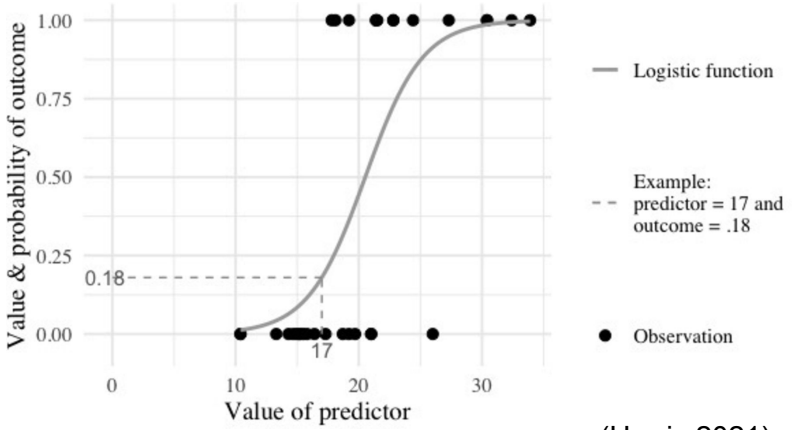
Optimization, learning and high-dimensional macroscopic limits

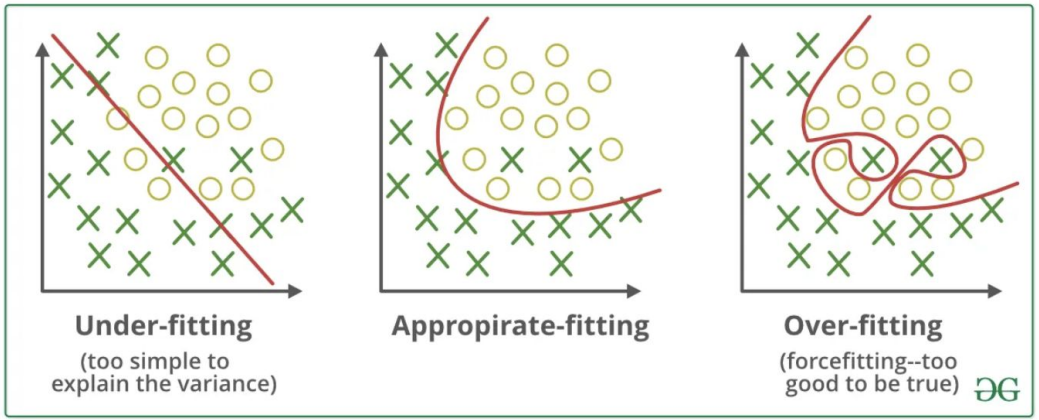
Iris Chang

Previous works

- Logistic Regression: models binary outcomes
  - Ex: Patient diagnosis (Salahi et al., 2019)
  - Regularization: adding a penalty to prevent overfitting
- When p fixed and n → ∞, MLE has nice properties (e.g. unbiased)
- In high dimensions, properties break down for unregularized, L_1 and L_2 regularized logistic regression (Candès & Sur, 2018; Salahi et al., 2019)



(Harris 2021)

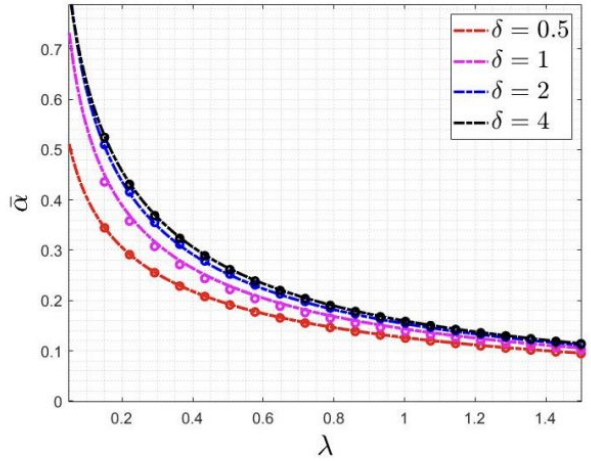


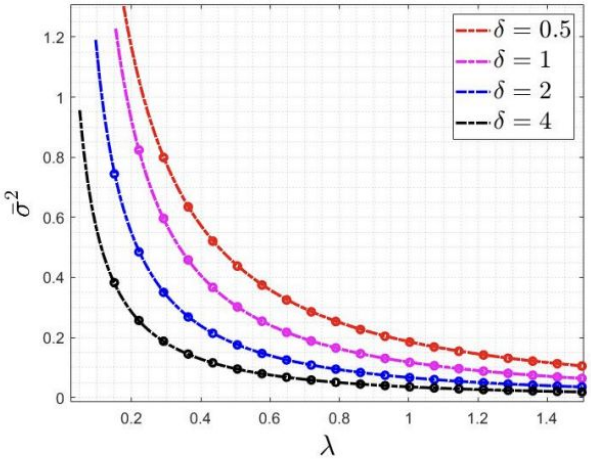
(Karas 2023)

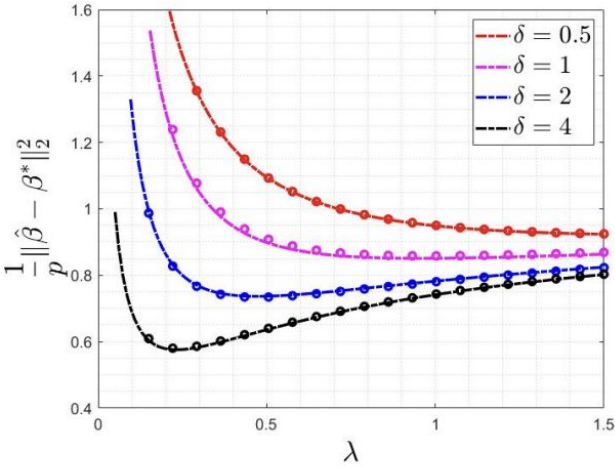
The Impact of Regularization on High-dimensional Logistic Regression

(Salehi et al. 2019)

- Need for new method to find performance metrics
- Use CGMT to find a system of six nonlinear equations
  - Simplifies to three nonlinear equations under L_2
- Solution allows performance metrics (e.g. MSE) to be calculated
- Ultimately useful when finding reg param (λ) that optimizes performance

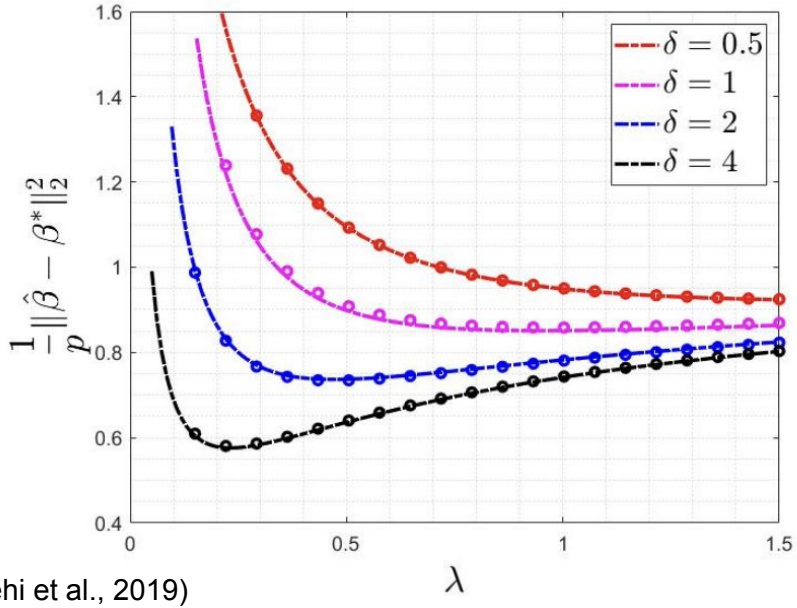




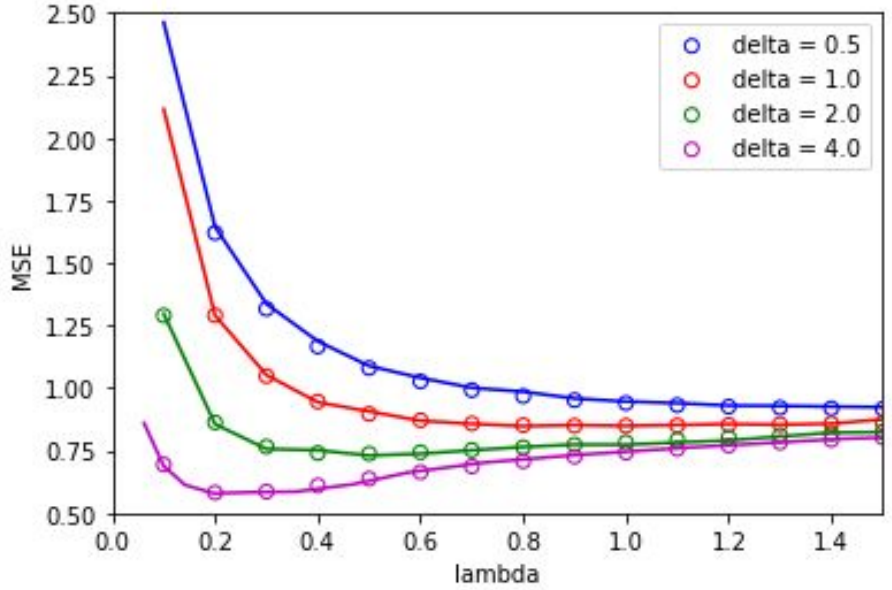


Replicating Results from Salehi et al.

- Points correspond to randomly generated data points
- Lines correspond to the
- λ = regularization parameter; δ = n / p

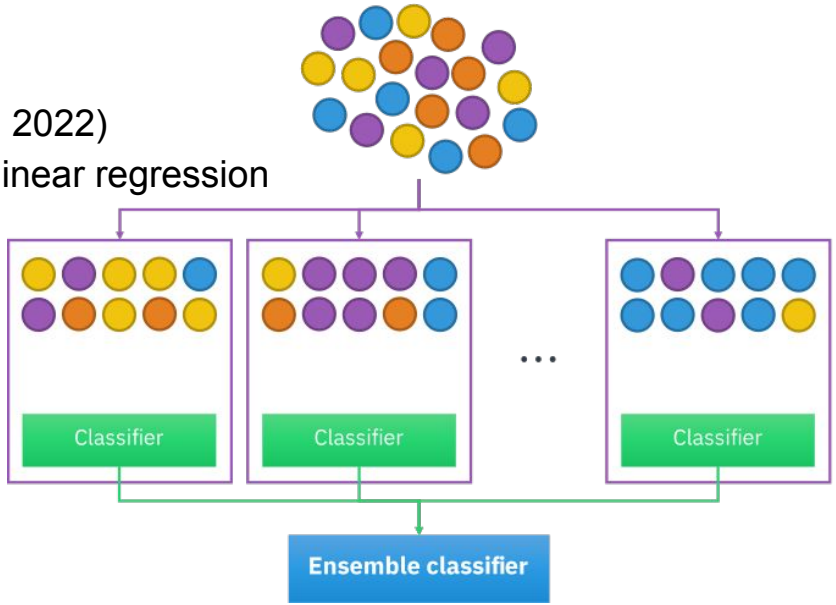


(Salehi et al., 2019)



Bagging Problem

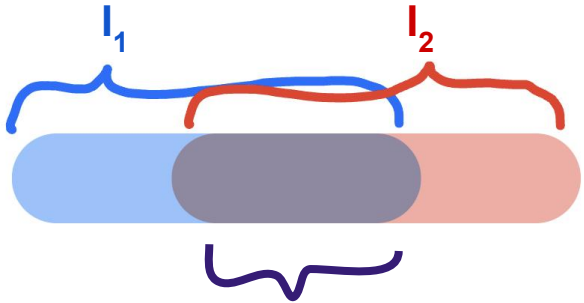
- Bagging:
  - Train classifier on subsets of whole dataset and then aggregate models from each
- Approaches to problem:
  - Same approach as Salehi et al.
  - Replica Method from Statistical Physics (Loureiro et al. 2022)
  - Borrowing from similar setting in linear regression



(Lamyai, 2014)

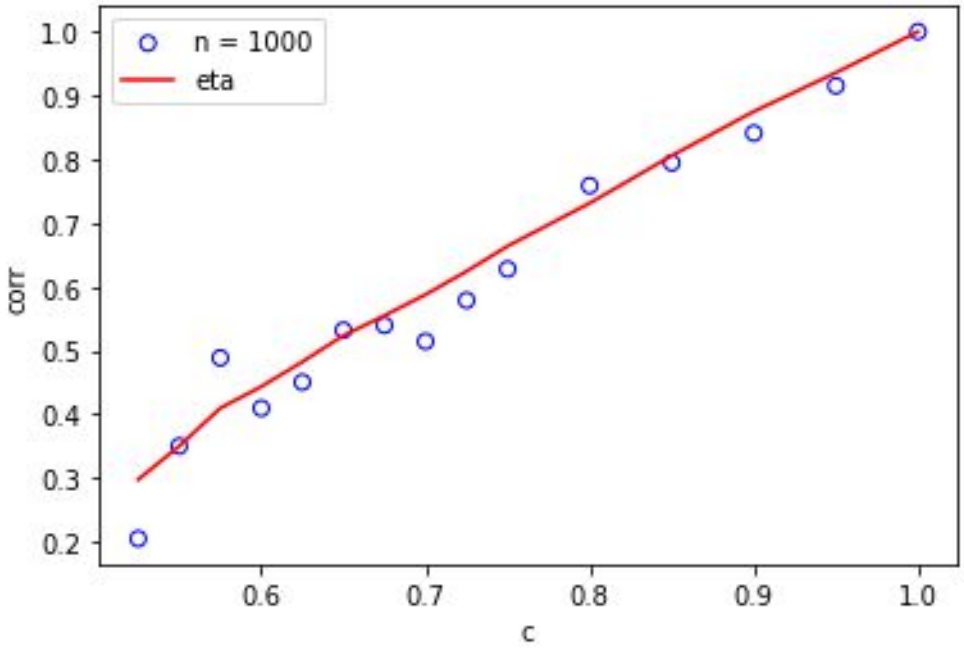
Inference from Linear Regression

- Restrictions:
  - Model parameters = 0
    → response variable (y's) gives no info about predictor (x)
  - Unregularized ($\lambda = 0$)
  - Divide into two equal subsets of equal size with some set amount of overlap
- Same setting in linear reg finds single term (η) that predicts correlation btwn estimators of I's

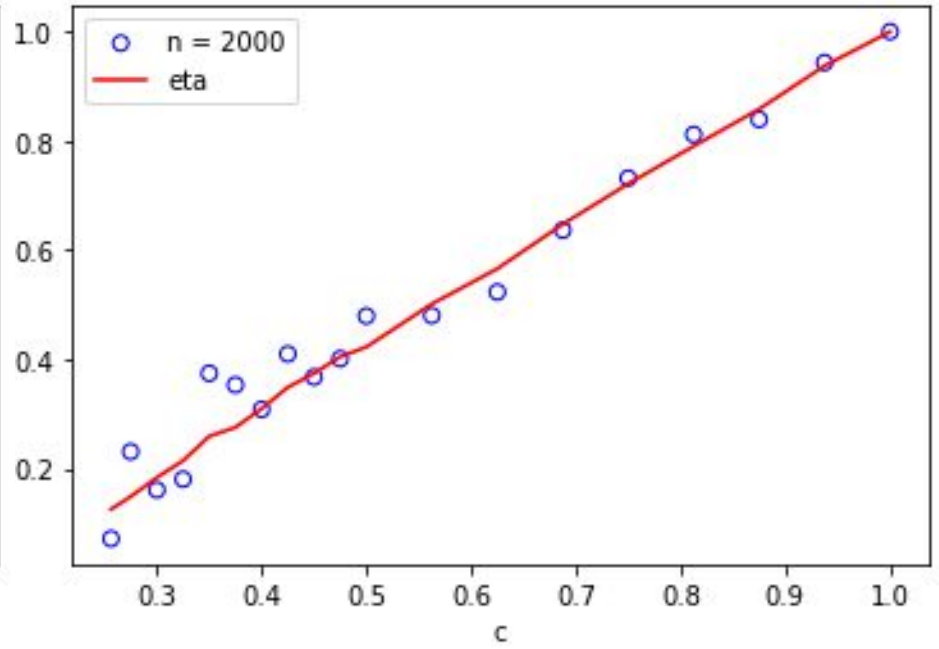


$|I_1| = |I_2| = c * n$
For constant $c \in (0,1)$

$| I_1 \cap I_2 | = c * | I |$



For p = 250 and n = 1000



For p = 250 and n = 2000

c ($\propto$ size of data subset) over correlation between estimators of I_1 and I_2

Future Questions

- Heavy restraints → more general model in future
- Exploring other approaches that were previously mentioned

Acknowledgements

Thank you to my mentor Pierre Bellec for his guidance throughout!

This work was carried out as a part of the 2023 DIMACS REU program at Rutgers University, supported by NSF grant CNS-2150186

Works Cited

Emmanuel J Candès and Pragya Sur. (2018). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. arXiv preprint arXiv:1804.09753.

Harris JK. (2021). Primer on binary logistic regression. *Family Medicine and Community Health.* doi: 10.1136/fmch-2021-001290

Karas, P. (2023). *Under, over and appropriate fitting in logistic regression*. L1 (Lasso) and L2 (Ridge) regularizations in logistic regression. Medium . Retrieved from https://ai.plainenglish.io/l1-lasso-and-l2-ridge-regularizations-in-logistic-regression-53ab6c952f15 .

Lamyai, S. (2014). *An illustration for the concept of bootstrap aggregating*. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Bootstrap_aggregating#/media/File:Ensemble_Bagging.svg.

Loureiro, Bruno, et al. “Learning Curves of Generic Features Maps for Realistic Datasets with a Teacher-Student Model*.” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2022, no. 11, 2022, p. 114001, https://doi.org/10.1088/1742-5468/ac9825.

Salehi, F., Abbasi, E., & Hassibi, B. (2019). *The Impact of Regularization on High-Dimensional Logistic Regression*. https://doi.org/ https://doi.org/10.48550/arXiv.1906.03761