Weierstraß-Institut für
Angewandte Analysis und Stochastik

High dimensional logistic regression

Vladimir Spokoiny ,

HSE University and IITP RAS Moscow,

WIAS and HU Berlin

ICOMP 2025

Observed $(Y_i, \boldsymbol{\Psi}_i)$ with binary outputs (labels) Y_i and the corresponding feature vectors $\boldsymbol{\Psi}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$.

Aim: predict (estimate) $f(\boldsymbol{\psi}) = \mathbb{E}(Y \,\big|\, \boldsymbol{\psi})$.

Logistic regression (a linear model for the canonical parameter):

$$f(\boldsymbol{\psi}) = \phi'(\langle \boldsymbol{\psi}, \boldsymbol{\upsilon} \rangle), \quad \phi(t) = \log(1 + \mathrm{e}^t), \quad \phi'(t) = \frac{\mathrm{e}^t}{1 + \mathrm{e}^t}\,.$$

MLE:

$$\widetilde{\boldsymbol{\upsilon}} = \operatorname*{argmin}_{\boldsymbol{\upsilon} \in \mathbb{R}^p} L(\boldsymbol{\upsilon}) = \operatorname*{argmin}_{\boldsymbol{\upsilon} \in \mathbb{R}^p} \sum_{i=1}^{n} \bigl\{ \phi(\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon} \rangle) - Y_i \langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon} \rangle \bigr\}\,.$$

In the limit of large samples in which p is fixed and $n \to \infty$, the MLE $\widetilde{\boldsymbol{\upsilon}}$ obeys

$$\sqrt{n}\big(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}\big) \xrightarrow{w} \mathcal{N}(0, F_{\boldsymbol{\upsilon}}) \,,$$

where $F_{\boldsymbol{\upsilon}}$ is the Fisher information matrix at $\boldsymbol{\upsilon}$

$$F_{\boldsymbol{\upsilon}} \stackrel{\mathrm{def}}{=} \mathbb{E}\big\{\boldsymbol{\Psi}\boldsymbol{\Psi}^{\top}\phi''(\langle\boldsymbol{\Psi}, \boldsymbol{\upsilon}\rangle)\big\} \,, \qquad \phi(t) = \frac{\mathrm{e}^t}{(1+\mathrm{e}^t)^2} \,.$$

see e.g.

– Lehmann and Romano. Testing statistical hypotheses. 2006.

– McCullagh and Nelder. Generalized linear models. 1989.

– Van der Vaart. Asymptotic statistics. 2000.

- [Sur and Candès, 2019] A modern maximum-likelihood theory for high-dimensional logistic regression,
- [Candès and Sur, 2020] The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression
- [Montanari et al., 2025] The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime
- [Kuchelmeister and van de Geer, 2024] Finite Sample Rates for Logistic Regression with Small Noise or Few Samples
- [Chardon et al., 2024] Finite-sample performance of the maximum likelihood estimator in logistic regression
- [Bach, 2024] High-Dimensional Analysis of Double Descent for Linear Regression with Random Projections

Main issues:

- high dimension $p \gg n$; need of regularization;
- random and ill-posed design $\boldsymbol{\Psi}\boldsymbol{\Psi}^\top$;
- finite-sample and dimension-free accuracy guarantees.

Tools:

- Perturbed optimization;
- random matrix theory;
- deviation bounds for fourth order random tensors;
- regularization for nonlinear inverse problems.

Let $L(\boldsymbol{\upsilon})$ be a random function (loss, empirical risk, negative log-likelihood). Consider

$$\widetilde{\boldsymbol{\upsilon}} = \operatorname*{argmin}_{\boldsymbol{\upsilon}} L(\boldsymbol{\upsilon}); \quad \boldsymbol{\upsilon}^* = \operatorname*{argmin}_{\boldsymbol{\upsilon}} \mathbb{E}\, L(\boldsymbol{\upsilon});$$

Includes MLE, LSE, LAD, Minimum Contrast, and many others procedures.

Aim: describe $\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*$ (estimation loss) and $L(\widetilde{\boldsymbol{\upsilon}}) - L(\boldsymbol{\upsilon}^*)$ (excess).

Perturbed optimization: $L(\boldsymbol{\upsilon})$ is a perturbation of $f(\boldsymbol{\upsilon}) = \mathbb{E}L(\boldsymbol{\upsilon})$ by a stochastic component

$$\zeta(\boldsymbol{\upsilon}) = L(\boldsymbol{\upsilon}) - \mathbb{E}L(\boldsymbol{\upsilon}).$$

Consider a penalized MLE for a penalty function $\mathrm{pen}_G(\boldsymbol{\upsilon})$

$$\widetilde{\boldsymbol{\upsilon}}_G = \operatorname*{argmin}_{\boldsymbol{\upsilon}} L_G(\boldsymbol{\upsilon}) = \operatorname*{argmin}_{\boldsymbol{\upsilon}} \bigl\{ L(\boldsymbol{\upsilon}) + \mathrm{pen}_G(\boldsymbol{\upsilon}) \bigr\} .$$

Typical examples: $\mathrm{pen}_G(\boldsymbol{\upsilon}) = \|G\boldsymbol{\upsilon}\|^2/2$ and $\mathrm{pen}_\lambda(\boldsymbol{\upsilon}) = \lambda\|\boldsymbol{\upsilon}\|_1$.

Consider three optimization problems

$$\widetilde{\boldsymbol{\upsilon}}_G = \operatorname*{argmin}_{\boldsymbol{\upsilon}} L_G(\boldsymbol{\upsilon}),$$

$$\boldsymbol{\upsilon}^*_G = \operatorname*{argmin}_{\boldsymbol{\upsilon}} \mathbb{E} L_G(\boldsymbol{\upsilon}),$$

$$\boldsymbol{\upsilon}^* = \operatorname*{argmin}_{\boldsymbol{\upsilon}} \mathbb{E} L(\boldsymbol{\upsilon}).$$

The penalized MLE $\widetilde{\boldsymbol{\upsilon}}_G$ estimates rather $\boldsymbol{\upsilon}^*_G$ than $\boldsymbol{\upsilon}^*$.

Let $f(\boldsymbol{\upsilon})$ be a smooth convex function,

$$\boldsymbol{\upsilon}^* = \operatorname*{argmin}_{\boldsymbol{\upsilon}} f(\boldsymbol{\upsilon}), \quad \mathbb{F} = \nabla^2 f(\boldsymbol{\upsilon}^*).$$

Let another function $g(\boldsymbol{\upsilon})$ satisfy for some vector $\boldsymbol{A}$

$$g(\boldsymbol{\upsilon}) - g(\boldsymbol{\upsilon}^*) = \langle \boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*, \boldsymbol{A} \rangle + f(\boldsymbol{\upsilon}) - f(\boldsymbol{\upsilon}^*).$$

Define

$$\boldsymbol{\upsilon}^\circ \stackrel{\mathrm{def}}{=} \operatorname*{argmin}_{\boldsymbol{\upsilon}} g(\boldsymbol{\upsilon}), \qquad g(\boldsymbol{\upsilon}^\circ) = \min_{\boldsymbol{\upsilon}} g(\boldsymbol{\upsilon}).$$

Aim: evaluate the quantities $\boldsymbol{\upsilon}^\circ - \boldsymbol{\upsilon}^*$ and $g(\boldsymbol{\upsilon}^\circ) - g(\boldsymbol{\upsilon}^*)$.

Consider $\boldsymbol{v}^* \stackrel{\text{def}}{=} \operatorname{argmin} f(\boldsymbol{v})\,, \boldsymbol{v}^\circ \stackrel{\text{def}}{=} \operatorname{argmin}\{f(\boldsymbol{v}) + \langle \boldsymbol{v}, \boldsymbol{A} \rangle\}.$

$(\boldsymbol{\mathcal{T}_3^*})$ *$f(\boldsymbol{v})$ is strongly convex, $D^2 \le \mathbb{F} \stackrel{\text{def}}{=} \nabla^2 f(\boldsymbol{v}^*)$, and*

$$\sup_{\boldsymbol{u}:\, \|D\boldsymbol{u}\| \le \mathtt{r}} \sup_{\boldsymbol{z} \in \mathbb{R}^p} \frac{\bigl|\langle \nabla^3 f(\boldsymbol{v}^* + \boldsymbol{u}), \boldsymbol{z}^{\otimes k} \rangle\bigr|}{\|D\boldsymbol{z}\|^3} \le \tau_3\,.$$

Proposition ([Spokoiny, 2025b, Spokoiny, 2025a])

Assume $(\boldsymbol{\mathcal{T}_3^*})$ at $\boldsymbol{v}^*$ with D^2, $\mathtt{r}$, and τ_3 s.t.

$$\mathtt{r} \ge \frac{3}{2}\|D\,\mathbb{F}^{-1}\boldsymbol{A}\|\,, \quad \tau_3 \|D\,\mathbb{F}^{-1}\boldsymbol{A}\| < \frac{4}{9}\,.$$

Then $\|D(\boldsymbol{v}^\circ - \boldsymbol{v}^*)\| \le (3/2)\|D\,\mathbb{F}^{-1}\boldsymbol{A}\|$ and moreover,

$$\|D^{-1}\mathbb{F}(\boldsymbol{v}^\circ - \boldsymbol{v}^* - \mathbb{F}^{-1}\boldsymbol{A})\| \le \frac{3}{4}\,\tau_3 \|D\,\mathbb{F}^{-1}\boldsymbol{A}\|^2\,.$$

Model:

$$H = \mathbb{E}\,\Psi\Psi^\top\,, \qquad H_n = \frac{1}{n}\sum_{i=1}^{n}\Psi_i\Psi_i^\top\,.$$

Whenever Ψ is subgaussian in all directions, and

$$n \geq K(d + \log(1/\delta)),$$

with probability at least $1-\delta$, it holds

$$\frac{1}{2}\|\boldsymbol{u}\|_H^2 \leq \|\boldsymbol{u}\|_{H_n}^2 \leq 2\|\boldsymbol{u}\|_H^2, \qquad \forall \boldsymbol{u} \in \mathbb{R}^p\,.$$

[Koltchinskii and Lounici, 2017], [Vershynin, 2018], [Tropp, 2015].

Let $\boldsymbol{\Psi}, \boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, \ldots$ be i.i.d. in $\mathbb{R}^p$ and satisfy

$$n \log \mathbb{E} \exp\langle \boldsymbol{\Psi}, \boldsymbol{u} \rangle \leq \frac{\mathtt{C}_\Psi}{2} \|\mathbb{D}\boldsymbol{u}\|^2 , \qquad \boldsymbol{u} \in \mathbb{R}^p ,$$

$$\operatorname{tr}(\mathbb{B}_4) \ll n^{1/2} ,$$

where $\mathbb{B}_4 \stackrel{\text{def}}{=} n\mathbb{D}^{-1}\mathbb{E}\big\{ \big|\phi^{(4)}(\langle \boldsymbol{\Psi}, \boldsymbol{\upsilon}_\lambda^* \rangle)\big|^{1/2} \boldsymbol{\Psi}\boldsymbol{\Psi}^\top \big\} \mathbb{D}^{-1}$.
Then on a set Ω_n with $\mathbb{P}\big(\Omega_n\big) \geq 1 - 1/n$

$$\sup_{\|\mathbb{D}\boldsymbol{z}\|=1} \left| \sum_{i=1}^n \phi^{(4)}(\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon}_\lambda^* \rangle) \langle \boldsymbol{\Psi}_i, \boldsymbol{z} \rangle^4 \right| \leq \frac{\mathtt{C}_4 + \varepsilon_4}{n} ,$$

where

$$\varepsilon_4 \lesssim \sqrt{\frac{\operatorname{tr}(\mathbb{B}_4)}{n}} + \frac{\operatorname{tr}^2(\mathbb{B}_4)}{n} + \sqrt{\frac{\log n}{n}} + \frac{\log^2 n}{n} .$$

[Zhivotovskiy, 2024] and [Al-Ghattas et al., 2025].

Penalized MLE:

$$\widetilde{\boldsymbol{\upsilon}}_G = \operatorname*{argmin}_{\boldsymbol{\upsilon}} L_G(\boldsymbol{\upsilon}) = \operatorname*{argmin}_{\boldsymbol{\upsilon}} \bigl\{ L(\boldsymbol{\upsilon}) + \mathrm{pen}_G(\boldsymbol{\upsilon}) \bigr\} .$$

An important example of penalty choice is a ridge penalty $G^2 = g^2 \mathbb{I}_p$. It is basis and coordinate-free and enforces the "benign overfitting" phenomenon in high-dimensional regression; see [Bartlett et al., 2020], [Cheng and Montanari, 2022], [Noskov et al., 2025] and references therein.

A phase transition effect: if the operator $\mathbb{F} = \nabla^2 L(\boldsymbol{\upsilon}^*)$ is smoother than the signal then one can achieve optimal accuracy.

Does not apply if the operator $\mathbb{F}$ is not smoother than the signal.

Observed $(Y_i, \boldsymbol{\Psi}_i)$ with binary labels Y_i and features $\boldsymbol{\Psi}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$.

Logistic regression: with $\phi(t) = \log(1 + \mathrm{e}^t)$,

$$\mathbb{E}(Y \,|\, \boldsymbol{\psi}) = f(\boldsymbol{\psi}) = \phi'(\langle \boldsymbol{\psi}, \boldsymbol{\upsilon} \rangle) .$$

MLE:

$$\widetilde{\boldsymbol{\upsilon}} = \operatorname*{argmin}_{\boldsymbol{\upsilon} \in \mathbb{R}^p} L(\boldsymbol{\upsilon}) = \operatorname*{argmin}_{\boldsymbol{\upsilon} \in \mathbb{R}^p} \sum_{i=1}^{n} \left\{ \phi(\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon} \rangle) - Y_i \langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon} \rangle \right\} .$$

Penalized MLE: for the quadratic penalty $\|G\boldsymbol{\upsilon}\|^2/2$:

$$\widetilde{\boldsymbol{\upsilon}}_G = \operatorname*{argmin}_{\boldsymbol{\upsilon} \in \mathbb{R}^p} L_G(\boldsymbol{\upsilon}) = \operatorname*{argmin}_{\boldsymbol{\upsilon} \in \mathbb{R}^p} L(\boldsymbol{\upsilon}) + \frac{1}{2} \|G\boldsymbol{\upsilon}\|^2 .$$

A penalized MLE $\widetilde{\boldsymbol{\upsilon}}_G$:

$$\widetilde{\boldsymbol{\upsilon}}_G = \operatorname*{argmin}_{\boldsymbol{\upsilon} \in \mathbb{R}^p} L_G(\boldsymbol{\upsilon}).$$

The truth $\boldsymbol{\upsilon}^*$:

$$\boldsymbol{\upsilon}^* = \operatorname*{argmin}_{\boldsymbol{\upsilon} \in \mathbb{R}^p} \mathbb{E} L(\boldsymbol{\upsilon}),$$

The penalized truth:

$$\boldsymbol{\upsilon}_G^* = \operatorname*{argmin}_{\boldsymbol{\upsilon} \in \mathbb{R}^p} \mathbb{E} L_G(\boldsymbol{\upsilon}).$$

The study is based on the decomposition

$$\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}^* = \underbrace{\widetilde{\boldsymbol{\upsilon}}_G - \boldsymbol{\upsilon}_G^*}_{\text{stochastic term}} + \underbrace{\boldsymbol{\upsilon}_G^* - \boldsymbol{\upsilon}^*}_{\text{bias}} .$$

The Fisher matrix $\mathbb{F}(\boldsymbol{\upsilon})$ at $\boldsymbol{\upsilon}$ is given by

$$\mathbb{F}(\boldsymbol{\upsilon}) = \nabla^2 \mathbb{E} L(\boldsymbol{\upsilon}) = \sum_{i=1}^{n} \phi''\bigl(\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon} \rangle\bigr) \boldsymbol{\Psi}_i \boldsymbol{\Psi}_i^\top .$$

We also write

$$\mathbb{F}_G(\boldsymbol{\upsilon}) = \mathbb{F}(\boldsymbol{\upsilon}) + G^2, \qquad \mathbb{F}_G = \mathbb{F}_G(\boldsymbol{\upsilon}_G^*) = \mathbb{F}(\boldsymbol{\upsilon}_G^*) + G^2.$$

Alternatively one can define $\mathbb{F}_G = \mathbb{F}_G(\boldsymbol{\upsilon}^*)$.

Our conditions rely on a metric tensor D which defines a local vicinity of $\boldsymbol{\upsilon}_G^*$. We assume $\mathbb{F} \le D^2 \le \mathbb{F}_G$ and $D^2 = \mathbb{F}_G$ is a default choice.

For checking the conditions, we need some regularity of the design $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_n$.

$(\boldsymbol{\Psi})$ *With a metric tensor* D *satisfying*

$$\mathbb{F}(\boldsymbol{\upsilon}_G^*) \leq D^2 \leq \mathbb{F}_G(\boldsymbol{\upsilon}_G^*) \text{ , for some } \delta, \delta_0 > 0$$

$$\sup_{\boldsymbol{z} \in \mathbb{R}^p} \frac{1}{\|D\boldsymbol{z}\|^4} \left| \sum_{i=1}^n \langle \boldsymbol{\Psi}_i, \boldsymbol{z} \rangle^4 \ \phi^{(4)}\left(\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon}_G^* \rangle \right) \right| \leq \delta^2 \,,$$

$$\max_{i \leq n} \| D^{-1} \boldsymbol{\Psi}_i \| \leq \delta_0 \,.$$

Even for a Gaussian regression, dimension-free guarantees are quite involved; e.g. [Bartlett et al., 2020], [Cheng and Montanari, 2022], [Noskov et al., 2025] and references therein.

As in these papers, consider a ridge (Tikhonov) penalization $\lambda\|\boldsymbol{\upsilon}\|^2/2$.

Let observations (labels) Y_i follow logistic regression with parameter $\boldsymbol{\upsilon}^*$, that is, the pairs $(Y_i, \boldsymbol{\Psi}_i)$ are i.i.d. with

$$\mathbb{P}(Y_i = 1 \,\big|\, \boldsymbol{\Psi}_i) = \sigma(\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon}^* \rangle)\,, \qquad \sigma(t) = \phi'(t) = \mathrm{e}^t/(1+\mathrm{e}^t)\,.$$

We assume $\boldsymbol{\Psi}, \boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, \ldots \in \mathbb{R}^p$ to be i.i.d. zero mean random vectors drawn from some distribution, e.g. $\boldsymbol{\Psi} \sim \mathcal{N}(0, \Sigma)$.

The central object of our study is the empirical Fisher information

$$\mathbb{F}_\lambda(\boldsymbol{\upsilon}) = \sum_{i=1}^n \phi''\big(\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon} \rangle\big)\, \boldsymbol{\Psi}_i\, \boldsymbol{\Psi}_i^\top + \lambda \mathbb{I}_p\,,$$

which is now random.

If the dimension p becomes larger than the sample size n, The properties of this random matrix become crucial; see e.g. [Bartlett et al., 2020], [Cheng and Montanari, 2022], [Sur and Candès, 2019], [Kuchelmeister and van de Geer, 2024], [Bach, 2024].

In our results, we fix a metric tensor $\mathbb{D}$. A canonical choice is

$$\mathbb{D}^2 = \mathbb{D}_\lambda^2 \stackrel{\mathrm{def}}{=} n\mathbb{E}\{\boldsymbol{\Psi}\boldsymbol{\Psi}^\top \phi''(\langle \boldsymbol{\Psi}, \boldsymbol{v}_\lambda^* \rangle)\} + \lambda \mathbb{I}_p \,.$$

We impose a sub-Gaussian condition on $\boldsymbol{\Psi}$: for some fixed constant $\mathtt{C}_\Psi$

$$n \log \exp\langle \boldsymbol{\Psi}, \boldsymbol{u} \rangle \le \frac{\mathtt{C}_\Psi}{2} \|\mathbb{D}_\lambda \boldsymbol{u}\|^2 \,, \qquad \boldsymbol{u} \in \mathbb{R}^p \,. \tag{1}$$

The estimator $\widetilde{\boldsymbol{\upsilon}}_\lambda$ solves the empirical problem

$$\widetilde{\boldsymbol{\upsilon}}_\lambda = \operatorname*{argmin}_{\boldsymbol{\upsilon}\in\mathbb{R}^p} L_\lambda(\boldsymbol{\upsilon}) = \operatorname*{argmin}_{\boldsymbol{\upsilon}\in\mathbb{R}^p} \left[-\sum_{i=1}^{n} \left\{ Y_i \left\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon} \right\rangle - \phi\big(\left\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon} \right\rangle\big) \right\} + \frac{\lambda}{2} \|\boldsymbol{\upsilon}\|^2 \right]$$

The whole analysis is based on the decomposition

$$\widetilde{\boldsymbol{\upsilon}}_\lambda - \boldsymbol{\upsilon}^* = (\widetilde{\boldsymbol{\upsilon}}_\lambda - \boldsymbol{\upsilon}_\lambda^*) + (\boldsymbol{\upsilon}_\lambda^* - \boldsymbol{\upsilon}^*) \tag{3}$$

for the penalized truth

$$\boldsymbol{\upsilon}_\lambda^* = \operatorname*{argmin}_{\boldsymbol{\upsilon}\in\mathbb{R}^p} \left[-\mathbb{E} \sum_{i=1}^{n} \left\{ \mathbb{E}(Y_i \,|\, \boldsymbol{\Psi}_i) \left\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon} \right\rangle - \phi\big(\left\langle \boldsymbol{\Psi}_i, \boldsymbol{\upsilon} \right\rangle\big) \right\} + \frac{\lambda}{2} \|\boldsymbol{\upsilon}\|^2 \right].$$

The first term $\widetilde{\boldsymbol{\upsilon}}_\lambda - \boldsymbol{\upsilon}_\lambda^*$ in (3) describes the stochastic error caused by using the noisy data Y_i, $i \le n$. The bias term $\boldsymbol{\upsilon}_\lambda^* - \boldsymbol{\upsilon}^*$ is due to the ridge penalty $\lambda\|\boldsymbol{\upsilon}\|^2/2$ in (2).

Later, we evaluate each difference using general results on perturbed optimization. Given λ, define

$$\bar{I\!F}_\lambda(\boldsymbol{\upsilon}) \stackrel{\mathrm{def}}{=} n\mathbb{E}\{\phi''(\langle \boldsymbol{\Psi}, \boldsymbol{\upsilon} \rangle)\, \boldsymbol{\Psi}\boldsymbol{\Psi}^\top\} + \lambda I\!I_p\,.$$

Also, set $\bar{I\!F}_\lambda \stackrel{\mathrm{def}}{=} \bar{I\!F}_\lambda(\boldsymbol{\upsilon}^*_\lambda)$ and define

$$\boldsymbol{S} \stackrel{\mathrm{def}}{=} -\sum_{i=1}^n (Y_i - \mathbb{E}Y_i)\boldsymbol{\Psi}_i\,.$$

We show that

$$\widetilde{\boldsymbol{\upsilon}}_\lambda - \boldsymbol{\upsilon}^*_\lambda \approx -\bar{I\!F}_\lambda^{-1}\boldsymbol{S}\,, \qquad \boldsymbol{\upsilon}^*_\lambda - \boldsymbol{\upsilon}^* \approx -\lambda\bar{I\!F}_\lambda^{-1}\boldsymbol{\upsilon}^*\,,$$

and provide some finite sample accuracy guarantees.

With $\mathbb{D}_\lambda^2 \stackrel{\mathrm{def}}{=} \bar{\mathbb{F}}_\lambda$, define

$$\mathbb{B}_2 \stackrel{\mathrm{def}}{=} n\mathbb{D}_\lambda^{-1}\mathbb{E}\{\phi''(\langle \boldsymbol{\Psi}, \boldsymbol{\upsilon}_\lambda^* \rangle)\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\}\mathbb{D}_\lambda^{-1}.$$

Proposition

Let the design distribution of the $\boldsymbol{\Psi}_i$'s satisfy (1). Assume $\operatorname{tr}\mathbb{B}_2 \ll n$. Then on a random set Ω_n with $\mathbb{P}(\Omega_n) \geq 1 - 1/n$, it holds

$$\bar{\mathbb{F}}_\lambda(1-\varepsilon_\lambda) \leq \mathbb{F}_\lambda \leq \bar{\mathbb{F}}_\lambda(1+\varepsilon_\lambda),$$

where

$$\varepsilon_\lambda \lesssim \sqrt{\frac{\mathtt{c}_\Psi \operatorname{tr}\mathbb{B}_2}{n}} + \sqrt{\frac{\log(n+p)}{n}} \ll 1.$$

Theorem

With $\xi_\lambda \stackrel{\mathrm{def}}{=} \|\mathbb{D}_\lambda^{-1}\boldsymbol{S}\|$, $b_\lambda \stackrel{\mathrm{def}}{=} \|\mathbb{D}_\lambda(\boldsymbol{\upsilon}_\lambda^* - \boldsymbol{\upsilon}^*)\|$, *and* $\tau_3 = \mathsf{c}_3 n^{-1/2}$ *for a constant* c_3 *depending on* c_Ψ *only, it holds on* Ω_n

$$\|\xi_\lambda\| \le \sqrt{\operatorname{tr} \mathbb{B}_2} + 2\sqrt{\log n}\,.$$

Furthermore, $b_\lambda \le \frac{3}{2}\|\lambda \bar{\mathbb{F}}_\lambda^{-1/2}\boldsymbol{\upsilon}^*\|$, *and*

$$\begin{aligned}&\left\|\mathbb{D}_\lambda(\widetilde{\boldsymbol{\upsilon}}_\lambda - \boldsymbol{\upsilon}^*) + \mathbb{D}_\lambda^{-1}\boldsymbol{S} + \lambda\bar{\mathbb{F}}_\lambda^{-1}\boldsymbol{\upsilon}^*\right\| \\ &\quad\le \frac{\varepsilon_\lambda}{1-\varepsilon_\lambda}\left(\xi_\lambda + b_\lambda\right) + \frac{\tau_3}{(1-\varepsilon_\lambda)^3}\left(\xi_\lambda^2 + \frac{3}{4}\|\lambda\bar{\mathbb{F}}_\lambda^{-1/2}\boldsymbol{\upsilon}^*\|^2\right).\end{aligned}$$

Distribution of the leading term and the remainder

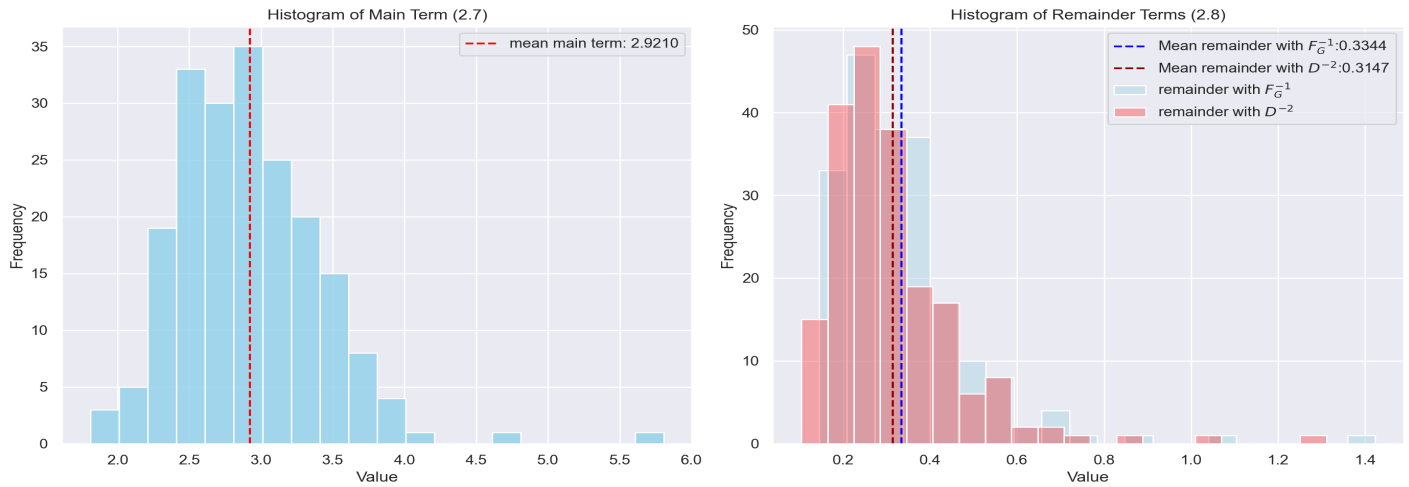

Abbildung: Distribution of the leading term and the remainder for $n = 100$.

Left: the norm of the leading term $\|\mathbb{F}_G^{-1}\nabla\zeta\|_\infty$ and $\|D^{-2}\nabla\zeta\|_\infty$.
Right: the errors $\|\widetilde{\boldsymbol{v}}_G - \boldsymbol{v}^* + \mathbb{F}_G^{-1}\nabla\zeta\|_\infty$ and
$\|\widetilde{\boldsymbol{v}}_G - \boldsymbol{v}^* + D^{-2}\nabla\zeta\|_\infty$ for $n = 100$.
The magnitude of the remainder is much smaller than the leading term, thus indicating a very good quality of the provided expansions.

Results for $n \in \{100, 200, 500, 1000, 2000\}$

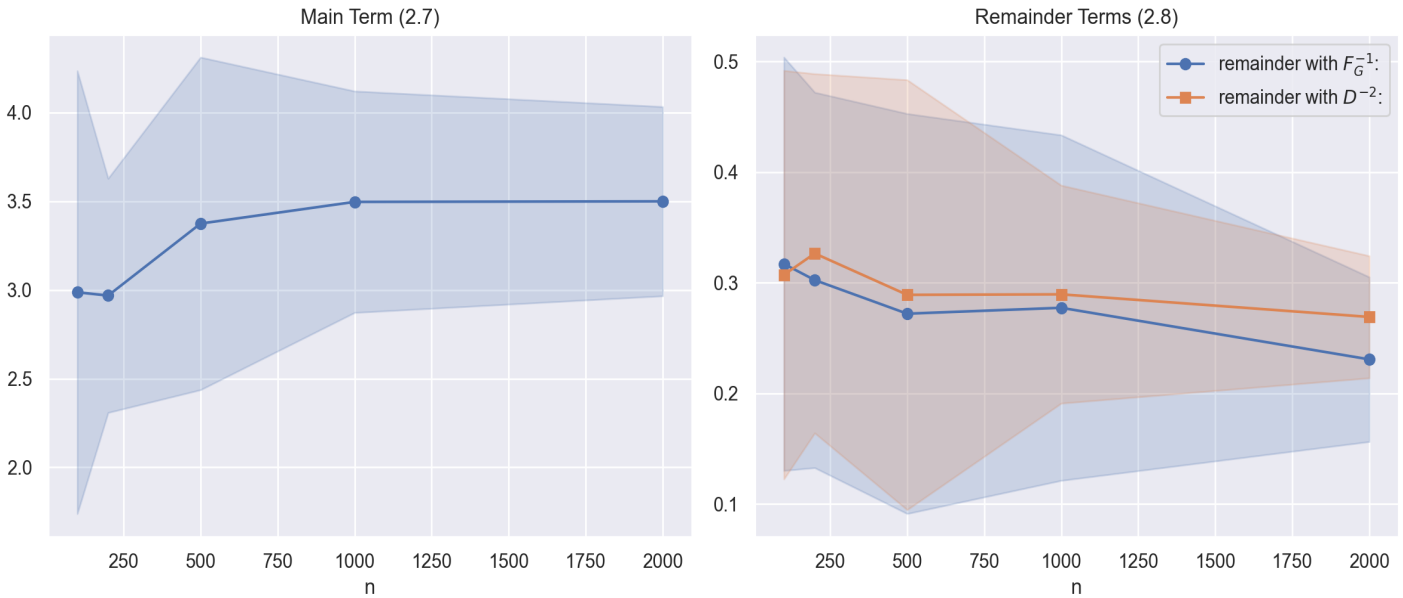

Abbildung: Comparison of the leading term and the remainder for different n.

- (Parametric) lower bounds.
- LASSO-type procedures. Issue $\|\boldsymbol{v}\|_1$ is non-smooth.
  Idea: replace by a smooth penalty ensuring $\mathbb{p} \asymp \#(\mathrm{active\,set})$.
- Other models including
  - dimension reduction, manifold learning
  - density estimation
  - error-in-operator models
  - matrix completion
  - inference for stochastic processes
  - statistical nonlinear inverse problems, PDEs
  - ...
- adaptation, parameter/penalty choice
- higher-order expansions, estimation of smooth functionals.

References

Al-Ghattas, O., Chen, J., and Sanz-Alonso, D. (2025).
Sharp concentration of simple random tensors.
arxiv.org/abs/2502.16916.

Bach, F. (2024).
High-dimensional analysis of double descent for linear regression with random projections.
SIAM Journal on Mathematics of Data Science, 6(1):26–50.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020).
Benign overfitting in linear regression.
Proceedings of the National Academy of Sciences, 117(48):30063–30070.

Candès, E. J. and Sur, P. (2020).
The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression.
The Annals of Statistics, 48(1):27–42.

Chardon, H., Lerasle, M., and Mourtada, J. (2024).
Finite-sample performance of the maximum likelihood estimator in logistic regression.

Cheng, C. and Montanari, A. (2022).
Dimension free ridge regression.
https://arxiv.org/abs/2210.08571.

Koltchinskii, V. and Lounici, K. (2017).
Concentration inequalities and moment bounds for sample covariance operators.
Bernoulli, 23(1):110 – 133.

References

Kuchelmeister, F. and van de Geer, S. (2024).
Finite sample rates for logistic regression with small noise or few samples.
Sankhya A.

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. (2025).
The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime.
The Annals of Statistics, 53(2):822 – 853.

Noskov, F., Puchkin, N., and Spokoiny, V. (2025).
Dimension-free bounds in high-dimensional linear regression via error-in-operator approach.
arxiv.org/abs/2502.15437.

Spokoiny, V. (2025a).
Marginal minimization and sup-norm expansions in perturbed optimization.
arxiv.org/2505.02562.

Spokoiny, V. (2025b).
Sharp bounds in perturbed smooth optimization.
arxiv.org/2504.11834.

Sur, P. and Candès, E. J. (2019).
A modern maximum-likelihood theory for high-dimensional logistic regression.
Proceedings of the National Academy of Sciences, 116(29):14516–14525.

References

Tropp, J. A. (2015).
An introduction to matrix concentration inequalities.
Foundations and Trends in Machine Learning, 8(1-2):1–230.

Vershynin, R. (2018).
High-Dimensional Probability: An Introduction with Applications in Data Science.
Number 47 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Zhivotovskiy, N. (2024).
Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle.
Electronic Journal of Probability, 29(none):1 – 28.