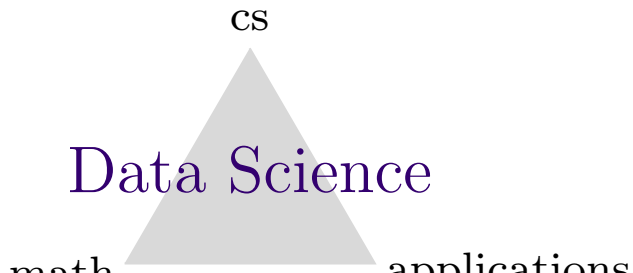
Data Science:
Statistics Meets Optimization



Johannes Lederer

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

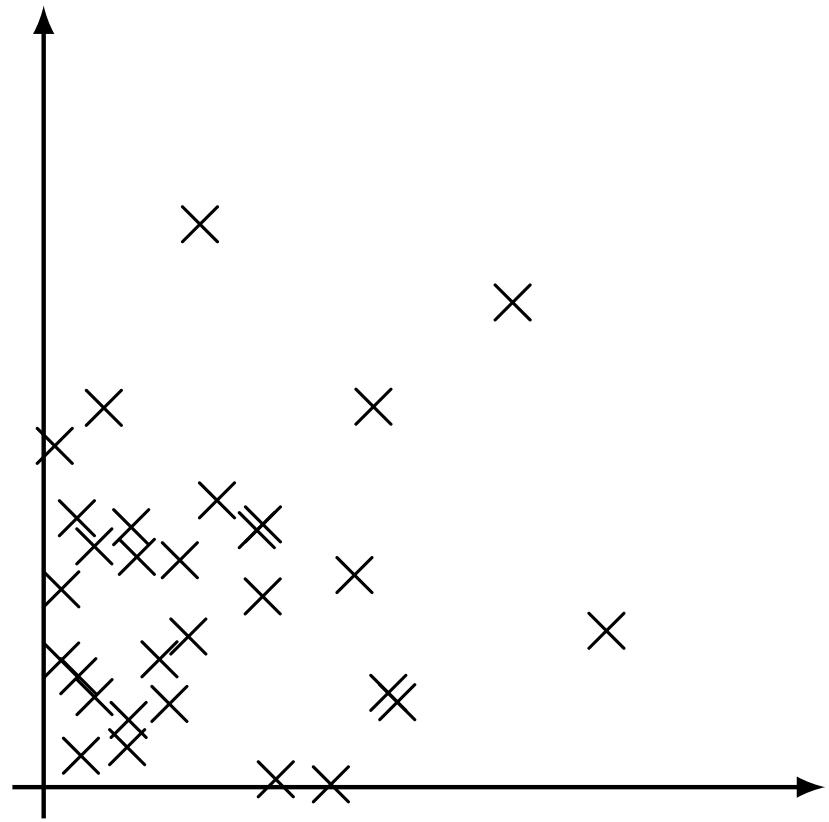
Funded through GINN

Transregio 391


Lambro, one year ago

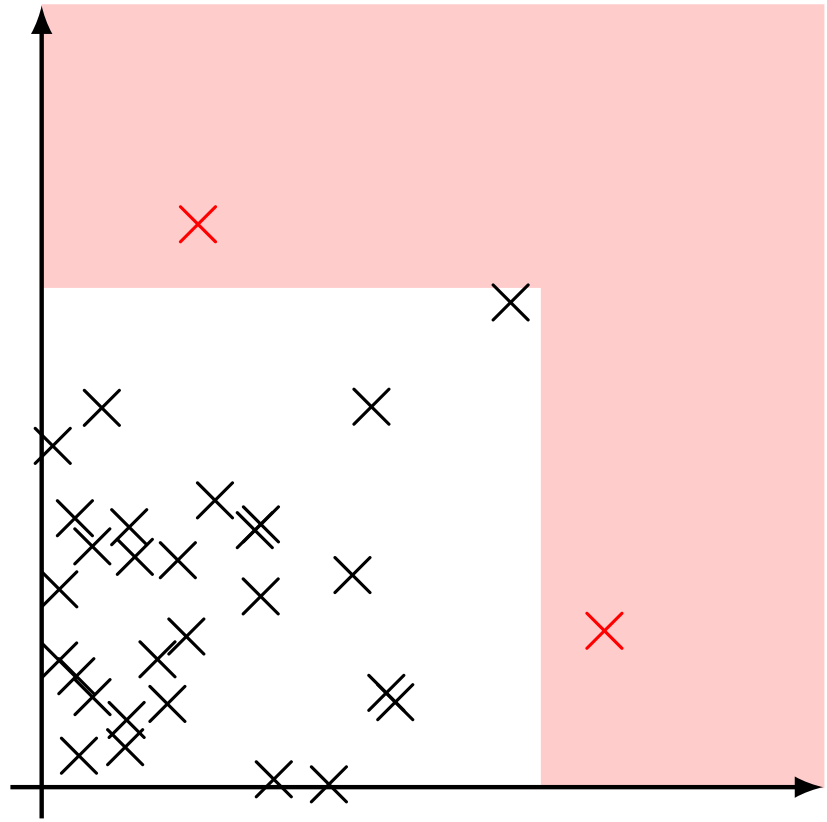
Extreme-Value Theory Is Not Extrapolation

(Extremes in high dimensions: methods and scalable algorithms, 2023)



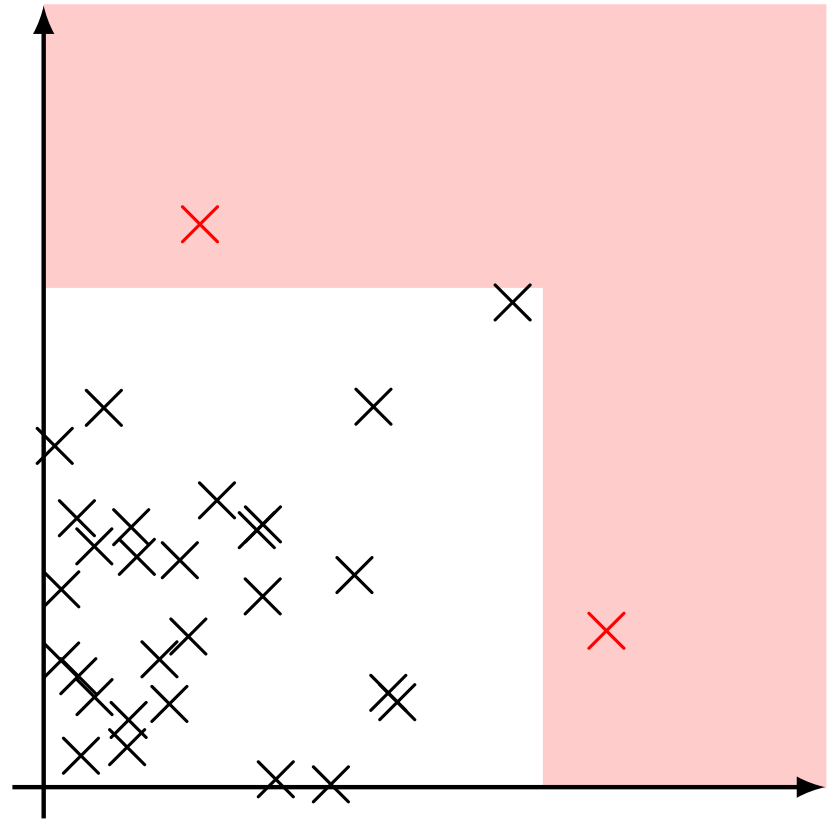
Extreme-Value Theory Is Not Extrapolation

(Extremes in high dimensions: methods and scalable algorithms, 2023)



Extreme-Value Theory Is Not Extrapolation

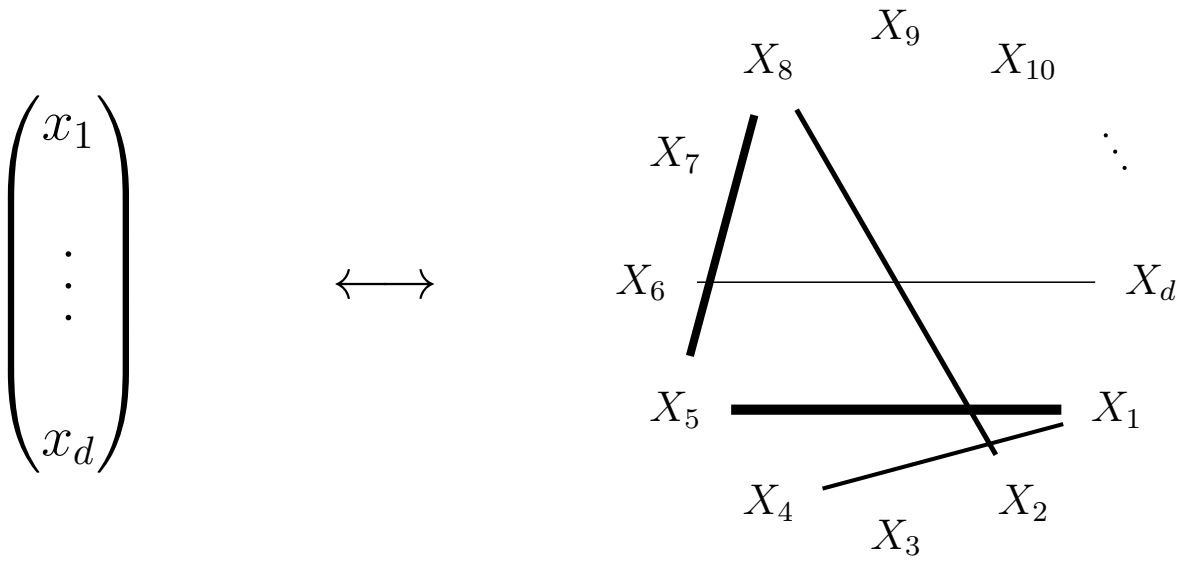
(Extremes in high dimensions: methods and scalable algorithms, 2023)



$$\lim_{t\to\infty} \mathbb{P}\{X/t \in \mathcal{A} \mid \|X\| > t\} \;\sim\; \mathfrak{m}[\mathcal{A}]$$

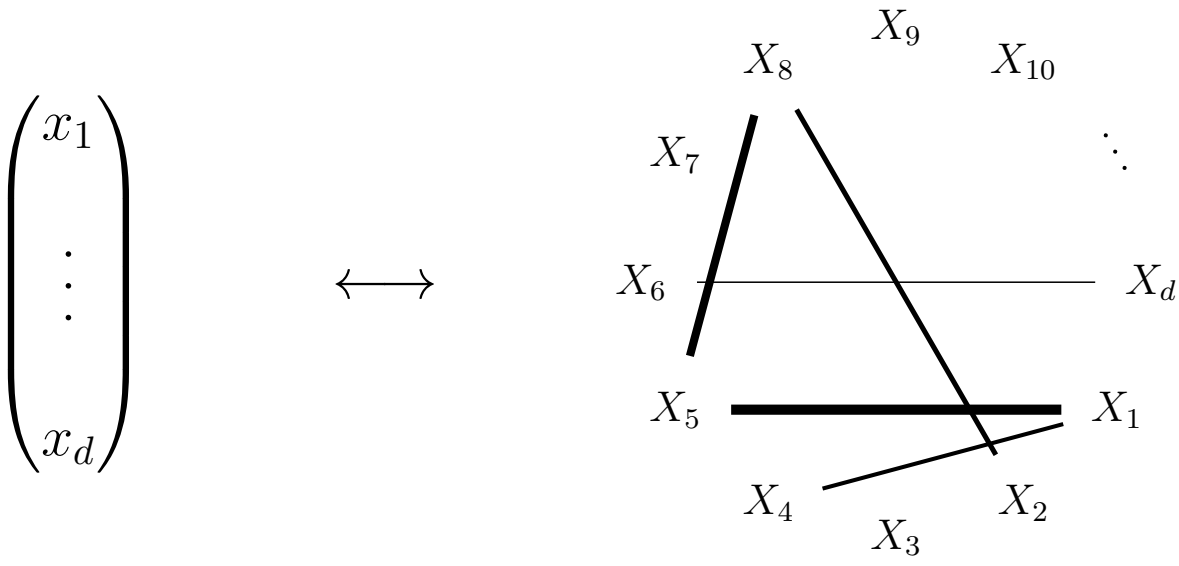
Networks Are High Dimensional

(Extremes in high dimensions: methods and scalable algorithms, 2023)



Networks Are High Dimensional

(Extremes in high dimensions: methods and scalable algorithms, 2023)



Key challenge: $p \approx d^2$ large, n usually small

Starting Point: Hüsler-Reiss Densities

(Extremes in high dimensions: methods and scalable algorithms, 2023)

$$\frac{1}{c_{\boldsymbol{\Gamma}}} \frac{1}{x_m} \left(\prod_{j=1}^{d} \frac{1}{x_j} \right) \mathfrak{n}_{d-1} \left[\log[\boldsymbol{x}_{-m}/x_m]; -\boldsymbol{\Gamma}_{-m,m}/2 \, , \quad \boldsymbol{\Sigma} \; \right]$$

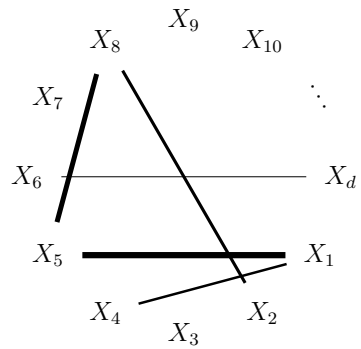
- $\boldsymbol{x} \in \{\boldsymbol{a} \in (0,\infty)^d \, : \, \|\boldsymbol{a}\| > 1\}$
- $\mathfrak{n}_{d-1}$: Gauss density
- $\boldsymbol{\Gamma} \in \mathbb{R}^{d \times d}$: variogram matrix
- $\boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}[\boldsymbol{\Gamma}, m] \in \mathbb{R}^{(d-1)\times(d-1)}$: covariance matrix
- $c_{\boldsymbol{\Gamma}}$: normalizing constant

Starting Point: Hüsler-Reiss Densities

(Extremes in high dimensions: methods and scalable algorithms, 2023)

$$\frac{1}{c_{\boldsymbol{\Gamma}}} \frac{1}{x_m} \left(\prod_{j=1}^{d} \frac{1}{x_j} \right) \mathfrak{n}_{d-1} \left[\log[\boldsymbol{x}_{-m}/x_m]; -\boldsymbol{\Gamma}_{-m,m}/2 \, , \quad \boldsymbol{\Sigma} \right]$$

- Almost like Gaussian graphical models



Starting Point: Hüsler-Reiss Densities

(Extremes in high dimensions: methods and scalable algorithms, 2023)

$$\frac{1}{c_{\boldsymbol{\Gamma}}} \frac{1}{x_m} \left(\prod_{j=1}^{d} \frac{1}{x_j} \right) \mathfrak{n}_{d-1} \left[\log[\boldsymbol{x}_{-m}/x_m]; \underbrace{-\boldsymbol{\Gamma}_{-m,m}/2}_{\rightsquigarrow \boldsymbol{\mu}}, \underbrace{\boldsymbol{\Sigma}}_{\rightsquigarrow \boldsymbol{\Lambda}^{-1}} \right]$$

- ☺ Almost like Gaussian graphical models
- ☹ Not really like Gaussian graphical models
  $\longrightarrow$ we establish generalized "densities" ☺

Starting Point: Hüsler-Reiss Densities

(Extremes in high dimensions: methods and scalable algorithms, 2023)

$$\frac{1}{c_{\boldsymbol{\Gamma}}} \frac{1}{x_m} \left(\prod_{j=1}^{d} \frac{1}{x_j} \right) \mathfrak{n}_{d-1} \big[\log[\boldsymbol{x}_{-m}/x_m]; \underbrace{-\boldsymbol{\Gamma}_{-m,m}/2}_{\rightsquigarrow \boldsymbol{\mu}}, \underbrace{\boldsymbol{\Sigma}}_{\rightsquigarrow \boldsymbol{\Lambda}^{-1}} \big]$$

- ☺ Almost like Gaussian graphical models
- ☹ Not really like Gaussian graphical models
  $\longrightarrow$ we establish generalized "densities" ☺
- ☹ Statistically and computationally challenging
  $\longrightarrow$ we establish a scalable estimator ☺

Score Matching Circumvents Normalizations

(Extremes in high dimensions: methods and scalable algorithms, 2023)

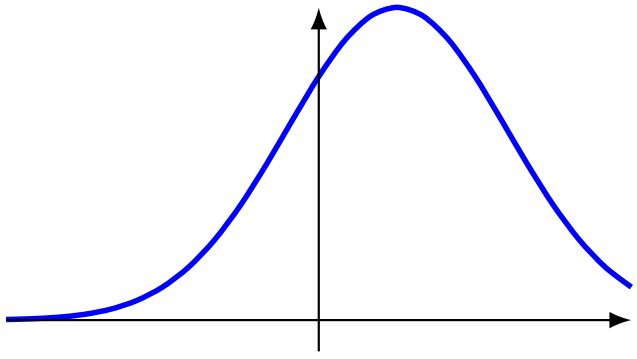
Our densities are of the form

$$\mathfrak{h}[\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Lambda}] \;=\; e^{\mathfrak{g}[\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Lambda}]}/c_{\boldsymbol{\mu},\boldsymbol{\Lambda}}$$

with explicit $\mathfrak{g}[\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Lambda}]$ and integral normalization $c_{\boldsymbol{\mu},\boldsymbol{\Lambda}}$.

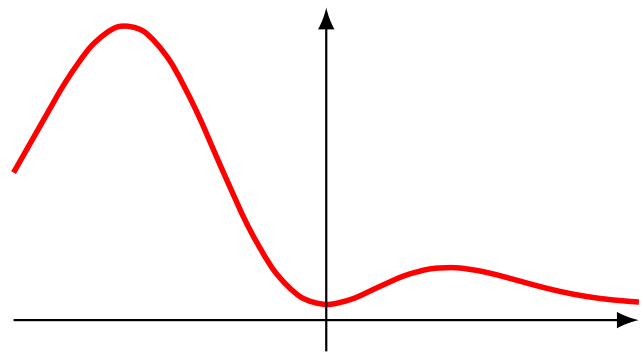
"density"

$\boldsymbol{x} \mapsto \mathfrak{h}[\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Lambda}]$



"likelihood"

$(\boldsymbol{\mu},\boldsymbol{\Lambda}) \mapsto \mathfrak{h}[\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Lambda}]$



Score Matching Circumvents Normalizations

(Extremes in high dimensions: methods and scalable algorithms, 2023)

Our densities are of the form

$$\mathfrak{h}[\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Lambda}] \;=\; e^{\mathfrak{g}[\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Lambda}]} / c_{\boldsymbol{\mu}, \boldsymbol{\Lambda}}$$

with explicit $\mathfrak{g}[\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Lambda}]$ and integral normalization $c_{\boldsymbol{\mu}, \boldsymbol{\Lambda}}$.

Maximum likelihood

$$\operatorname*{argmax}_{\boldsymbol{\mu}, \boldsymbol{\Lambda}} \mathbb{E}\Big[\log \mathfrak{h}[\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Lambda}]\Big]$$

$$= \operatorname*{argmax}_{\boldsymbol{\mu}, \boldsymbol{\Lambda}} \mathbb{E}\Big[\mathfrak{g}[\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Lambda}] - \log c_{\boldsymbol{\mu}, \boldsymbol{\Lambda}}\Big]$$

is computationally infeasible.

Score Matching Circumvents Normalizations

(Extremes in high dimensions: methods and scalable algorithms, 2023)

Our densities are of the form

$$\mathfrak{h}[\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Lambda}] \;=\; e^{\mathfrak{g}[\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Lambda}]}/c_{\boldsymbol{\mu},\boldsymbol{\Lambda}}$$

with explicit $\mathfrak{g}[\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Lambda}]$ and integral normalization $c_{\boldsymbol{\mu},\boldsymbol{\Lambda}}$.

Score matching

$$\operatorname*{argmin}_{\boldsymbol{\mu},\boldsymbol{\Lambda}} \mathbb{E}\Big[\big\|\underbrace{\nabla_{\boldsymbol{x}} \log \mathfrak{h}[\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Lambda}]}_{\text{“score”}} - \nabla_{\boldsymbol{x}} \log \mathfrak{h}^*[\boldsymbol{x}]\big\|_2^2\Big]$$

$$= \operatorname*{argmin}_{\boldsymbol{\mu},\boldsymbol{\Lambda}} \mathbb{E}\Big[\big\|\nabla_{\boldsymbol{x}}\, \mathfrak{g}[\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Lambda}] - \nabla_{\boldsymbol{x}} \log \mathfrak{h}^*[\boldsymbol{x}]\big\|_2^2\Big]$$

looks much more promising.

Score Matching Provides Feasible Objectives

(Extremes in high dimensions: methods and scalable algorithms, 2023)

Proposition (Score Matching): Under *some* assumptions

$$\begin{aligned} &\operatorname*{argmin}_{\boldsymbol{\mu},\boldsymbol{\Lambda}} \mathbb{E}\left[\left\|\nabla_{\boldsymbol{x}}\, \mathfrak{g}[\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Lambda}] - \nabla_{\boldsymbol{x}} \log \mathfrak{h}^*[\boldsymbol{x}]\right\|_2^2\right] \\ = \quad &\operatorname*{argmin}_{\boldsymbol{\mu},\boldsymbol{\Lambda}} \mathbb{E}\left[\sum_{j=1}^{d} \partial_{x_j} \mathfrak{g}[\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Lambda}] + 2\sum_{j=1}^{d}\left(\partial_{x_j}^2 \mathfrak{g}[\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Lambda}]\right)^2\right]. \end{aligned}$$

The Final Estimator Is Sparse and Scalable

(Extremes in high dimensions: methods and scalable algorithms, 2023)

$$
\begin{aligned}
(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Lambda}}) \in \operatorname*{argmin}_{\boldsymbol{\mu}, \boldsymbol{\Lambda}} \Bigg\{ & \sum_{i=1}^{n} \Big\| \big(\boldsymbol{\mu} - \mathbf{1} - (\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^\top - \operatorname{diag}[\boldsymbol{\Lambda}\mathbf{1} + \boldsymbol{\Lambda}^\top \mathbf{1}]) \log[\boldsymbol{x}_i]\big) \otimes \mathfrak{f}_1[\boldsymbol{x}_i] \Big\|_2^2 \\
& + \sum_{i=1}^{n} \big(\boldsymbol{\mu} - \mathbf{1} - (\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^\top - \operatorname{diag}[\boldsymbol{\Lambda}\mathbf{1} + \boldsymbol{\Lambda}^\top \mathbf{1}]) \log[\boldsymbol{x}_i]\big) \otimes \mathfrak{f}_2[\boldsymbol{x}_i] \\
& - \sum_{i=1}^{n} \operatorname{trace}\Big[(\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^\top - \operatorname{diag}[\boldsymbol{\Lambda}\mathbf{1} + \boldsymbol{\Lambda}^\top \mathbf{1}]) \boldsymbol{F}[\boldsymbol{x}_i]\Big] + r \operatorname{prior}[\boldsymbol{\mu}, \boldsymbol{\Lambda}] \Bigg\}
\end{aligned}
$$

Can be computed within seconds ($n \approx 1000$, $d \approx 20$, $p \approx 200$) to minutes ($n \approx 100\,000$, $d \approx 100$, $p \approx 5000$) on a standard laptop.

The Estimator Satisfies Finite-Sample Guarantees Even in High Dimensions

(Extremes in high dimensions: methods and scalable algorithms, 2023)

Theorem (Extremes): Under reasonable assumptions, such as r large enough, it holds that

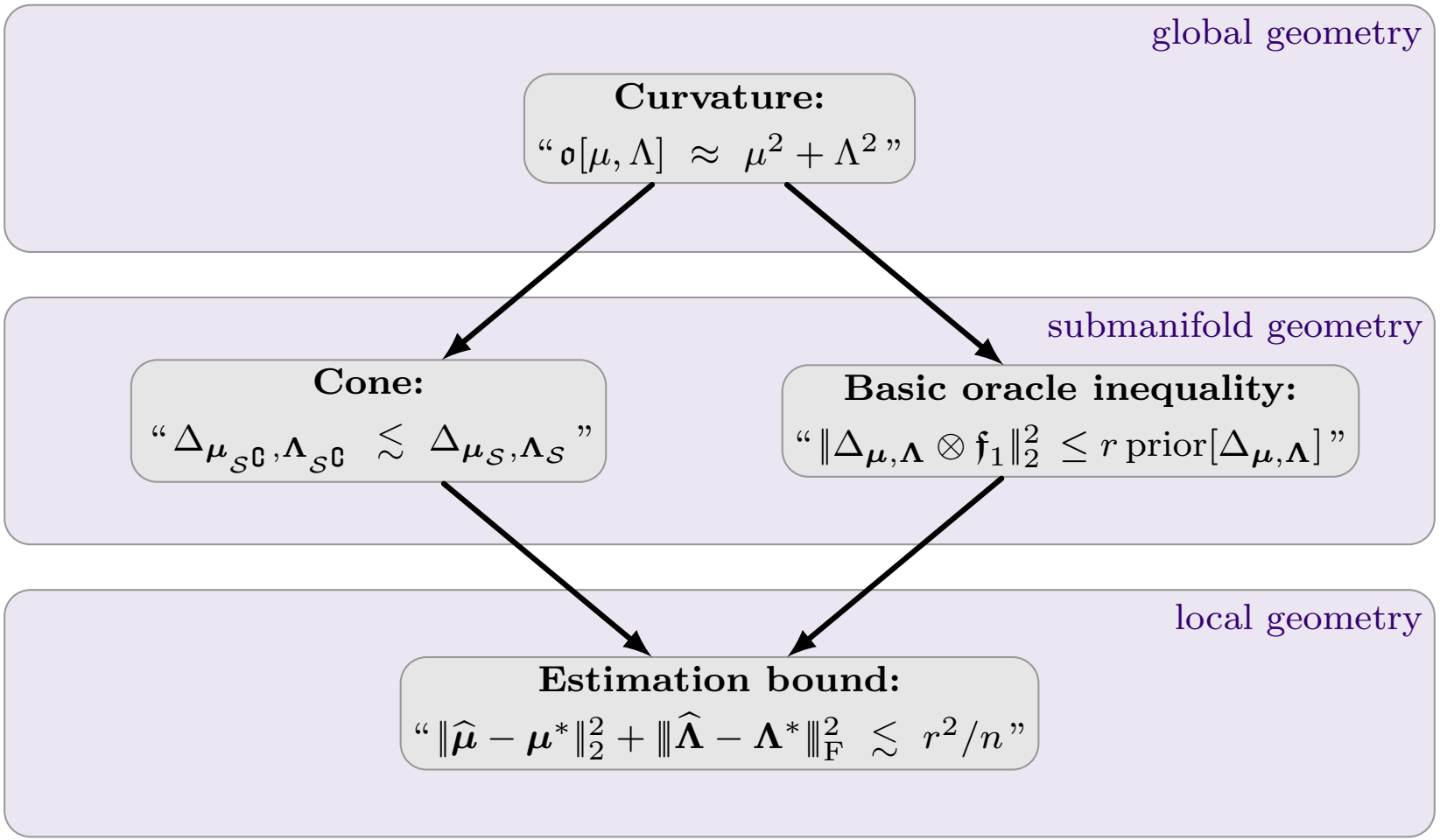
$$\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\|_2^2 + \|\widehat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}^*\|_{\mathrm{F}}^2 \lesssim \frac{\big(|\mathcal{S}_{\boldsymbol{\mu}^*}| + |\mathcal{S}_{\boldsymbol{\Lambda}^*}|\big) r^2}{n}$$

with high probability.

$\Longrightarrow$ "sparsistency"

The Statistical Proofs Are "Geometric"

(Extremes in high dimensions: methods and scalable algorithms, 2023)



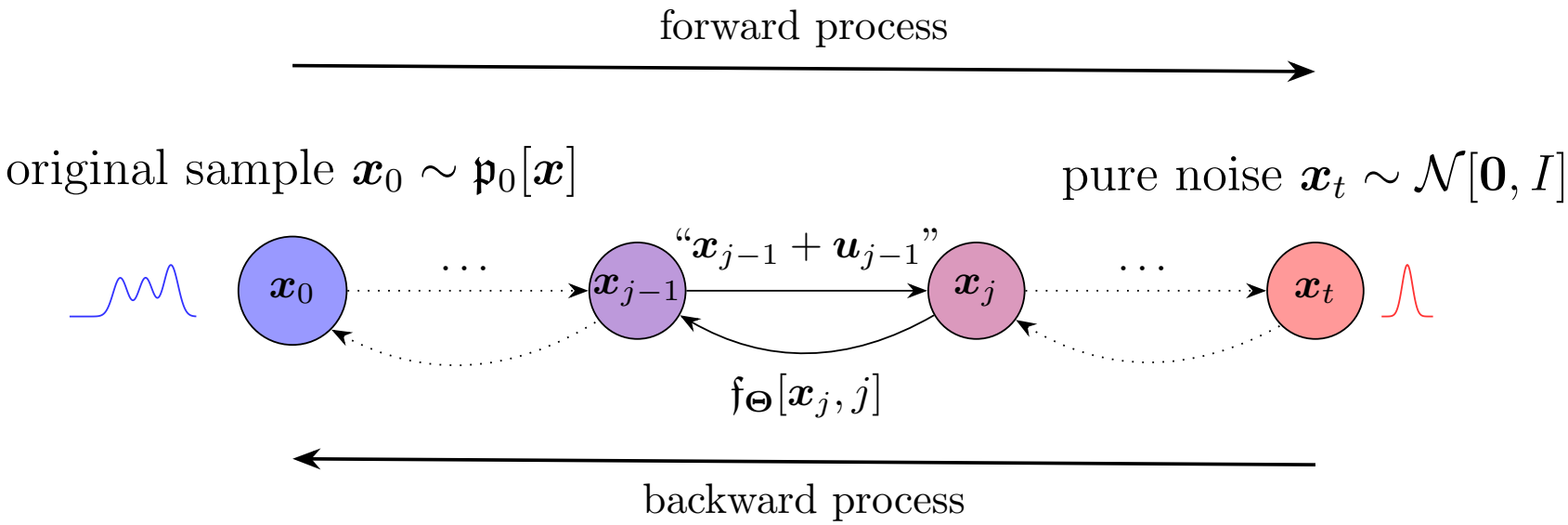
Donald J. Trump
@realDonaldTrump

7.26k ReTruths **31.5k** Likes

May 02, 2025, 10:29 PM

Diffusion Models Destroy and Rebuild

(Regularization can make diffusion models more efficient, 2025)



The Forward Process Injects Noise

(Regularization can make diffusion models more efficient, 2025)

$$\boldsymbol{x}_j \;:=\; \sqrt{1-\beta_j}\,\boldsymbol{x}_{j-1} + \sqrt{\beta_j}\,\boldsymbol{u}_{j-1}$$

with white noise $\boldsymbol{u}_j \sim \mathcal{N}[\mathbf{0}_d, I_d]$ and noise schedule $\beta_j \in (0,1)$

The distribution of $\boldsymbol{x}_j$ given an "original" sample $\boldsymbol{x}_0$ is

$$\mathbb{Q}_j[\boldsymbol{x}_j \mid \boldsymbol{x}_0] \;=\; \mathcal{N}\left[\boldsymbol{x}_j; \left(\prod_{l=1}^{j}\sqrt{1-\beta_l}\right)\boldsymbol{x}_0, \left(1-\prod_{l=1}^{j}(1-\beta_l)\right)I_d\right]$$

The Backward Process Recovers an Original

(Regularization can make diffusion models more efficient, 2025)

$$\boldsymbol{x}_{j-1} \;=\; \frac{1}{\sqrt{1-\beta_j}}\Big(\boldsymbol{x}_j + \beta_j \underbrace{\nabla_{\boldsymbol{x}_j} \log \mathfrak{q}_j[\boldsymbol{x}_j]}_{\text{“score”}}\Big) + \sqrt{\frac{\beta_j}{1-\beta_j}}\,\boldsymbol{z}_j$$

with the marginal density $\mathfrak{q}_j[\boldsymbol{x}_j] = \int \mathfrak{q}_j[\boldsymbol{x}_j \mid \boldsymbol{x}]\mathfrak{p}_0[\boldsymbol{x}]\,d\boldsymbol{x}$

and white noise $\boldsymbol{z}_j \sim \mathcal{N}[\mathbf{0}_d, I_d]$

Standard Estimation Suffers From the Curse of Dimensionality

(Regularization can make diffusion models more efficient, 2025)

$$\widehat{\boldsymbol{\Theta}} \;\in\; \operatorname*{argmin}_{\boldsymbol{\Theta}\in\mathcal{A}} \left\{ \mathbb{E}_{\substack{j\sim\mathcal{U}_{[0,t]} \\ X_j\sim\mathbb{Q}_j}} \left[\left\| \boldsymbol{s}_{\boldsymbol{\Theta}}[X_j, j] - \nabla_{X_j} \log \mathfrak{q}_j[X_j] \right\|_2^2 \right] \right\}$$

Theorem ("Classical"): It holds with high probability that

$$\mathfrak{d}_{\mathrm{KL}}[\mathfrak{p}_0, \widehat{\mathfrak{p}}_0] \;\lesssim\; \frac{d^2}{t}\,.$$

Sparsity Breaks the Curse

(Regularization can make diffusion models more efficient, 2025)

$$(\widehat{\boldsymbol{\Theta}}, \widehat{\kappa}) \;\in\; \underset{\substack{\boldsymbol{\Theta}\in\mathcal{A}_1 \\ \kappa\in(0,\infty)}}{\operatorname{argmin}} \left\{ \mathbb{E}_{\substack{j\sim\mathcal{U}_{[0,t]} \\ X_j\sim\mathbb{Q}_j}} \left[\left\| \kappa \boldsymbol{s}_{\boldsymbol{\Theta}}[X_j, j] - \nabla_{X_j} \log \mathfrak{q}_j[X_j] \right\|_2^2 \right] + r\kappa \right\}$$

Theorem (Regularized): It holds with high probability that

$$\mathfrak{d}_{\mathrm{KL}}[\mathfrak{p}_0, \widehat{\mathfrak{p}}_0] \;\lesssim\; \frac{s^2}{t}\,.$$

Regularization Makes a Visible Difference

(Regularization can make diffusion models more efficient, 2025)

Vanilla Diffusion

Training Data

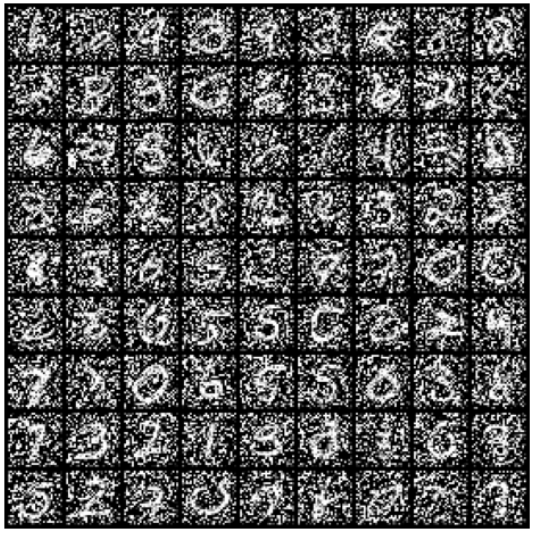
Regularized Diffusion

$t = 500$

Regularization Makes a Visible Difference

(Regularization can make diffusion models more efficient, 2025)

Vanilla Diffusion



Training Data

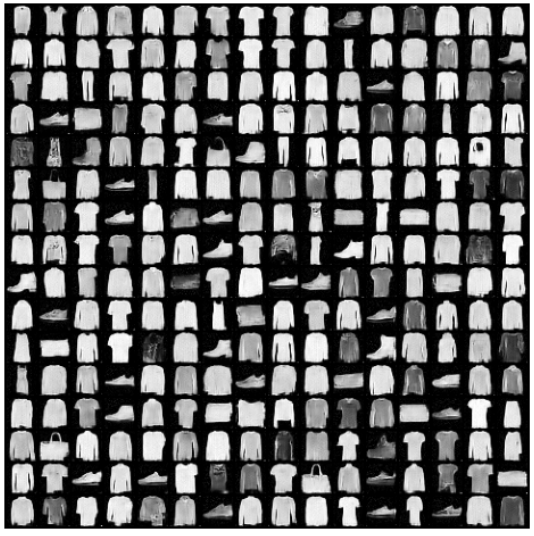
Regularized Diffusion

$t = 50$

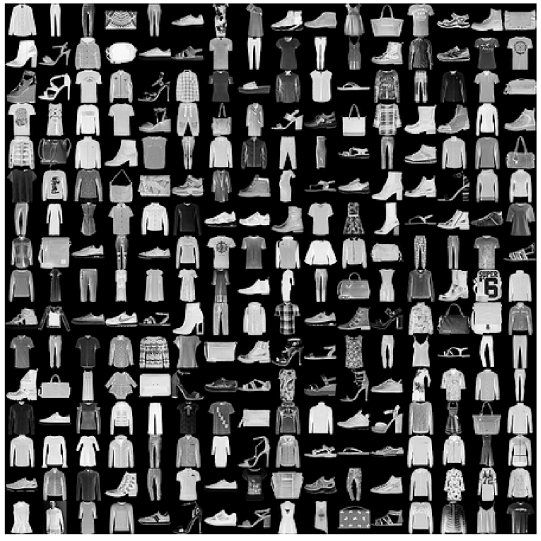
Regularization Makes a Visible Difference

(Regularization can make diffusion models more efficient, 2025)

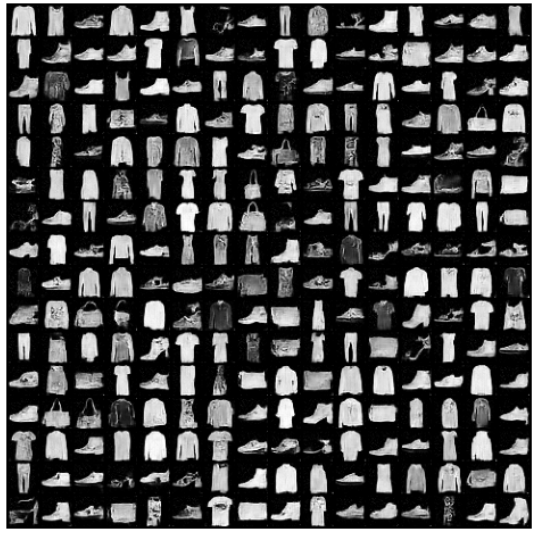
Vanilla Diffusion



Training Data



Regularized Diffusion

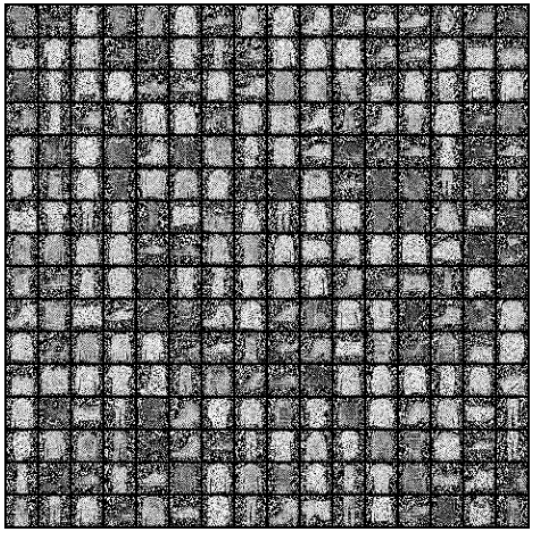


$t = 500$

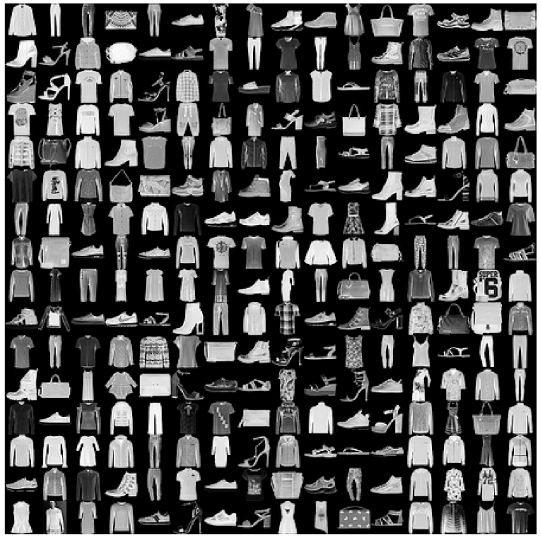
Regularization Makes a Visible Difference

(Regularization can make diffusion models more efficient, 2025)

Vanilla Diffusion

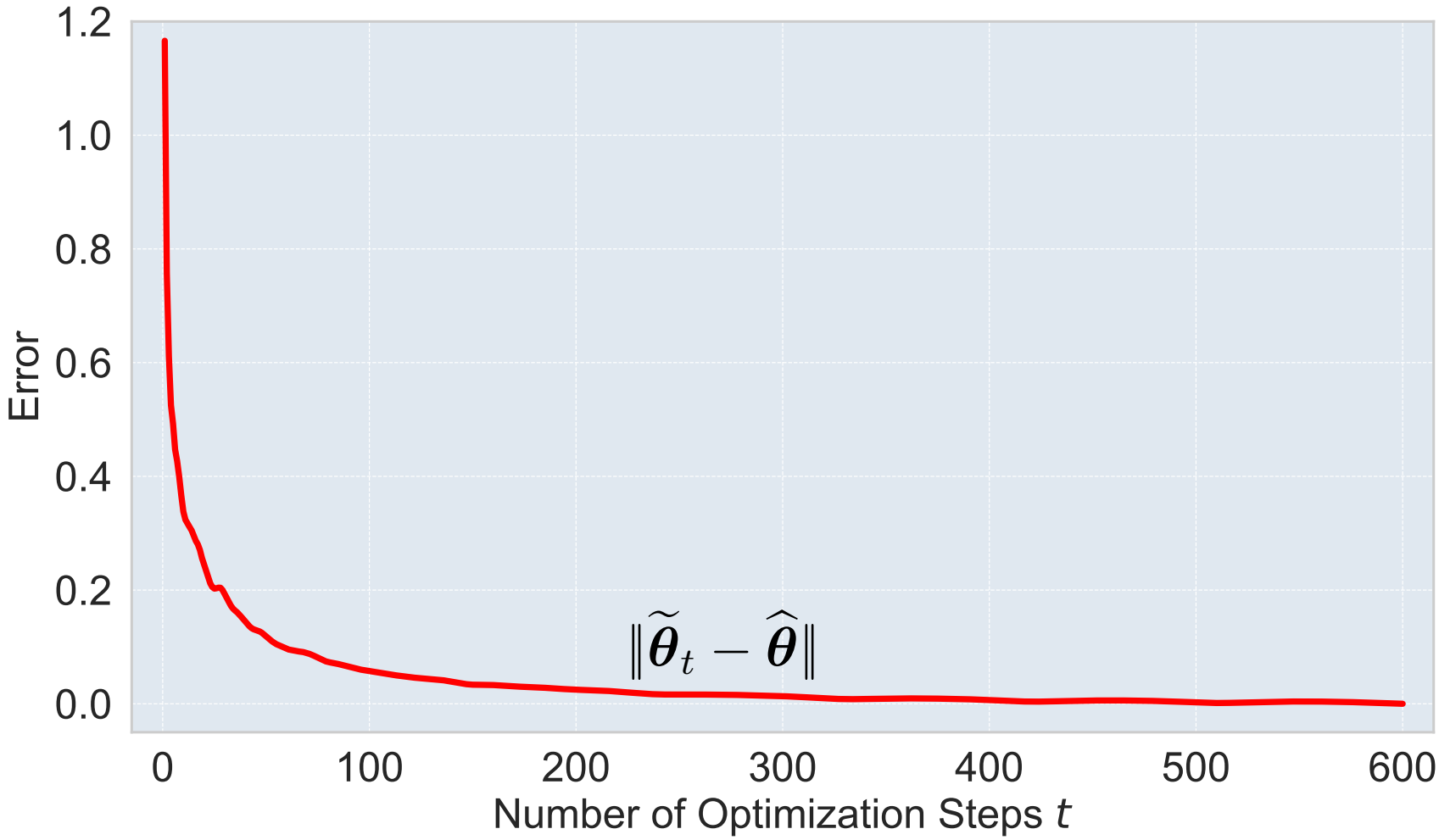


Training Data

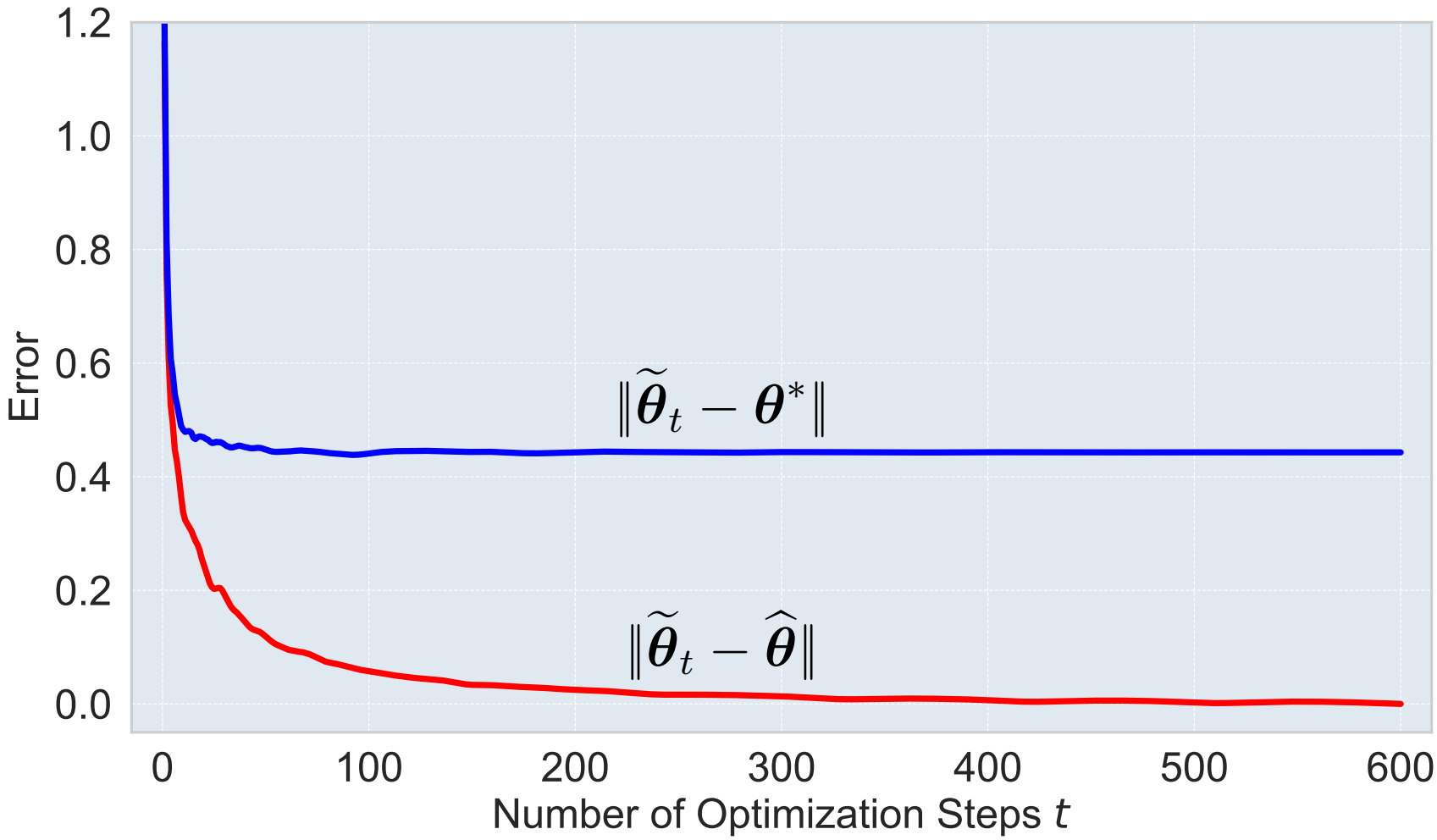


Regularized Diffusion

$t = 50$



Example: $\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\operatorname{argmin}} \{ \|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + r\|\boldsymbol{\theta}\|_1 \}$



Example: $\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\operatorname{argmin}} \{ \|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + r\|\boldsymbol{\theta}\|_1 \}$; $\boldsymbol{y} = X\boldsymbol{\theta}^* + \boldsymbol{u}$

Goal: Select t and r simultaneously

(Balancing statistical and computational precision: a general theory and applications to sparse regression, 2023)

$\widehat{\mathcal{F}}_{\mathcal{R}} = \{\widehat{\boldsymbol{\theta}}^{r} \, : \, r \in \mathcal{R}\}$ theoretical estimators

$$\mathfrak{d}[\widehat{\boldsymbol{\theta}}^{r}, \boldsymbol{\theta}^{*}] \; \leq \; \mathfrak{f}[r] \quad \forall \, r \geq r^{*}$$

$\widetilde{\mathcal{F}}_{\mathcal{R}} = \{\widetilde{\boldsymbol{\theta}}_{t}^{r} \, : \, r \in \mathcal{R}, t \in \{1, 2, \dots\}\}$ practical estimators

$$\mathfrak{d}[\widetilde{\boldsymbol{\theta}}_{t}^{r}, \widehat{\boldsymbol{\theta}}^{r}] \; \leq \; \mathfrak{g}[r, t] \quad \forall \, r \in \mathcal{R}$$

A Surprisingly Concise Algorithm

(Balancing statistical and computational precision: a general theory and applications to sparse regression, 2023)

Inputs : $\mathcal{R}$, $\mathfrak{f}$, $\mathfrak{g}$
Outputs: $\widetilde{\boldsymbol{\theta}} \in \widetilde{\mathcal{F}}_{\mathcal{R}}$, $\hat{r} \in \mathcal{R}$

Initialize $\hat{r} = \max \mathcal{R}$
Compute $\widetilde{\boldsymbol{\theta}}^{\hat{r}} = \widetilde{\boldsymbol{\theta}}^{\hat{r}}_{\hat{t}}$ with $\hat{t} = \min\{t : \mathfrak{g}[r,t] \leq \mathfrak{f}[r]\}$
Initialize $\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}^{\hat{r}}_{\hat{t}}$
while $\hat{r} \neq \min \mathcal{R}$ *AND*
    $\nexists\, r \in \mathcal{R} \setminus (-\infty, \hat{r}] : \mathfrak{d}[\widetilde{\boldsymbol{\theta}}^{r}_{\hat{t}}, \widetilde{\boldsymbol{\theta}}^{\hat{r}}_{\hat{t}}] > \mathfrak{f}[r] + \mathfrak{g}[r,\hat{t}] + \mathfrak{f}[\hat{r}] + \mathfrak{g}[\hat{r},\hat{t}]$
do
    Update $\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}^{\hat{r}}_{\hat{t}}$
    Go to $\hat{r} = \max\{\mathcal{R} \setminus [\hat{r}, \infty)\}$
    Compute $\widetilde{\boldsymbol{\theta}}^{\hat{r}} = \widetilde{\boldsymbol{\theta}}^{\hat{r}}_{\hat{t}}$ with $\hat{t} = \min\{t : \mathfrak{g}[r,t] \leq \mathfrak{f}[r]\}$
end

We Have Optimal Theoretical Guarantees

(Balancing statistical and computational precision: a general theory and applications to sparse regression, 2023)

Theorem (Optimality): It holds that

$$\mathfrak{d}[\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*] \;\leq\; 6\,\mathfrak{f}[r^*]\,.$$

The Theorem Has a Short, Elementary Proof

Consider $\hat{r}$ with $\widetilde{\boldsymbol{\theta}}^{\hat{r}} = \widetilde{\boldsymbol{\theta}}$ (we omit $\tilde{t}$ for ease of notation)

If $\hat{r} > r^*$, then $\exists\, r', r'' \geq r^*$ such that

$$\mathfrak{d}[\widetilde{\boldsymbol{\theta}}^{r'}, \widetilde{\boldsymbol{\theta}}^{r''}] > \mathfrak{f}[r'] + \mathfrak{g}[r'] + \mathfrak{f}[r''] + \mathfrak{g}[r''] \,.$$

The Theorem Has a Short, Elementary Proof

Consider $\hat{r}$ with $\widetilde{\boldsymbol{\theta}}^{\hat{r}} = \widetilde{\boldsymbol{\theta}}$ (we omit $\tilde{t}$ for ease of notation)

If $\hat{r} > r^*$, then $\exists\, r', r'' \geq r^*$ such that

$$\mathfrak{d}[\widetilde{\boldsymbol{\theta}}^{r'}, \widetilde{\boldsymbol{\theta}}^{r''}] > \mathfrak{f}[r'] + \mathfrak{g}[r'] + \mathfrak{f}[r''] + \mathfrak{g}[r''] .$$

But

$$\begin{aligned}
&\mathfrak{d}[\widetilde{\boldsymbol{\theta}}^{r'}, \widetilde{\boldsymbol{\theta}}^{r''}] \\
&\leq \quad \mathfrak{d}[\widetilde{\boldsymbol{\theta}}^{r'}, \boldsymbol{\theta}^*] \quad + \quad \mathfrak{d}[\widetilde{\boldsymbol{\theta}}^{r''}, \boldsymbol{\theta}^*] \\
&\leq \underbrace{\mathfrak{d}[\widehat{\boldsymbol{\theta}}^{r'}, \boldsymbol{\theta}^*]}_{\text{stat}} + \underbrace{\mathfrak{d}[\widetilde{\boldsymbol{\theta}}^{r'}, \widehat{\boldsymbol{\theta}}^{r'}]}_{\text{opti}} + \underbrace{\mathfrak{d}[\widehat{\boldsymbol{\theta}}^{r''}, \boldsymbol{\theta}^*]}_{\text{stat}} + \underbrace{\mathfrak{d}[\widetilde{\boldsymbol{\theta}}^{r''}, \widehat{\boldsymbol{\theta}}^{r''}]}_{\text{opti}} \\
&\leq \quad \mathfrak{f}[r'] \quad + \quad \mathfrak{g}[r'] \quad + \quad \mathfrak{f}[r''] \quad + \quad \mathfrak{g}[r''] .
\end{aligned}$$

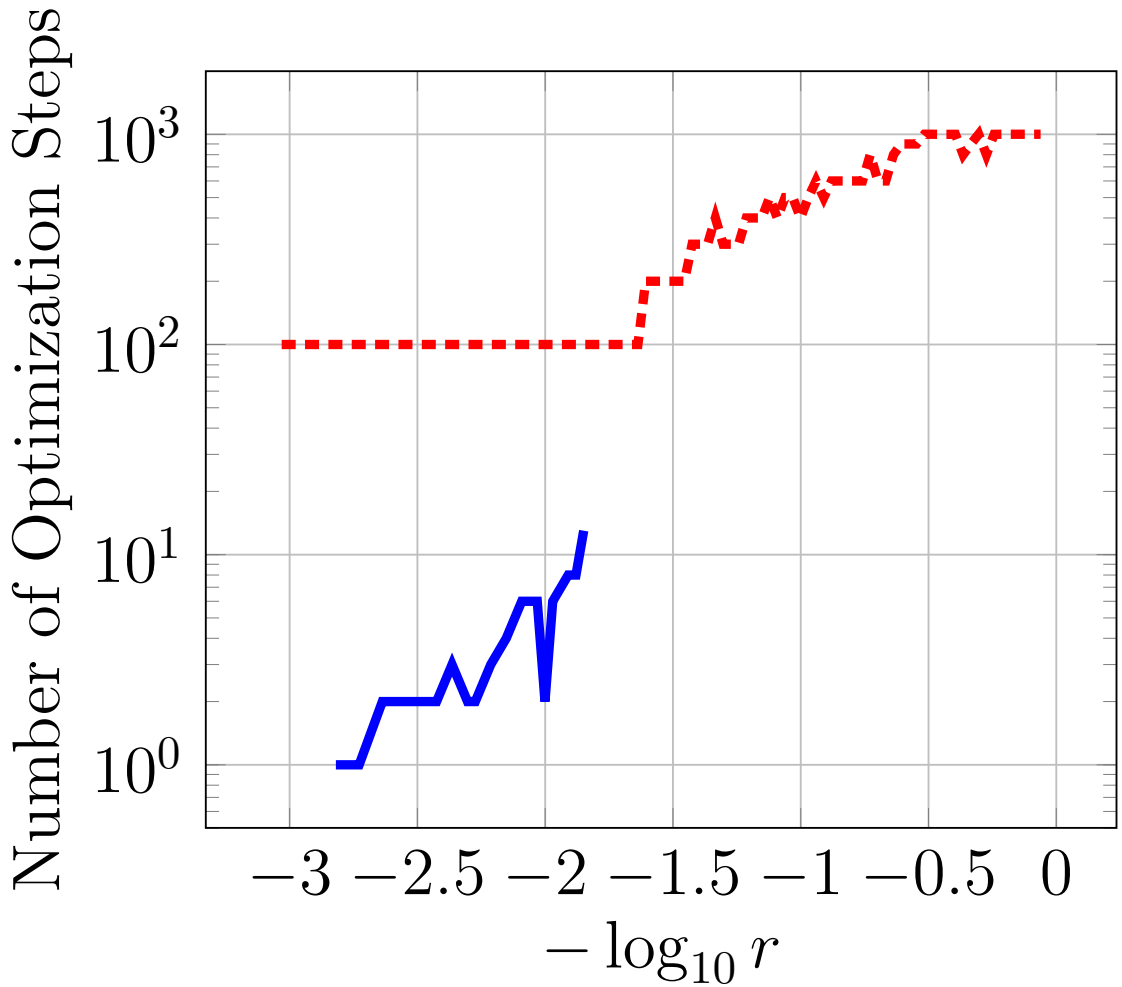
The Theorem Has a Short, Elementary Proof

Recall that $\hat{r} \leq r^*$. Hence,

$$
\begin{aligned}
& \mathfrak{d}[\widetilde{\boldsymbol{\theta}}^{\hat{r}}, \boldsymbol{\theta}^*] \\
\leq \quad & \underbrace{\mathfrak{d}[\widetilde{\boldsymbol{\theta}}^{\hat{r}}, \widetilde{\boldsymbol{\theta}}^{r^*}]}_{\text{algorithm}} + \underbrace{\mathfrak{d}[\widetilde{\boldsymbol{\theta}}^{r^*}, \boldsymbol{\theta}^*]}_{\text{stat\&opti}} \\
\leq \quad & \mathfrak{f}[\hat{r}] + \mathfrak{g}[\hat{r}] + \mathfrak{f}[r^*] + \mathfrak{g}[r^*] + \mathfrak{f}[r^*] + \mathfrak{g}[r^*] \\
\leq \quad & 3\,\mathfrak{f}[r^*] + 3\,\mathfrak{g}[r^*] \leq 6\,\mathfrak{f}[r^*]\,.
\end{aligned}
$$

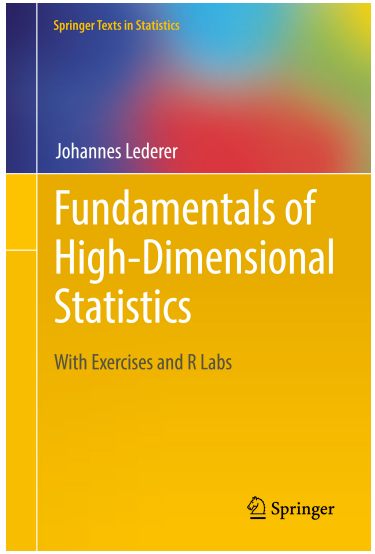
Works Like a Charm for the Lasso

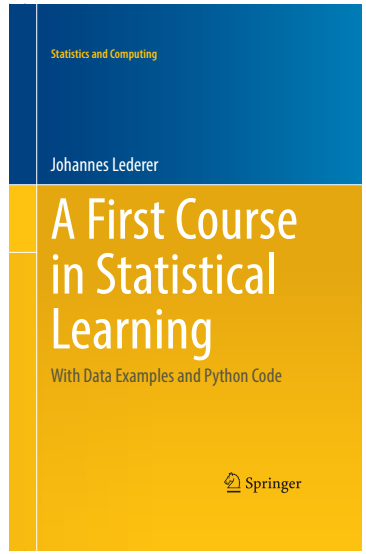
(Balancing statistical and computational precision: a general theory and applications to sparse regression, 2023)



Stats Is a Cornerstone of Data Science and AI

Stats Is a Cornerstone of Data Science and AI





Workshop August 2025: `datascienceminds.com`

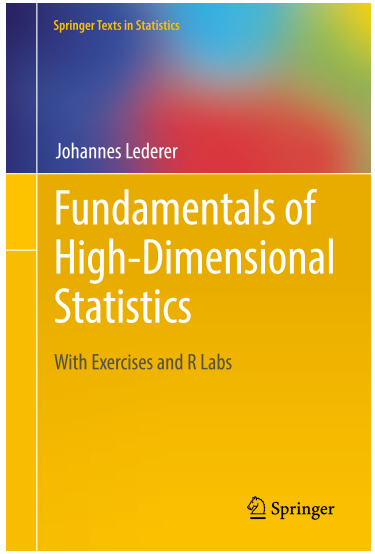
`johanneslederer.com`

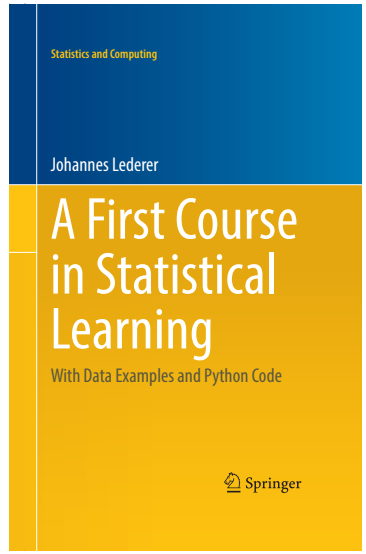
LedererLab

johannesledererer

thedata.scienceguy

Stats Is a Cornerstone of Data Science and AI





Workshop August 2025: `datascienceminds.com`

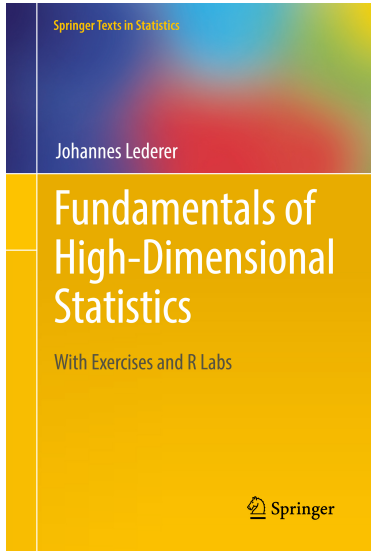
`johanneslederer.com`

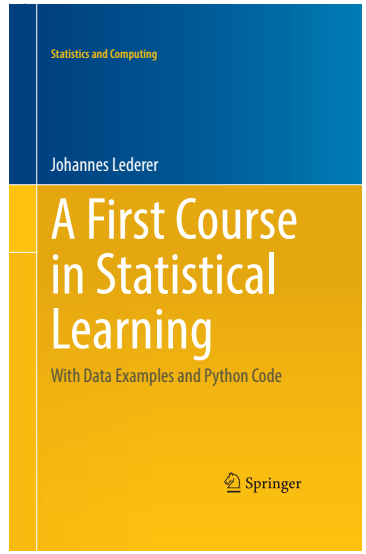
LedererLab

johannesledererer

thedata.scienceguy

Stats Is a Cornerstone of Data Science and AI





Workshop August 2025: `datascienceminds.com`

`johanneslederer.com`

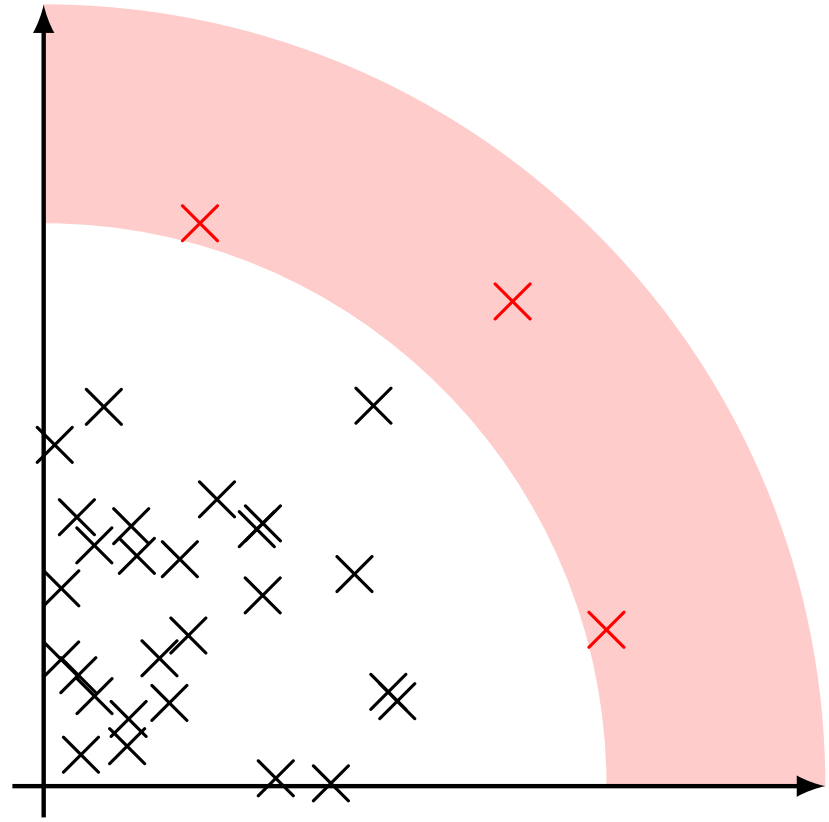
LedererLab

johannesledererer

thedata.scienceguy

Extreme-Value Theory Is Not Extrapolation

(Extremes in high dimensions: methods and scalable algorithms, 2023)



$$\lim_{t\to\infty} \mathbb{P}\{X/t \in \mathcal{A} \mid \|X\| > t\} \;\sim\; \mathfrak{m}[\mathcal{A}]$$

New Densities Disentangle the Parameters

(Extremes in high dimensions: methods and scalable algorithms, 2023)

Theorem: Every Hüsler-Reiss density is of the form

$$\frac{1}{c_{\boldsymbol{\mu},\boldsymbol{\Lambda}}} \left(\prod_{j=1}^{d} \frac{1}{x_j} \right) \exp\Big[\boldsymbol{\mu}^\top \log[\boldsymbol{x}] - \frac{1}{2} \log[\boldsymbol{x}] \big(\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^\top - \operatorname{diag}(\boldsymbol{\Lambda}\mathbf{1} + \boldsymbol{\Lambda}^\top \mathbf{1}) \big) \log[\boldsymbol{x}] \Big] .$$

- $\boldsymbol{\mu} \in \mathbb{R}^d$: shifts
- $\boldsymbol{\Lambda} \in \mathbb{R}^{d \times d}$ (upper triangular): dependences
- $c_{\boldsymbol{\mu},\boldsymbol{\Lambda}}$: normalization

We Develop a Score-Matching Approach That Satisfies the Boundary Conditions

(Extremes in high dimensions: methods and scalable algorithms, 2023)

$$\operatorname{argmin}_{\boldsymbol{\mu},\boldsymbol{\Lambda}} \mathbb{E}\Big[\big\| \big(\nabla_{\boldsymbol{x}} \log \mathfrak{h}[\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Lambda}] - \nabla_{\boldsymbol{x}} \log \mathfrak{h}[\boldsymbol{x}]\big) \otimes \underbrace{\boldsymbol{x} \otimes \log[\boldsymbol{x}]}_{\mathfrak{w}[\boldsymbol{x}]} \big\|_2^2\Big]$$

$\log[x]$ $\qquad$ $x \in \mathcal{X}$

The Estimator Behaves as Intended and can Be Computed Super Fast

(Extremes in high dimensions: methods and scalable algorithms, 2023)

$\mathbf{n = 500},\ \mathbf{d = 20},\ \mathbf{p = 190}$

r	$1000\sqrt{\frac{\log[d]}{n}}$	$100\sqrt{\frac{\log[d]}{n}}$	$10\sqrt{\frac{\log[d]}{n}}$	$1\sqrt{\frac{\log[d]}{n}}$	$0.1\sqrt{\frac{\log[d]}{n}}$	$0.01\sqrt{\frac{\log[d]}{n}}$	0
$\|\widehat{\mathbf{\Lambda}}\|_0/(d(d-1)/2)$	$33.9\,\% \pm 2.7\,\%$	$63.1\,\% \pm 3.7\,\%$	$96.0\,\% \pm 1.3\,\%$	$99.7\,\% \pm 0.5\,\%$	$100.0\,\% \pm 0.1\,\%$	$100.0\,\% \pm 0.1\,\%$	$100.0\,\% \pm 0.1\,\%$
$t_{\text{pre}} + t_{\text{opt}}$ (s)	$< 0.01\,s \pm 0.01\,s \quad + \quad 0.83\,s \pm 0.11\,s$						

$\mathbf{n = 50\,000},\ \mathbf{d = 80},\ \mathbf{p = 3160}$

r	$1000\sqrt{\frac{\log[d]}{n}}$	$100\sqrt{\frac{\log[d]}{n}}$	$10\sqrt{\frac{\log[d]}{n}}$	$1\sqrt{\frac{\log[d]}{n}}$	$0.1\sqrt{\frac{\log[d]}{n}}$	$0.01\sqrt{\frac{\log[d]}{n}}$	0
$\|\widehat{\mathbf{\Lambda}}\|_0/(d(d-1)/2)$	$79.9\,\% \pm 0.7\,\%$	$98.0\,\% \pm 0.3\,\%$	$99.8\,\% \pm 0.1\,\%$	$100.0\,\% \pm 0.0\,\%$	$100.0\,\% \pm 0.0\,\%$	$100.0\,\% \pm 0.0\,\%$	$100.0\,\% \pm 0.0\,\%$
$t_{\text{pre}} + t_{\text{opt}}$	$35.65\,s \pm 2.72\,s \quad + \quad 199.14\,s \pm 17.59\,s$						

Plenary talk at ICORS, May 21, 2025

Image credits:
River Lambro May 2024: LaPress/REX/Shutterstock
realDonaldTrump on Truth Social