banburski, and Q. Liao (2020). Proceedings of the National Academy of Sciences 117(48), 30039–30045.

Morse-Bott properties around risk minimizers

Lemma

If g is Morse–Bott function on an open neighborhood $\mathcal{U}$ of a critical point p in a Banach space, then it obeys a first Lojasiewicz inequality with the (optimal) exponent 2.
That is, there exists constant $C > 0$ and a neighborhood $\mathcal{V} \subset \mathcal{U}$ of p such that

$$\|g(x) - g(\mathcal{H})\| \geq C \; dist(x, \mathcal{H})^2 \qquad \forall x \in \mathcal{V}$$

Feehan, P. (2019). Resolution of singularities and geometric proofs of the Lojasiewicz inequalities. Geometry & Topology 23(7), 3273–3313.
Feehan, P. M. (2020). On the Morse–Bott property of analytic functions on Banach spaces with Lojasiewicz exponent one half. Calculus of Variations and Partial Differential Equations 59(2), 1–50.

Morse-Bott properties around risk minimizers

- Morse-Bott: Hessian at local minimizers are non-degenerate in the normal direction to the level set
- For shallow, irreducible, tanh networks, the global minimizers of $R(\alpha)$ are isolated with positive-definite Hessian
- Conjecture: For reducible generating networks, the risk function is Morse-Bott around local minimizers and the Lojasiewicz exponent of the risk function is optimal

Other directions

- High-dimensional setting ($n \ll p$)
- ReLU and other non-analytic activations
- Local feature importance
- Conjecture: Group Lasso is inconsitent for neural networks