Exact expressions for double descent and implicit regularization via surrogate random design

Michael W. Mahoney
ICSI and Department of Statistics, UC Berkeley

Joint work with Michał Dereziński and Feynman Liang

December 2019

Supervised learning

Input: $\mathbf{x} \sim \mu$,

Label: $y = f^*(\mathbf{x}) + \xi$, ξ - noise

Supervised learning

Input: $\mathbf{x} \sim \mu$,

Label: $y = f^*(\mathbf{x}) + \xi$, ξ - noise

Training data: $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$

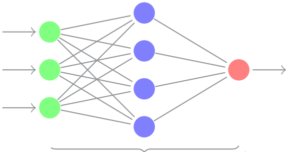
Supervised learning

Input: $\mathbf{x} \sim \mu,$

Label: $y = f^*(\mathbf{x}) + \xi,$ ξ - noise

Training data: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

Learning model: $\mathbf{x}_i$



$f_{\mathbf{w}}(\mathbf{x}_i)$

parameters: $\mathbf{w} \in \mathbb{R}^d$

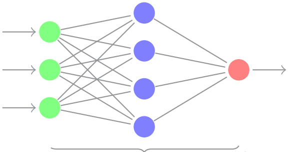
Supervised learning

Input: $\mathbf{x} \sim \mu$,

Label: $y = f^*(\mathbf{x}) + \xi$, ξ - noise

Training data: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

Learning model: $\mathbf{x}_i$



$f_{\mathbf{w}}(\mathbf{x}_i)$

parameters: $\mathbf{w} \in \mathbb{R}^d$

Error: $\mathrm{MSE}\left[f_{\mathbf{w}}\right] = \mathbb{E}\,\|f_{\mathbf{w}} - f^*\|^2$

When do supervised models learn?

Goal: $\mathrm{MSE}\big[f_{\mathbf{w}}\big] \ll \mathrm{MSE}\big[f_{\mathrm{null}}\big], \qquad f_{\mathrm{null}} \equiv 0$

When do supervised models learn?

Goal: $\mathrm{MSE}\left[f_{\mathbf{w}}\right] \ll \mathrm{MSE}\left[f_{\mathrm{null}}\right], \qquad f_{\mathrm{null}} \equiv 0$

"Classical" answer (e.g., VC theory): use $n \gg d$

Models learn when there is more <u>data</u> than <u>parameters</u>.

When do supervised models learn?

Goal: $\mathrm{MSE}\big[f_{\mathbf{w}}\big] \ll \mathrm{MSE}\big[f_{\mathrm{null}}\big], \qquad f_{\mathrm{null}} \equiv 0$

"Classical" answer (e.g., VC theory): use $n \gg d$

Models learn when there is more <u>data</u> than <u>parameters</u>.

"Modern" answer (e.g., deep learning): use $d \gg n$

Models learn when there is more <u>parameters</u> than <u>data</u>.

When do supervised models learn?

Goal: $\mathrm{MSE}\big[f_{\mathbf{w}}\big] \ll \mathrm{MSE}\big[f_{\mathrm{null}}\big], \qquad f_{\mathrm{null}} \equiv 0$

"Classical" answer (e.g., VC theory): use $n \gg d$

Models learn when there is more <u>data</u> than <u>parameters</u>.

"Modern" answer (e.g., deep learning): use $d \gg n$

Models learn when there is more <u>parameters</u> than <u>data</u>.

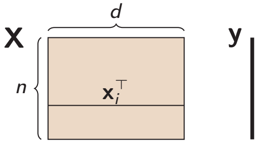
How to reconcile the two paradigms?

Simple model: linear regression

Standard i.i.d. random design

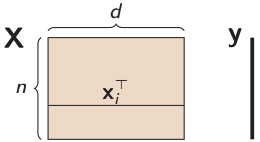
$\mathbf{X} \sim \mu^n$

$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\xi} \qquad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$



Simple model: linear regression

Standard i.i.d. random design
$\mathbf{X} \sim \mu^n$
$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\xi} \qquad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$

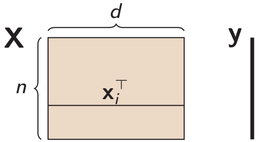


Moore-Penrose estimator:

$$\mathbf{X}^\dagger\mathbf{y} = \begin{cases} \text{minimum norm solution}, & \text{for } n \le d, \\ \text{least squares solution}, & \text{for } n > d. \end{cases}$$

Simple model: linear regression

Standard i.i.d. random design
$\mathbf{X} \sim \mu^n$
$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\xi} \qquad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$



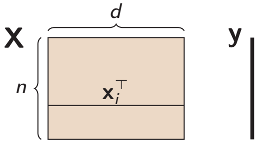
Moore-Penrose estimator:

$$\mathbf{X}^\dagger \mathbf{y} = \begin{cases} \text{minimum norm solution}, & \text{for } n \le d, \\ \text{least squares solution}, & \text{for } n > d. \end{cases}$$

Goal: find $\quad \mathrm{MSE}\left[\mathbf{X}^\dagger \mathbf{y}\right] = \mathbb{E}\,\|\mathbf{X}^\dagger \mathbf{y} - \mathbf{w}\|^2$

Simple model: linear regression

Standard i.i.d. random design
$\mathbf{X} \sim \mu^n$
$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\xi} \qquad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$



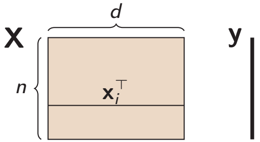
Moore-Penrose estimator:

$$\mathbf{X}^\dagger \mathbf{y} = \begin{cases} \text{minimum norm solution}, & \text{for } n \le d, \\ \text{least squares solution}, & \text{for } n > d. \end{cases}$$

Goal: find $\quad \mathrm{MSE}\left[\mathbf{X}^\dagger \mathbf{y}\right] = \mathbb{E}\,\|\mathbf{X}^\dagger \mathbf{y} - \mathbf{w}\|^2$
Prior work: asymptotics [HMRT19] and upper bounds [BLLT19]

Simple model: linear regression

Standard i.i.d. random design
$\mathbf{X} \sim \mu^n$
$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\xi} \qquad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$



Moore-Penrose estimator:

$$\mathbf{X}^\dagger \mathbf{y} = \begin{cases} \text{minimum norm solution}, & \text{for } n \le d, \\ \text{least squares solution}, & \text{for } n > d. \end{cases}$$

Goal: find $\mathrm{MSE}\left[\mathbf{X}^\dagger \mathbf{y}\right] = \mathbb{E}\, \|\mathbf{X}^\dagger \mathbf{y} - \mathbf{w}\|^2$
Prior work: asymptotics [HMRT19] and upper bounds [BLLT19]

No closed form expressions, even for $\mu = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$!

Main result I: exact non-asymptotic MSE

Idea: replace standard i.i.d. design with a surrogate design

$$\underbrace{\mathbf{X} \sim \mu^n}_{\text{i.i.d.}} \qquad \Longrightarrow \qquad \underbrace{\bar{\mathbf{X}} \sim S_\mu^n}_{\text{surrogate}}$$

Main result I: exact non-asymptotic MSE

Idea: replace standard i.i.d. design with a surrogate design

$$\underbrace{\mathbf{X} \sim \mu^n}_{\text{i.i.d.}} \qquad \Longrightarrow \qquad \underbrace{\bar{\mathbf{X}} \sim S_\mu^n}_{\text{surrogate}}$$

Theorem

Let $\bar{\mathbf{X}} \sim S_\mu^n$, $\bar{y}_i = \bar{\mathbf{x}}_i^\top \mathbf{w} + \xi$ *and* $\mathbf{\Sigma}_\mu = \mathbb{E}_\mu[\mathbf{x}\mathbf{x}^\top]$. *Then,*

Main result I: exact non-asymptotic MSE

Idea: replace standard i.i.d. design with a surrogate design

$$\underbrace{\mathbf{X} \sim \mu^n}_{\text{i.i.d.}} \quad \Longrightarrow \quad \underbrace{\bar{\mathbf{X}} \sim S_\mu^n}_{\text{surrogate}}$$

Theorem

Let $\bar{\mathbf{X}} \sim S_\mu^n$, $\bar{y}_i = \bar{\mathbf{x}}_i^\top \mathbf{w} + \xi$ *and* $\boldsymbol{\Sigma}_\mu = \mathbb{E}_\mu[\mathbf{x}\mathbf{x}^\top]$. *Then,*

$$\mathrm{MSE}\big[\bar{\mathbf{X}}^\dagger \bar{\mathbf{y}}\big] =$$

$$\begin{cases} \sigma^2 \,\mathrm{tr}\big((\boldsymbol{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1}\big)\frac{1-\alpha_n}{d-n} \; + \; \frac{\mathbf{w}^\top(\boldsymbol{\Sigma}_\mu+\lambda_n\mathbf{I})^{-1}\mathbf{w}}{\mathrm{tr}((\boldsymbol{\Sigma}_\mu+\lambda_n\mathbf{I})^{-1})}(d-n), & (n<d), \\ \sigma^2 \,\mathrm{tr}(\boldsymbol{\Sigma}_\mu^{-1}), & (n=d), \\ \sigma^2 \,\mathrm{tr}(\boldsymbol{\Sigma}_\mu^{-1})\frac{1-\beta_n}{n-d}, & (n>d), \end{cases}$$

Main result I: exact non-asymptotic MSE

Idea: replace standard i.i.d. design with a surrogate design

$$\underbrace{\mathbf{X} \sim \mu^n}_{\text{i.i.d.}} \qquad \Longrightarrow \qquad \underbrace{\bar{\mathbf{X}} \sim S_\mu^n}_{\text{surrogate}}$$

Theorem

Let $\bar{\mathbf{X}} \sim S_\mu^n$, $\bar{y}_i = \bar{\mathbf{x}}_i^\top \mathbf{w} + \xi$ *and* $\mathbf{\Sigma}_\mu = \mathbb{E}_\mu[\mathbf{x}\mathbf{x}^\top]$. *Then,*

$$\mathrm{MSE}\big[\bar{\mathbf{X}}^\dagger \bar{\mathbf{y}}\big] =$$

$$\begin{cases} \sigma^2 \operatorname{tr}\big((\mathbf{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1}\big)\frac{1-\alpha_n}{d-n} \;+\; \frac{\mathbf{w}^\top(\mathbf{\Sigma}_\mu+\lambda_n\mathbf{I})^{-1}\mathbf{w}}{\operatorname{tr}((\mathbf{\Sigma}_\mu+\lambda_n\mathbf{I})^{-1})}(d-n), & (n<d), \\ \sigma^2 \operatorname{tr}(\mathbf{\Sigma}_\mu^{-1}), & (n=d), \\ \sigma^2 \operatorname{tr}(\mathbf{\Sigma}_\mu^{-1})\frac{1-\beta_n}{n-d}, & (n>d), \end{cases}$$

where $n = \operatorname{tr}((\mathbf{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1}\mathbf{\Sigma}_\mu)$, $\alpha_n = \frac{\det(\mathbf{\Sigma}_\mu)}{\det(\mathbf{\Sigma}_\mu + \lambda_n \mathbf{I})}$, $\beta_n = \mathrm{e}^{d-n}$.

Isotropic features: double descent curve

$\mathbf{X} \sim \mu^n$ - standard Gaussian design, $\mu = \mathcal{N}(\mathbf{0}, \mathbf{I})$, $d = 100$

Isotropic features: double descent curve

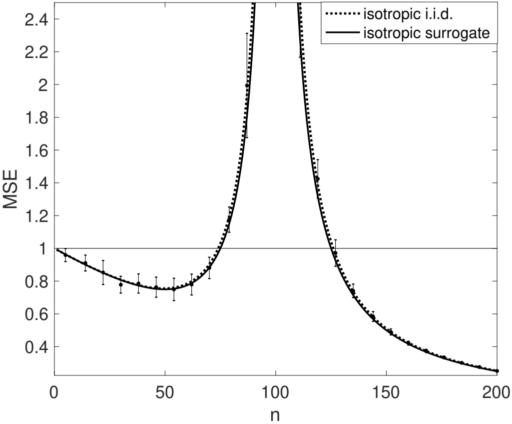
$\mathbf{X} \sim \mu^n$ - standard Gaussian design, $\mu = \mathcal{N}(\mathbf{0}, \mathbf{I})$, $d = 100$

$$\mathrm{MSE}\big[\mathbf{X}^{\dagger}\mathbf{y}\big] = \begin{cases} \frac{\sigma^2 n}{d-n-1} + \|\mathbf{w}\|^2 \frac{d-n}{d}, & (n < d-1) \\ \frac{\sigma^2 d}{n-d-1}, & (n > d+1) \end{cases} \qquad (\text{let } \|\mathbf{w}\| = 1)$$

Isotropic features: double descent curve

$\mathbf{X} \sim \mu^n$ - standard Gaussian design, $\mu = \mathcal{N}(\mathbf{0}, \mathbf{I})$, $d = 100$

$$\mathrm{MSE}[\mathbf{X}^\dagger \mathbf{y}] = \begin{cases} \frac{\sigma^2 n}{d-n-1} + \|\mathbf{w}\|^2 \frac{d-n}{d}, & (n < d-1) \\ \frac{\sigma^2 d}{n-d-1}, & (n > d+1) \end{cases} \quad (\text{let } \|\mathbf{w}\| = 1)$$



Gaussian features: effect of spectral decay

$\mathbf{X} \sim \mu^n$ - multivariate Gaussian design, $\mu = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, $d = 100$

$\mathbf{\Sigma}$ - exponentially decaying eigenvalues, condition number κ

Gaussian features: effect of spectral decay

$\mathbf{X} \sim \mu^n$ - multivariate Gaussian design, $\mu = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, $d = 100$

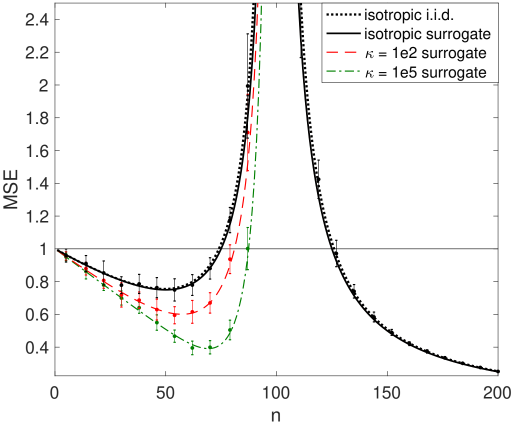
$\mathbf{\Sigma}$ - exponentially decaying eigenvalues, condition number κ

$$\mathrm{MSE}\big[\mathbf{X}^{\dagger}\mathbf{y}\big] = ?$$

Gaussian features: effect of spectral decay

$\mathbf{X} \sim \mu^n$ - multivariate Gaussian design, $\mu = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, $d = 100$
$\boldsymbol{\Sigma}$ - exponentially decaying eigenvalues, condition number κ

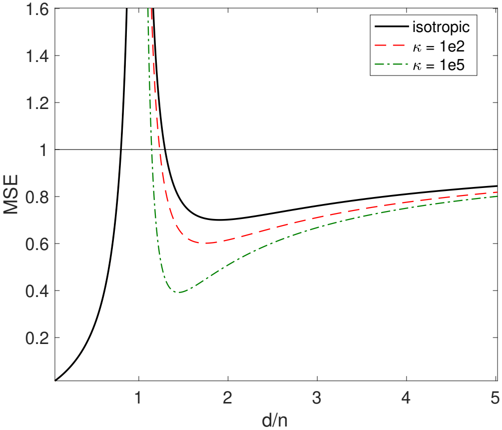
$$\mathrm{MSE}\left[\mathbf{X}^{\dagger}\mathbf{y}\right] = ?$$



Gaussian features: effect of model complexity

$\mathbf{X} \sim \mu^n$ - multivariate Gaussian design, $\mu = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, $n = 100$
$\mathbf{\Sigma}$ - exponentially decaying eigenvalues, condition number κ

$$\mathrm{MSE}\left[\mathbf{X}^{\dagger}\mathbf{y}\right] = ?$$



Main result II: implicit regularization of minimum norm

Why does this *unregularized* model learn when $n < d$?

Main result II: implicit regularization of minimum norm

Why does this *unregularized* model learn when $n < d$?
Because taking minimum norm induces ℓ_2*-regularization.*

Main result II: implicit regularization of minimum norm

Why does this *unregularized* model learn when $n < d$?
Because taking minimum norm induces ℓ_2-regularization.

Theorem

Let $\bar{\mathbf{X}} \sim S_\mu^n$, $\bar{y}_i = y(\mathbf{x})$ *and* $\boldsymbol{\Sigma}_\mu = \mathbb{E}_\mu[\mathbf{x}\mathbf{x}^\top]$. *Then,*

Main result II: implicit regularization of minimum norm

Why does this *unregularized* model learn when $n < d$?
Because taking minimum norm induces ℓ_2-regularization.

Theorem

Let $\bar{\mathbf{X}} \sim S_\mu^n$, $\bar{y}_i = y(\mathbf{x})$ *and* $\mathbf{\Sigma}_\mu = \mathbb{E}_\mu[\mathbf{x}\mathbf{x}^\top]$. *Then,*

$$\mathbb{E}\big[\bar{\mathbf{X}}^\dagger \bar{\mathbf{y}}\big] = \begin{cases} (\mathbf{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1} \mathbf{v}_{\mu,y} & \text{for } n < d, \\ \mathbf{\Sigma}_\mu^{-1} \mathbf{v}_{\mu,y} & \text{for } n \geq d, \end{cases}$$

where $n = \mathrm{tr}\big((\mathbf{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1} \mathbf{\Sigma}_\mu\big)$ *and* $\mathbf{v}_{\mu,y} = \mathbb{E}_\mu[y(\mathbf{x})\,\mathbf{x}]$.

Main result II: implicit regularization of minimum norm

Why does this *unregularized* model learn when $n < d$?
Because taking minimum norm induces ℓ_2-regularization.

Theorem

Let $\bar{\mathbf{X}} \sim S_\mu^n$, $\bar{y}_i = y(\mathbf{x})$ *and* $\mathbf{\Sigma}_\mu = \mathbb{E}_\mu[\mathbf{x}\mathbf{x}^\top]$. *Then,*

$$\mathbb{E}\big[\bar{\mathbf{X}}^\dagger \bar{\mathbf{y}}\big] = \begin{cases} (\mathbf{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1} \mathbf{v}_{\mu,y} & \text{for } n < d, \\ \mathbf{\Sigma}_\mu^{-1} \mathbf{v}_{\mu,y} & \text{for } n \geq d, \end{cases}$$

where $n = \mathrm{tr}((\mathbf{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1} \mathbf{\Sigma}_\mu)$ *and* $\mathbf{v}_{\mu,y} = \mathbb{E}_\mu[y(\mathbf{x})\, \mathbf{x}]$.

$$(\mathbf{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1} \mathbf{v}_{\mu,y} \; = \; \operatorname*{argmin}_{\widehat{\mathbf{w}}} \; \mathbb{E}_{\mu,y}\Big[\big(\mathbf{x}^\top \widehat{\mathbf{w}} - y(\mathbf{x})\big)^2\Big] + \lambda_n \|\widehat{\mathbf{w}}\|^2$$

Implicit regularization

Our observations:

- Minimum norm induces an ℓ_2-regularizer: $\lambda_n \|\widehat{\mathbf{w}}\|^2$

Implicit regularization

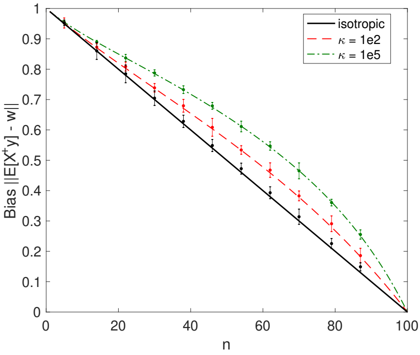
Our observations:

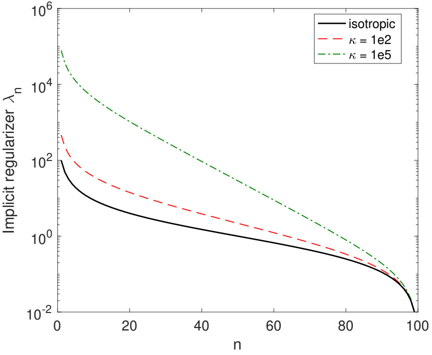
- Minimum norm induces an ℓ_2-regularizer: $\lambda_n \|\widehat{\mathbf{w}}\|^2$
- Sample size is effective dimension: $n = \mathrm{tr}((\mathbf{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1} \mathbf{\Sigma}_\mu)$

Implicit regularization

Our observations:

- Minimum norm induces an ℓ_2-regularizer: $\lambda_n \|\widehat{\mathbf{w}}\|^2$
- Sample size is effective dimension: $n = \mathrm{tr}((\boldsymbol{\Sigma}_\mu + \lambda_n \mathbf{I})^{-1} \boldsymbol{\Sigma}_\mu)$





Consistency of surrogate expressions

$$\mathrm{MSE}\big[\mathbf{X}^\dagger\mathbf{y}\big] = \sigma^2\mathbb{E}\big[\mathrm{tr}\big((\mathbf{X}^\top\mathbf{X})^\dagger\big)\big] + \mathbf{w}^\top\mathbb{E}\big[\mathbf{I} - \mathbf{X}^\dagger\mathbf{X}\big]\mathbf{w}.$$

Consistency of surrogate expressions

$$\mathrm{MSE}\big[\mathbf{X}^\dagger\mathbf{y}\big] = \sigma^2\mathbb{E}\big[\mathrm{tr}\big((\mathbf{X}^\top\mathbf{X})^\dagger\big)\big] + \mathbf{w}^\top\mathbb{E}\big[\mathbf{I} - \mathbf{X}^\dagger\mathbf{X}\big]\mathbf{w}.$$

$$\mu = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \quad \Longrightarrow \quad \begin{cases} \mathbf{X}^\top\mathbf{X} & - \quad \text{Wishart distribution} \\ \mathbf{X}^\dagger\mathbf{X} & - \quad \text{Gaussian projection} \end{cases}$$

Consistency of surrogate expressions

$$\mathrm{MSE}\big[\mathbf{X}^\dagger\mathbf{y}\big] = \sigma^2\mathbb{E}\big[\mathrm{tr}\big((\mathbf{X}^\top\mathbf{X})^\dagger\big)\big] + \mathbf{w}^\top\mathbb{E}\big[\mathbf{I}-\mathbf{X}^\dagger\mathbf{X}\big]\mathbf{w}.$$

$$\mu = \mathcal{N}(\mathbf{0},\mathbf{\Sigma}) \quad \Longrightarrow \quad \begin{cases} \mathbf{X}^\top\mathbf{X} & - \quad \text{Wishart distribution} \\ \mathbf{X}^\dagger\mathbf{X} & - \quad \text{Gaussian projection} \end{cases}$$

Conjecture

Fix $n/d < 1$ and let $\mu = \mathcal{N}(\mathbf{0},\mathbf{\Sigma})$, where $c\mathbf{I} \preceq \mathbf{\Sigma} \preceq C\mathbf{I}$. Then:

Consistency of surrogate expressions

$$\mathrm{MSE}\big[\mathbf{X}^\dagger\mathbf{y}\big] = \sigma^2\mathbb{E}\big[\mathrm{tr}\big((\mathbf{X}^\top\mathbf{X})^\dagger\big)\big] + \mathbf{w}^\top\mathbb{E}\big[\mathbf{I}-\mathbf{X}^\dagger\mathbf{X}\big]\mathbf{w}.$$

$$\mu = \mathcal{N}(\mathbf{0},\mathbf{\Sigma}) \quad\Longrightarrow\quad \begin{cases} \mathbf{X}^\top\mathbf{X} & - \quad \text{Wishart distribution} \\ \mathbf{X}^\dagger\mathbf{X} & - \quad \text{Gaussian projection} \end{cases}$$

Conjecture

Fix $n/d < 1$ and let $\mu = \mathcal{N}(\mathbf{0},\mathbf{\Sigma})$, where $c\mathbf{I} \preceq \mathbf{\Sigma} \preceq C\mathbf{I}$. Then:

$$\left|\frac{\mathbb{E}\big[\mathrm{tr}((\mathbf{X}^\top\mathbf{X})^\dagger)\big]}{\mathcal{V}(\mathbf{\Sigma},n)} - 1\right| = O(1/d) \quad \textit{for} \quad \mathcal{V}(\mathbf{\Sigma},n) = \frac{1-\alpha_n}{\lambda_n},$$

Consistency of surrogate expressions

$$\mathrm{MSE}\big[\mathbf{X}^\dagger\mathbf{y}\big] = \sigma^2\mathbb{E}\big[\mathrm{tr}\big((\mathbf{X}^\top\mathbf{X})^\dagger\big)\big] + \mathbf{w}^\top\mathbb{E}\big[\mathbf{I} - \mathbf{X}^\dagger\mathbf{X}\big]\mathbf{w}.$$

$$\mu = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \quad \Longrightarrow \quad \begin{cases} \mathbf{X}^\top\mathbf{X} & - \quad \text{Wishart distribution} \\ \mathbf{X}^\dagger\mathbf{X} & - \quad \text{Gaussian projection} \end{cases}$$

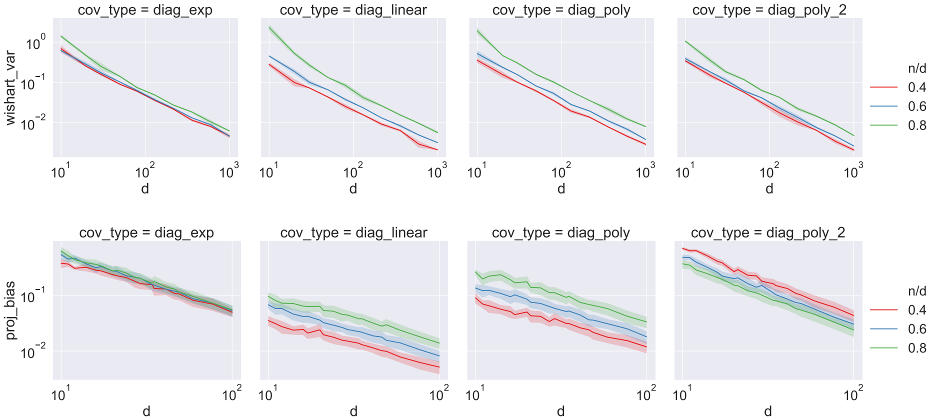
Conjecture

Fix $n/d < 1$ and let $\mu = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $c\mathbf{I} \preceq \boldsymbol{\Sigma} \preceq C\mathbf{I}$. Then:

$$\left|\frac{\mathbb{E}\big[\mathrm{tr}((\mathbf{X}^\top\mathbf{X})^\dagger)\big]}{\mathcal{V}(\boldsymbol{\Sigma}, n)} - 1\right| = O(1/d) \quad \textit{for} \quad \mathcal{V}(\boldsymbol{\Sigma}, n) = \frac{1-\alpha_n}{\lambda_n},$$

$$\left|\frac{\mathbf{w}^\top\mathbb{E}[\mathbf{I} - \mathbf{X}^\dagger\mathbf{X}]\mathbf{w}}{\mathbf{w}^\top\mathcal{B}(\boldsymbol{\Sigma}, n)\mathbf{w}} - 1\right| = O(1/d) \quad \textit{for} \quad \mathcal{B}(\boldsymbol{\Sigma}, n) = \lambda_n(\boldsymbol{\Sigma} + \lambda_n\mathbf{I})^{-1}.$$

Empirical evidence for the conjecture



Surrogate design: rescaling by pseudo-determinant

Definition

Let K be a random variable over non-negative integers.

A determinantal surrogate design $\bar{\mathbf{X}} \sim \mathrm{Det}(\mu, K)$ is defined so that

$$\mathbb{E}\big[F(\bar{\mathbf{X}})\big] \;\propto\; \mathbb{E}[\mathrm{pdet}(\mathbf{X}\mathbf{X}^\top)F(\mathbf{X})] \quad \text{for} \quad \mathbf{X} \sim \mu^K.$$

Surrogate design: rescaling by pseudo-determinant

Definition

Let K be a random variable over non-negative integers.

A determinantal surrogate design $\bar{\mathbf{X}} \sim \mathrm{Det}(\mu, K)$ is defined so that

$$\mathbb{E}\big[F(\bar{\mathbf{X}})\big] \;\propto\; \mathbb{E}[\mathrm{pdet}(\mathbf{X}\mathbf{X}^\top)F(\mathbf{X})] \quad \text{for} \quad \mathbf{X} \sim \mu^K.$$

- The proportionality constant is $1/\mathbb{E}[\mathrm{pdet}(\mathbf{X}\mathbf{X}^\top)]$.

Surrogate design: rescaling by pseudo-determinant

Definition

Let K be a random variable over non-negative integers.

A determinantal surrogate design $\bar{\mathbf{X}} \sim \mathrm{Det}(\mu, K)$ is defined so that

$$\mathbb{E}\big[F(\bar{\mathbf{X}})\big] \;\propto\; \mathbb{E}[\mathrm{pdet}(\mathbf{X}\mathbf{X}^\top)F(\mathbf{X})] \quad \text{for} \quad \mathbf{X} \sim \mu^K.$$

- The proportionality constant is $1/\mathbb{E}[\mathrm{pdet}(\mathbf{X}\mathbf{X}^\top)]$.
- To compute it, we let K be a Poisson random variable.

Surrogate design: rescaling by pseudo-determinant

Definition

Let K be a random variable over non-negative integers.

A determinantal surrogate design $\bar{\mathbf{X}} \sim \mathrm{Det}(\mu, K)$ is defined so that

$$\mathbb{E}\big[F(\bar{\mathbf{X}})\big] \;\propto\; \mathbb{E}[\mathrm{pdet}(\mathbf{X}\mathbf{X}^\top)F(\mathbf{X})] \quad \text{for} \quad \mathbf{X} \sim \mu^K.$$

- The proportionality constant is $1/\mathbb{E}[\mathrm{pdet}(\mathbf{X}\mathbf{X}^\top)]$.
- To compute it, we let K be a Poisson random variable.
- New expectation formulas for $K \sim \text{Poisson}(\gamma)$ and $\mathbf{X} \sim \mu^K$:

$$\mathbb{E}\big[\det(\mathbf{X}\mathbf{X}^\top)\big] = \mathrm{e}^{-\gamma}\det(\mathbf{I} + \gamma\boldsymbol{\Sigma}_\mu)$$

$$\mathbb{E}\big[\det(\mathbf{X}^\top\mathbf{X})\big] = \det(\gamma\boldsymbol{\Sigma}_\mu)$$

New technique: *determinant preserving random matrices*

Definition

A random $d \times d$ matrix $\mathbf{A}$ is determinant preserving (d.p.) if

$$\mathbb{E}\big[\det(\mathbf{A}_{\mathcal{I},\mathcal{J}})\big] = \det\big(\mathbb{E}[\mathbf{A}_{\mathcal{I},\mathcal{J}}]\big) \quad \text{for all } \mathcal{I},\mathcal{J} \subseteq [d] \text{ s.t. } |\mathcal{I}| = |\mathcal{J}|.$$

New technique: *determinant preserving random matrices*

Definition

A random $d \times d$ matrix $\mathbf{A}$ is determinant preserving (d.p.) if

$$\mathbb{E}\big[\det(\mathbf{A}_{\mathcal{I},\mathcal{J}})\big] = \det\big(\mathbb{E}[\mathbf{A}_{\mathcal{I},\mathcal{J}}]\big) \quad \text{for all } \mathcal{I},\mathcal{J} \subseteq [d] \text{ s.t. } |\mathcal{I}| = |\mathcal{J}|.$$

Examples:

- $\mathbf{A}$ has i.i.d. Gaussian entries $a_{ij} \sim \mathcal{N}(0,1)$

New technique: *determinant preserving random matrices*

Definition

A random $d \times d$ matrix $\mathbf{A}$ is determinant preserving (d.p.) if

$$\mathbb{E}\big[\det(\mathbf{A}_{\mathcal{I},\mathcal{J}})\big] = \det\big(\mathbb{E}[\mathbf{A}_{\mathcal{I},\mathcal{J}}]\big) \quad \text{for all } \mathcal{I},\mathcal{J} \subseteq [d] \text{ s.t. } |\mathcal{I}| = |\mathcal{J}|.$$

Examples:

- $\mathbf{A}$ has i.i.d. Gaussian entries $a_{ij} \sim \mathcal{N}(0,1)$
- $\mathbf{A} = s\mathbf{Z}$, where s is random and $\mathbf{Z}$ is a fixed, rank-1 matrix

New technique: *determinant preserving random matrices*

Definition

A random $d \times d$ matrix $\mathbf{A}$ is determinant preserving (d.p.) if

$$\mathbb{E}\big[\det(\mathbf{A}_{\mathcal{I},\mathcal{J}})\big] = \det\big(\mathbb{E}[\mathbf{A}_{\mathcal{I},\mathcal{J}}]\big) \quad \text{for all } \mathcal{I}, \mathcal{J} \subseteq [d] \text{ s.t. } |\mathcal{I}| = |\mathcal{J}|.$$

Examples:

- $\mathbf{A}$ has i.i.d. Gaussian entries $a_{ij} \sim \mathcal{N}(0,1)$
- $\mathbf{A} = s\mathbf{Z}$, where s is random and $\mathbf{Z}$ is a fixed, rank-1 matrix
- $\mathbf{A} = \mathbf{X}^\top \mathbf{X}$, where $\mathbf{X} \sim \mu^K$ and $K \sim \mathrm{Poisson}(\gamma)$

New technique: *determinant preserving random matrices*

Definition

A random $d \times d$ matrix $\mathbf{A}$ is determinant preserving (d.p.) if

$$\mathbb{E}\big[\det(\mathbf{A}_{\mathcal{I},\mathcal{J}})\big] = \det\big(\mathbb{E}[\mathbf{A}_{\mathcal{I},\mathcal{J}}]\big) \quad \text{for all } \mathcal{I},\mathcal{J} \subseteq [d] \text{ s.t. } |\mathcal{I}| = |\mathcal{J}|.$$

Examples:

- $\mathbf{A}$ has i.i.d. Gaussian entries $a_{ij} \sim \mathcal{N}(0,1)$
- $\mathbf{A} = s\mathbf{Z}$, where s is random and $\mathbf{Z}$ is a fixed, rank-1 matrix
- $\mathbf{A} = \mathbf{X}^\top \mathbf{X}$, where $\mathbf{X} \sim \mu^K$ and $K \sim \text{Poisson}(\gamma)$

Theorem (closure properties)

If $\mathbf{A}, \mathbf{B}$ are d.p. and independent, then $\mathbf{A} + \mathbf{B}$ and $\mathbf{AB}$ are also d.p.

Next steps

- Consistency of surrogate expressions

Next steps

- Consistency of surrogate expressions
- Random feature models

Next steps

- Consistency of surrogate expressions
- Random feature models
- Non-linear and kernelized models

Next steps

- Consistency of surrogate expressions
- Random feature models
- Non-linear and kernelized models
- Prediction error

References

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler.
Benign overfitting in linear regression.
Technical Report Preprint: arXiv:1906.11300, 2019.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani.
Surprises in high-dimensional ridgeless least squares interpolation.
Technical Report Preprint: arXiv:1903.08560, 2019.