Asymptotics of the Sketched Pseudoinverse

December 2022

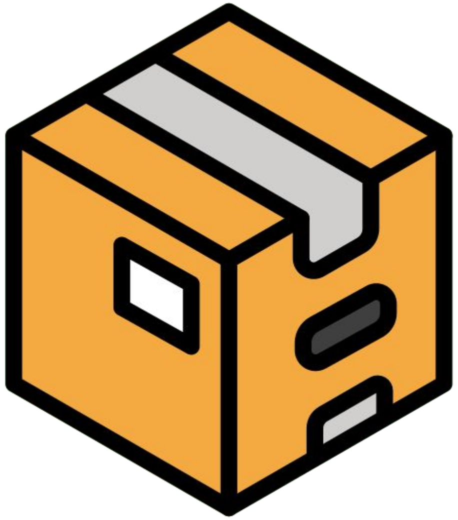
Daniel LeJeune
Rice University
Department of Electrical and Computer Engineering

Pratik Patil
Carnegie Mellon University
Department of Statistics and Machine Learning

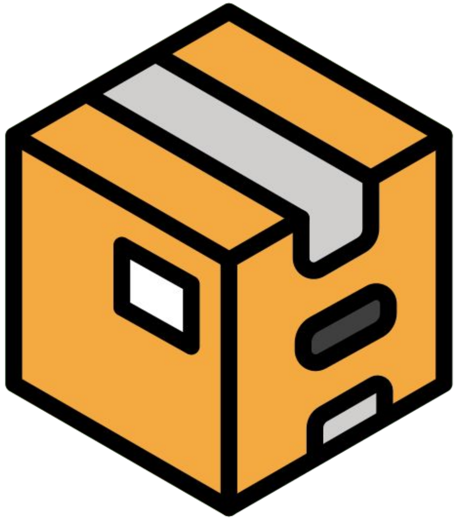
Joint work with: Hamid Javadi, Richard Baraniuk, Ryan Tibshirani

Too Much Data

Too Much Data



Too Much Data

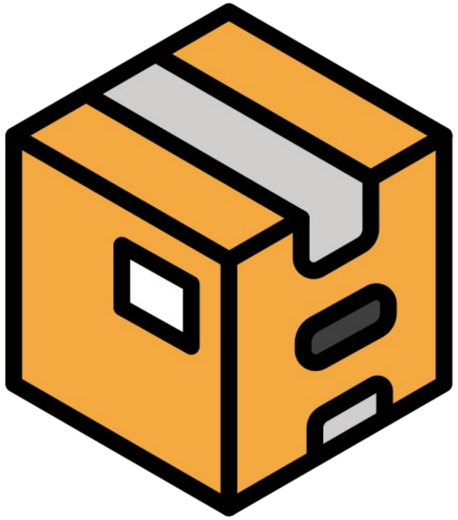


Age=27, Height=5’11”, ABO=A, ALT=36, Glucose=78, Creatinine=0.99, Sodium=132, Carbon Dioxide=21…

Age=56, Height=5’3”, ABO=AB, ALT=40, Glucose=98, Creatinine=0.63, Sodium=182, Carbon Dioxide=25…

Age=41, Height=6’4”, ABO=O, ALT=36, Glucose=84, Creatinine=0.79, Sodium=156, Carbon Dioxide=22…

Too Much Data



Age=27, Height=5’11”, ABO=A, ALT=36, Glucose=78, Creatinine=0.99, Sodium=132, Carbon Dioxide=21…

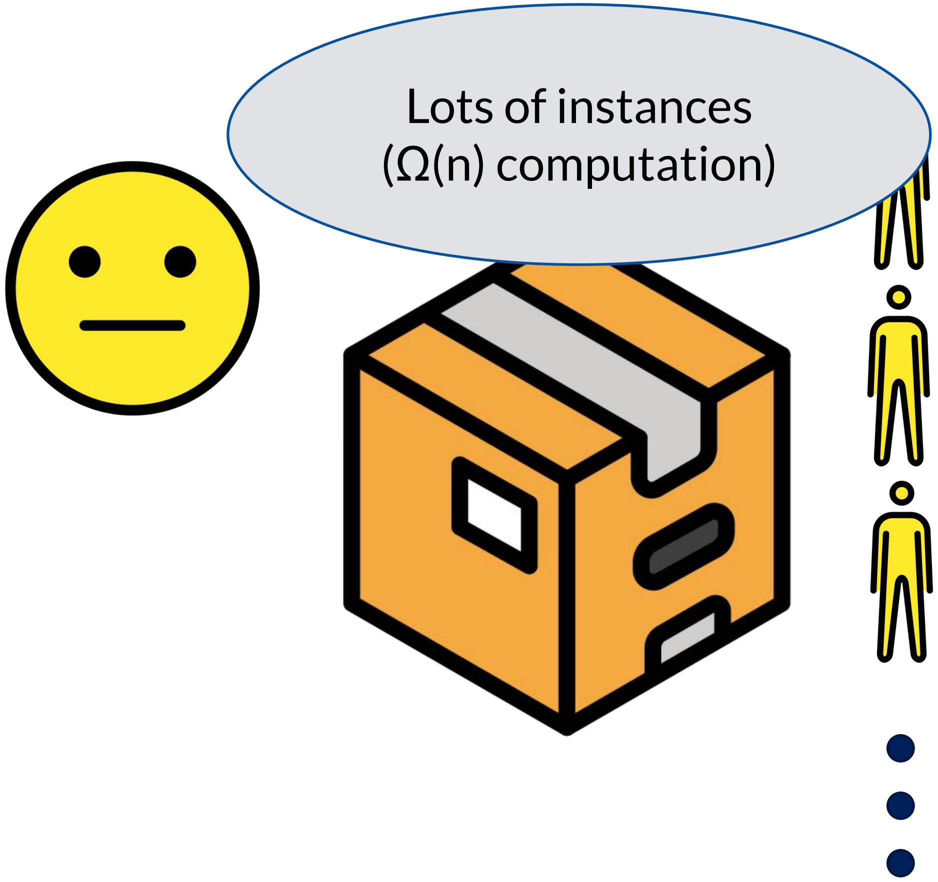
Age=56, Height=5’3”, ABO=AB, ALT=40, Glucose=98, Creatinine=0.63, Sodium=182, Carbon Dioxide=25…

Age=41, Height=6’4”, ABO=O, ALT=36, Glucose=84, Creatinine=0.79, Sodium=156, Carbon Dioxide=22…

⋮

Q: Odds of cancer in next 5 years?

Too Much Data



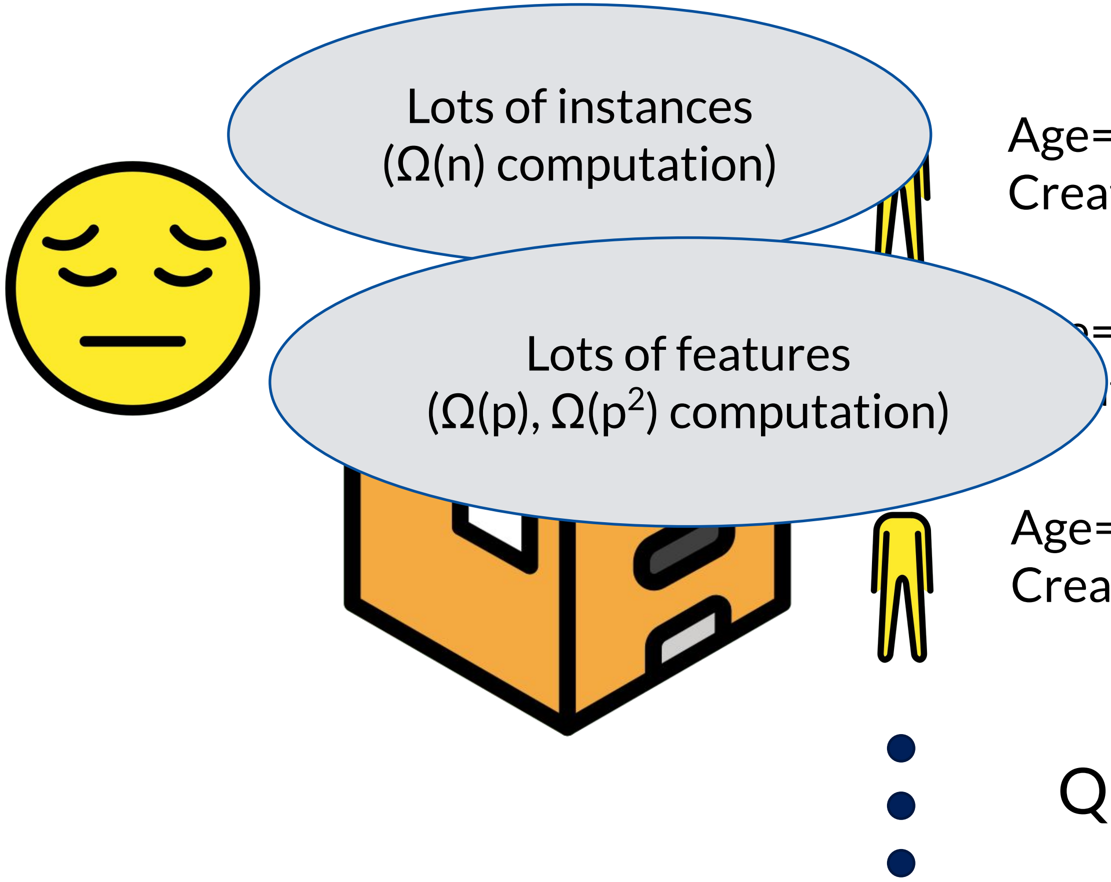
Age=27, Height=5’11”, ABO=A, ALT=36, Glucose=78, Creatinine=0.99, Sodium=132, Carbon Dioxide=21…

Age=56, Height=5’3”, ABO=AB, ALT=40, Glucose=98, Creatinine=0.63, Sodium=182, Carbon Dioxide=25…

Age=41, Height=6’4”, ABO=O, ALT=36, Glucose=84, Creatinine=0.79, Sodium=156, Carbon Dioxide=22…

Q: Odds of cancer in next 5 years?

Too Much Data



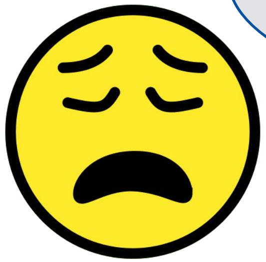
Age=27, Height=5’11”, ABO=A, ALT=36, Glucose=78,
Creatinine=0.99, Sodium=132, Carbon Dioxide=21…

=56, Height=5’3”, ABO=AB, ALT=40, Glucose=98,
tinine=0.63, Sodium=182, Carbon Dioxide=25…

Age=41, Height=6’4”, ABO=O, ALT=36, Glucose=84,
Creatinine=0.79, Sodium=156, Carbon Dioxide=22…

Q: Odds of cancer in next 5 years?

Too Much Data



Lots of instances
($\Omega(n)$ computation)

Lots of features
($\Omega(p)$, $\Omega(p^2)$ computation)

Curse of dimensionality
(Too many solutions)

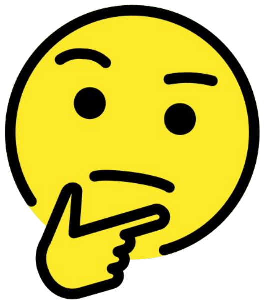
Age=27, Height=5’11”, ABO=A, ALT=36, Glucose=78,
Creatinine=0.99, Sodium=132, Carbon Dioxide=21…

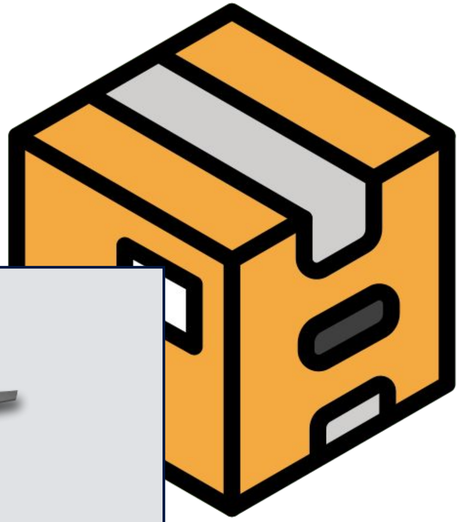
=56, Height=5’3”, ABO=AB, ALT=40, Glucose=98,
tinine=0.63, Sodium=182, Carbon Dioxide=25…

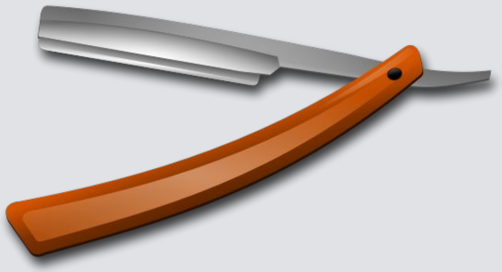
Age=41, Height=6’4”, ABO=O, ALT=36, Glucose=84,
inine=0.79, Sodium=156, Carbon Dioxide=22…

Q: Odds of cancer in next 5 years?

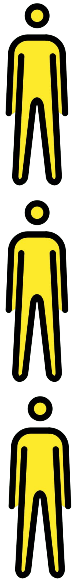
Too Much Data







Occam's razor:
The true function should not be too complicated.



Age=27, Height=5'11", ABO=A, ALT=36, Glucose=78, Creatinine=0.99, Sodium=132, Carbon Dioxide=21…

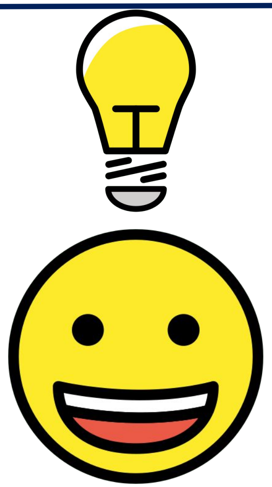
Age=56, Height=5'3", ABO=AB, ALT=40, Glucose=98, Creatinine=0.63, Sodium=182, Carbon Dioxide=25…

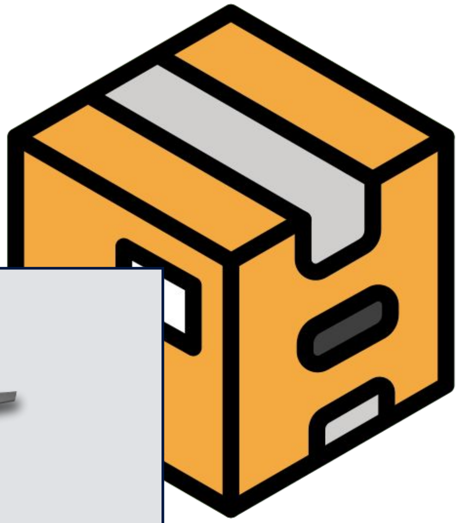
Age=41, Height=6'4", ABO=O, ALT=36, Glucose=84, Creatinine=0.79, Sodium=156, Carbon Dioxide=22…

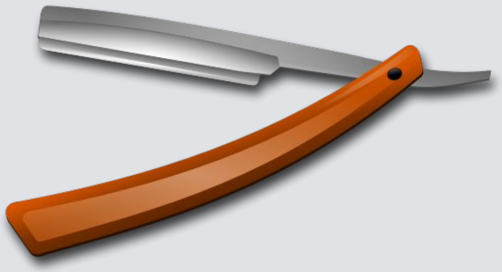
⋮

Q: Odds of cancer in next 5 years?

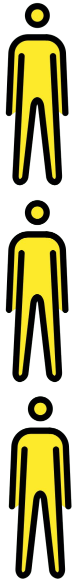
Too Much Data







Occam's razor:
The true function should not be too complicated.



Age=27, He… …ucose=78,
Creatinin… …xide=21…

Age, ABO, Carbon Dioxide…

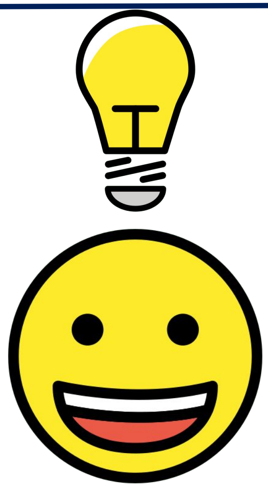
Age=56, Height=5'3", ABO=AB, ALT=40, Glucose=98, Creatinine=0.63, Sodium=182, Carbon Dioxide=25…

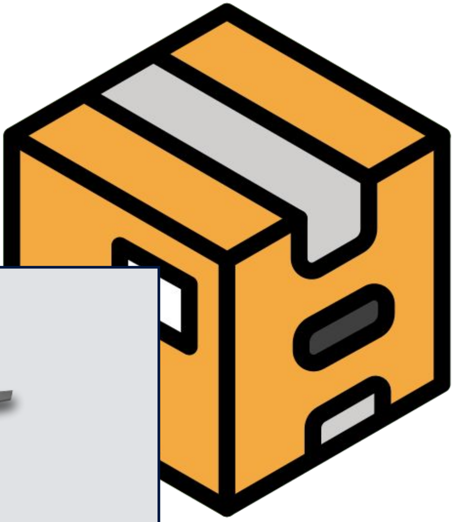
Age=41, Height=6'4", ABO=O, ALT=36, Glucose=84, Creatinine=0.79, Sodium=156, Carbon Dioxide=22…

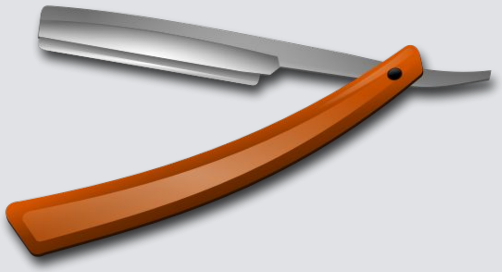
⋮

Q: Odds of cancer in next 5 years?

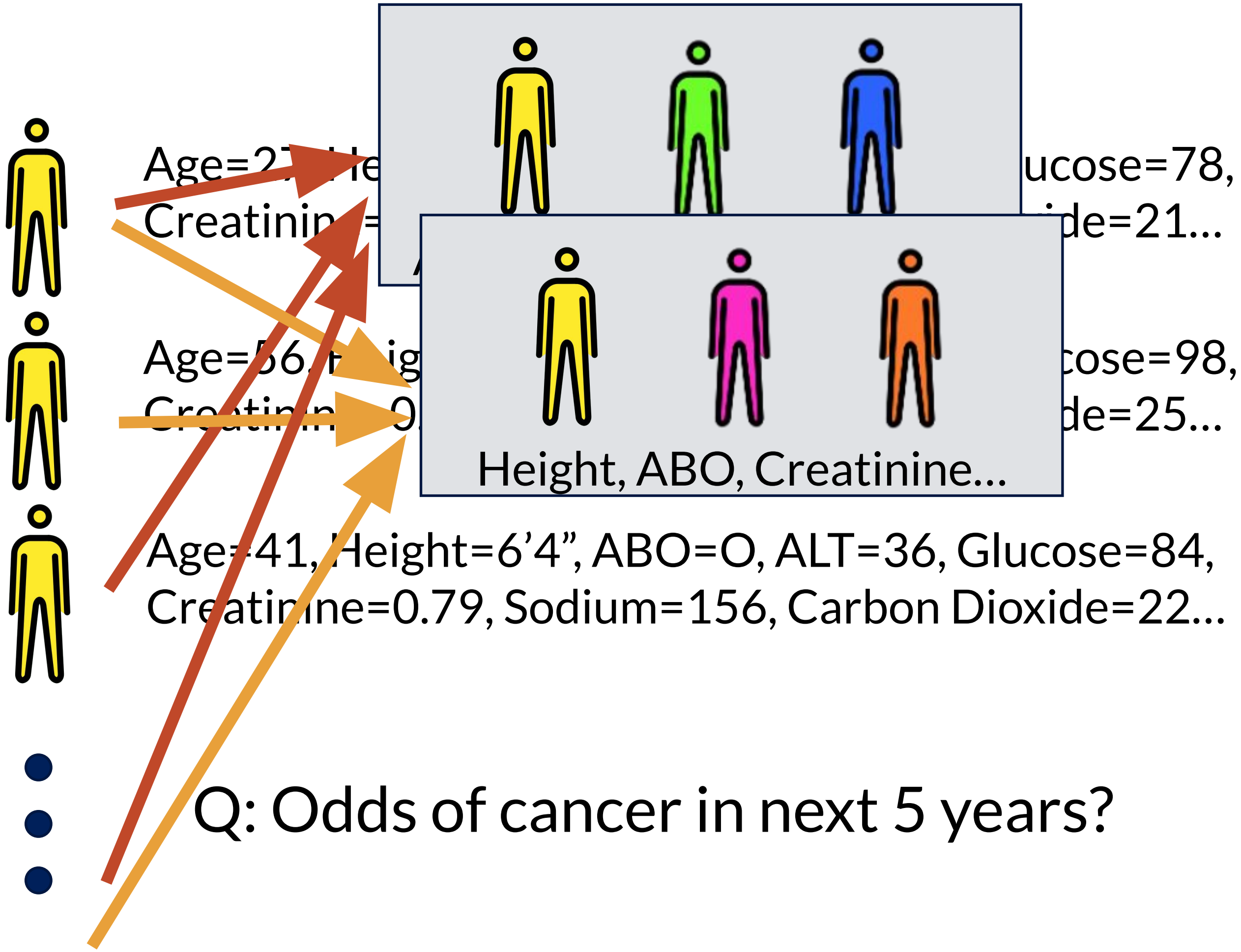
Too Much Data





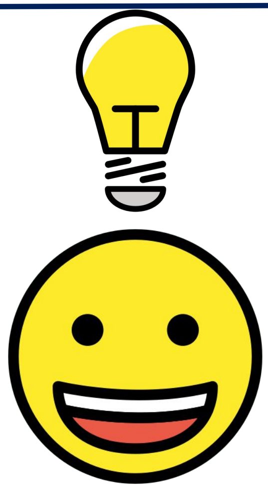


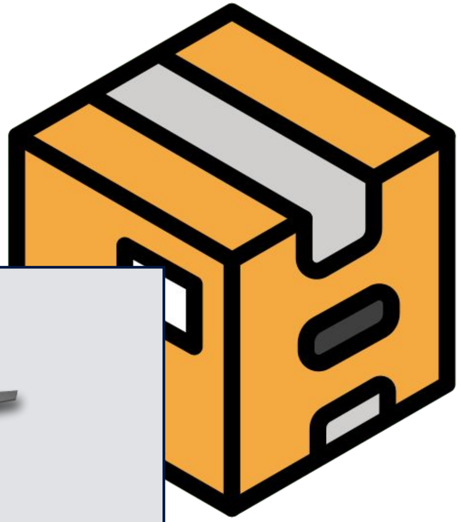
Occam's razor:
The true function should not be too complicated.

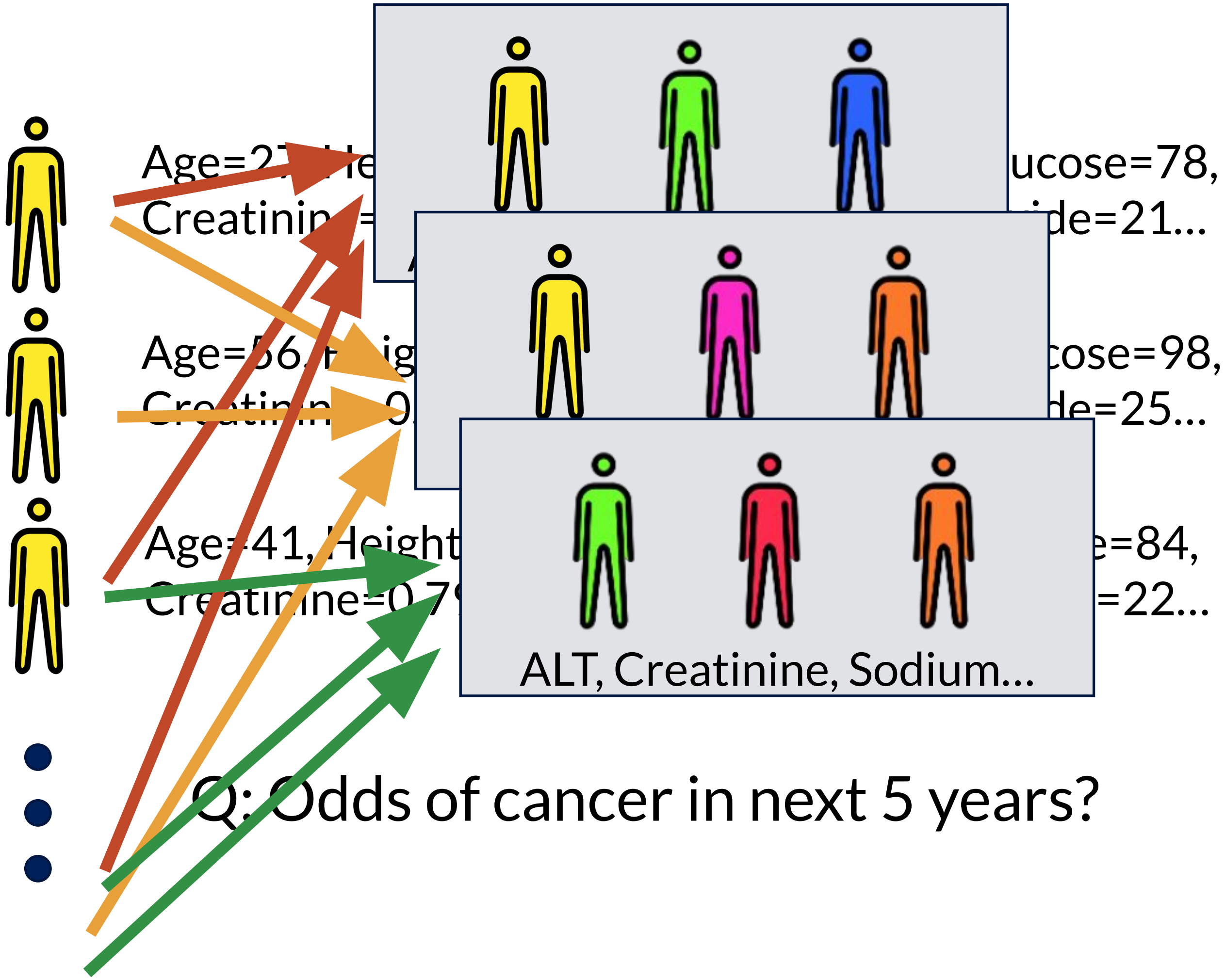


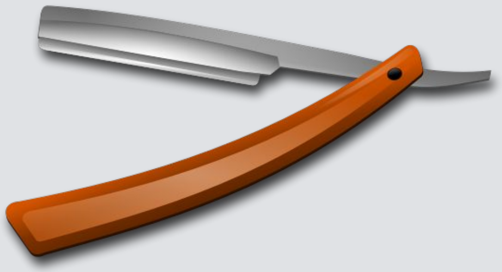
Q: Odds of cancer in next 5 years?

Too Much Data



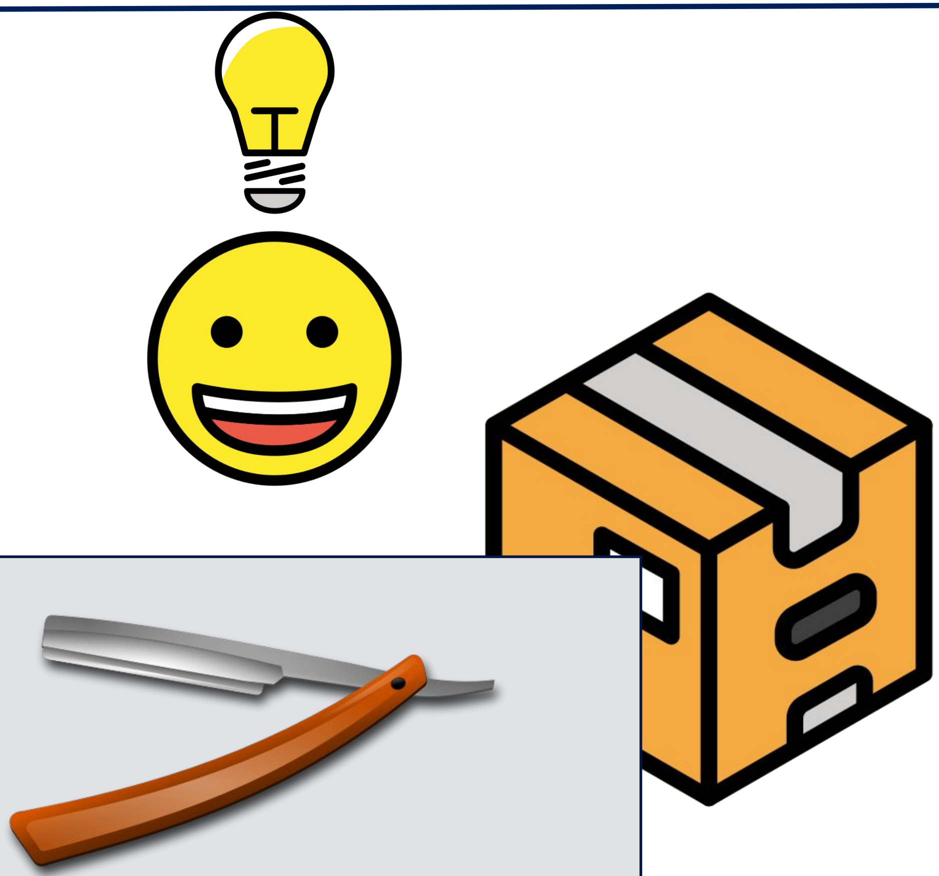


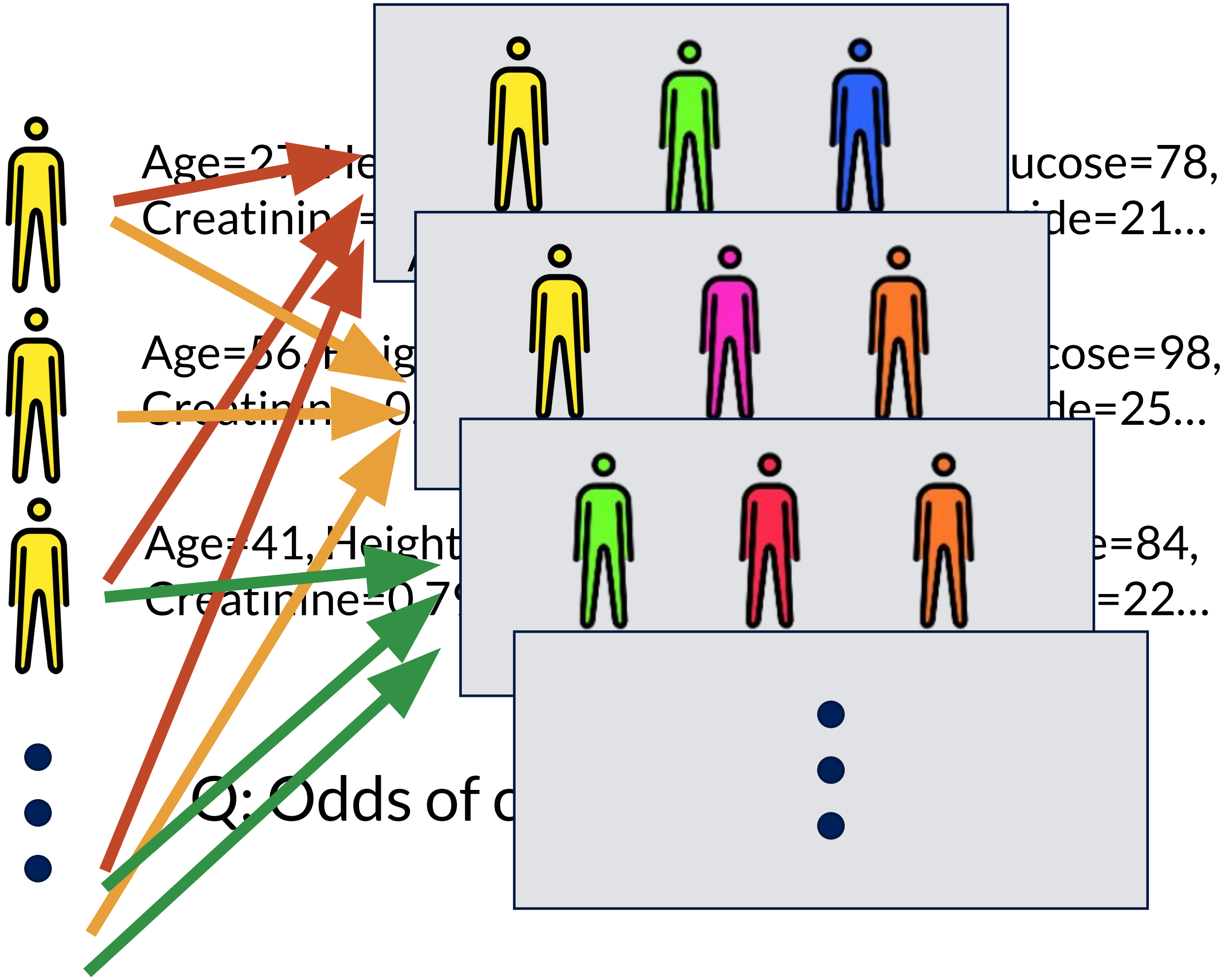




Occam's razor:
The true function should not be too complicated.

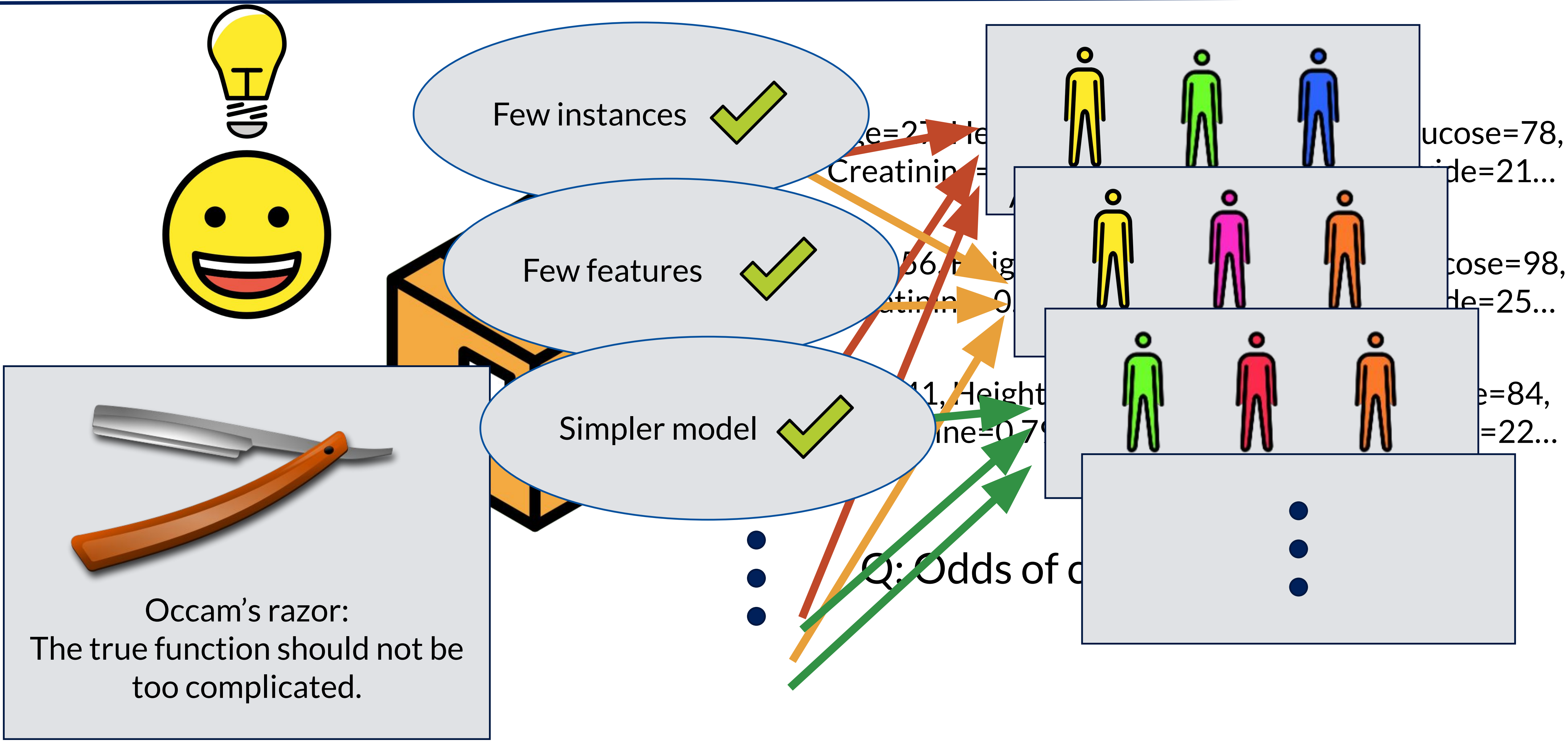
Too Much Data

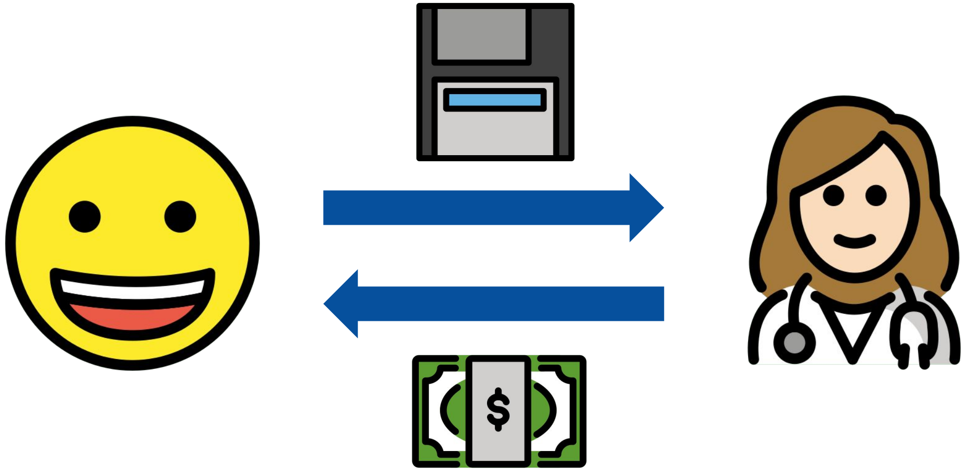


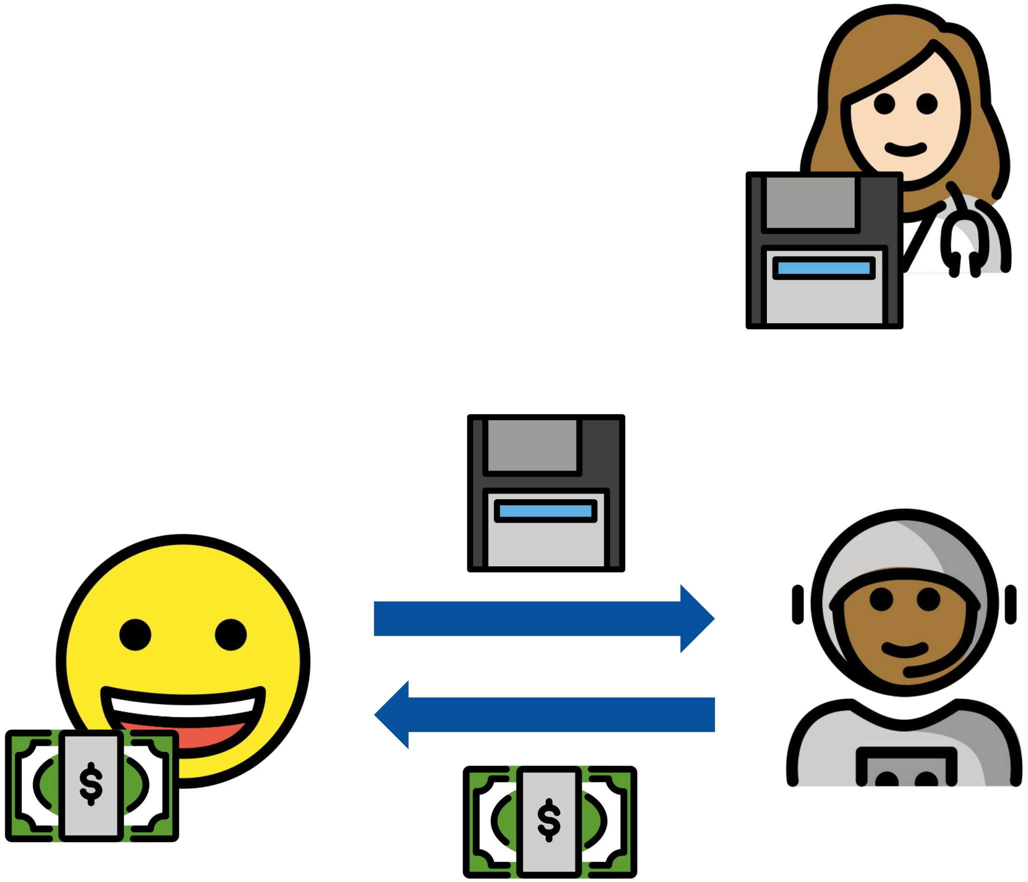


Occam's razor:
The true function should not be too complicated.

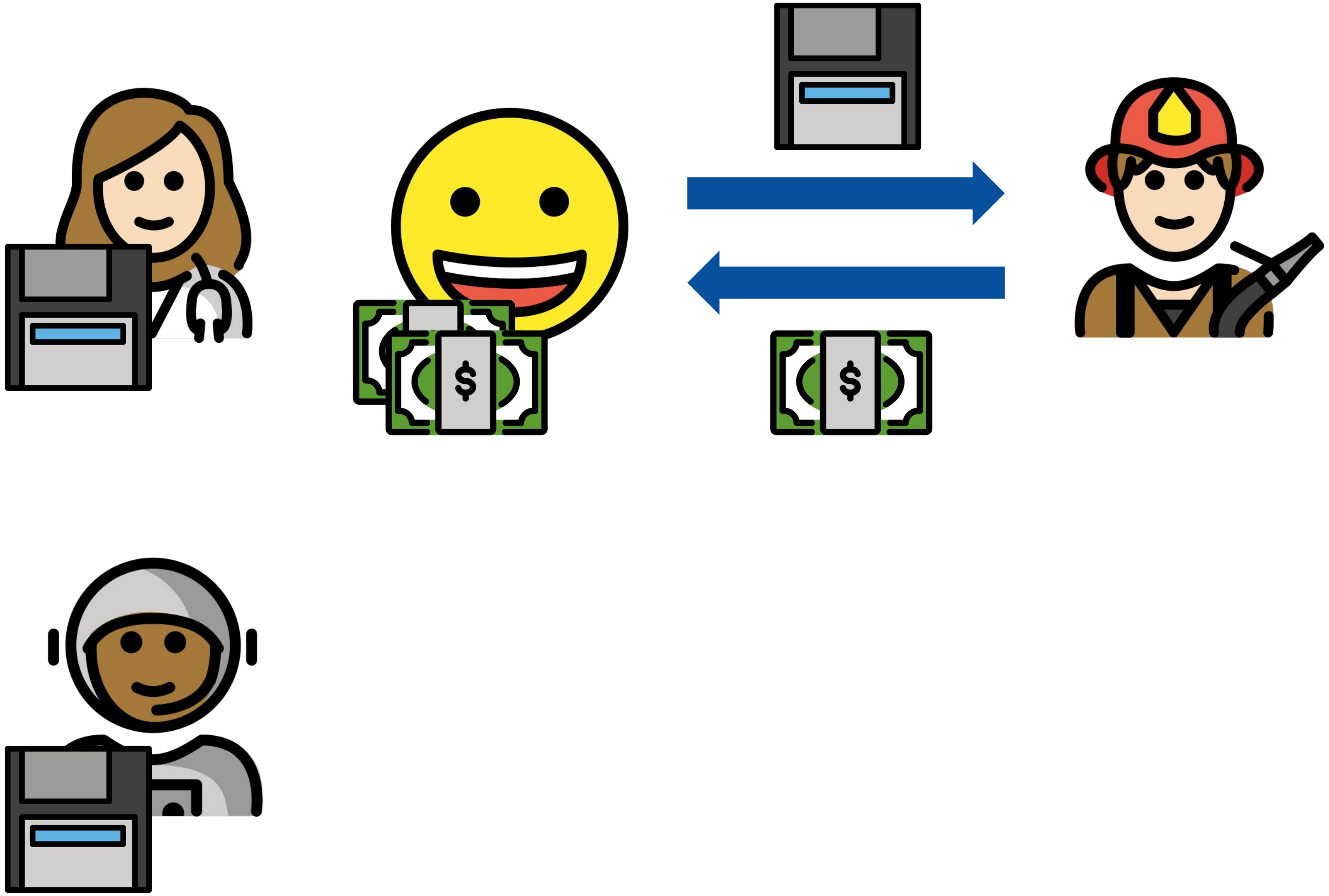
Too Much Data

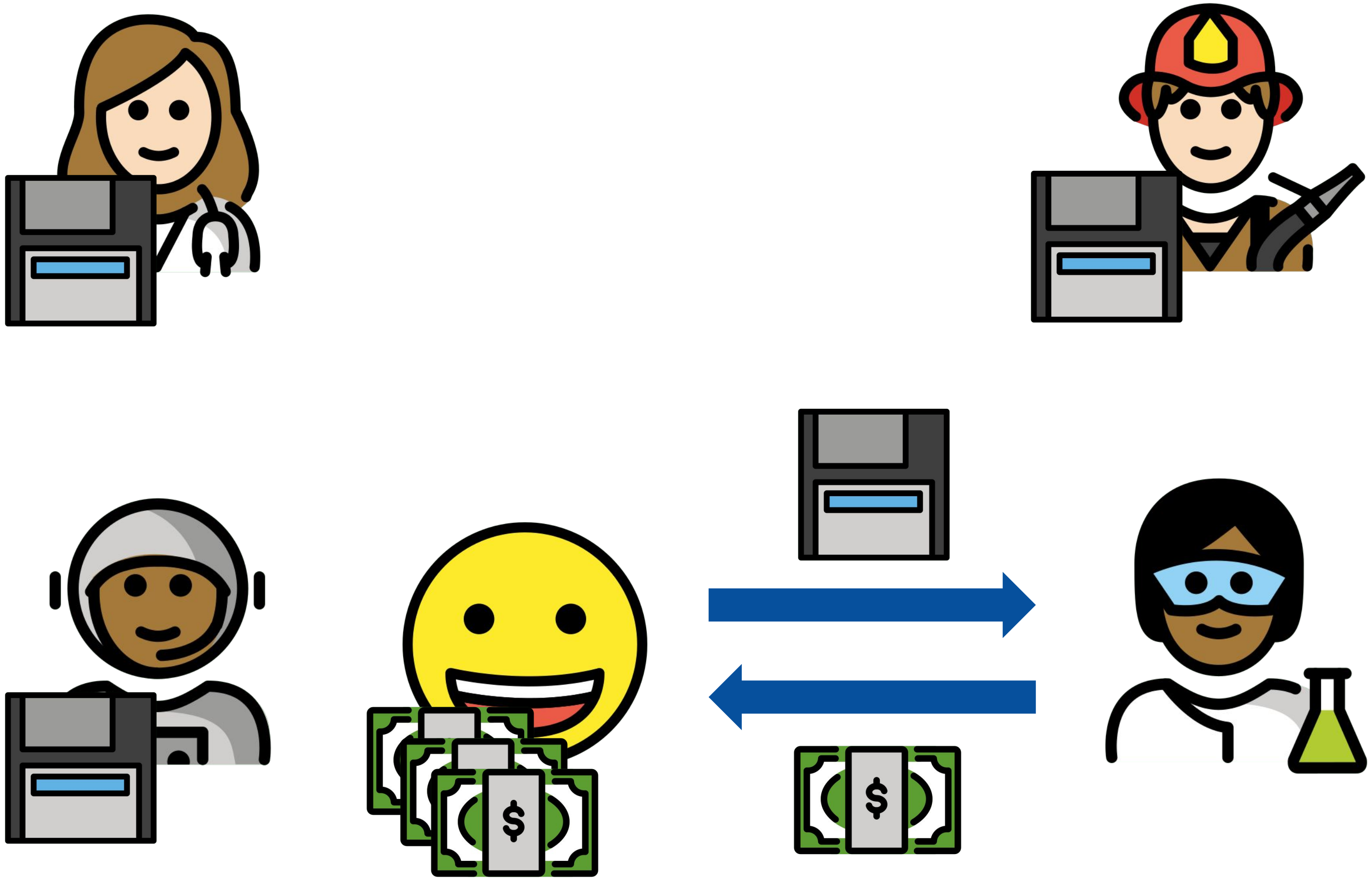




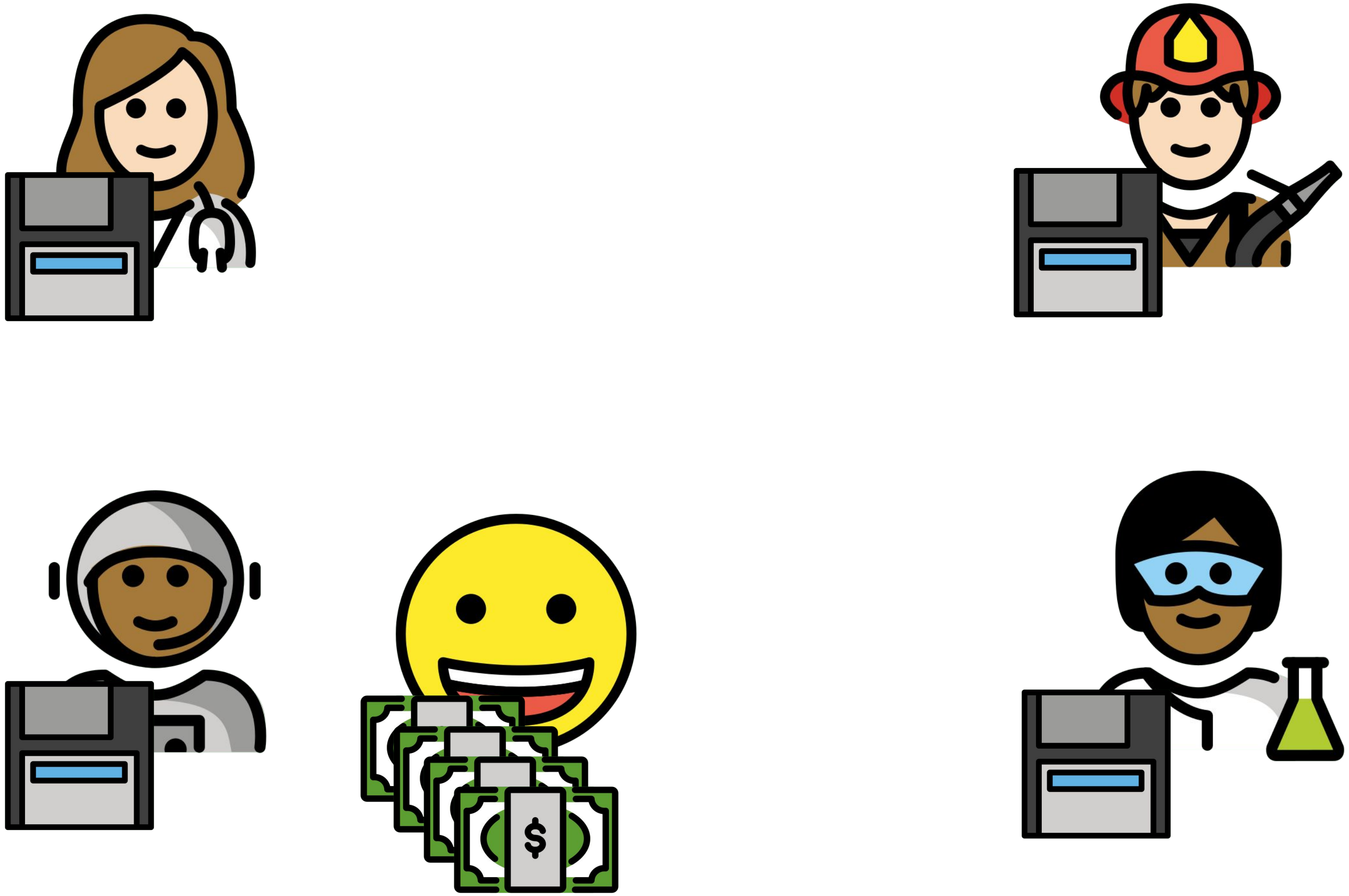


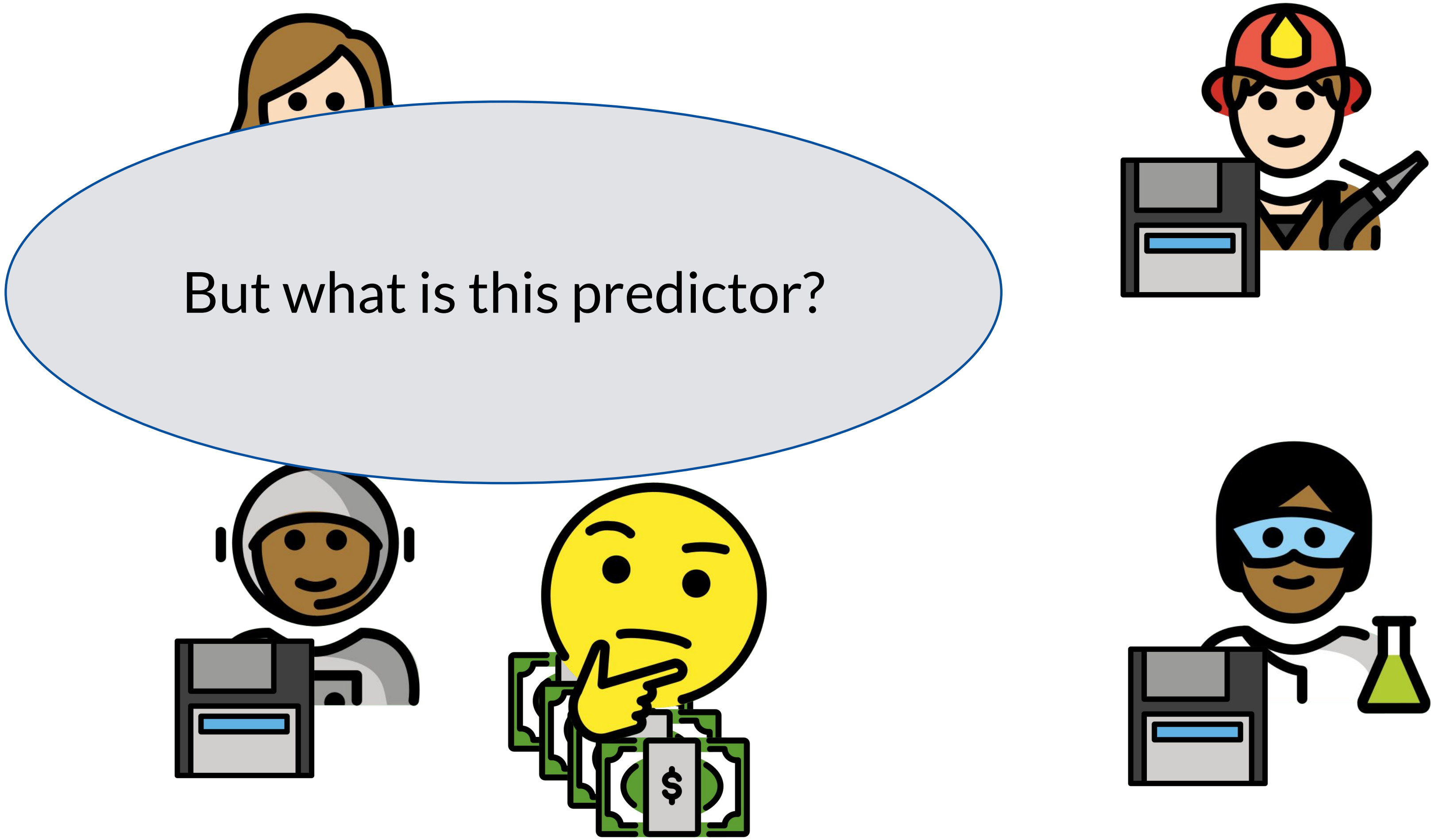
The State of Data Science



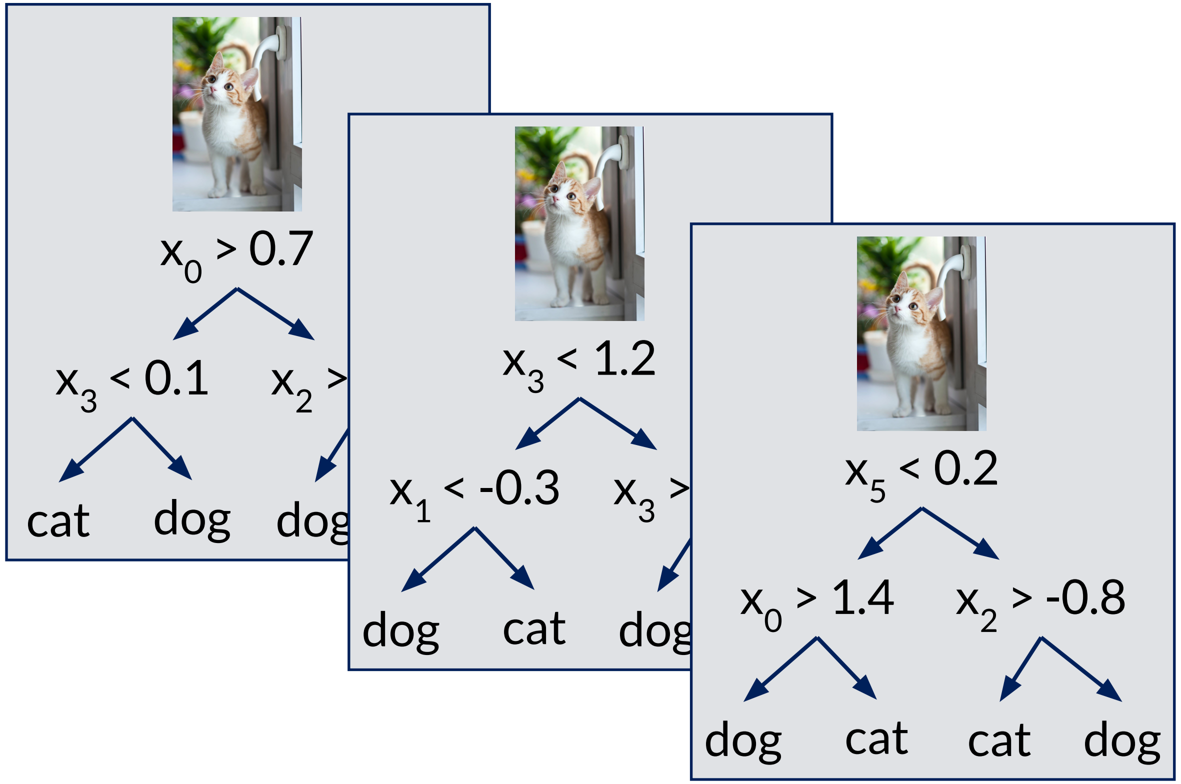


The State of Data Science




But what is this predictor?

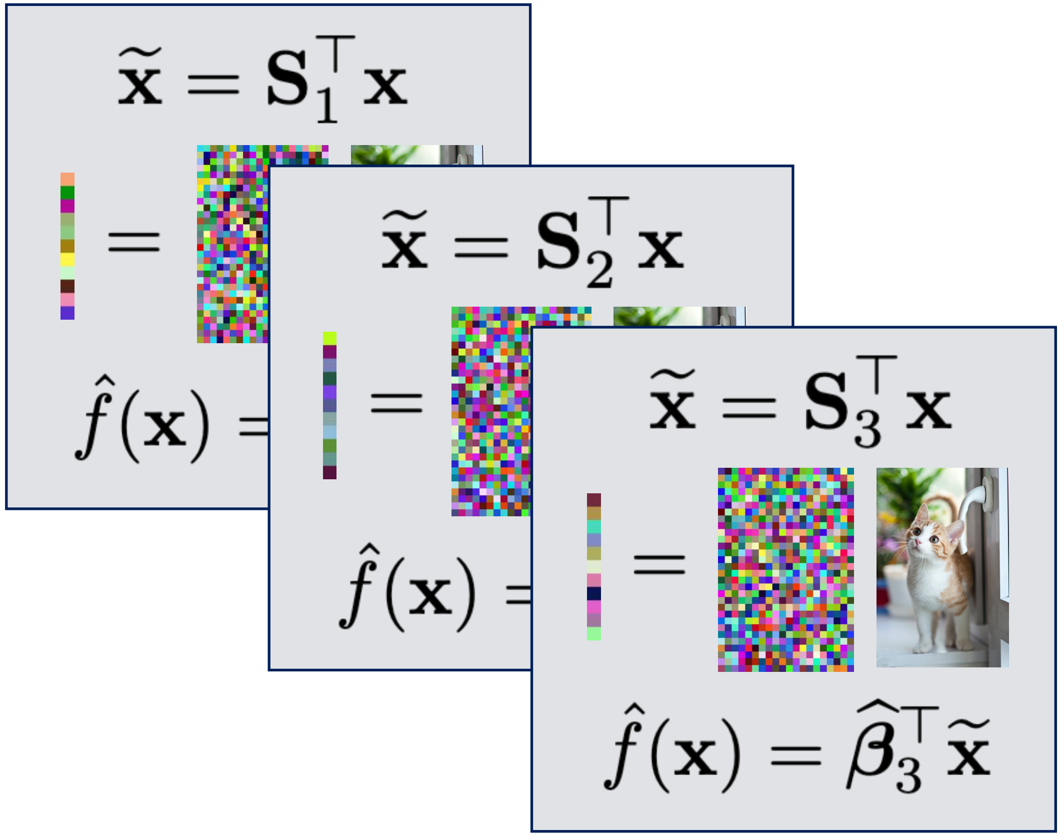
- Random forests [1]
  - Ensemble of decision trees
    - Trained on bootstrap samples
    - Branch on random feature subsets



[1] L Breiman. "Random forests." Machine Learning 45, 2001.

- Random forests [1]
    - Ensemble of decision trees
        - Trained on bootstrap samples
        - Branch on random feature subsets

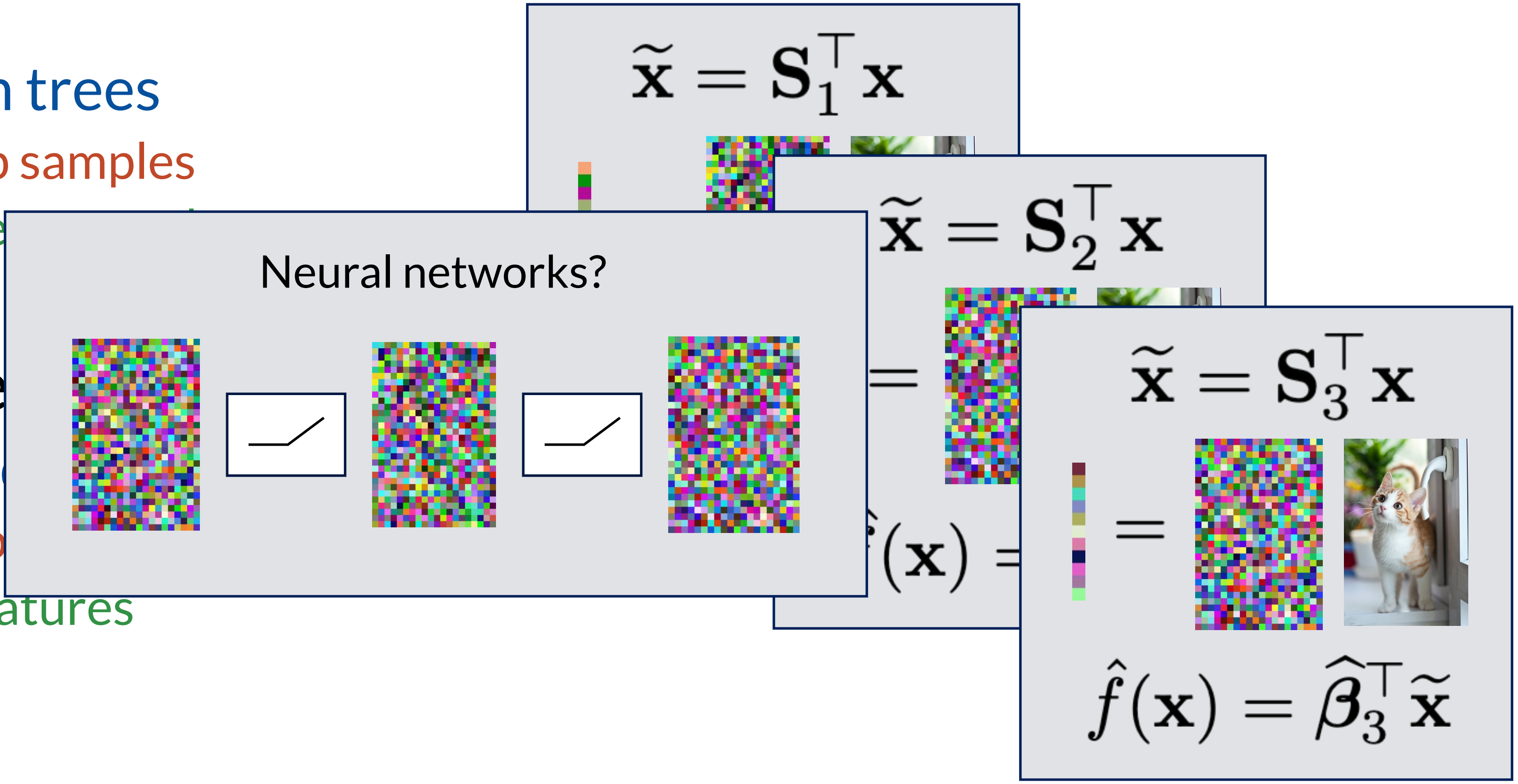
- Random projection ensembles [2]
    - Ensemble of sketched regressors
        - Randomly project observations
        - Randomly project features



[1] L Breiman. "Random forests." Machine Learning 45, 2001.
[2] GA Thanei, C Heinze, N Meinshausen. "Random projections for large-scale regression." Big and Complex Data Analysis, 2017.

Existing Practice

- Random forests [1]
  - Ensemble of decision trees
    - Trained on bootstrap samples
    - Branch on random fe
- Random projection e
  - Ensemble of sketche
    - Randomly project ob
    - Randomly project features



[1] L Breiman. “Random forests.” Machine Learning 45, 2001.
[2] GA Thanei, C Heinze, N Meinshausen. “Random projections for large-scale regression.” Big and Complex Data Analysis, 2017.

Theoretical Questions

- Prior work: What performance do ensembles provably achieve?

- Prior work: What performance do ensembles provably achieve?
  - Random forests are difficult to analyze
    - E.g., purely random forests [3] make analysis tractable
  - Simplified models [e.g., 2, 3] have proven upper bounds
    - Consistent estimation
    - Ensembles strictly better than individual models

[3] S Arlot, R Genuer. "Analysis of purely random forests bias." arXiv preprint arXiv:1407.3939.

Theoretical Questions

- Prior work: What performance do ensembles provably achieve?
  - Random forests are difficult to analyze
    - E.g., purely random forests [3] make analysis tractable
  - Simplified models [e.g., 2, 3] have proven upper bounds
    - Consistent estimation
    - Ensembles strictly better than individual models

- New question: What predictions do ensembles provably make?

[3] S Arlot, R Genuer. "Analysis of purely random forests bias." arXiv preprint arXiv:1407.3939.

Theoretical Questions

- Prior work: What performance do ensembles provably achieve?
  - Random forests are difficult to analyze
    - E.g., purely random forests [3] make analysis tractable
  - Simplified models [e.g., 2, 3] have proven upper bounds
    - Consistent estimation
    - Ensembles strictly better than individual models

But what is this predictor?

- New question: What predictions do ensembles provably make?

[3] S Arlot, R Genuer. "Analysis of purely random forests bias." arXiv preprint arXiv:1407.3939.

Theoretical Questions

- Prior work: What performance do ensembles provably achieve?
  - Random forests are difficult to analyze
    - E.g., purely random forests [3] make analysis tractable
  - Simplified models [e.g., 2, 3] have proven upper bounds
    - Consistent estimation
    - Ensembles strictly better than individual models

But what is this predictor?

- New question: What predictions do ensembles provably make?
  - If there is implicit regularization in ensembles, can we make it explicit?
  - Stronger than consistency, tighter than upper bounds
    - Understand both good/optimal as well as suboptimal ensemble models

[3] S Arlot, R Genuer. "Analysis of purely random forests bias." arXiv preprint arXiv:1407.3939.

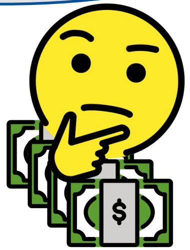
- Prior work: What performance do ensembles provably achieve?
  - Random forests are difficult to analyze
    - E.g., purely random forests [3] make analysis tractable
  - Simplified models [e
    - Consistent estimati
    - Ensembles strictly
- New question: Wha vably make?
  - If there is implicit regularization in ensembles, can we make it explicit?
  - Stronger than consistency, tighter than upper bounds
    - Understand both good/optimal as well as suboptimal ensemble models

Spoiler:

Randomized least squares = ridge + noise

Randomized ensembles = ridge

But what is this predictor?



[3] S Arlot, R Genuer. "Analysis of purely random forests bias." arXiv preprint arXiv:1407.3939.

Notation

- Ground truth function $f : \mathcal{X} \to \mathbb{R}$
- Ensemble of estimators $\hat{f}_1, \ldots, \hat{f}_K : \mathcal{X} \to \mathbb{R}$
- Squared error

$$R(\hat{f}_1, \ldots, \hat{f}_K) \triangleq \mathbb{E}\left[R(\hat{f}_1, \ldots, \hat{f}_K; x)\right]$$

$$R(\hat{f}_1, \ldots, \hat{f}_K; x) \triangleq \mathbb{E}\left[\left(\frac{1}{K}\sum_{k=1}^{K}\hat{f}_k(x) - f(x)\right)^2\right]$$

Early Work: Linear Regression Setting

- Training data: $\mathbf{X} = \left[\mathbf{x}_1 \ldots \mathbf{x}_n\right]^\top \in \mathbb{R}^{n \times p}$

$$\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$$

[4] **DL**, H Javadi, RG Baraniuk. "The implicit regularization of ordinary least squares ensembles." AISTATS, 2020.

Early Work: Linear Regression Setting

- Training data: $\mathbf{X} = \left[\mathbf{x}_1 \ldots \mathbf{x}_n\right]^\top \in \mathbb{R}^{n \times p}$

$$\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$$

- Data model: $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, $f(\mathbf{x}) = \boldsymbol{\beta}^{*\top}\mathbf{x}$, $\|\boldsymbol{\beta}^*\|_2 = 1$

$$y_i \sim \mathcal{N}\left(f(\mathbf{x}_i), \sigma^2\right)$$

[4] **DL**, H Javadi, RG Baraniuk. "The implicit regularization of ordinary least squares ensembles." AISTATS, 2020.

Early Work: Linear Regression Setting

- Training data: $\mathbf{X} = \left[\mathbf{x}_1 \ldots \mathbf{x}_n\right]^\top \in \mathbb{R}^{n \times p}$

  $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$

- Data model: $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mathbf{0}, \mathbf{I}_p\right), \ \ f(\mathbf{x}) = \boldsymbol{\beta}^{*\top} \mathbf{x}\ , \ \ \|\boldsymbol{\beta}^*\|_2 = 1$

  $y_i \sim \mathcal{N}\left(f(\mathbf{x}_i), \sigma^2\right)$

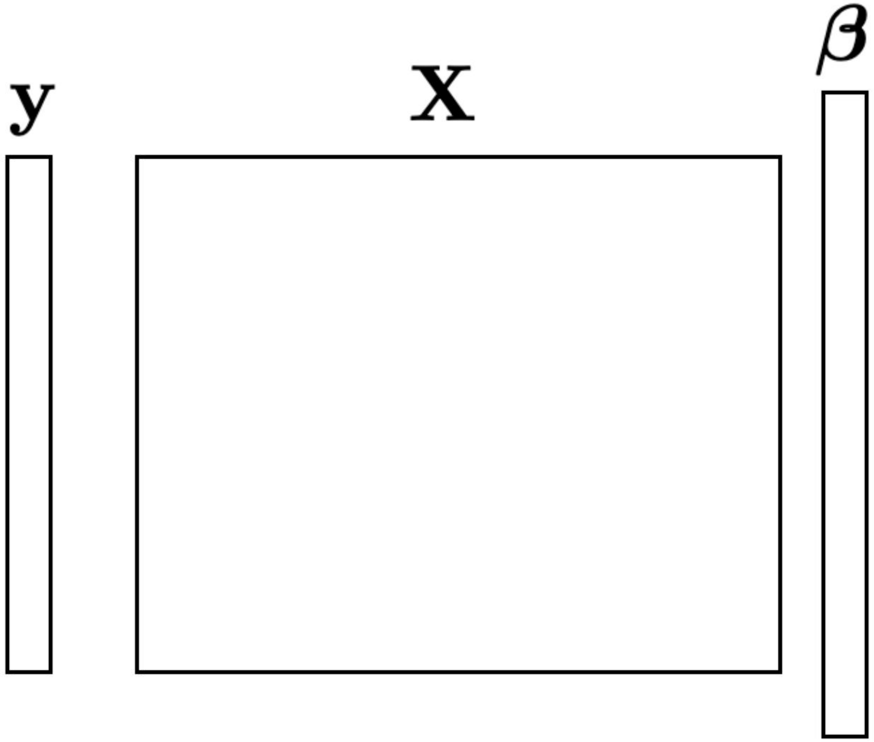
- Estimators: $\hat{f}_k(\mathbf{x}) = \widehat{\boldsymbol{\beta}}_k^\top \mathbf{x}$

  $\widehat{\boldsymbol{\beta}}_k = \mathbf{S}_k \cdot \underset{\mathbf{b}}{\arg\min}\ \frac{1}{2} \left\| \mathbf{T}_k^\top \left(\mathbf{y} - \mathbf{X}\mathbf{S}_k \mathbf{b}\right) \right\|_2^2$

[4] **DL**, H Javadi, RG Baraniuk. "The implicit regularization of ordinary least squares ensembles." AISTATS, 2020.

Random Subsampling

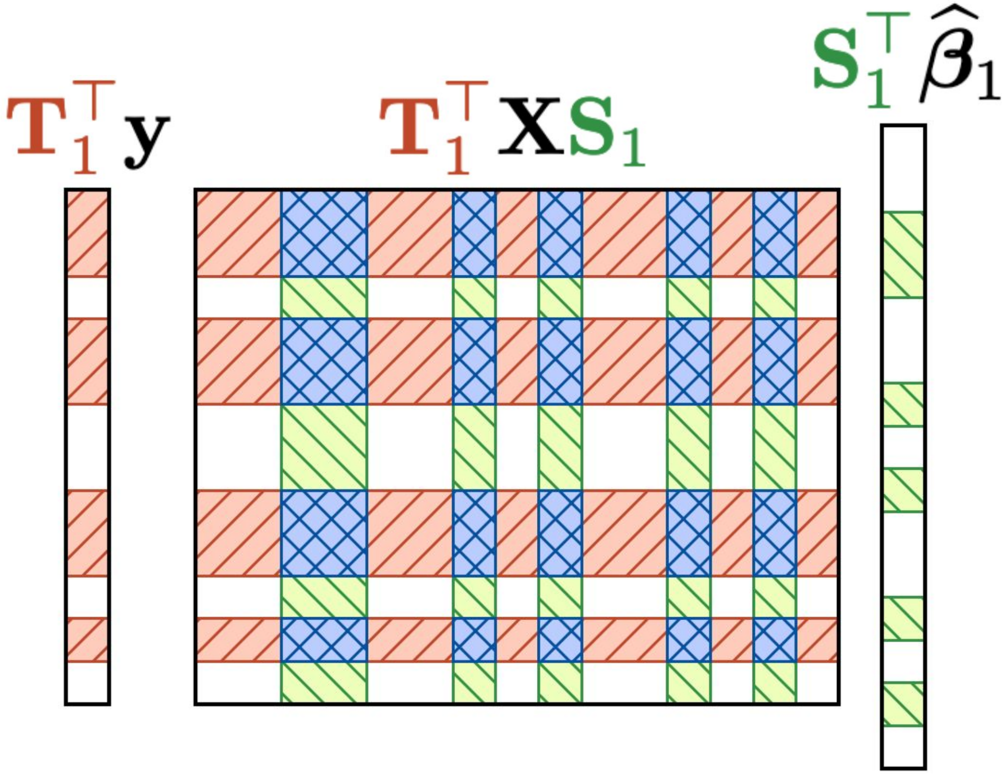
- Subsampling operators $\mathbf{S}_k \in \mathbb{R}^{p \times q}$, $\mathbf{T}_k \in \mathbb{R}^{n \times m}$, $m > q + 1$
  - Uniformly random columns of identity matrix



[4] **DL**, H Javadi, RG Baraniuk. "The implicit regularization of ordinary least squares ensembles." AISTATS, 2020.

Random Subsampling

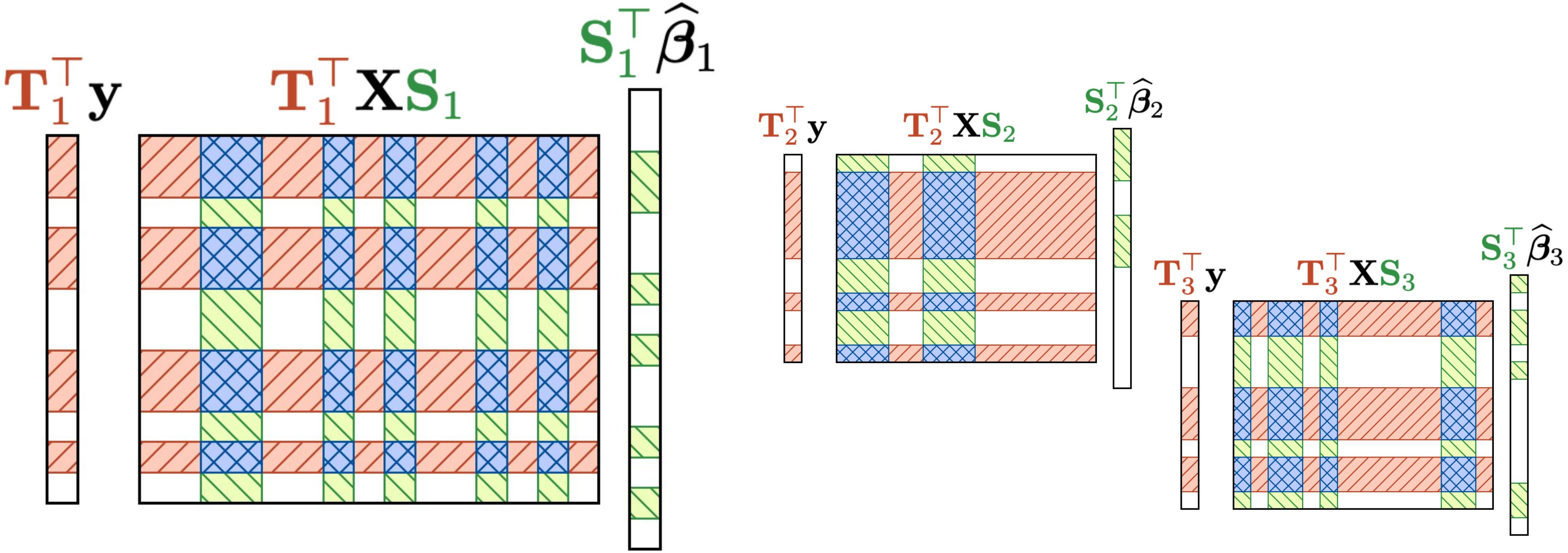
- Subsampling operators $\mathbf{S}_k \in \mathbb{R}^{p\times q}$, $\mathbf{T}_k \in \mathbb{R}^{n\times m}$, $m > q+1$
  - Uniformly random columns of identity matrix



[4] **DL**, H Javadi, RG Baraniuk. “The implicit regularization of ordinary least squares ensembles.” AISTATS, 2020.

Random Subsampling

- Subsampling operators $\mathbf{S}_k \in \mathbb{R}^{p\times q}$, $\mathbf{T}_k \in \mathbb{R}^{n\times m}$, $m > q+1$
  - Uniformly random columns of identity matrix



[4] **DL**, H Javadi, RG Baraniuk. “The implicit regularization of ordinary least squares ensembles.” AISTATS, 2020.

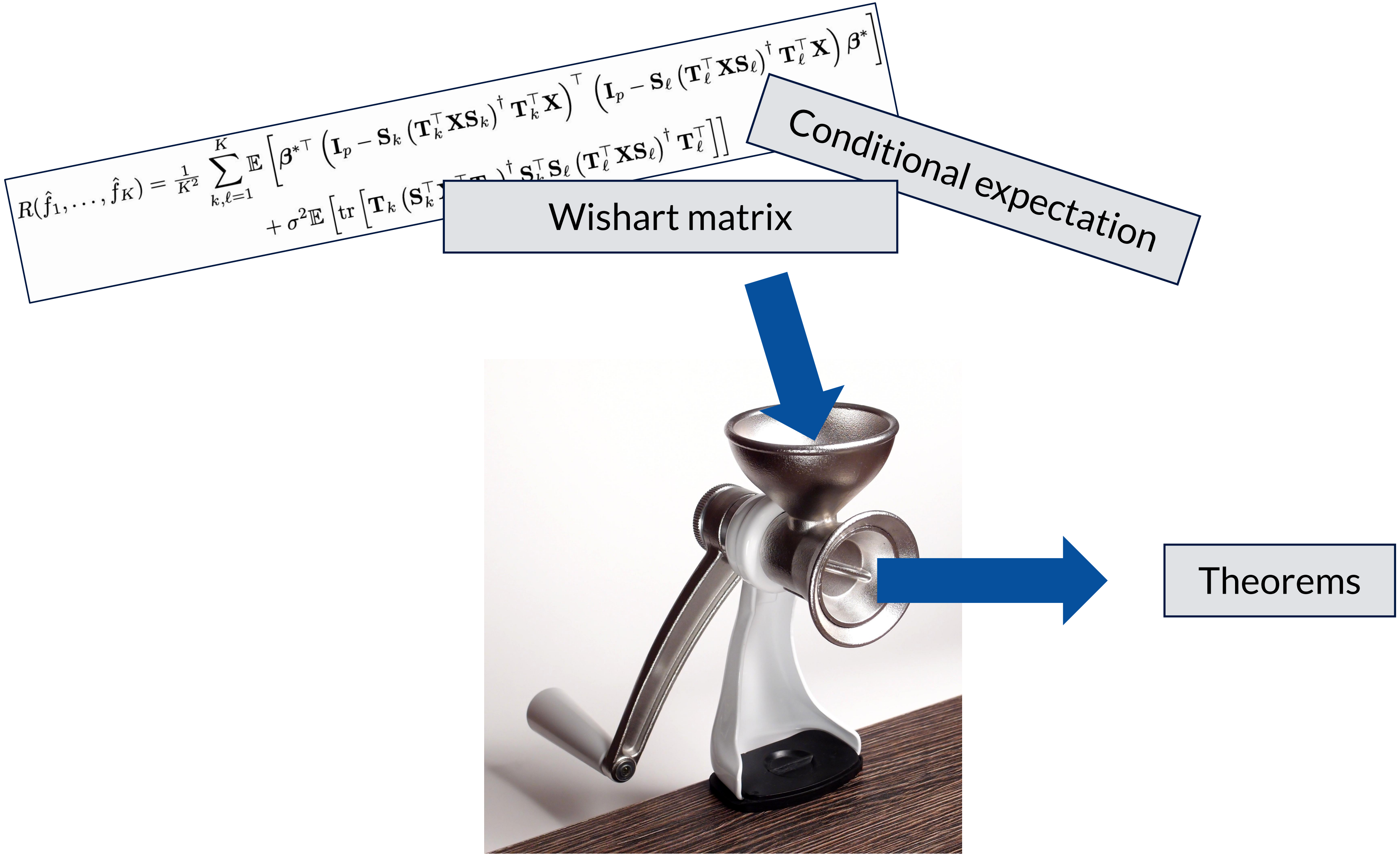
Turn the Crank...

$$R(\hat{f}_1, \ldots, \hat{f}_K) = \frac{1}{K^2} \sum_{k,\ell=1}^{K} \mathbb{E}\left[\boldsymbol{\beta}^{*\top} \left(\mathbf{I}_p - \mathbf{S}_k \left(\mathbf{T}_k^\top \mathbf{X} \mathbf{S}_k\right)^\dagger \mathbf{T}_k^\top \mathbf{X}\right)^\top \left(\mathbf{I}_p - \mathbf{S}_\ell \left(\mathbf{T}_\ell^\top \mathbf{X} \mathbf{S}_\ell\right)^\dagger \mathbf{T}_\ell^\top \mathbf{X}\right) \boldsymbol{\beta}^*\right]$$
$$+ \sigma^2 \mathbb{E}\left[\operatorname{tr}\left[\mathbf{T}_k \left(\mathbf{S}_k^\top \mathbf{X}^\top \mathbf{T}_k\right)^\dagger \mathbf{S}_k^\top \mathbf{S}_\ell \left(\mathbf{T}_\ell^\top \mathbf{X} \mathbf{S}_\ell\right)^\dagger \mathbf{T}_\ell^\top\right]\right]$$

Conditional expectation

Wishart matrix

Turn the Crank...



Error under Proportional Asymptotics

- Exact error expression for finite ensembles

Theorem 1. *In the limit as* $(n, p, m, q) \to \infty$ *with* $p/n \to \gamma$, $m/n \to \eta$, $q/p \to \alpha$, *if* $\eta > \alpha\gamma$,

$$R(\hat{f}_1, \ldots, \hat{f}_K) = \frac{K-1}{K}\left(\frac{(1-\alpha)^2 + \sigma^2\alpha^2\gamma}{1-\alpha^2\gamma}\right) + \frac{1}{K}\left(\frac{\eta(1-\alpha) + \sigma^2\alpha\gamma}{\eta - \alpha\gamma}\right).$$

[4] **DL**, H Javadi, RG Baraniuk. "The implicit regularization of ordinary least squares ensembles." AISTATS, 2020.

Error under Proportional Asymptotics

- Exact error expression for finite ensembles

Theorem 1. *In the limit as* $(n, p, m, q) \to \infty$ *with* $p/n \to \gamma$, $m/n \to \eta$, $q/p \to \alpha$, *if* $\eta > \alpha\gamma$,

$$R(\hat{f}_1, \ldots, \hat{f}_K) = \frac{K-1}{K}\left(\frac{(1-\alpha)^2 + \sigma^2\alpha^2\gamma}{1-\alpha^2\gamma}\right) + \frac{1}{K}\left(\frac{\eta(1-\alpha) + \sigma^2\alpha\gamma}{\eta - \alpha\gamma}\right).$$

- Infinite ensemble error depends only on feature subsampling

$$R_{\text{ens}}^{\infty}(\alpha) \triangleq \lim_{K\to\infty} R(\hat{f}_1, \ldots, \hat{f}_K) = \frac{(1-\alpha)^2 + \sigma^2\alpha^2\gamma}{1-\alpha^2\gamma}$$

[4] **DL**, H Javadi, RG Baraniuk. "The implicit regularization of ordinary least squares ensembles." AISTATS, 2020.

- Ridge regression: $\hat{f}_\lambda(\mathbf{x}) = \widehat{\boldsymbol{\beta}}_\lambda^\top \mathbf{x},\ \widehat{\boldsymbol{\beta}}_\lambda = \arg\min_{\mathbf{b}} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \frac{\lambda}{2}\|\mathbf{b}\|_2^2$

[4] **DL**, H Javadi, RG Baraniuk. "The implicit regularization of ordinary least squares ensembles." AISTATS, 2020.

Optimality Theorem

- Ridge regression: $\hat{f}_\lambda(\mathbf{x}) = \widehat{\boldsymbol{\beta}}_\lambda^\top \mathbf{x},\ \widehat{\boldsymbol{\beta}}_\lambda = \arg\min_{\mathbf{b}} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \frac{\lambda}{2}\|\mathbf{b}\|_2^2$
- Tuning feature subsampling is optimal

Theorem 2. *Under the conditions of Theorem 1 and if* $\boldsymbol{\beta}^* \sim \mathcal{N}(\mathbf{0}, p^{-1}\mathbf{I})$,

$$\inf_{\alpha < \gamma^{-1}} R_{\text{ens}}^{\infty}(\alpha) = \inf_{\lambda > 0} R(\hat{f}_\lambda).$$

[4] **DL**, H Javadi, RG Baraniuk. "The implicit regularization of ordinary least squares ensembles." AISTATS, 2020.

Optimality Theorem

- Ridge regression: $\hat{f}_\lambda(\mathbf{x}) = \widehat{\boldsymbol{\beta}}_\lambda^\top \mathbf{x}, \; \widehat{\boldsymbol{\beta}}_\lambda = \arg\min_{\mathbf{b}} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \frac{\lambda}{2}\|\mathbf{b}\|_2^2$
- Tuning feature subsampling is optimal

Theorem 2. *Under the conditions of Theorem 1 and if* $\boldsymbol{\beta}^* \sim \mathcal{N}(\mathbf{0}, p^{-1}\mathbf{I})$,

$$\inf_{\alpha<\gamma^{-1}} R_{\text{ens}}^{\infty}(\alpha) = \inf_{\lambda>0} R(\hat{f}_\lambda).$$

Spoiler:

Randomized least squares = ridge + noise

Randomized ensembles = ridge

[4] **DL**, H Javadi, RG Baraniuk. "The implicit regularization of ordinary least squares ensembles." AISTATS, 2020.

Optimality Theorem

- Ridge regression: $\hat{f}_\lambda(\mathbf{x}) = \widehat{\boldsymbol{\beta}}_\lambda^\top \mathbf{x},\ \widehat{\boldsymbol{\beta}}_\lambda = \arg\min_{\mathbf{b}} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \frac{\lambda}{2}\|\mathbf{b}\|_2^2$
- Tuning feature subsampling is optimal

Theorem 2. *Under the conditions of Theorem 1 and if* $\boldsymbol{\beta}^* \sim \mathcal{N}(\mathbf{0}, p^{-1}\mathbf{I})$,

$$\inf_{\alpha<\gamma^{-1}} R_{\text{ens}}^{\infty}(\alpha) = \inf_{\lambda>0} R(\hat{f}_\lambda).$$

- However, proof sheds little insight

$$\inf_{\alpha<\gamma^{-1}} R_{\text{ens}}^{\infty}(\alpha) = \frac{1}{2}\left(\frac{\gamma-1}{\gamma} - \sigma^2 + \sqrt{\left(\sigma^2 - \frac{\gamma-1}{\gamma}\right)^2 + 4\sigma^2}\right) = \inf_{\lambda>0} R(\hat{f}_\lambda)$$

[4] **DL**, H Javadi, RG Baraniuk. "The implicit regularization of ordinary least squares ensembles." AISTATS, 2020.

Interpretation of Results

- Since ridge regression is optimal, ensemble is optimal
  - Ridge regression is the minimum mean squared error (MMSE) estimator

Interpretation of Results

- Since ridge regression is optimal, ensemble is optimal
  - Ridge regression is the minimum mean squared error (MMSE) estimator

- Does this imply ensemble converges to ridge regression?
  - For finite dimensions, MMSE is unique
  - For infinite dimensions?
  - Optimality theorem suggests convergence to ridge, but not rigorous

Interpretation of Results

- Since ridge regression is optimal, ensemble is optimal
  - Ridge regression is the minimum mean squared error (MMSE) estimator

- Does this imply ensemble converges to ridge regression?
  - For finite dimensions, MM
  - For infinite dimensio
  - Optimality theorem su … ut not rigorous

But what is this predictor?

- Infinite ensemble with only feature subsampling:

$$\hat{f}_{\text{ens}}^{\infty}(\mathbf{x}) = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \widehat{\boldsymbol{\beta}}_k^\top \mathbf{x} = \mathbf{y}^\top \mathbf{X} \left(\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \mathbf{S}_k \left(\mathbf{S}_k^\top \mathbf{X}^\top \mathbf{X} \mathbf{S}_k \right)^{-1} \mathbf{S}_k^\top \right) \mathbf{x}$$

- Infinite ensemble with only feature subsampling:

$$\hat{f}^{\infty}_{\text{ens}}(\mathbf{x}) = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \widehat{\boldsymbol{\beta}}_k^\top \mathbf{x} = \mathbf{y}^\top \mathbf{X} \left(\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \mathbf{S}_k \left(\mathbf{S}_k^\top \mathbf{X}^\top \mathbf{X} \mathbf{S}_k \right)^{-1} \mathbf{S}_k^\top \right) \mathbf{x}$$

- Determinantal Point Process pseudoinverse [7]:

Theorem (Mutny et al. 2020). *If* $\mathbf{M} \succ \mathbf{0}$ *and* $\mathbf{S} \sim \mathcal{DPP}(\frac{1}{\lambda}\mathbf{M})$,

$$\mathbb{E}\left[\mathbf{S}\left(\mathbf{S}^\top \mathbf{M} \mathbf{S}\right)^{-1} \mathbf{S}^\top\right] = (\mathbf{M} + \lambda \mathbf{I})^{-1},$$

where λ *is the solution to* $\mathbb{E}\left[\frac{q_{\mathrm{S}}}{p}\right] = \mathrm{tr}(\mathbf{M}(\mathbf{M}+\lambda\mathbf{I})^{-1})$.

[5] M Mutny, M Dereziński, A Krause. "Convergence analysis of block coordinate algorithms with determinantal sampling." AISTATS, 2020.

Moving to Sketched Ensembles

- Problem: strict isotropic assumption on data
  - Solution: let subsampling operators be isotropic sketches instead

$$[\mathbf{S}]_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tfrac{1}{q}), \; [\mathbf{T}]_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \frac{1}{m})$$

Moving to Sketched Ensembles

- Problem: strict isotropic assumption on data
  - Solution: let subsampling operators be isotropic sketches instead

$$[\mathbf{S}]_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tfrac{1}{q}), \; [\mathbf{T}]_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tfrac{1}{m})$$

- Problem: marginal error results
  - Solution: use asymptotic equivalences from random matrix theory (RMT)

Moving to Sketched Ensembles

- Problem: strict isotropic assumption on data
  - Solution: let subsampling operators be isotropic sketches instead

$$[\mathbf{S}]_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tfrac{1}{q}), \ \ [\mathbf{T}]_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tfrac{1}{m})$$

- Problem: marginal error results
  - Solution: use asymptotic equivalences from random matrix theory (RMT)
- Problem: estimators use ordinary least squares, $m > q + 1$
  - Solution: consider ridge regression for arbitrary sketch sizes

Moving to Sketched Ensembles

- Problem: strict isotropic assumption on data
  - Solution: let subsampling operators be isotropic sketches instead

$$[\mathbf{S}]_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tfrac{1}{\ldots}) \quad [\mathbf{T}]_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tfrac{1}{\ldots})$$

- Problem: marginal e…
  - Solution: use asymp… trix theory (RMT)

> Spoiler:
>
> Randomized least squares = ridge + noise
>
> Randomized ensembles = ridge

- Problem: estimators use ordinary least squares, $m > q + 1$
  - Solution: consider ridge regression for arbitrary sketch sizes

Asymptotic Equivalences

- Asymptotic equivalences [6]:

Definition. *Two sequences of matrices* $\mathbf{A}_n, \mathbf{B}_n$ *are* ***asymptotically equivalent****, written* $\mathbf{A}_n \simeq \mathbf{B}_n$*, if for every sequence* $\boldsymbol{\Theta}_n$ *having uniformly bounded trace norm, almost surely*

$$\lim_{n \to \infty} \operatorname{tr}\left[\boldsymbol{\Theta}_n(\mathbf{A}_n - \mathbf{B}_n)\right] = 0.$$

[6] E Dobriban, Y Sheng. "Distributed linear regression by averaging." Annals of Statistics, 2021.

Asymptotic Equivalences

- Asymptotic equivalences [6]:

Definition. *Two sequences of matrices* $\mathbf{A}_n, \mathbf{B}_n$ *are* ***asymptotically equivalent****, written* $\mathbf{A}_n \simeq \mathbf{B}_n$*, if for every sequence* $\mathbf{\Theta}_n$ *having uniformly bounded trace norm, almost surely*

$$\lim_{n\to\infty} \operatorname{tr}\left[\mathbf{\Theta}_n(\mathbf{A}_n - \mathbf{B}_n)\right] = 0.$$

- Admits a calculus:
  - Addition $\mathbf{A}_n \simeq \mathbf{B}_n,\ \mathbf{C}_n \simeq \mathbf{D}_n \implies \mathbf{A}_n + \mathbf{C}_n \simeq \mathbf{B}_n + \mathbf{D}_n$
  - Multiplication $\mathbf{A}_n \simeq \mathbf{B}_n \implies \mathbf{C}_n\mathbf{A}_n\mathbf{D}_n \simeq \mathbf{C}_n\mathbf{B}_n\mathbf{D}_n$
  - Elements $\mathbf{A}_n \simeq \mathbf{B}_n \implies [\mathbf{A}_n]_{ij} - [\mathbf{B}_n]_{ij} \xrightarrow{\text{a.s.}} 0$
  - Differentiation [7] $f(\mathbf{A}_n; z) \simeq g(\mathbf{B}_n; z) \implies f'(\mathbf{A}_n; z) \simeq g'(\mathbf{B}_n; z)$

[6] E Dobriban, Y Sheng. "Distributed linear regression by averaging." Annals of Statistics, 2021.
[7] E Dobriban, Y Sheng. "WONDER: Weighted one-shot distributed ridge regression in high dimensions." JMLR, 2020.

An Asymptotic Equivalence of Resolvents

Theorem (Rubio & Mestre 2011). *Let* $\mathbf{Z} \in \mathbb{C}^{n\times p}$ *be a random matrix consisting of i.i.d. random variables that have mean 0, variance 1, and finite absolute moment of order* $8+\delta$ *for some* $\delta > 0$. *Let* $\boldsymbol{\Sigma} \in \mathbb{C}^{p\times p}$ *be a positive semidefinite matrix with operator norm uniformly bounded in* p, *and let* $\mathbf{X} = \mathbf{Z}\boldsymbol{\Sigma}^{1/2}$. *Then, for* $z \in \mathbb{C}^+$, *as* $n, p \to \infty$ *such that* $0 < \liminf \frac{p}{n} \leq \limsup \frac{p}{n} < \infty$, *we have*

$$\left(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p\right)^{-1} \simeq \left(c(z)\boldsymbol{\Sigma} - z\mathbf{I}_p\right)^{-1}, \tag{32}$$

where $c(z)$ *is the unique solution in* $\mathbb{C}^-$ *to the fixed point equation*

$$\frac{1}{c(z)} - 1 = \tfrac{1}{n}\mathrm{tr}\left[\boldsymbol{\Sigma}\left(c(z)\boldsymbol{\Sigma} - z\mathbf{I}_p\right)^{-1}\right]. \tag{33}$$

Furthermore, $\frac{1}{p}\mathrm{tr}\left[\boldsymbol{\Sigma}(c(z)\boldsymbol{\Sigma} - z\mathbf{I}_p)^{-1}\right]$ *is a Stieltjes transform of a certain positive measure on* $\mathbb{R}_{\geq 0}$ *with total mass* $\frac{1}{p}\mathrm{tr}[\boldsymbol{\Sigma}]$.

[8] F Rubio, X Mestre. "Spectral convergence for a general class of random matrices." Statistics & Probability Letters, 2011.

An Asymptotic Equivalence of Resolvents

Theorem (Rubio & Mestre 2011). *Let* $\mathbf{Z} \in \mathbb{C}^{n\times p}$ *be a random matrix consisting of i.i.d. random variables that have mean 0, variance 1, and finite absolute moment of order* $8+\delta$ *for some* $\delta > 0$*. Let* $\mathbf{\Sigma} \in \mathbb{C}^{p\times p}$ *be a positive semidefinite matrix with operator norm uniformly bounded in* p*, and let* $\mathbf{X} = \mathbf{Z}\mathbf{\Sigma}^{1/2}$*. Then, for* $z \in \mathbb{C}^{+}$*, as* $n, p \to \infty$ *such that* $0 < \liminf \frac{p}{n} \leq \limsup \frac{p}{n} < \infty$*, we have*

$$\left(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p\right)^{-1} \simeq (c(z)\mathbf{\Sigma} - z$$

What about real arguments?

where $c(z)$ *is the unique solution in* $\mathbb{C}^{-}$ *to the fixed point equation*

$$\frac{1}{c(z)} - 1 = \frac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}\left(c(z)\mathbf{\Sigma} - z\mathbf{I}_p\right)^{-1}\right]. \tag{33}$$

Furthermore, $\frac{1}{p}\mathrm{tr}\left[\mathbf{\Sigma}(c(z)\mathbf{\Sigma} - z\mathbf{I}_p)^{-1}\right]$ *is a Stieltjes transform of a certain positive measure on* $\mathbb{R}_{\geq 0}$ *with total mass* $\frac{1}{p}\mathrm{tr}[\mathbf{\Sigma}]$*.*

[8] F Rubio, X Mestre. "Spectral convergence for a general class of random matrices." Statistics & Probability Letters, 2011.

Real-valued Asymptotic Equivalence

Theorem (Rubio & Mestre 2011). *Let* $\mathbf{Z} \in \mathbb{C}^{n\times p}$ *be a random matrix consisting of i.i.d. random variables that have mean 0, variance 1, and finite absolute moment of order* $8+\delta$ *for some* $\delta > 0$. *Let* $\mathbf{\Sigma} \in \mathbb{C}^{p\times p}$ *be a positive semidefinite matrix with operator norm uniformly bounded in* p, *and let* $\mathbf{X} = \mathbf{Z}\mathbf{\Sigma}^{1/2}$. *Then, for* $z \in \mathbb{C}^+$, *as* $n, p \to \infty$ *such that* $0 < \liminf \frac{p}{n} \leq \limsup \frac{p}{n} < \infty$, *we have*

$$\left(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p\right)^{-1} \simeq \left(c(z)\mathbf{\Sigma} - z\mathbf{I}_p\right)^{-1}, \tag{32}$$

where $c(z)$ *is the unique solution in* $\mathbb{C}^-$ *to the fixed point equation*

$$\frac{1}{c(z)} - 1 = \tfrac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}\left(c(z)\mathbf{\Sigma} - z\mathbf{I}_p\right)^{-1}\right]. \tag{33}$$

Furthermore, $\frac{1}{p}\mathrm{tr}\left[\mathbf{\Sigma}(c(z)\mathbf{\Sigma} - z\mathbf{I}_p)^{-1}\right]$ *is a Stieltjes transform of a certain positive measure on* $\mathbb{R}_{\geq 0}$ *with total mass* $\frac{1}{p}\mathrm{tr}[\mathbf{\Sigma}]$.

Real-valued Asymptotic Equivalence

Theorem (Rubio & Mestre 2011). *Let $\mathbf{Z} \in \mathbb{C}^{n\times p}$ be a random matrix consisting of i.i.d. random variables that have mean 0, variance 1, and finite absolute moment of order $8+\delta$ for some $\delta > 0$. Let $\mathbf{\Sigma} \in \mathbb{C}^{p\times p}$ be a positive semidefinite matrix with operator norm uniformly bounded in p, and let $\mathbf{X} = \mathbf{Z}\mathbf{\Sigma}^{1/2}$. Then, for $z \in \mathbb{C}^+$, as $n, p \to \infty$ such that $0 < \liminf \frac{p}{n} \leq \limsup \frac{p}{n} < \infty$, we have*

$$\left(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p\right)$$

where $c(z)$ is the u

Analytic continuation →

Furthermore, $\frac{1}{p}\mathrm{tr}\left[\mathbf{\Sigma}(c(z)\mathbf{\Sigma} - z\mathbf{I}_p)^{-1}\right]$ is a Stieltjes transform of a certain positive measure on $\mathbb{R}_{\geq 0}$ with total mass $\frac{1}{p}\mathrm{tr}[\mathbf{\Sigma}]$.

Theorem 3. *Let $\zeta_0, z_0 \in \mathbb{R}$ be the unique solutions, satisfying $\zeta_0 < \lambda^+_{\min}(\mathbf{\Sigma})$, to system of equations*

$$1 = \frac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}^2\left(\mathbf{\Sigma} - \zeta_0\mathbf{I}_p\right)^{-2}\right], \quad z_0 = \zeta_0\left(1 - \frac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}\left(\mathbf{\Sigma} - \zeta_0\mathbf{I}_p\right)^{-1}\right]\right). \tag{34}$$

Then, for each $z \in \mathbb{R}$ satisfying $z < \liminf z_0$, as $n, p \to \infty$ such that $0 < \liminf \frac{p}{n} \leq \limsup \frac{p}{n} < \infty$, we have

$$z\left(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p\right)^{-1} \simeq \zeta\left(\mathbf{\Sigma} - \zeta\mathbf{I}_p\right)^{-1}, \tag{35}$$

where $\zeta \in \mathbb{R}$ is the unique solution in $(-\infty, \zeta_0)$ to the fixed-point equation

$$z = \zeta\left(1 - \frac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}\left(\mathbf{\Sigma} - \zeta\mathbf{I}_p\right)^{-1}\right]\right). \tag{36}$$

Furthermore, as $n, p \to \infty$, $|\zeta + \frac{1}{v(z)}| \xrightarrow{\text{a.s.}} 0$, where $v(z)$ is the companion Stieltjes transform of the spectrum of $\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X}$ given by

$$v(z) = \frac{1}{n}\mathrm{tr}\left[\left(\tfrac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}} - z\mathbf{I}_n\right)^{-1}\right],$$

and $|z_0 - \lambda^+_{\min}(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X})| \xrightarrow{\text{a.s.}} 0$.

Real-valued Asymptotic Equivalence

Limits of negative regularization

Theorem (Rubio & Mestre 2011)**.** *Let $\mathbf{Z} \in \mathbb{C}^{n\times p}$ be a random matrix consisting of i.i.d. random variables that have mean 0, variance 1, and finite absolute moment of order $8+\delta$ for some $\delta > 0$. Let $\mathbf{\Sigma} \in \mathbb{C}^{p\times p}$ be a positive semidefinite matrix with operator norm uniformly bounded in p, and let $\mathbf{X} = \mathbf{Z}\mathbf{\Sigma}^{1/2}$. Then, for $z \in \mathbb{C}^+$, as $n, p \to \infty$ such that $0 < \liminf \frac{p}{n} \leq \limsup \frac{p}{n} < \infty$, we have*

$$\left(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p\right)^{-1} \simeq \left(c(z)\mathbf{\Sigma} - z\mathbf{I}_p\right)^{-1}, \tag{32}$$

where $c(z)$ is the unique solution in $\mathbb{C}^-$ to the fixed point equation

$$\frac{1}{c(z)} - 1 = \tfrac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}\left(c(z)\mathbf{\Sigma} - z\mathbf{I}_p\right)^{-1}\right]. \tag{33}$$

Furthermore, $\frac{1}{p}\mathrm{tr}\left[\mathbf{\Sigma}(c(z)\mathbf{\Sigma} - z\mathbf{I}_p)^{-1}\right]$ is a Stieltjes transform of a certain positive measure on $\mathbb{R}_{\geq 0}$ with total mass $\frac{1}{p}\mathrm{tr}[\mathbf{\Sigma}]$.

Theorem 3. *Let $\zeta_0, z_0 \in \mathbb{R}$ be the unique solutions, satisfying $\zeta_0 < \lambda^+_{\min}(\mathbf{\Sigma})$, to system of equations*

$$1 = \tfrac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}^2\left(\mathbf{\Sigma} - \zeta_0\mathbf{I}_p\right)^{-2}\right], \quad z_0 = \zeta_0\left(1 - \tfrac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}\left(\mathbf{\Sigma} - \zeta_0\mathbf{I}_p\right)^{-1}\right]\right). \tag{34}$$

Then, for each $z \in \mathbb{R}$ satisfying $z < \liminf z_0$, as $n, p \to \infty$ such that $0 < \liminf \frac{p}{n} \leq \limsup \frac{p}{n} < \infty$, we have

$$z\left(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p\right)^{-1} \simeq \zeta\left(\mathbf{\Sigma} - \zeta\mathbf{I}_p\right)^{-1}, \tag{35}$$

where $\zeta \in \mathbb{R}$ is the unique solution in $(-\infty, \zeta_0)$ to the fixed-point equation

$$z = \zeta\left(1 - \tfrac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}\left(\mathbf{\Sigma} - \zeta\mathbf{I}_p\right)^{-1}\right]\right). \tag{36}$$

Furthermore, as $n, p \to \infty$, $|\zeta + \frac{1}{v(z)}| \xrightarrow{\text{a.s.}} 0$, where $v(z)$ is the companion Stieltjes transform of the spectrum of $\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X}$ given by

$$v(z) = \tfrac{1}{n}\mathrm{tr}\left[\left(\tfrac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}} - z\mathbf{I}_n\right)^{-1}\right],$$

and $|z_0 - \lambda^+_{\min}(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X})| \xrightarrow{\text{a.s.}} 0.$

Real-valued Asymptotic Equivalence

Theorem (Rubio & Mestre 2011). *Let $\mathbf{Z} \in \mathbb{C}^{n \times p}$ be a random matrix consisting of i.i.d. random variables that have mean 0, variance 1, and finite absolute moment of order $8+\delta$ for some $\delta > 0$. Let $\mathbf{\Sigma} \in \mathbb{C}^{p \times p}$ be a positive semidefinite matrix with operator norm uniformly bounded in p, and let $\mathbf{X} = \mathbf{Z}\mathbf{\Sigma}^{1/2}$. Then, for $z \in \mathbb{C}^+$, as $n, p \to \infty$ such that $0 < \liminf \frac{p}{n} \leq$*

$$\left(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p\right)^{-1} \simeq \left(c(z)\mathbf{\Sigma} - z\mathbf{I}_p\right)^{-1}$$

$$\frac{1}{c(z)} - 1 = \tfrac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}\left(c(z)\mathbf{\Sigma} - z\mathbf{I}_p\right)^{-1}\right]. \tag{33}$$

Furthermore, $\frac{1}{p}\mathrm{tr}\left[\mathbf{\Sigma}(c(z)\mathbf{\Sigma} - z\mathbf{I}_p)^{-1}\right]$ is a Stieltjes transform of a cert… sitive measure on $\mathbb{R}_{\geq 0}$ with total mass $\frac{1}{p}\mathrm{tr}[\mathbf{\Sigma}]$.

Reparameterization

Theorem 3. *Let $\zeta_0, z_0 \in \mathbb{R}$ be the unique solutions, satisfying $\zeta_0 < \lambda^+_{\min}(\mathbf{\Sigma})$, to system of equations*

$$1 = \tfrac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}^2\left(\mathbf{\Sigma} - \zeta_0\mathbf{I}_p\right)^{-2}\right], \quad z_0 = \zeta_0\left(1 - \tfrac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}\left(\mathbf{\Sigma} - \zeta_0\mathbf{I}_p\right)^{-1}\right]\right). \tag{34}$$

Then, for each $z \in \mathbb{R}$ satisfying $z < \liminf z_0$, as $n, p \to \infty$ such that $0 < \liminf \frac{p}{n} \leq \limsup \frac{p}{n} < \infty$, we have

$$z\left(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p\right)^{-1} \simeq \zeta\left(\mathbf{\Sigma} - \zeta\mathbf{I}_p\right)^{-1}, \tag{35}$$

where $\zeta \in \mathbb{R}$ is the unique solution in $(-\infty, \zeta_0)$ to the fixed-point equation

$$z = \zeta\left(1 - \tfrac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}\left(\mathbf{\Sigma} - \zeta\mathbf{I}_p\right)^{-1}\right]\right). \tag{36}$$

rthermore, as $n, p \to \infty$, $|\zeta + \frac{1}{v(z)}| \xrightarrow{\text{a.s.}} 0$, where $v(z)$ is the companion Stieltjes transform of the spectrum of $\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X}$ given by

$$v(z) = \tfrac{1}{n}\mathrm{tr}\left[\left(\tfrac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}} - z\mathbf{I}_n\right)^{-1}\right],$$

and $|z_0 - \lambda^+_{\min}(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X})| \xrightarrow{\text{a.s.}} 0$.

Real-valued Asymptotic Equivalence

Theorem (Rubio & Mestre 2011)**.** *Let* $\mathbf{Z} \in \mathbb{C}^{n \times p}$ *be a random matrix consisting of i.i.d. random variables that have mean 0, variance 1, and finite absolute moment of order* $8+\delta$ *for some* $\delta > 0$*. Let* $\mathbf{\Sigma} \in \mathbb{C}^{p \times p}$ *be a positive semidefinite matrix with operator norm uniformly bounded in* p*, and let* $\mathbf{X} = \mathbf{Z}\mathbf{\Sigma}^{1/2}$*. Then, for* $z \in \mathbb{C}^+$*, as* $n, p \to \infty$ *such that* $0 < \liminf \frac{p}{n} \leq \limsup \frac{p}{n} < \infty$*, we have*

$$\left(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p\right)^{-1} \simeq \left(c(z)\mathbf{\Sigma} - z\mathbf{I}_p\right)^{-1}, \tag{32}$$

where $c(z)$ *is the unique solution in* $\mathbb{C}^-$ *to the fixed point equation*

$$\frac{1}{c(z)} - 1 = \tfrac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}\left(c(z)\mathbf{\Sigma} - z\mathbf{I}_p\right)^{-1}\right]. \tag{33}$$

Furthermore, $\frac{1}{p}\mathrm{tr}\left[\mathbf{\Sigma}(c(z)\mathbf{\Sigma} - z\mathbf{I}_p)^{-1}\right]$ *is a Stieltjes transform of a certain positive measure on* $\mathbb{R}_{\geq 0}$ *with total mass* $\frac{1}{p}\mathrm{tr}[\mathbf{\Sigma}]$.

Theorem 3. *Let* $\zeta_0, z_0 \in \mathbb{R}$ *be the unique solutions, satisfying* $\zeta_0 < \lambda^+_{\min}(\mathbf{\Sigma})$*, to system of equations*

$$1 = \tfrac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}^2\left(\mathbf{\Sigma} - \zeta_0\mathbf{I}_p\right)^{-2}\right], \quad z_0 = \zeta_0\left(1 - \tfrac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}\left(\mathbf{\Sigma} - \zeta_0\mathbf{I}_p\right)^{-1}\right]\right). \tag{34}$$

Then, for each $z \in \mathbb{R}$ *satisfying* $z < \liminf z_0$*, as* $n, p \to \infty$ *such that* $0 < \liminf \frac{p}{n} \leq \limsup \frac{p}{n} < \infty$*, we have*

$$z\left(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p\right)^{-1} \simeq \zeta\left(\mathbf{\Sigma} - \zeta\mathbf{I}_p\right)^{-1}, \tag{35}$$

where $\zeta \in \mathbb{R}$ *is the unique solution in* $(-\infty$

Explicit form of implicit value

$$z = \zeta\left(1 - \tfrac{1}{n}\mathrm{tr}\left[\mathbf{\Sigma}\left(\mathbf{\Sigma} - \zeta\mathbf{I}_p\right)^{-1}\right]\right). \tag{36}$$

Furthermore, as $n, p \to \infty$, $|\zeta + \frac{1}{v(z)}| \xrightarrow{\text{a.s.}} 0$, *where* $v(z)$ *is the companion Stieltjes transform of the spectrum of* $\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X}$ *given by*

$$v(z) = \tfrac{1}{n}\mathrm{tr}\left[\left(\tfrac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}} - z\mathbf{I}_n\right)^{-1}\right]$$

and $|z_0 - \lambda^+_{\min}(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X})| \xrightarrow{\text{a.s.}} 0$.

Obtaining a Sketching Equivalence

$$z\left(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p\right)^{-1} \simeq \zeta\left(\mathbf{\Sigma} - \zeta\mathbf{I}_p\right)^{-1}$$

Obtaining a Sketching Equivalence

$$z\left(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p\right)^{-1} \simeq \zeta\left(\boldsymbol{\Sigma} - \zeta\mathbf{I}_p\right)^{-1}$$

$$\mathbf{X} = \sqrt{q}\mathbf{S}^{\mathsf{H}}\mathbf{A}^{1/2}$$

$$\lambda = -z$$

$$\mu = -\zeta$$

Obtaining a Sketching Equivalence

$$z\left(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p\right)^{-1} \simeq \zeta\left(\boldsymbol{\Sigma} - \zeta\mathbf{I}_p\right)^{-1}$$

$$\mathbf{X} = \sqrt{q}\mathbf{S}^{\mathsf{H}}\mathbf{A}^{1/2}$$

$$\lambda = -z$$

$$\mu = -\zeta$$

$$\mathbf{I}_p - \lambda\left(\mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^{\mathsf{H}}\mathbf{A}^{1/2} + \lambda\mathbf{I}_p\right)^{-1} \simeq \mathbf{I}_p - \mu\left(\mathbf{A} + \mu\mathbf{I}_p\right)^{-1}$$

$$\implies \mathbf{A}^{1/2}\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}}\mathbf{A}^{1/2} \simeq \mathbf{A}^{1/2}\left(\mathbf{A} + \mu\mathbf{I}_p\right)^{-1}\mathbf{A}^{1/2}$$

Obtaining a Sketching Equivalence

$$z\left(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} - z\mathbf{I}_p\right)^{-1} \simeq \zeta\left(\mathbf{\Sigma} - \zeta\mathbf{I}_p\right)^{-1}$$

Uniform convergence

$$\mathbf{X} = \sqrt{q}\mathbf{S}^{\mathsf{H}}\mathbf{A}^{1/2}$$
$$\lambda = -z$$
$$\mu = -\zeta$$

$$\mathbf{I}_p - \lambda\left(\mathbf{A}^{1/2}\mathbf{S}\mathbf{S}^{\mathsf{H}}\mathbf{A}^{1/2} + \lambda\mathbf{I}_p\right)^{-1} \simeq \mathbf{I}_p - \mu\left(\mathbf{A} + \mu\mathbf{I}_p\right)^{-1}$$

$$\implies \cancel{\mathbf{A}^{1/2}}\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}}\cancel{\mathbf{A}^{1/2}} \simeq \cancel{\mathbf{A}^{1/2}}\left(\mathbf{A} + \mu\mathbf{I}_p\right)^{-1}\cancel{\mathbf{A}^{1/2}}$$

$$\hat{f}_{\mathrm{ens}}(\mathbf{x}) = \mathbf{y}^{\top}\mathbf{X}\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{S}_k\left(\mathbf{S}_k^{\top}\mathbf{X}^{\top}\mathbf{X}\mathbf{S}_k\right)^{-1}\mathbf{S}_k^{\top}\right)\mathbf{x}$$

First-order Sketching Equivalence

Theorem 4. *For each* $\lambda > \limsup \lambda_0$, *as* $q, p \to \infty$,

$$\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}} \simeq \left(\mathbf{A} + \mu \mathbf{I}_p\right)^{-1},$$

where μ *is the unique solution in* (μ_0, ∞) *to the fixed point equation*

$$\lambda = \mu\left(1 - \tfrac{1}{q}\operatorname{tr}\left[\mathbf{A}\left(\mathbf{A} + \mu \mathbf{I}_p\right)^{-1}\right]\right).$$

First-order Sketching Equivalence

Theorem 4. *For each* $\lambda > \limsup \lambda_0$*, as* $q, p \to \infty$,

$$\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}} \simeq \left(\mathbf{A} + \mu \mathbf{I}_p\right)^{-1},$$

where μ *is the unique solution in* (μ_0, ∞) *to the fixed point equation*

$$\lambda = \mu\left(1 - \tfrac{1}{q}\mathrm{tr}\left[\mathbf{A}\left(\mathbf{A} + \mu \mathbf{I}_p\right)^{-1}\right]\right).$$

- That is, sketching + ridge = another ridge without sketching.

First-order Sketching Equivalence

Theorem 4. *For each* $\lambda > \limsup \lambda_0$*, as* $q, p \to \infty$*,*

$$\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}} \simeq (\mathbf{A} + \mu\mathbf{I}_p)^{-1}$$

where μ *is the unique solution in* (μ_0, ∞) *to the fixed point*

$$\lambda = \mu\left(1 - \frac{1}{q}\mathrm{tr}\left[\mathbf{A}\left(\mathbf{A} + \mu\mathbf{I}_p\right)^{-1}\right]\right).$$

$$\mathbb{E}\left[\frac{q_{\mathcal{S}}}{p}\right] = \mathrm{tr}(\mathbf{M}\left(\mathbf{M} + \mu\mathbf{I}\right)^{-1})$$

Same as DPP when $\lambda = 0$

- That is, sketching + ridge = another ridge without sketching.

First-order Sketching Equivalence

Theorem 4. *For each $\lambda > \limsup \lambda_0$, as $q, p \to \infty$,*

$$\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}} \simeq \left(\mathbf{A} + \mu \mathbf{I}_p\right)^{-1},$$

where μ is the unique solution in (μ_0, ∞) to the fixed point equation

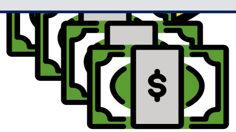
$$\lambda = \mu\left(1 - \tfrac{1}{q}\operatorname{tr}\left[\mathbf{A}\left(\mathbf{A} + \mu \mathbf{I}_p\right)^{-1}\right]\right).$$

- That is, sketching + ketching.

Spoiler:

Randomized least squares = ridge + noise

Randomized ensembles = ridge



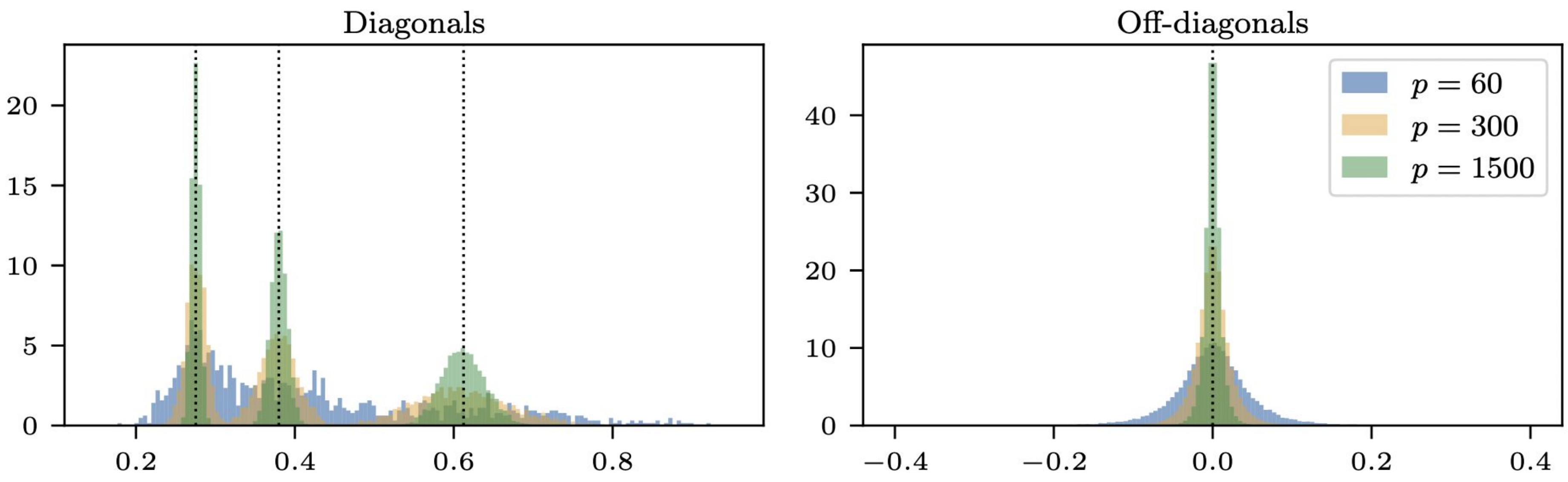
Element-wise Convergence

$\mathbf{A} = \mathrm{diag}(0\ldots,1\ldots,2\ldots)$

$\alpha = 0.8$

$\lambda = 1$

$\mu \approx 1.63$



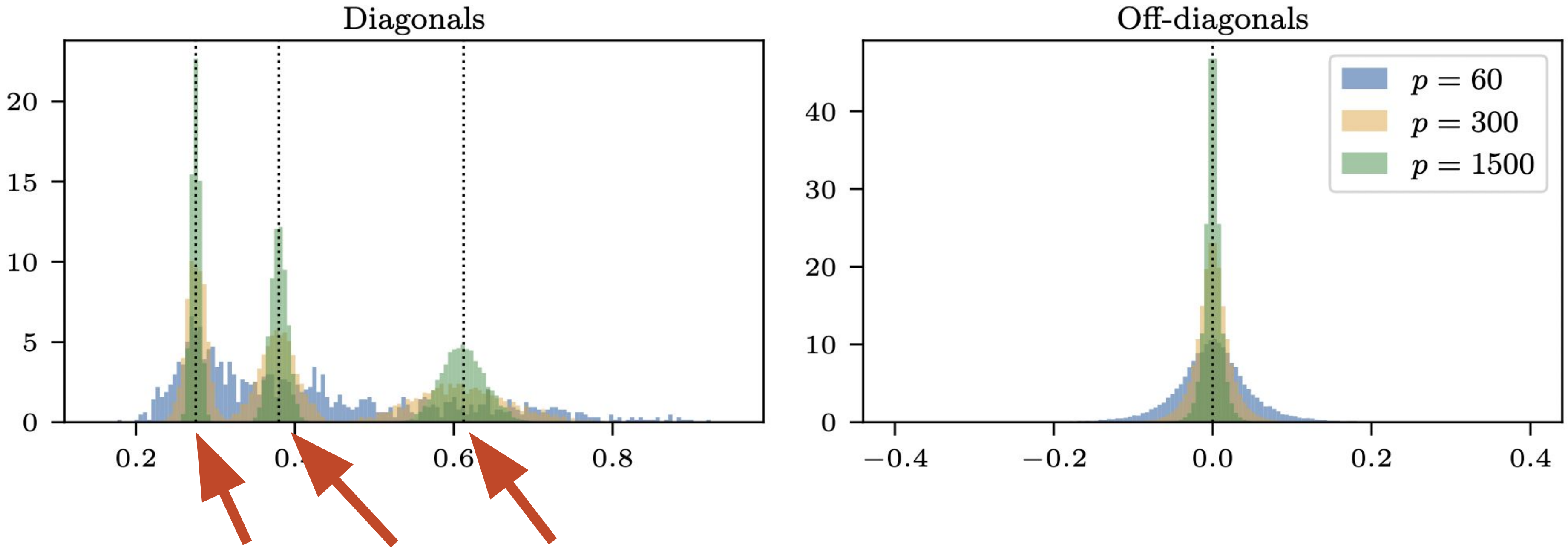
Element-wise Convergence

$\mathbf{A} = \text{diag}(0\ldots, 1\ldots, 2\ldots)$

$\alpha = 0.8$

$\lambda = 1$

$\mu \approx 1.63$



$$\frac{1}{a+\mu} = \frac{1}{3.63}, \quad \frac{1}{2.63}, \quad \frac{1}{1.63}$$

A Blurring Analogy

Original Image

256x256

A Blurring Analogy

$\mathbf{S}^{\mathrm{H}}\mathbf{AS}$
Downsample

Original Image

256x256

64x64

32x32

16x16

A Blurring Analogy

Original Image

256x256

	$\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S}$ Downsample	$\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q\right)^{-1}$ Blur (1.6%)
64x64		
32x32		
16x16		

A Blurring Analogy

Original Image

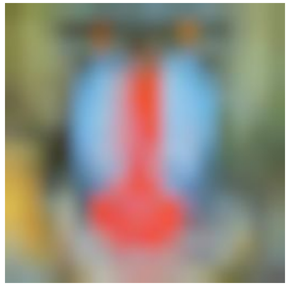
256x256

	$\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S}$ Downsample	$\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S}+\lambda\mathbf{I}_q\right)^{-1}$ Blur (1.6%)	$\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S}+\lambda\mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}}$ Upsample
64x64			
32x32			
16x16			

A Blurring Analogy

Original Image

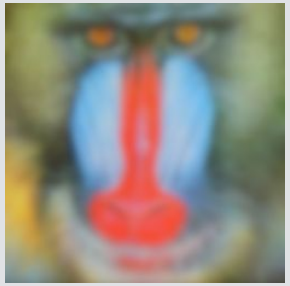
256x256

	$\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S}$ Downsample	$(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q)^{-1}$ Blur (1.6%)	$\mathbf{S}(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q)^{-1}\mathbf{S}^{\mathsf{H}}$ Upsample
64x64			
32x32			
16x16			

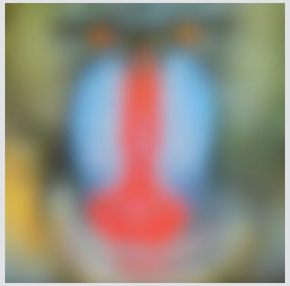
$(\mathbf{A} + \mu\mathbf{I}_p)^{-1}$

Original Image

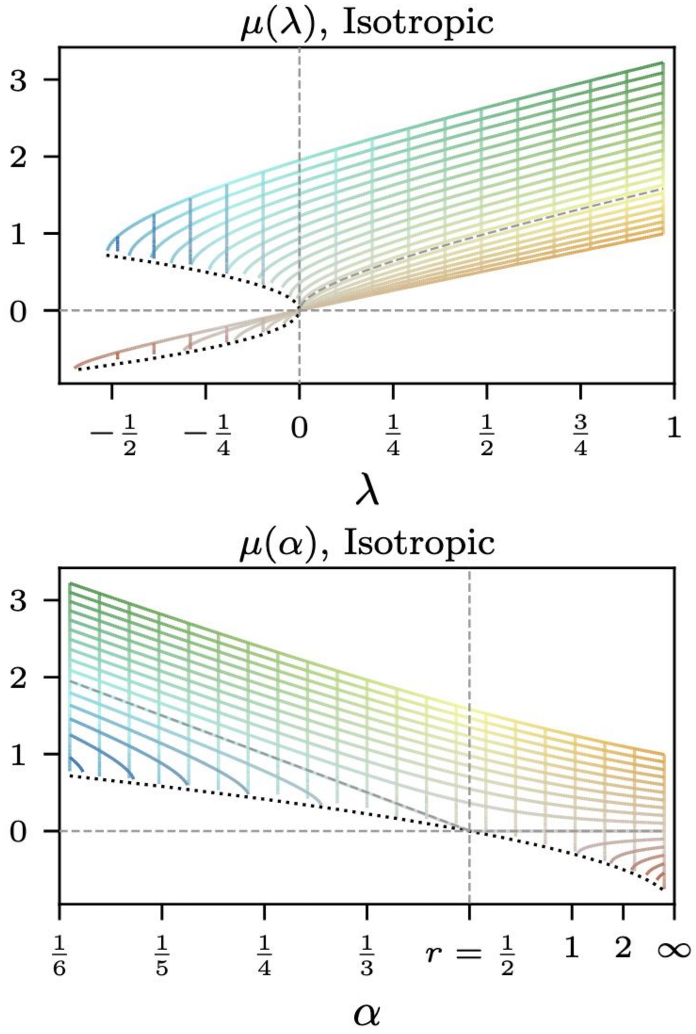
Blur (1.6%)



Blur (3.1%)

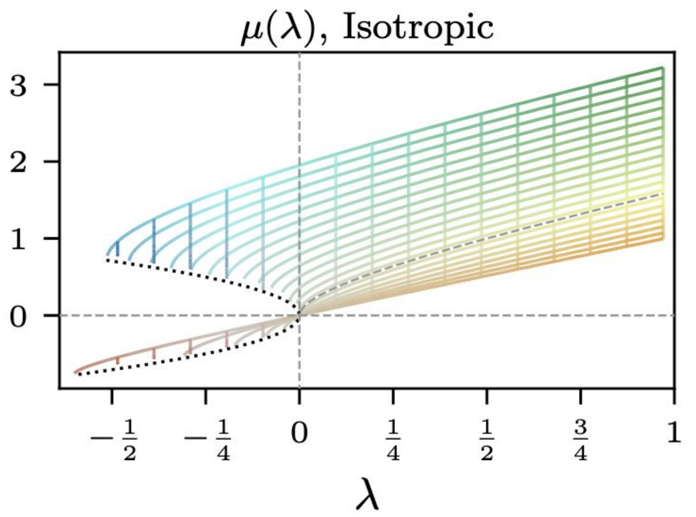


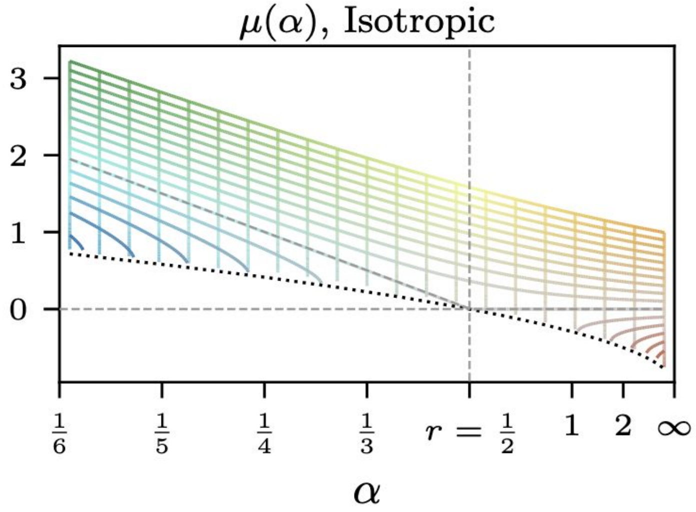
Implicit Regularization Behavior



- Example: isotropic spectrum with $r(\mathbf{A}) = \frac{1}{2}$

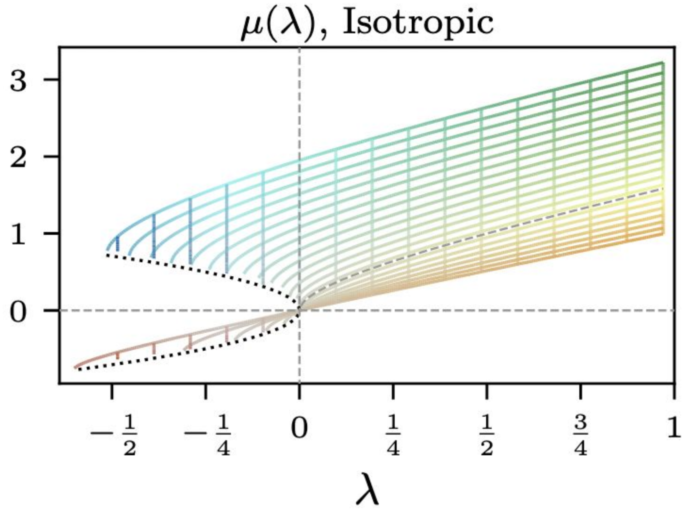
Implicit Regularization Behavior

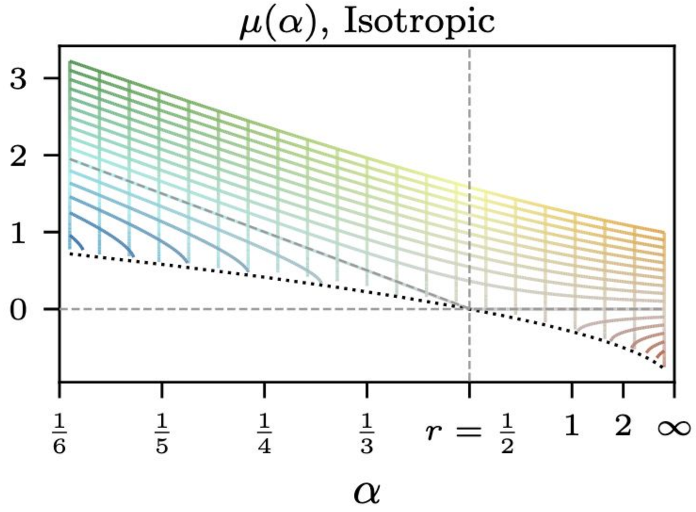




- Example: isotropic spectrum with $r(\mathbf{A}) = \frac{1}{2}$
- $\lambda \mapsto \mu$ is increasing and concave
- $\alpha \mapsto \mu$ is decreasing unless $\alpha > r(\mathbf{A})$ and $\lambda < 0$

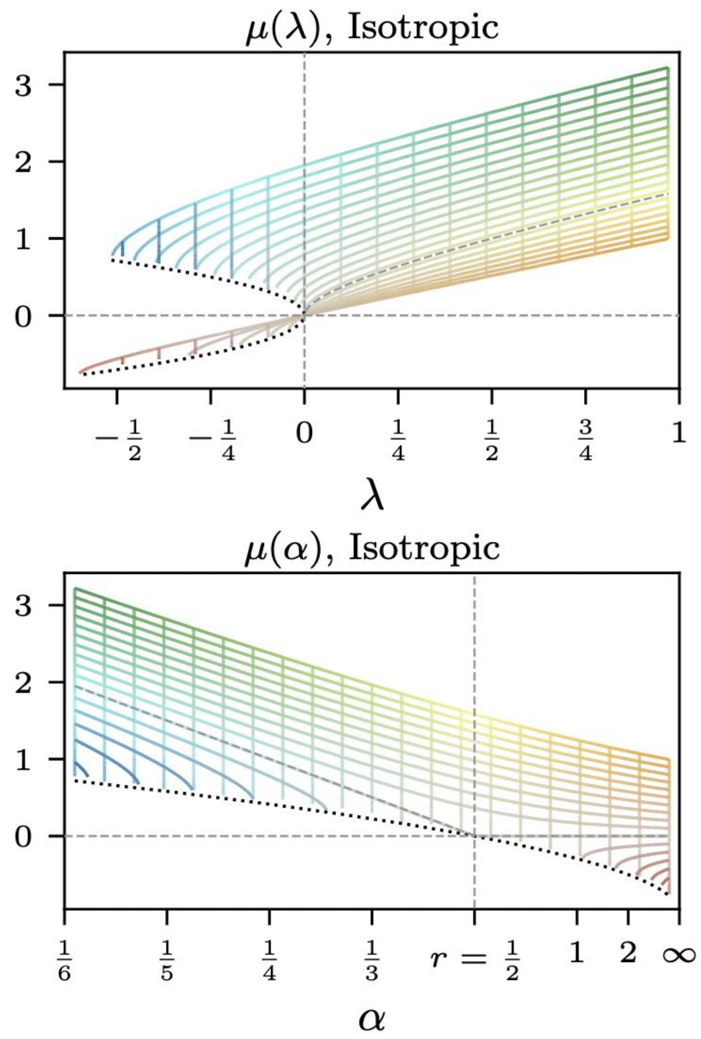
Implicit Regularization Behavior





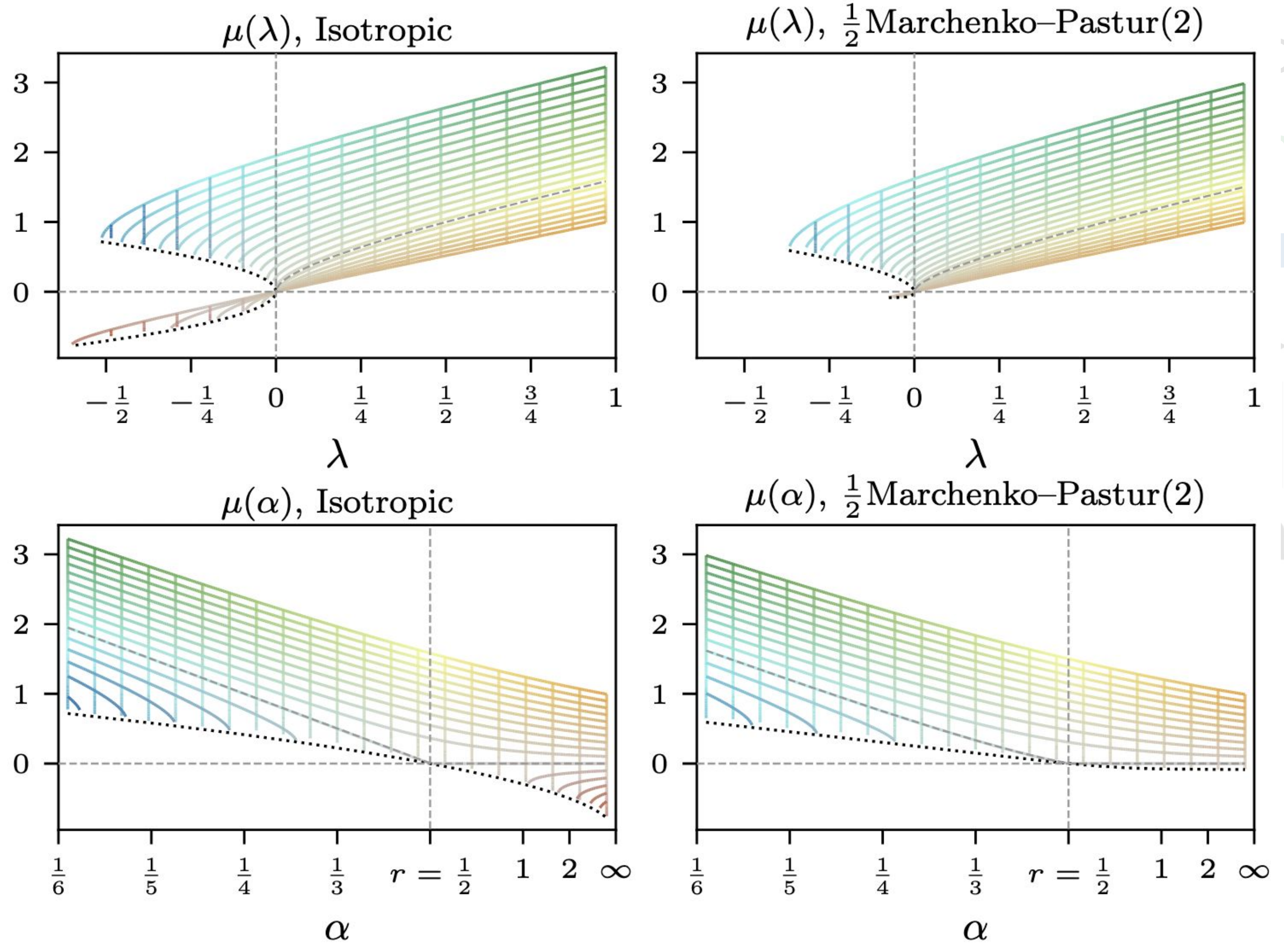
- Example: isotropic spectrum with $r(\mathbf{A}) = \frac{1}{2}$
- $\lambda \mapsto \mu$ is increasing and concave
- $\alpha \mapsto \mu$ is decreasing unless $\alpha > r(\mathbf{A})$ and $\lambda < 0$
- $\mu \geq \lambda$ unless $\alpha > r(\mathbf{A})$ and $\lambda < 0$
- $\mu \leq \lambda + \frac{1}{q}\mathrm{tr}[\mathbf{A}]$

Implicit Regularization Behavior



- Example: isotropic spectrum with $r(\mathbf{A}) = \frac{1}{2}$
- $\lambda \mapsto \mu$ is increasing and concave
- $\alpha \mapsto \mu$ is decreasing unless $\alpha > r(\mathbf{A})$ and $\lambda < 0$
- $\mu \geq \lambda$ unless $\alpha > r(\mathbf{A})$ and $\lambda < 0$
- $\mu \leq \lambda + \frac{1}{q}\mathrm{tr}[\mathbf{A}]$
- $\mathrm{sign}(\mu) = \mathrm{sign}(\lambda)$ if $\alpha > r(\mathbf{A})$
- else $\mu \geq 0$

Implicit Regularization Behavior



: spectrum with $r(\mathbf{A}) = \frac{1}{2}$

ng and concave

ng unless $\alpha > r(\mathbf{A})$ and $\lambda < 0$

$r(\mathbf{A})$ and $\lambda < 0$

(λ) if $\alpha > r(\mathbf{A})$

First-Order is Not Enough

- First-order equivalence is similar to expectation equivalence

First-Order is Not Enough

- First-order equivalence is similar to expectation equivalence
- $\mathbb{E}[X] = \mathbb{E}[Y]$ does not imply that $\mathbb{E}[X^2] = \mathbb{E}[Y^2]$
- Similarly, products of equivalences do not compose if not independent

First-Order is Not Enough

- First-order equivalence is similar to expectation equivalence
- $\mathbb{E}\left[X\right] = \mathbb{E}\left[Y\right]$ does not imply that $\mathbb{E}\left[X^2\right] = \mathbb{E}\left[Y^2\right]$
- Similarly, products of equivalences do not compose if not independent
- Solution: derivative rule of asymptotic equivalence

$$\frac{d}{dz}\left(\mathbf{A} - z\mathbf{I}\right)^{-1} = -\left(\mathbf{A} - z\mathbf{I}\right)^{-2}$$

Second-order Sketching Equivalence

Theorem 5. *If $\mathbf{\Psi} \in \mathbb{C}^{p \times p}$ is a deterministic or random positive semidefinite matrix independent of $\mathbf{S}$ with $\|\mathbf{\Psi}\|_{\text{op}}$ uniformly bounded in p, then*

$$\mathbf{S}\big(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q\big)^{-1}\mathbf{S}^{\mathsf{H}}\mathbf{\Psi}\mathbf{S}\big(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q\big)^{-1}\mathbf{S}^{\mathsf{H}} \simeq \left(\mathbf{A} + \mu \mathbf{I}_p\right)^{-1}\left(\mathbf{\Psi} + \mu' \mathbf{I}_p\right)\left(\mathbf{A} + \mu \mathbf{I}_p\right)^{-1},$$

where

$$\mu' = \frac{\frac{1}{q}\operatorname{tr}\left[\mu^3\left(\mathbf{A} + \mu \mathbf{I}_p\right)^{-1}\mathbf{\Psi}\left(\mathbf{A} + \mu \mathbf{I}_p\right)^{-1}\right]}{\lambda + \frac{1}{q}\operatorname{tr}\left[\mu^2 \mathbf{A}\left(\mathbf{A} + \mu \mathbf{I}_p\right)^{-2}\right]} \geq 0.$$

Second-order Sketching Equivalence

Theorem 5. *If* $\boldsymbol{\Psi} \in \mathbb{C}^{p\times p}$ *is a deterministic or random positive semidefinite matrix independent of* $\mathbf{S}$ *with* $\|\boldsymbol{\Psi}\|_{\mathrm{op}}$ *uniformly bounded in* p*, then*

$$\mathbf{S}\big(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q\big)^{-1}\mathbf{S}^{\mathsf{H}}\boldsymbol{\Psi}\mathbf{S}\big(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q\big)^{-1}\mathbf{S}^{\mathsf{H}} \simeq (\mathbf{A} + \mu\mathbf{I}_p)^{-1}\left(\boldsymbol{\Psi} + \boxed{\mu'\mathbf{I}_p}\right)(\mathbf{A} + \mu\mathbf{I}_p)^{-1},$$

where

$$\mu' = \frac{\frac{1}{q}\mathrm{tr}\left[\mu^3(\mathbf{A} + \mu\mathbf{I}_p)^{-1}\boldsymbol{\Psi}(\mathbf{A} + \mu\mathbf{I}_p)^{-1}\right]}{\lambda + \frac{1}{q}\mathrm{tr}\left[\mu^2\mathbf{A}(\mathbf{A} + \mu\mathbf{I}_p)^{-2}\right]} \geq 0.$$

- That is, second-order adds an isotropic inflation factor
- Depends significantly on the choice of $\boldsymbol{\Psi}$

Second-order Sketching Equivalence

Theorem 5. *If $\boldsymbol{\Psi} \in \mathbb{C}^{p \times p}$ is a deterministic or random positive semidefinite matrix independent of $\mathbf{S}$ with $\|\boldsymbol{\Psi}\|_{\mathrm{op}}$ uniformly bounded in p, then*

$$\mathbf{S}(\mathbf{S}^{\mathsf{H}}\mathbf{A} \ \ldots \ \simeq (\mathbf{A} + \mu \mathbf{I}_p)^{-1} (\boldsymbol{\Psi} + \mu' \mathbf{I}_p) (\mathbf{A} + \mu \mathbf{I}_p)^{-1},$$

where

$$\ldots \frac{\ldots \boldsymbol{\Psi} (\mathbf{A} + \mu \mathbf{I}_p)^{-1}\big]}{\lambda + \frac{1}{q} \mathrm{tr}\big[\mu^2 \mathbf{A} (\mathbf{A} + \mu \mathbf{I}_p)^{-2}\big]} \geq 0.$$

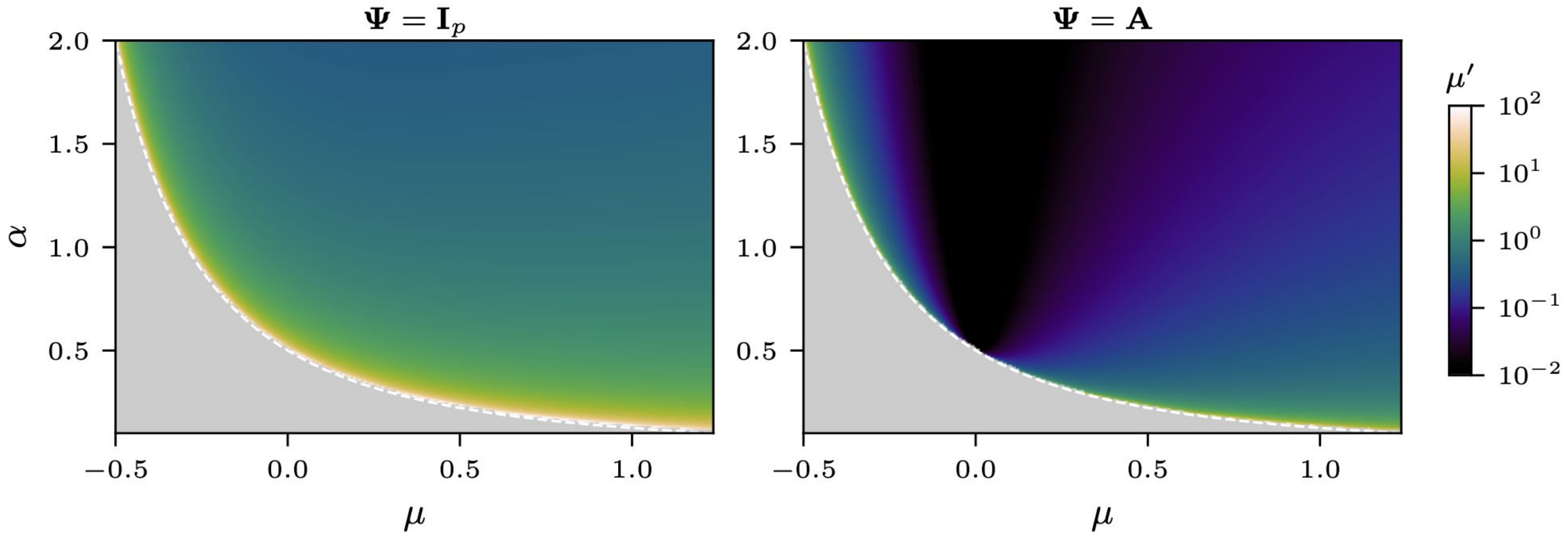
Spoiler:

Randomized least squares = ridge + noise

Randomized ensembles = ridge

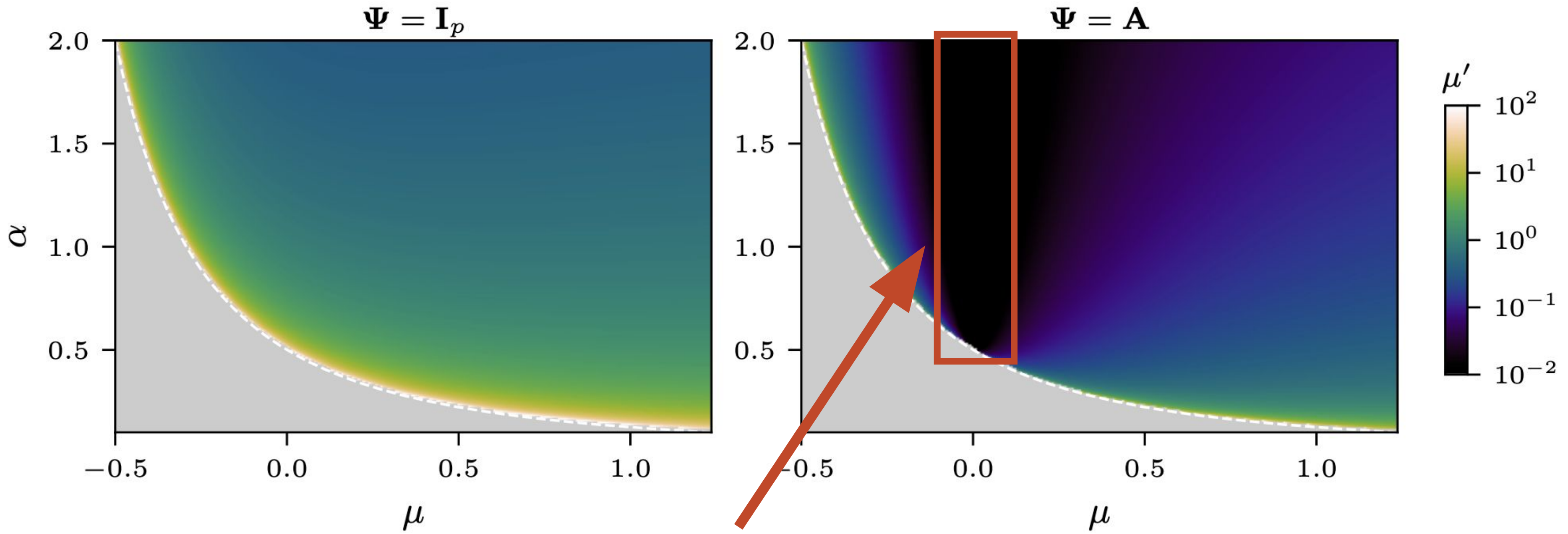
- That is, second-order adds an isotropic inflation factor
- Depends significantly on the choice of $\boldsymbol{\Psi}$

- Example: isotropic spectrum with $r(\mathbf{A}) = \frac{1}{2}$



One Incredible Regime

- Example: isotropic spectrum with $r(\mathbf{A}) = \frac{1}{2}$



- If $\mathbf{\Psi} \in \mathrm{Range}(\mathbf{A})$, $\alpha > r(\mathbf{A})$, and $\mu = \lambda = 0$, there is no inflation
  - Agrees with classical sketching results: sketch larger than rank
  - Sketching is ideal for benign overfitting

Application: Ridge Regression

How about a machine
learning problem?

- Primal (observations) and dual (features) sketching:

$$\widehat{\boldsymbol{\beta}}_P \triangleq \arg\min_{\mathbf{b}} \frac{1}{n}\left\|\mathbf{T}^\top(\mathbf{y} - \mathbf{X}\mathbf{b})\right\|_2^2 + \lambda\|\mathbf{b}\|_2^2$$

$$\widehat{\boldsymbol{\beta}}_D \triangleq \mathbf{S} \cdot \arg\min_{\mathbf{b}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\mathbf{S}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_2^2$$

$$\widehat{\boldsymbol{\beta}}_{PD} \triangleq \mathbf{S} \cdot \arg\min_{\mathbf{b}} \frac{1}{n}\left\|\mathbf{T}^\top(\mathbf{y} - \mathbf{X}\mathbf{S}\mathbf{b})\right\|_2^2 + \lambda\|\mathbf{b}\|_2^2$$

Preparing for Equivalences

- Express in terms of the sketched pseudoinverse:

$$\widehat{\boldsymbol{\beta}}_\psi = \frac{1}{\sqrt{n}}\mathbf{X}_\psi^\ddagger \mathbf{y},\ \psi \in \{P, D, PD\}$$

$$\mathbf{X}_P^\ddagger \triangleq \frac{1}{\sqrt{n}}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{T}\mathbf{T}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{T}\mathbf{T}^\top$$

$$\mathbf{X}_D^\ddagger \triangleq \frac{1}{\sqrt{n}}\mathbf{S}\left(\frac{1}{n}\mathbf{S}^\top\mathbf{X}^\top\mathbf{X}\mathbf{S} + \lambda\mathbf{I}\right)^{-1}\mathbf{S}^\top\mathbf{X}^\top$$

$$\mathbf{X}_{PD}^\ddagger \triangleq \frac{1}{\sqrt{n}}\mathbf{S}\left(\frac{1}{n}\mathbf{S}^\top\mathbf{X}^\top\mathbf{T}\mathbf{T}^\top\mathbf{X}\mathbf{S} + \lambda\mathbf{I}\right)^{-1}\mathbf{S}^\top\mathbf{X}^\top\mathbf{T}\mathbf{T}^\top$$

First-order Data Pseudoinverse Equivalence

Theorem 6. *If* $\left\|\frac{1}{\sqrt{n}}\mathbf{X}\right\|_{\mathrm{op}}$ *is uniformly bounded in* p*, then as* $m, n, q, p \to \infty$*,*

$$\mathbf{X}_{\psi}^{\ddagger} \simeq \frac{1}{\sqrt{n}}\left(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} + \mu_{\psi}\mathbf{I}\right)^{-1}\mathbf{X}^{\mathsf{H}},$$

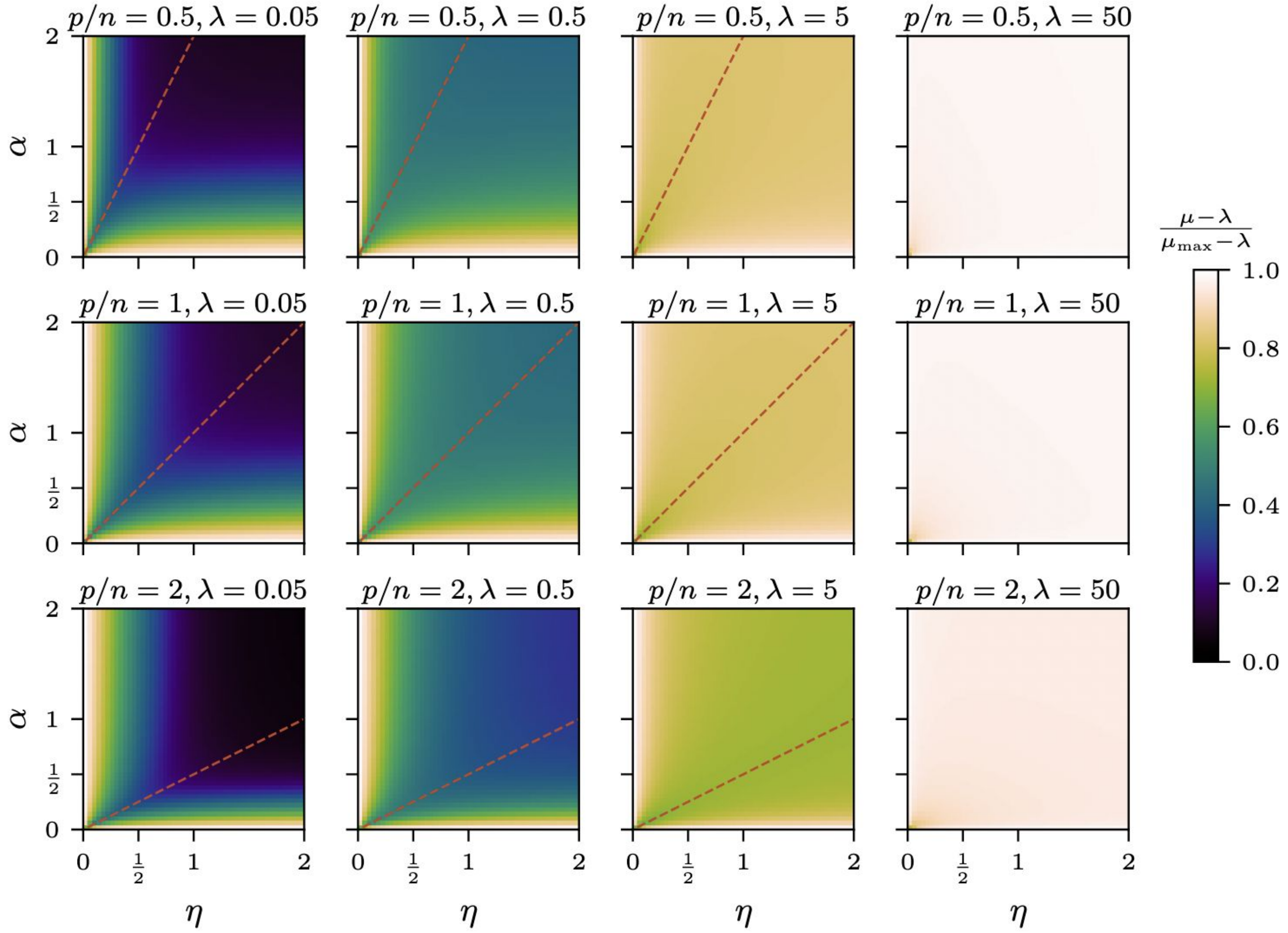
where μ_{ψ} *are is the most positive solutions to the equations*

$$\lambda = \mu_{P}\left(1 - \frac{1}{m}\mathrm{tr}\left[\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X}\left(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} + \mu_{P}\mathbf{I}\right)^{-1}\right]\right),$$

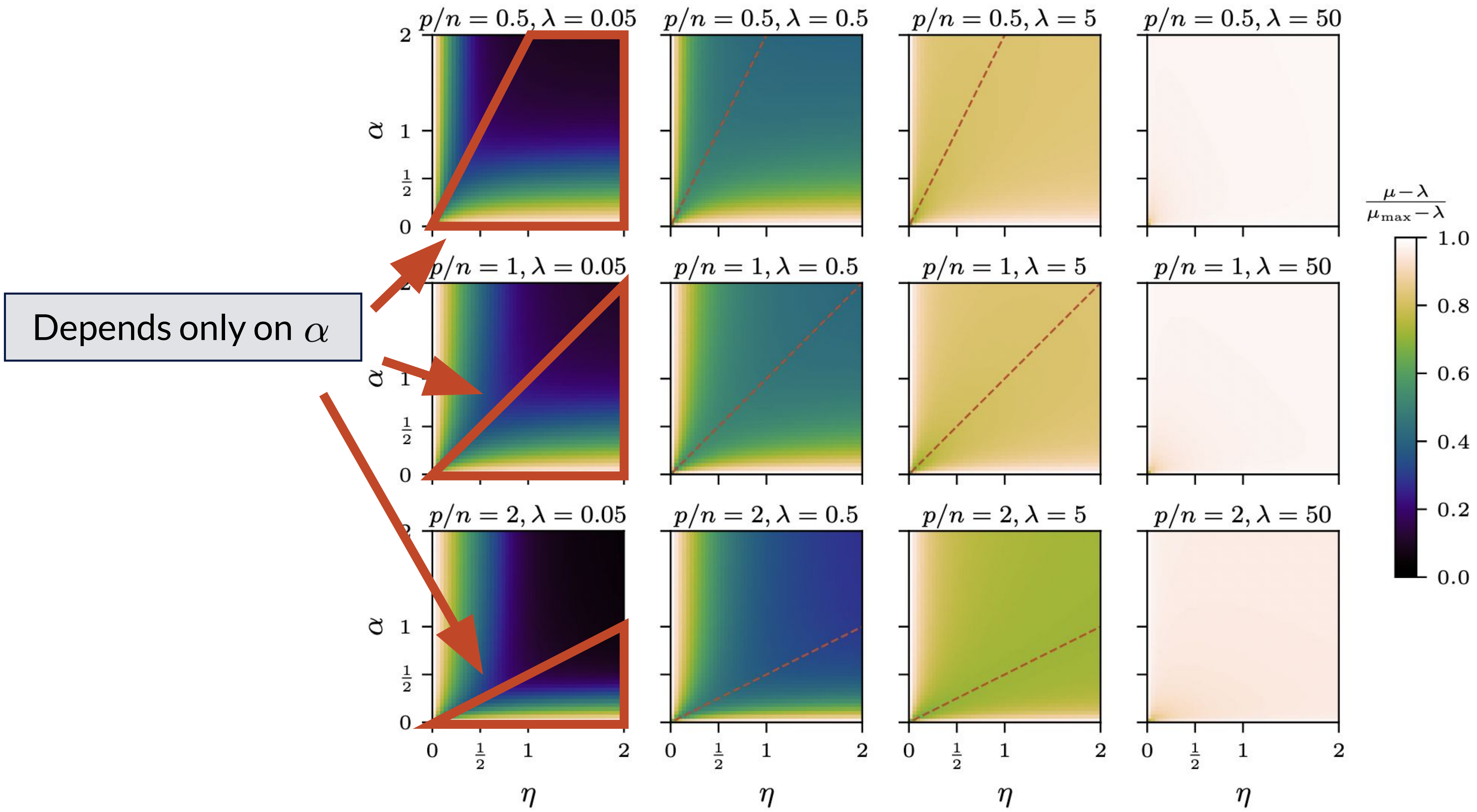
$$\lambda = \mu_{D}\left(1 - \frac{1}{q}\mathrm{tr}\left[\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X}\left(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} + \mu_{D}\mathbf{I}\right)^{-1}\right]\right),$$

$$\frac{\lambda}{\theta} - 1 = \frac{m}{q}\left(\frac{\theta}{\mu_{PD}} - 1\right), \quad \theta = \mu_{PD}\left(1 - \frac{1}{m}\mathrm{tr}\left[\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X}\left(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} + \mu_{PD}\mathbf{I}\right)^{-1}\right]\right).$$

Sharp Phase Transition for Joint Sketching



Sharp Phase Transition for Joint Sketching



Sketching Makes Same Predictions as Ridge

Corollary 7. *If* $\|\frac{1}{\sqrt{n}}\mathbf{y}\|_2$ *is uniformly bounded in* p *and* $\mathbf{w} \in \mathbb{C}^p$ *is independent of* $\mathbf{S}$ *and* $\mathbf{T}$ *such that* $\|\mathbf{w}\|_2$ *is uniformly bounded in* p*, then for any continuous function* $f : \mathbb{C} \to \mathbb{C}$*, as* $p \to \infty$,

$$f(\mathbf{w}^{\mathsf{H}} \widehat{\boldsymbol{\beta}}_{\psi}) - f(\mathbf{w}^{\mathsf{H}} \widehat{\boldsymbol{\beta}}^{\mathrm{ridge}}_{\mu_\psi}) \xrightarrow{\mathrm{a.s.}} 0.$$

Sketching Makes Same Predictions as Ridge

Corollary 7. *If* $\|\frac{1}{\sqrt{n}}\mathbf{y}\|_2$ *is uniformly bounded in* p *and* $\mathbf{w} \in \mathbb{C}^p$ *is independent of* $\mathbf{S}$ *and* $\mathbf{T}$ *such that* $\|\mathbf{w}\|_2$ *is uniformly bounded in* p*, then for any continuous function* $f\colon \mathbb{C} \to \mathbb{C}$*, as* $p \to \infty$*,*

$$f(\mathbf{w}^{\mathsf{H}}\widehat{\boldsymbol{\beta}}_{\psi}) - f(\mathbf{w}^{\mathsf{H}}\widehat{\boldsymbol{\beta}}^{\mathrm{ridge}}_{\mu_\psi}) \xrightarrow{\text{a.s.}} 0.$$

- Sketching asymptotically makes the same prediction as the equivalent ridge on any single test point

Sketching Makes Same Predictions as Ridge

Corollary 7. *If* $\|\frac{1}{\sqrt{n}}\mathbf{y}\|_2$ *is uniformly bounded in* p *and* $\mathbf{w} \in \mathbb{C}^p$ *is independent of* $\mathbf{S}$ *and* $\mathbf{T}$ *such that* $\|\mathbf{w}\|_2$ *is uniformly bounded in* p*, then for any continuous function* $f\colon \mathbb{C} \to \mathbb{C}$*, as* $p \to \infty$*,*

$$f(\mathbf{w}^{\mathsf{H}}\widehat{\boldsymbol{\beta}}_{\psi}) - f(\mathbf{w}^{\mathsf{H}}\widehat{\boldsymbol{\beta}}^{\mathrm{ridge}}_{\mu_\psi}) \xrightarrow{\text{a.s.}} 0.$$

- Sketching asymptotically makes the same prediction as the equivalent ridge on any single test point

Why this qualifier?

Sketching Makes Same Predictions as Ridge

Corollary 7. *If* $\|\frac{1}{\sqrt{n}}\mathbf{y}\|_2$ *is uniformly bounded in* p *and* $\mathbf{w} \in \mathbb{C}^p$ *is independent of* S *and* $\mathbf{T}$ *such that* $\|\mathbf{w}\|_2$ *is uniformly bounded in* p*, then for any continuous functi*… $\mathbb{C}$*, as* $p \to \infty$*,*

$$f(\mathbf{w}^{\mathsf{H}}\widehat{\boldsymbol{\beta}}_{\psi}) - f(\mathbf{w}^{\mathsf{H}}\widehat{\boldsymbol{\beta}}^{\mathrm{ridge}} \ldots$$

- Sketching asymptoticall… prediction as the equivalent ridge on any sin…

Pointwise convergence does not imply uniform convergence

Why this qualifier?

Quadratic Metrics of Sketched Ensembles

- Ensemble of independent sketches: $\widehat{\boldsymbol{\beta}}_{\psi}^{\mathrm{ens}} = \frac{1}{K} \sum_{k=1}^{K} \widehat{\boldsymbol{\beta}}_{\psi}^{(k)}$
- Quadratic error metrics: $\mathcal{E}_{\boldsymbol{\Psi}}\left(\boldsymbol{\beta}, \boldsymbol{\beta}'\right) = \left(\boldsymbol{\beta} - \boldsymbol{\beta}'\right)^{\mathsf{H}} \boldsymbol{\Psi} \left(\boldsymbol{\beta} - \boldsymbol{\beta}'\right)$
  - Includes test risk

Quadratic Metrics of Sketched Ensembles

Theorem 8. *If $\boldsymbol{\Psi} \in \mathbb{C}^{p \times p}$ is a positive semidefinite matrix and $\boldsymbol{\beta}' \in \mathbb{C}^p$ a vector such that $\|\boldsymbol{\Psi}\|_{\mathrm{op}}$ and $\|\boldsymbol{\beta}'\|_2$ are uniformly bounded in p and $(\boldsymbol{\Psi}, \boldsymbol{\beta})$ is independent of $(\mathbf{S}_k, \mathbf{T}_k)_{k=1}^K$, then for $\psi \in \{P, D\}$,*

$$\mathcal{E}_{\boldsymbol{\Psi}}\left(\widehat{\boldsymbol{\beta}}_P^{\mathrm{ens}}, \boldsymbol{\beta}'\right) - \left(\mathcal{E}_{\boldsymbol{\Psi}}\left(\widehat{\boldsymbol{\beta}}_{\mu_P}^{\mathrm{ridge}}, \boldsymbol{\beta}'\right) + \frac{\mu_P'}{Kn}\mathbf{y}^{\mathsf{H}}\left(\tfrac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}} + \mu_P\mathbf{I}\right)^{-2}\mathbf{y}\right) \xrightarrow{\text{a.s.}} 0,$$

$$\mathcal{E}_{\boldsymbol{\Psi}}\left(\widehat{\boldsymbol{\beta}}_D^{\mathrm{ens}}, \boldsymbol{\beta}'\right) - \left(\mathcal{E}_{\boldsymbol{\Psi}}\left(\widehat{\boldsymbol{\beta}}_{\mu_D}^{\mathrm{ridge}}, \boldsymbol{\beta}'\right) + \frac{\mu_D'}{Kn}\mathbf{y}^{\mathsf{H}}\mathbf{X}\left(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} + \mu_D\mathbf{I}\right)^{-2}\mathbf{X}^{\mathsf{H}}\mathbf{y}\right) \xrightarrow{\text{a.s.}} 0,$$

where

$$\mu_P' = \frac{\frac{1}{m}\mathrm{tr}\left[\mu_P^3\left(\tfrac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}} + \mu_P\mathbf{I}\right)^{-1}\tfrac{1}{n}\mathbf{X}\boldsymbol{\Psi}\mathbf{X}^{\mathsf{H}}\left(\tfrac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}} + \mu_P\mathbf{I}\right)^{-1}\right]}{\lambda + \frac{1}{m}\mathrm{tr}\left[\mu_P^2\tfrac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}}\left(\tfrac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}} + \mu_P\mathbf{I}\right)^{-2}\right]},$$

$$\mu_D' = \frac{\frac{1}{q}\mathrm{tr}\left[\mu_D^3\left(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} + \mu_D\mathbf{I}\right)^{-1}\boldsymbol{\Psi}\left(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} + \mu_D\mathbf{I}\right)^{-1}\right]}{\lambda + \frac{1}{q}\mathrm{tr}\left[\mu_D^2\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X}\left(\tfrac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} + \mu_D\mathbf{I}\right)^{-2}\right]}.$$

Quadratic Metrics of Sketched Ensembles

Theorem 8. *If* $\boldsymbol{\Psi} \in \mathbb{C}^{p \times p}$ *is a positive semidefinite matrix and* $\boldsymbol{\beta}' \in \mathbb{C}^p$ *a vector such that* $\|\boldsymbol{\Psi}\|_{\mathrm{op}}$ *and* $\|\boldsymbol{\beta}'\|_2$ *are uniformly bounded in* p *and* $(\boldsymbol{\Psi}, \boldsymbol{\beta})$ *is independent of* $(\mathbf{S}_k, \mathbf{T}_k)_{k=1}^K$, *then for* $\psi \in \{P, D\}$,

$$\mathcal{E}_{\boldsymbol{\Psi}}\left(\widehat{\boldsymbol{\beta}}_{\psi}^{\mathrm{ens}}, \boldsymbol{\beta}'\right) \xrightarrow{\text{a.s.}} \mathcal{E}_{\boldsymbol{\Psi}}\left(\widehat{\boldsymbol{\beta}}_{\mu_\psi}^{\mathrm{ridge}}, \boldsymbol{\beta}'\right) + O\left(\frac{\mu'_\psi}{K}\right)$$

where

$$\mu'_P = \frac{\frac{1}{m}\mathrm{tr}\left[\mu_P^3 \left(\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}} + \mu_P \mathbf{I}\right)^{-1} \frac{1}{n}\mathbf{X}\boldsymbol{\Psi}\mathbf{X}^{\mathsf{H}} \left(\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}} + \mu_P \mathbf{I}\right)^{-1}\right]}{\lambda + \frac{1}{m}\mathrm{tr}\left[\mu_P^2 \frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}} \left(\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}} + \mu_P \mathbf{I}\right)^{-2}\right]},$$

$$\mu'_D = \frac{\frac{1}{q}\mathrm{tr}\left[\mu_D^3 \left(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} + \mu_D \mathbf{I}\right)^{-1} \boldsymbol{\Psi} \left(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} + \mu_D \mathbf{I}\right)^{-1}\right]}{\lambda + \frac{1}{q}\mathrm{tr}\left[\mu_D^2 \frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} \left(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} + \mu_D \mathbf{I}\right)^{-2}\right]}.$$

Quadratic Metrics of Sketched Ensembles

Theorem 8. *If $\Psi \in \mathbb{C}^{p\times p}$ is a positive semidefinite matrix and $\beta' \in \mathbb{C}^p$ a vector such that $\|\Psi\|_{\mathrm{op}}$ and $\|\beta'\|_2$ are uniformly bounded in p and (Ψ, β) is independent of $(\mathbf{S}_k, \mathbf{T}_k)_{k=1}^K$, then for $\psi \in \{P, D\}$,*

$$\mathcal{E}_{\Psi}\left(\widehat{\beta}^{\mathrm{ens}}_{\psi}, \beta'\right) \xrightarrow{\text{a.s.}} \mathcal{E}_{\Psi}\left(\widehat{\beta}^{\mathrm{ridge}}_{\mu_\psi}, \beta'\right) + O\left(\frac{\mu'_\psi}{K}\right)$$

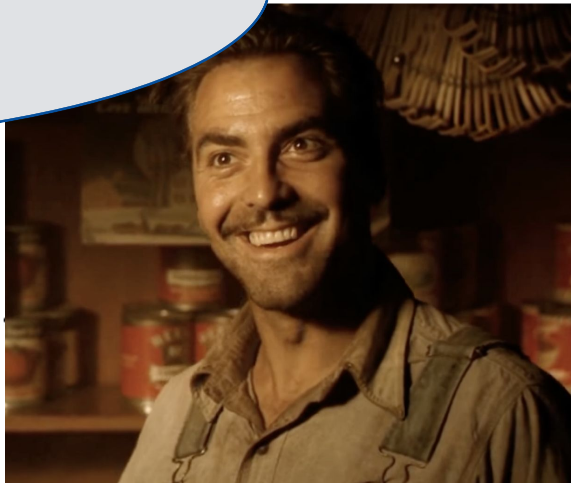
where

Well, ain't this a geometrical oddity.

$O\left(\frac{\mu'_\psi}{K}\right)$ from everywhere!

$$\mu'_P = \frac{\frac{1}{m}\mathrm{tr}\left[\mu_P^3 \left(\frac{1}{n}\mathbf{X}\mathbf{X}\cdots\right.\right.}{\lambda + \frac{1}{m}\mathrm{tr}\left[\mu_P^2 \frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}} \left(\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{H}} + \mu_P \mathbf{I}\right)\cdots\right]}$$

$$\mu'_D = \frac{\frac{1}{q}\mathrm{tr}\left[\mu_D^3 \left(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} + \mu_D \mathbf{I}\right)^{-1} \Psi \left(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} + \mu_D \mathbf{I}\right)^{-1}\right]}{\lambda + \frac{1}{q}\mathrm{tr}\left[\mu_D^2 \frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} \left(\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X} + \mu_D \mathbf{I}\right)^{-2}\right]}$$



Quadratic Conclusions

$$\mathcal{E}_{\boldsymbol{\Psi}}\left(\widehat{\boldsymbol{\beta}}_{\psi}^{\mathrm{ens}}, \boldsymbol{\beta}'\right) \xrightarrow{\text{a.s.}} \mathcal{E}_{\boldsymbol{\Psi}}\left(\widehat{\boldsymbol{\beta}}_{\mu_\psi}^{\mathrm{ridge}}, \boldsymbol{\beta}'\right) + O\left(\frac{\mu_\psi'}{K}\right)$$

- For infinite K, sketched ensemble = ridge

$$\mathcal{E}_{\boldsymbol{\Psi}}\left(\widehat{\boldsymbol{\beta}}_{\psi}^{\mathrm{ens}}, \boldsymbol{\beta}'\right) \xrightarrow{\mathrm{a.s.}} \mathcal{E}_{\boldsymbol{\Psi}}\left(\widehat{\boldsymbol{\beta}}_{\mu_\psi}^{\mathrm{ridge}}, \boldsymbol{\beta}'\right) + O\left(\frac{\mu'_\psi}{K}\right)$$

- For infinite K, sketched ensemble = ridge

Spoiler:

Randomized least squares = ridge + noise

Randomized ensembles = ridge

$$\mathcal{E}_{\boldsymbol{\Psi}}\left(\widehat{\boldsymbol{\beta}}_{\psi}^{\mathrm{ens}}, \boldsymbol{\beta}'\right) \xrightarrow{\text{a.s.}} \mathcal{E}_{\boldsymbol{\Psi}}\left(\widehat{\boldsymbol{\beta}}_{\mu_\psi}^{\mathrm{ridge}}, \boldsymbol{\beta}'\right) + O\left(\frac{\mu'_\psi}{K}\right)$$

- For infinite K, sketched ensemble = ridge
- For finite K, sketched ensemble is worse than ridge

$$\mathcal{E}_{\boldsymbol{\Psi}}\left(\widehat{\boldsymbol{\beta}}_{\psi}^{\text{ens}}, \boldsymbol{\beta}'\right) \xrightarrow{\text{a.s.}} \mathcal{E}_{\boldsymbol{\Psi}}\left(\widehat{\boldsymbol{\beta}}_{\mu_\psi}^{\text{ridge}}, \boldsymbol{\beta}'\right) + O\left(\frac{\mu'_\psi}{K}\right)$$

- For infinite K, sketched ensemble = ridge
- For finite K, sketched ensemble is worse than ridge
  - Unless $\boldsymbol{\Psi} \in \mathrm{Range}(\mathbf{A})$, $\alpha > r(\mathbf{A})$, and $\mu = \lambda = 0$!

A Sketched Ensemble Efficiency Experiment

- Setup: fixed $O(p^2 n)$ budget ensembles with $K = \lfloor \frac{1}{\alpha^2} \rfloor$, $r(\mathbf{\Sigma}) = \frac{1}{2}$

A Sketched Ensemble Efficiency Experiment

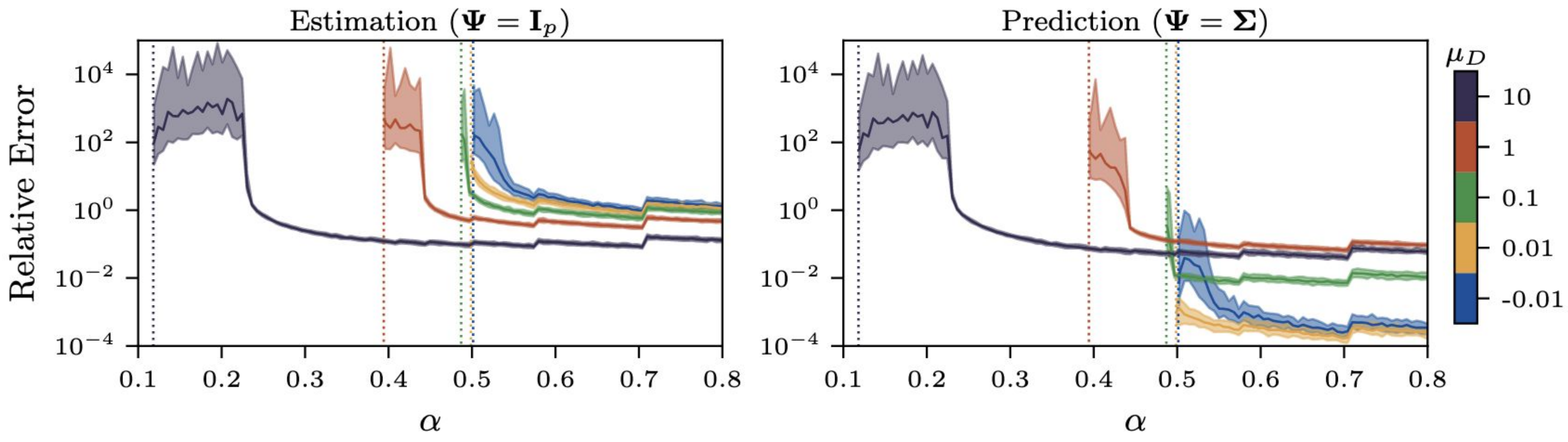
- Setup: fixed $O(p^2 n)$ budget ensembles with $K = \lfloor \frac{1}{\alpha^2} \rfloor$, $r(\mathbf{\Sigma}) = \frac{1}{2}$
- Fixed target μ_D, varying α, with λ uniquely determined

A Sketched Ensemble Efficiency Experiment

- Setup: fixed $O(p^2n)$ budget ensembles with $K = \lfloor \frac{1}{\alpha^2} \rfloor$, $r(\boldsymbol{\Sigma}) = \frac{1}{2}$
- Fixed target μ_D, varying α, with λ uniquely determined
- Error: relative error $\frac{\mathcal{E}_{\boldsymbol{\Psi}}\left(\widehat{\boldsymbol{\beta}}_D^{\text{ens}}, \widehat{\boldsymbol{\beta}}_{\mu_D}^{\text{ridge}}\right)}{\mathcal{E}_{\boldsymbol{\Psi}}\left(\mathbf{0}, \widehat{\boldsymbol{\beta}}_{\mu_D}^{\text{ridge}}\right)}$

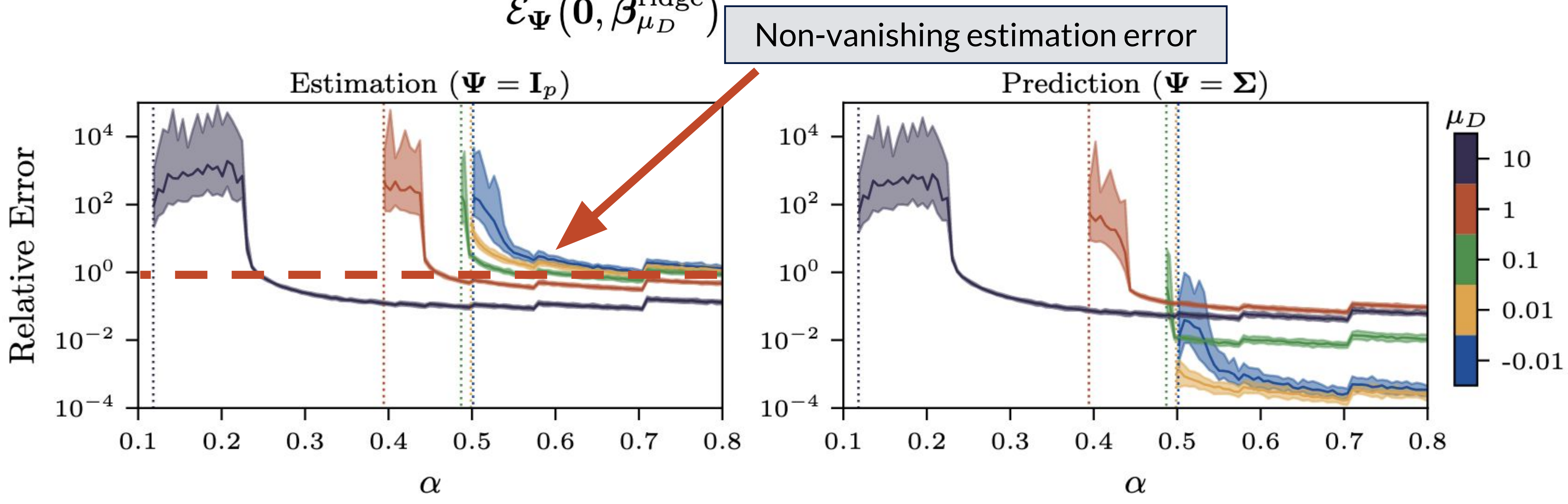
A Sketched Ensemble Efficiency Experiment

- Setup: fixed $O(p^2 n)$ budget ensembles with $K = \lfloor \frac{1}{\alpha^2} \rfloor$, $r(\mathbf{\Sigma}) = \frac{1}{2}$
- Fixed target μ_D, varying α, with λ uniquely determined
- Error: relative error $\frac{\mathcal{E}_{\mathbf{\Psi}}\left(\widehat{\boldsymbol{\beta}}_D^{\text{ens}}, \widehat{\boldsymbol{\beta}}_{\mu_D}^{\text{ridge}}\right)}{\mathcal{E}_{\mathbf{\Psi}}\left(\mathbf{0}, \widehat{\boldsymbol{\beta}}_{\mu_D}^{\text{ridge}}\right)}$



A Sketched Ensemble Efficiency Experiment

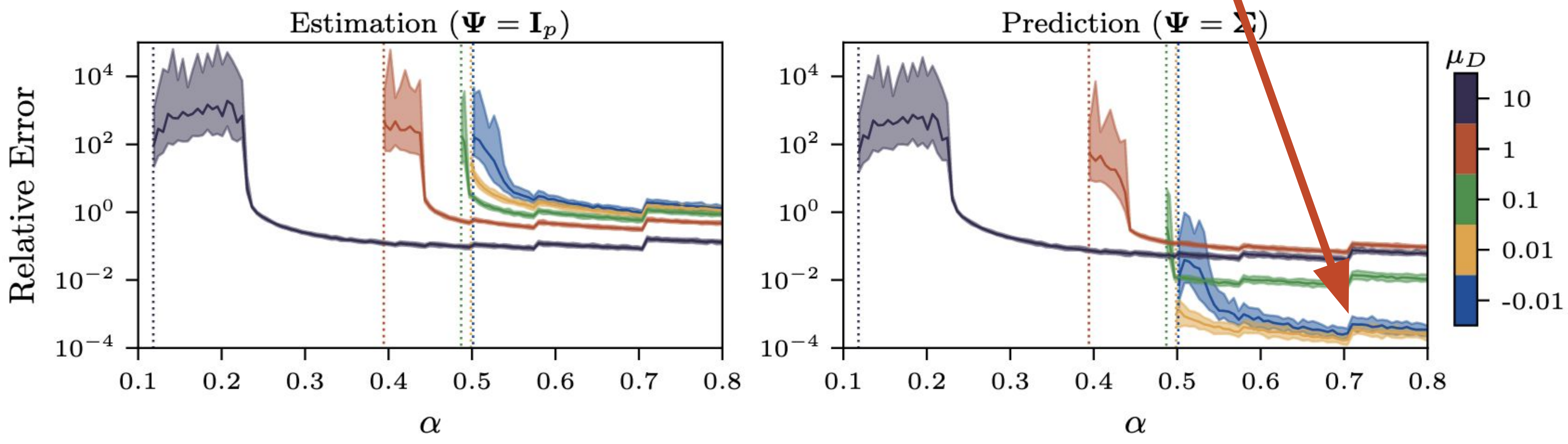
- Setup: fixed $O(p^2 n)$ budget ensembles with $K = \lfloor \frac{1}{\alpha^2} \rfloor$, $r(\boldsymbol{\Sigma}) = \frac{1}{2}$
- Fixed target μ_D, varying α, with λ uniquely determined
- Error: relative error $\dfrac{\mathcal{E}_{\boldsymbol{\Psi}}\left(\widehat{\boldsymbol{\beta}}_D^{\text{ens}}, \widehat{\boldsymbol{\beta}}_{\mu_D}^{\text{ridge}}\right)}{\mathcal{E}_{\boldsymbol{\Psi}}\left(\mathbf{0}, \widehat{\boldsymbol{\beta}}_{\mu_D}^{\text{ridge}}\right)}$



A Sketched Ensemble Efficiency Experiment

- Setup: fixed $O(p^2 n)$ budget ensembles with $K = \lfloor \frac{1}{\alpha^2} \rfloor$, $r(\mathbf{\Sigma}) = \frac{1}{2}$
- Fixed target μ_D, varying α, with λ uniquely determined
- Error: relative error $\frac{\mathcal{E}_{\mathbf{\Psi}}\left(\widehat{\boldsymbol{\beta}}_D^{\text{ens}}, \widehat{\boldsymbol{\beta}}_{\mu_D}^{\text{ridge}}\right)}{\mathcal{E}_{\mathbf{\Psi}}\left(\mathbf{0}, \widehat{\boldsymbol{\beta}}_{\mu_D}^{\text{ridge}}\right)}$

Vanishing prediction error!



Better Sketches?

What if I use a better/faster
sketch than i.i.d.?

A Free Sketching Conjecture

Conjecture 9. *Let* $\mathbf{S} \in \mathbb{C}^{p \times q}$ *be a norm-preserving sketch such that* $\mathbf{S}\mathbf{S}^{\mathsf{H}}$ *and* $\mathbf{A}$ *converge almost surely to operators that are free with respect to the average trace* $\frac{1}{p}\mathrm{tr}[\cdot]$. *Then there exists a monotonic mapping* $\lambda \mapsto \gamma$ *such that*

$$\mathbf{S}\big(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda\mathbf{I}_q\big)^{-1}\mathbf{S}^{\mathsf{H}} \simeq \big(\mathbf{A} + \gamma\mathbf{I}_p\big)^{-1}.$$

A Free Sketching Conjecture

Conjecture 9. *Let* $\mathbf{S} \in \mathbb{C}^{p \times q}$ *be a norm-preserving sketch such that* $\mathbf{S}\mathbf{S}^{\mathsf{H}}$ *and* $\mathbf{A}$ *converge almost surely to operators that are free with respect to the average trace* $\frac{1}{p}\mathrm{tr}[\cdot]$. *Then there exists a monotonic mapping* $\lambda \mapsto \gamma$ *such that*

$$\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}} \simeq \left(\mathbf{A} + \gamma \mathbf{I}_p\right)^{-1}.$$

- Includes i.i.d. sketching and more
  - Orthogonal sketching
  - Efficient sketches like CountSketch, FJLT, SRHT?

A Free Sketching Conjecture

Conjecture 9. *Let $\mathbf{S} \in \mathbb{C}^{p \times q}$ be a norm-preserving sketch such that $\mathbf{S}\mathbf{S}^{\mathsf{H}}$ and $\mathbf{A}$ converge almost surely to operators that are free with respect to the average trace $\frac{1}{p}\mathrm{tr}[\cdot]$. Then there exists a monotonic mapping $\lambda \mapsto \gamma$ such that*

$$\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}} \simeq \left(\mathbf{A} + \gamma \mathbf{I}_p\right)^{-1}.$$

- Includes i.i.d. sketching and more
  - Orthogonal sketching
  - Efficient sketches like CountSketch, FJLT, SRHT?
- Spectrum of sketch controls $\lambda \mapsto \gamma$

A Free Sketching Conjecture

Conjecture 9. *Let $\mathbf{S} \in \mathbb{C}^{p \times q}$ be a norm-preserving sketch such that $\mathbf{S}\mathbf{S}^{\mathsf{H}}$ and $\mathbf{A}$ converge almost surely to operators that are free with respect to the average trace $\frac{1}{p}\mathrm{tr}[\cdot]$. Then there exists a monotonic mapping $\lambda \mapsto \gamma$ such that*

$$\mathbf{S}\big(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q\big)^{-1}\mathbf{S}^{\mathsf{H}} \simeq \big(\mathbf{A} + \gamma \mathbf{I}_p\big)^{-1}.$$

- Includes i.i.d. sketching and more
  - Orthogonal sketching
  - Efficient sketches like CountSketch, FJLT, SRHT?
- Spectrum of sketch controls $\lambda \mapsto \gamma$
- Higher order equivalences naturally follow

A Free Sketching Conjecture

Conjecture 9. *Let $\mathbf{S} \in \mathbb{C}^{p \times q}$ be a norm-preserving sketch such that $\mathbf{S}\mathbf{S}^{\mathsf{H}}$ and $\mathbf{A}$ converge almost surely to operators that are free with respect to the average trace $\frac{1}{p}\mathrm{tr}[\cdot]$. Then there exists a monotonic mapping $\lambda \mapsto \gamma$ such that*

$$\mathbf{S}\left(\mathbf{S}^{\mathsf{H}}\mathbf{A}\mathbf{S} + \lambda \mathbf{I}_q\right)^{-1}\mathbf{S}^{\mathsf{H}} \simeq \left(\mathbf{A} + \gamma \mathbf{I}_p\right)^{-1}.$$

- Includes i.i.d. sketching and more
  - Orthogonal sketching
  - Efficient sketches like CountSketch, FJLT, SRHT?
- Spectrum of sketch controls $\lambda \mapsto \gamma$
- Higher order equivalences naturally follow

Conjecture 10. *For* $q \leq p$ *let* $\sqrt{\frac{q}{p}}\mathbf{Q} \in \mathbb{C}^{p \times q}$ *be a Haar-distributed matrix with orthonormal columns. Then*

$$\mathbf{Q}\left(\mathbf{Q}^{\mathsf{H}}\mathbf{A}\mathbf{Q} + \lambda\mathbf{I}_q\right)^{-1}\mathbf{Q}^{\mathsf{H}} \simeq \left(\mathbf{A} + \gamma\mathbf{I}_p\right)^{-1},$$

where γ *is the most positive solution to*

$$\frac{1}{p}\mathrm{tr}\left[\left(\mathbf{A} + \gamma\mathbf{I}_p\right)^{-1}\right](\gamma - \alpha\lambda) = 1 - \alpha.$$

Furthermore, for $\mu > 0$ *applied to the same* $(\mathbf{A}, \alpha, \lambda)$*, we have* $\gamma < \mu$.

Orthogonal Sketching

Conjecture 10. *For $q \le p$ let $\sqrt{\frac{q}{p}}\mathbf{Q} \in \mathbb{C}^{p\times q}$ be a Haar-distributed matrix with orthonormal columns. Then*

$$\mathbf{Q}\left(\mathbf{Q}^{\mathsf{H}}\mathbf{A}\mathbf{Q} + \lambda\mathbf{I}_q\right)^{-1}\mathbf{Q}^{\mathsf{H}} \simeq \left(\mathbf{A} + \gamma\mathbf{I}_p\right)^{-1},$$

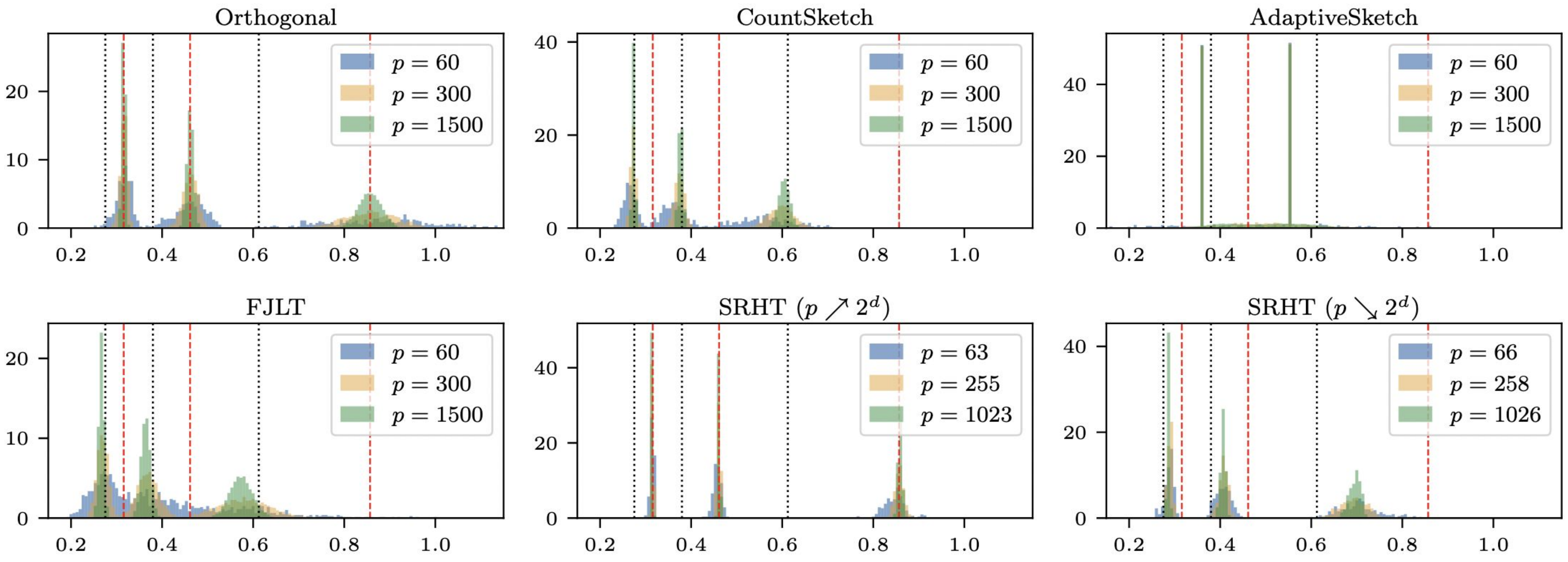
where γ is the most positive solution to

$$\frac{1}{p}\mathrm{tr}\left[(\mathbf{A} + \gamma\mathbf{I}_p)^{-1}\right](\gamma - \alpha\lambda) = 1 - \alpha.$$

Furthermore, for $\mu > 0$ applied to the same $(\mathbf{A}, \alpha, \lambda)$, we have $\gamma < \mu$.

- Same form as i.i.d. sketching, but with less regularization

Equivalence for Sketches Used in Practice?



- Early work:
  - Hints of deep connection between ensembles and ridge
    - Tuned ensembles with subsampling achieve same risk as optimal ridge

Summary

- Early work:
  - Hints of deep connection between ensembles and ridge
    - Tuned ensembles with subsampling achieve same risk as optimal ridge
- Current work:
  - Asymptotic equivalence between random projections and ridge
    - Ridge equivalence on a weak level even for single learners
    - Convergence in quadratic metrics to ridge regression for ensembles
    - Sufficiently large sketches enable accurate ridgeless regression even without ensembles

Summary

- Early work:
  - Hints of deep connection between ensembles and ridge
    - Tuned ensembles with subsampling achieve same risk as optimal ridge
- Current work:
  - Asymptotic equivalence between random projections and ridge
    - Ridge equivalence on a weak level even for single learners
    - Convergence in quadratic metrics to ridge regression for ensembles
    - Sufficiently large sketches enable accurate ridgeless regression even without ensembles
- Future work:
  - More asymptotic equivalences
    - Generalized cross-validation with sketching
    - General linear models via leave-one-dimension-out
    - Asymptotics of PCA

Questions?

Corollary 11. *Let* $\mathbf{W}$ *be an invertible* $p \times p$ *positive semidefnite matrix, either deterministic or random but independent of* $\mathbf{S}$ *with* $\limsup \|\mathbf{W}\|_{\mathrm{op}} < \infty$*. Let* $\widetilde{\mathbf{S}} = \mathbf{W}^{1/2}\mathbf{S}$*. Then for each* $\lambda > -\liminf \lambda_{\min}^{+}(\widetilde{\mathbf{S}}^\top \mathbf{A} \widetilde{\mathbf{S}})$ *as* $p, q \to \infty$ *such that* $0 < \liminf \frac{q}{p} \leq \limsup \frac{q}{p} < \infty$,

$$\widetilde{\mathbf{S}}\big(\widetilde{\mathbf{S}}^\top \mathbf{A} \widetilde{\mathbf{S}} + \lambda \mathbf{I}_q\big)^{-1} \widetilde{\mathbf{S}}^\top \simeq \big(\mathbf{A} + \mu \mathbf{W}^{-1}\big)^{-1},$$

where μ *most positive solution to*

$$\lambda = \mu \left(1 - \tfrac{1}{q} \mathrm{tr}\left[\mathbf{A}\left(\mathbf{A} + \mu \mathbf{W}^{-1}\right)^{-1}\right]\right).$$

Theorem 12. *Let* $\mathbf{X} = \mathbf{Z}\mathbf{\Sigma}^{1/2}$ *and* $\mathbf{\Sigma} \in \mathbb{C}^{p \times p}$ *have eigenvalue decomposition* $\mathbf{U}\mathbf{D}\mathbf{U}^{\mathsf{H}}$*, and let* $\mathbf{\Pi}_{\mathcal{A}}$ *be the projection operator of the principal eigenspace corresponding to a set of eigenvalues* $\mathcal{A}$ *of the matrix* $\frac{1}{n}\mathbf{X}^{\mathsf{H}}\mathbf{X}$*. Then there exists a family of measures* μ_{σ^2} *for all* $\sigma^2 \geq 0$ *such that for any* $\mathcal{A} \subseteq \mathbb{R}_{\geq 0}$*, in the limit as* $p \to \infty$*,*

$$\mathbf{\Pi}_{\mathcal{A}} \simeq \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathsf{H}},$$

where $\mathbf{\Lambda}$ *is a diagonal matrix defined for by*

$$[\mathbf{\Lambda}]_{ii} = \mu_{[\mathbf{\Sigma}]_{ii}}(\mathcal{A}).$$