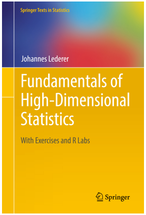
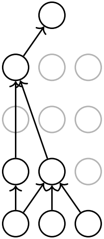
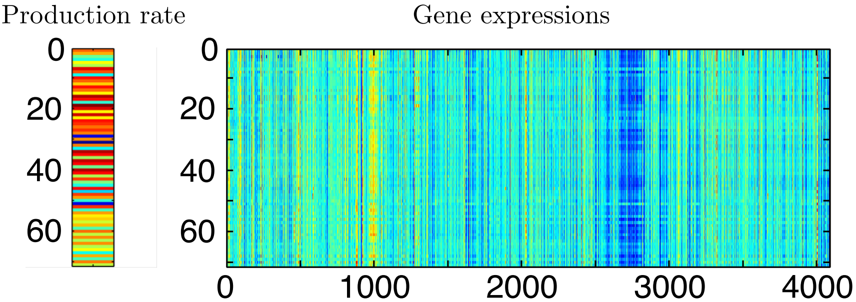
Sparse Deep Learning: From High-Dimensional Statistics to Neural Networks



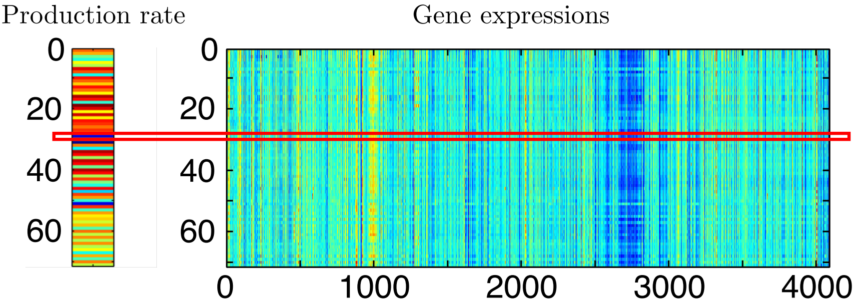


Johannes Lederer

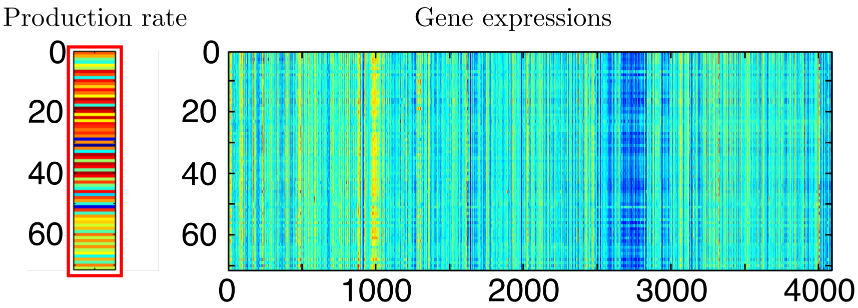
High-Dimensionality Is Common: Riboflavin Production in *B. subtilis*



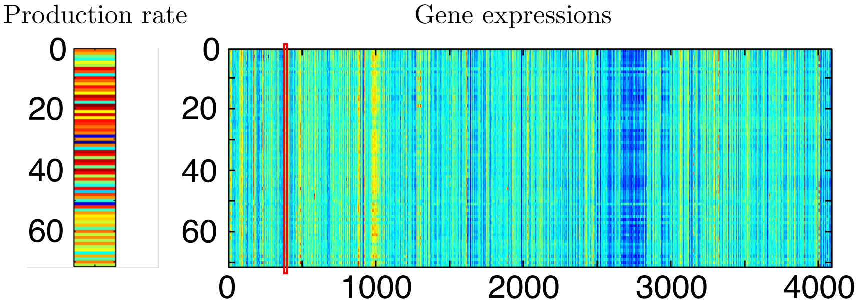
High-Dimensionality Is Common: Riboflavin Production in *B. subtilis*



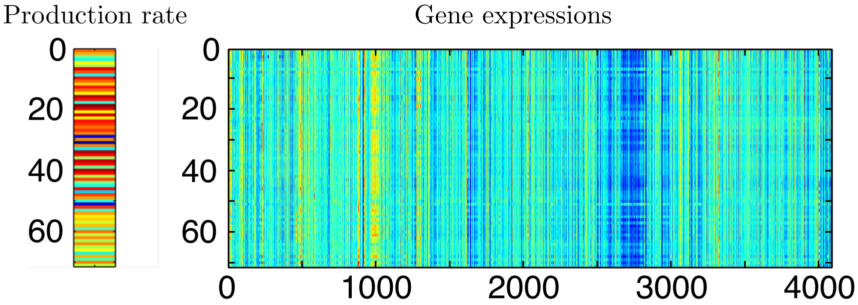
High-Dimensionality Is Common: Riboflavin Production in *B. subtilis*



High-Dimensionality Is Common: Riboflavin Production in *B. subtilis*



High-Dimensionality Is Common: Riboflavin Production in *B. subtilis*



$$\boldsymbol{y} = X\boldsymbol{\beta}^* + \boldsymbol{u}$$

Sparsity-Inducing Prior Terms Are Standard Tools in Statistics

(Fundamentals of High-Dimensional Statistics—With Exercises and R Labs, 2021)

$$\boldsymbol{y} \;=\; X\boldsymbol{\beta}^* + \boldsymbol{u}$$

$$\widehat{\boldsymbol{\beta}}_{\text{lasso}} \;\in\; \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\Big\{\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + r\|\boldsymbol{\beta}\|_1\Big\}$$

$$\widehat{\boldsymbol{\beta}}_{\text{grplasso}} \;\in\; \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\Big\{\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + r\|\boldsymbol{\beta}\|_{2,1}\Big\}$$

$$\text{where} \quad \|\boldsymbol{\beta}\|_1 \;:=\; \sum_{j=1}^{p}|\beta_j| \quad \text{and} \quad \|\boldsymbol{\beta}\|_{2,1} \;:=\; \sum_{k=1}^{g}\|\boldsymbol{\beta}_{\mathcal{G}_k}\|_2$$

Theory Exists: Oracle Inequalities

(Oracle inequalities for high-dimensional prediction, 2019)

Theorem (Power-One Bound): It holds with probability at least $1 - 1/n$ that

$$\frac{\|X\boldsymbol{\beta}^* - X\widehat{\boldsymbol{\beta}}_{\text{lasso}}\|_2^2}{n} \;\leq\; 6\sigma\|\boldsymbol{\beta}^*\|_1\|\boldsymbol{x}\|_n\sqrt{\frac{\log[np]}{n}}\,.$$

$\sigma \;\leftrightarrow\;$ noise

$\|\boldsymbol{x}\|_n \;=\; 1$ (for normalized inputs)

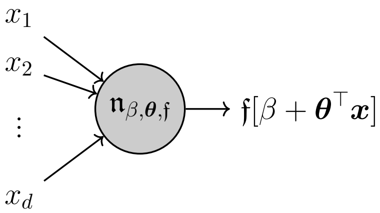
$n \;=\; \#$ samples

$p \;=\; \#$ parameters

Neural Networks Consist of Neurons

(Activation functions in artificial neural networks: a systematic overview, 2021)

$$
\begin{aligned}
\mathfrak{n}_{\beta,\boldsymbol{\theta},\mathfrak{f}} \; : \; \mathbb{R}^d &\to \mathbb{R} \\
\boldsymbol{x} &\mapsto \mathfrak{f}[\beta + \boldsymbol{\theta}^\top \boldsymbol{x}] = \mathfrak{f}\left[\beta + \sum_{j=1}^{d} \theta_j x_j\right]
\end{aligned}
$$

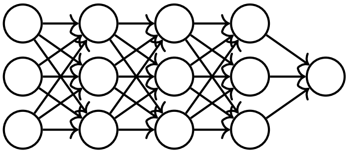


Neurons Can Be Readily Combined

(Activation functions in artificial neural networks: a systematic overview, 2021)

$$
\begin{aligned}
\mathfrak{g}_\Theta \;:\; \mathbb{R}^d &\to \mathbb{R} \\
\boldsymbol{x} &\mapsto \mathfrak{g}_\Theta[\boldsymbol{x}] := W^L \mathfrak{f}^L\big[\ldots W^1 \mathfrak{f}^1[W^0 \boldsymbol{x}]\big]
\end{aligned}
$$

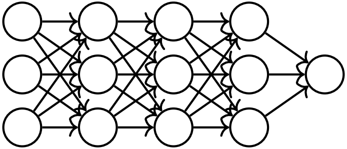
$$
\mathcal{A} := \left\{ \Theta = (W^L, \ldots, W^0) \;:\; W^l \in \mathbb{R}^{p_{l+1} \times p_l} \right\}
$$



Neural Networks Are High-Dimensional

(Statistical guarantees for regularized neural networks, 2021)

$$\mathcal{A} := \left\{ \Theta = (W^L, \ldots, W^0) \;:\; W^l \in \mathbb{R}^{p_{l+1} \times p_l} \right\}$$



Number of parameters in the toy network: $P = 30$

Number of parameters in general: $P = \sum_{l=0}^{L} p_{l+1} p_l$

We Invoke Sparsity I: Motivation

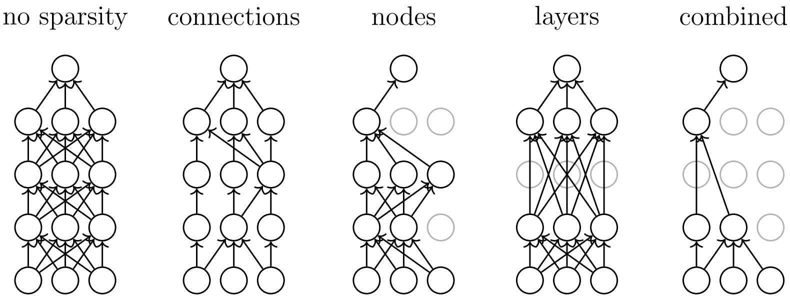
(Layer sparsity in neural networks, 2020)

Sparsity is well established in statistics and machine learning:

- Avoid overfitting
- Save computations and memory
- Improve interpretability

We Invoke Sparsity II: Concepts

(Layer sparsity in neural networks, 2020)



We Invoke Sparsity III: Implementations

(Layer sparsity in neural networks, 2020)

$$\mathfrak{r}^{\text{con}}[\Theta] \;:=\; \sum_{l=0}^{L} |||W^l|||_1 \;:=\; \sum_{l=0}^{L} \sum_{v=1}^{p_{l+1}} \sum_{w=1}^{p_l} |(W^l)_{vw}|$$

$$\mathfrak{r}^{\text{node}}[\Theta] \;:=\; \sum_{l=0}^{L} |||W^l|||_{2,1} \;:=\; \sum_{l=0}^{L} \sum_{v=1}^{p_{l+1}} \sqrt{\sum_{w=1}^{p_l} |(W^l)_{vw}|^2}$$

$$\mathfrak{r}^{\text{lay}}[\Theta] \;:=\; \sum_{l=1}^{L} |||W^l|||_{2,-} \;:=\; \sum_{l=1}^{L} \sqrt{\sum_{v=1}^{p_{l+1}} \sum_{w=1}^{p_l} \big(\operatorname{neg}[(W^l)_{vw}]\big)^2}$$

Sparsity Leads to Accurate Deep Learning

(Statistical guarantees for regularized neural networks, 2021)

Theorem: It holds with probability at least $1 - 1/N$ that

$$\mathrm{err}^2[\hat{\kappa}_{\mathfrak{h}}\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}] \ \leq \ a\sigma\kappa_*\Big(\frac{2a_{\mathrm{Lip}}}{L}\Big)^L \|\boldsymbol{x}\|_N\sqrt{L\log(2P)}\frac{\log(2N)}{\sqrt{N}}\,.$$

$a \ = \ $ constant

$\sigma \ \leftrightarrow \ $ noise

$\kappa_* \ \leftrightarrow \ \|\boldsymbol{\beta}^*\|_1$

$a_{\mathrm{Lip}} \ = \ 1$ (for relu)

$\|\boldsymbol{x}\|_N \ = \ 1$ (for normalized inputs)

$N \ = \ $ # samples

$P \ = \ $ # parameters

$L \ = \ $ # hidden layers

We Use Standard Measures for the Accuracy

(Statistical guarantees for regularized neural networks, 2021)

$$\mathrm{err}[\hat{\kappa}_{\mathfrak{h}}\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}] \;:=\; \sqrt{\frac{1}{N}\sum_{i=1}^{N}\Bigl(\hat{\kappa}_{\mathfrak{h}}\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i] - \kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i]\Bigr)^2}$$

$$\mathrm{risk}[\hat{\kappa}_{\mathfrak{h}}\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}] \;:=\; \mathbb{E}_{(\boldsymbol{x},y)}\Bigl[\Bigl(\hat{\kappa}_{\mathfrak{h}}\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}] - y\Bigr)^2\Bigr]$$

Using a Parametric Model for Illustration

(Statistical guarantees for regularized neural networks, 2021)

$$y_i \;=\; \kappa_* \mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i] + u_i$$

$u_1, \ldots, u_N$ i.i.d. $\mathcal{N}[0, \sigma^2]$

Using a Parametric Model for Illustration

(Statistical guarantees for regularized neural networks, 2021)

$$y_i \;=\; \kappa_* \mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i] + u_i$$

Proposition: Assume relu activation and define for a norm $\mathfrak{h}$

$$\mathcal{A}_{\mathfrak{h}} \;:=\; \bigl\{\Theta \in \mathcal{A} \;:\; \mathfrak{h}[\Theta] \;\leq\; 1\bigr\}\,.$$

Then, for every $\Theta \in \mathcal{A}$, there exists a pair of $\kappa \in [0, \infty)$ and $\Omega \in \mathcal{A}_{\mathfrak{h}}$ such that

$$\mathfrak{g}_{\Theta}[\boldsymbol{x}] \;=\; \kappa \mathfrak{g}_{\Omega}[\boldsymbol{x}] \qquad\qquad \text{for all } \boldsymbol{x} \in \mathbb{R}^d\,;$$

and vice versa, for every pair of $\kappa \in [0, \infty)$ and $\Omega \in \mathcal{A}_{\mathfrak{h}}$, there exists a $\Theta \in \mathcal{A}$ such that the above equality holds.

Scale Regularization Reduces Ambiguity

(Statistical guarantees for regularized neural networks, 2021)

$$(\hat{\kappa}_{\mathfrak{h}}, \widehat{\Omega}_{\mathfrak{h}}) \;\in\; \operatorname*{argmin}_{\substack{\kappa\in[0,\infty)\\ \Omega\in\mathcal{A}_{\mathfrak{h}}}} \left\{ \frac{1}{N}\sum_{i=1}^{N}\bigl(y_i - \kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]\bigr)^2 + r\kappa \right\}$$

r is an "appropriate" tuning parameter

Our Proofs Connect Different Fields

(Statistical guarantees for regularized neural networks, 2021)

High-dimensional statistics

- Oracle inequalities and effective noise

Analysis

- “Scale-free” NNs are bounded and Lipschitz

Empirical-process theory

- Metric entropy and concentration inequalities

Oracle Inequality Due to Scale Trick

(Statistical guarantees for regularized neural networks, 2021)

$$\frac{1}{N}\sum_{i=1}^{N}\big(y_i-\hat{\kappa}_{\mathfrak{h}}\,\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i]\big)^2+r\hat{\kappa}_{\mathfrak{h}}\;\leq\;\frac{1}{N}\sum_{i=1}^{N}\big(y_i-\kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]\big)^2+r\kappa$$

Oracle Inequality Due to Scale Trick

(Statistical guarantees for regularized neural networks, 2021)

$$\frac{1}{N}\sum_{i=1}^{N}\big(y_i - \hat{\kappa}_{\mathfrak{h}}\,\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i]\big)^2 + r\hat{\kappa}_{\mathfrak{h}} \;\leq\; \frac{1}{N}\sum_{i=1}^{N}\big(y_i - \kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]\big)^2 + r\kappa$$

$$\frac{1}{N}\sum_{i=1}^{N}\big(\kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i] + u_i - \hat{\kappa}_{\mathfrak{h}}\,\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i]\big)^2 + r\hat{\kappa}_{\mathfrak{h}} \;\leq\; \frac{1}{N}\sum_{i=1}^{N}\big(\kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i] + u_i - \kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]\big)^2 + r\kappa$$

Oracle Inequality Due to Scale Trick

(Statistical guarantees for regularized neural networks, 2021)

$$\frac{1}{N}\sum_{i=1}^{N}\big(y_i - \hat{\kappa}_{\mathfrak{h}}\,\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i]\big)^2 + r\hat{\kappa}_{\mathfrak{h}} \;\leq\; \frac{1}{N}\sum_{i=1}^{N}\big(y_i - \kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]\big)^2 + r\kappa$$

$$\frac{1}{N}\sum_{i=1}^{N}\big(\kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i] + u_i - \hat{\kappa}_{\mathfrak{h}}\,\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i]\big)^2 + r\hat{\kappa}_{\mathfrak{h}} \;\leq\; \frac{1}{N}\sum_{i=1}^{N}\big(\kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i] + u_i - \kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]\big)^2 + r\kappa$$

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\big(\hat{\kappa}_{\mathfrak{h}}\,\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i] - \kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i]\big)^2 \;\leq\;& \frac{1}{N}\sum_{i=1}^{N}\big(\kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i] - \kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i]\big)^2 \\ &+ \frac{2}{N}\sum_{i=1}^{N}\hat{\kappa}_{\mathfrak{h}}\,\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i]u_i - \frac{2}{N}\sum_{i=1}^{N}\kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]u_i + r\kappa - r\hat{\kappa}_{\mathfrak{h}}\end{aligned}$$

Oracle Inequality Due to Scale Trick

(Statistical guarantees for regularized neural networks, 2021)

$$\frac{1}{N}\sum_{i=1}^{N}\big(y_i-\hat{\kappa}_{\mathfrak{h}}\,\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i]\big)^2+r\hat{\kappa}_{\mathfrak{h}} \;\le\; \frac{1}{N}\sum_{i=1}^{N}\big(y_i-\kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]\big)^2+r\kappa$$

$$\frac{1}{N}\sum_{i=1}^{N}\big(\kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i]+u_i-\hat{\kappa}_{\mathfrak{h}}\,\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i]\big)^2+r\hat{\kappa}_{\mathfrak{h}} \;\le\; \frac{1}{N}\sum_{i=1}^{N}\big(\kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i]+u_i-\kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]\big)^2+r\kappa$$

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\big(\hat{\kappa}_{\mathfrak{h}}\,\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i]-\kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i]\big)^2 \;\le\; &\frac{1}{N}\sum_{i=1}^{N}\big(\kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]-\kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i]\big)^2\\ &+\frac{2}{N}\sum_{i=1}^{N}\hat{\kappa}_{\mathfrak{h}}\,\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i]u_i-\frac{2}{N}\sum_{i=1}^{N}\kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]u_i+r\kappa-r\hat{\kappa}_{\mathfrak{h}}\end{aligned}$$

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\big(\hat{\kappa}_{\mathfrak{h}}\,\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i]-\kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i]\big)^2 \le &\frac{1}{N}\sum_{i=1}^{N}\big(\kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]-\kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i]\big)^2\\ &+\hat{\kappa}_{\mathfrak{h}}\sup_{\Omega\in\mathcal{A}_{\mathfrak{h}}}\Big|\frac{2}{N}\sum_{i=1}^{N}\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]u_i\Big|+\kappa\sup_{\Omega\in\mathcal{A}_{\mathfrak{h}}}\Big|\frac{2}{N}\sum_{i=1}^{N}\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]u_i\Big|+r\kappa-r\hat{\kappa}_{\mathfrak{h}}\,.\end{aligned}$$

Oracle Inequality Due to Scale Trick

(Statistical guarantees for regularized neural networks, 2021)

$$\frac{1}{N}\sum_{i=1}^{N}\big(y_i-\hat{\kappa}_{\mathfrak{h}}\,\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i]\big)^2+r\hat{\kappa}_{\mathfrak{h}} \;\leq\; \frac{1}{N}\sum_{i=1}^{N}\big(y_i-\kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]\big)^2+r\kappa$$

$$\frac{1}{N}\sum_{i=1}^{N}\big(\kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i]+u_i-\hat{\kappa}_{\mathfrak{h}}\,\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i]\big)^2+r\hat{\kappa}_{\mathfrak{h}} \;\leq\; \frac{1}{N}\sum_{i=1}^{N}\big(\kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i]+u_i-\kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]\big)^2+r\kappa$$

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\big(\hat{\kappa}_{\mathfrak{h}}\,\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i]-\kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i]\big)^2 \;\leq\;& \frac{1}{N}\sum_{i=1}^{N}\big(\kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]-\kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i]\big)^2\\ &+\frac{2}{N}\sum_{i=1}^{N}\hat{\kappa}_{\mathfrak{h}}\,\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i]u_i-\frac{2}{N}\sum_{i=1}^{N}\kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]u_i+r\kappa-r\hat{\kappa}_{\mathfrak{h}}\end{aligned}$$

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\big(\hat{\kappa}_{\mathfrak{h}}\,\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i]-\kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i]\big)^2 \leq\;& \frac{1}{N}\sum_{i=1}^{N}\big(\kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]-\kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i]\big)^2\\ &+\hat{\kappa}_{\mathfrak{h}}\sup_{\Omega\in\mathcal{A}_{\mathfrak{h}}}\Big|\frac{2}{N}\sum_{i=1}^{N}\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]u_i\Big|+\kappa\sup_{\Omega\in\mathcal{A}_{\mathfrak{h}}}\Big|\frac{2}{N}\sum_{i=1}^{N}\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]u_i\Big|+r\kappa-r\hat{\kappa}_{\mathfrak{h}}\,.\end{aligned}$$

$$\frac{1}{N}\sum_{i=1}^{N}\big(\hat{\kappa}_{\mathfrak{h}}\,\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i]-\kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i]\big)^2 \;\leq\; \frac{1}{N}\sum_{i=1}^{N}\big(\kappa\mathfrak{g}_{\Omega}[\boldsymbol{x}_i]-\kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i]\big)^2+2r\kappa$$

Connecting Statistics and Computations: Stationary Points Instead of Global Optima

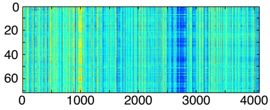
(Upcoming)

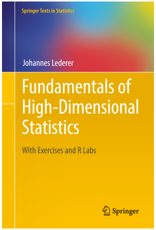
$$y_i \; = \; \underbrace{\alpha\beta}_{=:\gamma} x_i + u_i$$

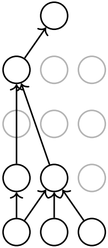
$$(\widehat{\alpha}, \widehat{\beta}) \; \in \; \operatorname*{argmin}_{\alpha,\beta\in\mathbb{R}} \left\{ \sum_{i=1}^{n} (y_i - \alpha\beta x_i)^2 \right\}$$

$$\widehat{\gamma} \; \in \; \operatorname*{argmin}_{\gamma\in\mathbb{R}} \left\{ \sum_{i=1}^{n} (y_i - \gamma x_i)^2 \right\}$$

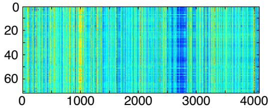
Sparsity Is a Promising Concept in Deep Learning

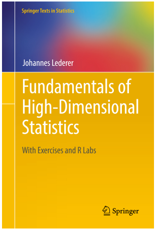


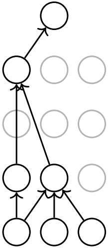




Sparsity Is a Promising Concept in Deep Learning

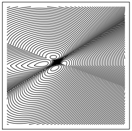




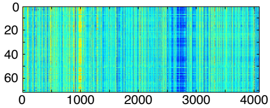


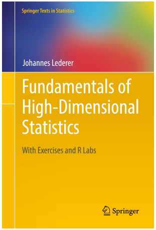
johanneslederer.com

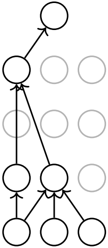
github.com/LedererLab



Sparsity Is a Promising Concept in Deep Learning

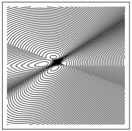




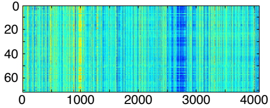


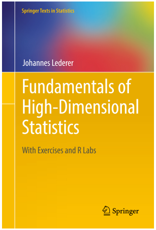
johanneslederer.com

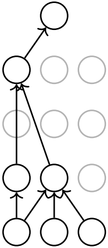
github.com/LedererLab



Sparsity Is a Promising Concept in Deep Learning

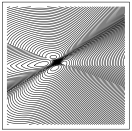


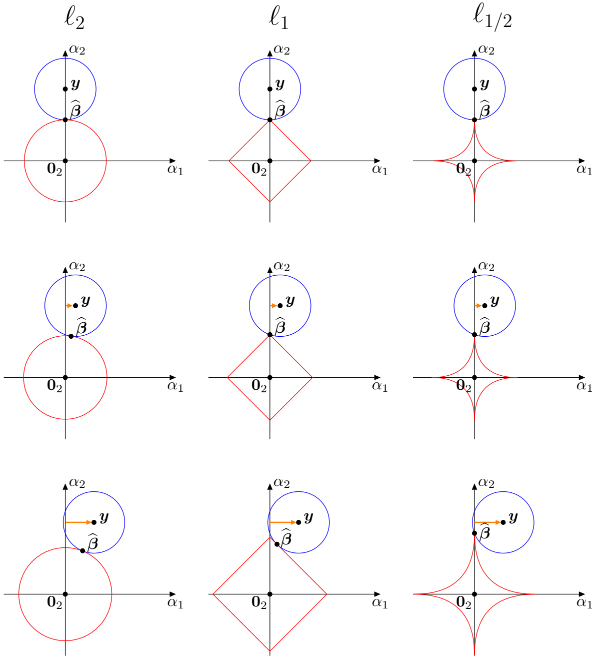




johanneslederer.com

github.com/LedererLab





ℓ_2
ℓ_1
$\ell_{1/2}$
α_2
α_1
$\boldsymbol{y}$
$\widehat{\boldsymbol{\beta}}$
$\mathbf{0}_2$

We Use Standard Measures for the Accuracy

(Statistical guarantees for regularized neural networks, 2021)

$$\operatorname{err}[\hat{\kappa}_{\mathfrak{h}}\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}] \;:=\; \sqrt{\frac{1}{N}\sum_{i=1}^{N}\Bigl(\hat{\kappa}_{\mathfrak{h}}\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}_i] - \kappa_*\mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i]\Bigr)^2}$$

$$\operatorname{risk}[\hat{\kappa}_{\mathfrak{h}}\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}] \;:=\; \mathbb{E}_{(\boldsymbol{x},y)}\Bigl[\Bigl(\hat{\kappa}_{\mathfrak{h}}\mathfrak{g}_{\widehat{\Omega}_{\mathfrak{h}}}[\boldsymbol{x}] - y\Bigr)^2\Bigr]$$

Using a Parametric Model for Illustration

(Statistical guarantees for regularized neural networks, 2021)

$$y_i \;=\; \kappa_* \mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i] + u_i$$

$u_1, \ldots, u_N$ i.i.d. $\mathcal{N}[0, \sigma^2]$

Using a Parametric Model for Illustration

(Statistical guarantees for regularized neural networks, 2021)

$$y_i \;=\; \kappa_* \mathfrak{g}_{\Omega_*}[\boldsymbol{x}_i] + u_i$$

Proposition: Assume relu activation and define for a norm $\mathfrak{h}$

$$\mathcal{A}_{\mathfrak{h}} \;:=\; \left\{\Theta \in \mathcal{A} \;:\; \mathfrak{h}[\Theta] \;\leq\; 1\right\}.$$

Then, for every $\Theta \in \mathcal{A}$, there exists a pair of $\kappa \in [0, \infty)$ and $\Omega \in \mathcal{A}_{\mathfrak{h}}$ such that

$$\mathfrak{g}_{\Theta}[\boldsymbol{x}] \;=\; \kappa \mathfrak{g}_{\Omega}[\boldsymbol{x}] \qquad \text{for all } \boldsymbol{x} \in \mathbb{R}^d\,;$$

and vice versa, for every pair of $\kappa \in [0, \infty)$ and $\Omega \in \mathcal{A}_{\mathfrak{h}}$, there exists a $\Theta \in \mathcal{A}$ such that the above equality holds.

Scale Regularization Reduces Ambiguity

(Statistical guarantees for regularized neural networks, 2021)

$$(\hat{\kappa}_{\mathfrak{h}}, \widehat{\Omega}_{\mathfrak{h}}) \;\in\; \operatorname*{argmin}_{\substack{\kappa \in [0,\infty) \\ \Omega \in \mathcal{A}_{\mathfrak{h}}}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \big(y_i - \kappa \mathfrak{g}_{\Omega}[\boldsymbol{x}_i]\big)^2 + r\kappa \right\}$$

r is an "appropriate" tuning parameter

Neural Networks Have Lipschitz Properties

(Statistical guarantees for regularized neural networks, 2021)

Proposition: It holds for every $\boldsymbol{x} \in \mathbb{R}^d$ and $\Theta = (W^L, \dots, W^0), \Gamma = (V^L, \dots, V^0) \in \mathcal{A}$ that

$$|\mathfrak{g}_\Theta[\boldsymbol{x}] - \mathfrak{g}_\Gamma[\boldsymbol{x}]| \;\leq\; c_{\mathrm{Lip}}[\boldsymbol{x}] |\!|\!|\Theta - \Gamma|\!|\!|_{\mathrm{F}}$$

with

$$c_{\mathrm{Lip}}[\boldsymbol{x}] \;:=\; 2(a_{\mathrm{Lip}})^L \sqrt{L} \|\boldsymbol{x}\|_2 \max_{l \in \{0,\dots,L\}} \prod_{j \in \{0,\dots,L\}, j \neq l} (\|W^j\|_2 \vee \|V^j\|_2)\,.$$

Neural Networks Have Lipschitz Properties

(Statistical guarantees for regularized neural networks, 2021)

$$
\begin{aligned}
S_l\mathfrak{g}_\Theta \;:\; \mathbb{R}^d &\to \mathbb{R}^{p_l} \\
\boldsymbol{x} &\mapsto S_l\mathfrak{g}_\Theta[\boldsymbol{x}] := \mathfrak{f}^l\big[\ldots W^1\mathfrak{f}^1[W^0\boldsymbol{x}]\big]
\end{aligned}
$$

$$
\|S_{l-1}\mathfrak{g}_\Theta[\boldsymbol{x}]\|_2 \;\le\; (a_{\mathrm{Lip}})^{l-1}\|\boldsymbol{x}\|_2\prod_{j=0}^{l-2}\|W^j\|_2
$$

$$
\begin{aligned}
S^l\mathfrak{g}_\Theta \;:\; \mathbb{R}^{p_l} &\to \mathbb{R} \\
\boldsymbol{z} &\mapsto S^l\mathfrak{g}_\Theta[\boldsymbol{z}] := W^L\mathfrak{f}^L\big[\ldots W^l\mathfrak{f}^l[W^{l-1}\boldsymbol{z}]\big]
\end{aligned}
$$

$$
|S^{l+1}\mathfrak{g}_\Theta[\boldsymbol{z}_1] - S^{l+1}\mathfrak{g}_\Theta[\boldsymbol{z}_2]| \;\le\; (a_{\mathrm{Lip}})^{L-l}\|\boldsymbol{z}_1-\boldsymbol{z}_2\|_2\prod_{j=l}^{L}\|W^j\|_2
$$

The Relevant Dudley Integrals Are Bounded

(Statistical guarantees for regularized neural networks, 2021)

$\mathcal{G}_1 \ := \ \{\mathfrak{g}_\Omega : \Omega \in \mathcal{A}_{\|\cdot\|_1}\}$

$\|\mathfrak{g}_\Theta - \mathfrak{g}_\Gamma\|_N \ := \ \sqrt{\sum_{i=1}^N \big(\mathfrak{g}_\Theta[\boldsymbol{x}_i] - \mathfrak{g}_\Gamma[\boldsymbol{x}_i]\big)^2/N}$

H usual metric entropy and J corresponding $\delta/(8\sigma)$-Dudley integral

Proposition: It holds for every $v \in (0,\infty)$ and $\delta, \sigma \in (0,\infty)$ that satisfy $\delta \leq 8\sigma c_{\mathrm{Lip1}}$ that

$$H\big[v, \mathcal{G}_1, \|\cdot\|_N\big] \ \leq \ \frac{6(c_{\mathrm{Lip1}})^2}{v^2}\log\left[\frac{ePv^2}{(c_{\mathrm{Lip1}})^2} \vee 2e\right]$$

and

$$J\big[v, \mathcal{G}_1, \|\cdot\|_N\big] \ \leq \ \frac{5c_{\mathrm{Lip1}}}{2}\sqrt{\log\left[eP \vee 2e\right]}\log\left[\frac{8\sigma c_{\mathrm{Lip1}}}{\delta}\right],$$

where $c_{\mathrm{Lip1}} := 2(2a_{\mathrm{Lip}}/L)^L\sqrt{L}\|\boldsymbol{x}\|_N$.

The Relevant Dudley Integrals Are Bounded

(Statistical guarantees for regularized neural networks, 2021)

Step 1: $H[v, \mathcal{G}_1, \|\cdot\|_N]$ to $H[v/c_{\mathrm{Lip1}}, \mathcal{A}_1, |||\cdot|||_{\mathrm{F}}]$ via Lipschitzness

Step 2: "Sparse covering"

Step 3: Approximations in the Dudley integral

Theory Exists: Oracle Inequalities

(Oracle inequalities for high-dimensional prediction, 2019)

Theorem (Power-One Bound): As long as $r \geq 2\|X^\top u\|_\infty$, it holds that

$$\frac{\|X\Theta_* - X\widehat{\omega}_{\text{lasso}}\|_2^2}{n} \;\leq\; 2\|\Theta_*\|_1 \,\frac{r}{n}\,.$$

Theory Exists: Oracle Inequalities

(Oracle inequalities for high-dimensional prediction, 2019)

$$\|\boldsymbol{y} - X\widehat{\omega}_{\text{lasso}}\|_2^2 + r\|\widehat{\omega}_{\text{lasso}}\|_1 \ \leq \ \|\boldsymbol{y} - X\Theta_*\|_2^2 + r\|\Theta_*\|_1$$

$$\Rightarrow \quad \|X\Theta_* - X\widehat{\omega}_{\text{lasso}}\|_2^2 + 2\langle X^\top u, \Theta_* - \widehat{\omega}_{\text{lasso}}\rangle + r\|\widehat{\omega}_{\text{lasso}}\|_1 \ \leq \ r\|\Theta_*\|_1$$

$$\begin{aligned}\Rightarrow \quad \|X\Theta_* - X\widehat{\omega}_{\text{lasso}}\|_2^2 \ \leq \ & 2\|X^\top u\|_\infty\|\Theta_*\|_1 + 2\|X^\top u\|_\infty\|\widehat{\omega}_{\text{lasso}}\|_1 \\ & + r\|\Theta_*\|_1 - r\|\widehat{\omega}_{\text{lasso}}\|_1\end{aligned}$$

$$\Rightarrow \quad \|X\Theta_* - X\widehat{\omega}_{\text{lasso}}\|_2^2 \ \leq \ 2r\|\Theta_*\|_1$$

Theory Exists: Concentration Inequalities

(Fundamentals of High-Dimensional Statistics—With Exercises and R Labs, 2021)

Theorem (Effective Noise): Consider a fixed design matrix $X \in \mathbb{R}^{n\times p}$ with $\max_j (X^\top X)_{jj}/n = 1$, and consider Gauss-distributed noise $u \sim \mathcal{N}_n[\mathbf{0}_n, \sigma^2\, \mathrm{I}_{n\times n}]$. Then, for all $t \in (0,1)$, it holds that

$$\mathbb{P}\Big\{2\|X^\top u\|_\infty/n \,\leq\, \sigma\sqrt{8\log[p/t]/n}\Big\} \;\geq\; 1-t\,.$$

Theory Exists: Concentration Inequalities

(Fundamentals of High-Dimensional Statistics—With Exercises and R Labs, 2021)

$$\mathbb{P}\{2\|X^\top u\|_\infty > r\} \leq \sum_{j=1}^{p} \mathbb{P}\{2|(X^\top u)_j| > r\}.$$

$$\mathbb{P}\{2\|X^\top u\|_\infty > r\} \leq p \max_{j\in\{1,\dots,p\}} \mathbb{P}\left\{\frac{|(X^\top u)_j|}{\sigma\sqrt{(X^\top X)_{jj}}} > \frac{r}{2\sigma\sqrt{n}}\right\}$$

$$\mathbb{P}\{|z| \geq a\} \leq e^{-\frac{a^2}{2}} \qquad \text{for all } a \in [0,\infty)$$