Some statistical questions around linear prediction

Jaouad Mourtada (CREST, ENSAE)

Based in part on joint works with S. Gaïffas (Paris Diderot), T. Vaškevičius (EPFL) and N. Zhivotovskiy (ETH Zürich).

Journées MAS 2022
Rouen, 29-31 août

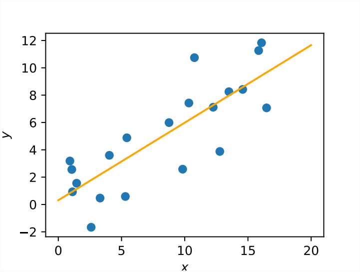
Outline

Linear regression: “universal” lower bound

Linear regression: distribution-free guarantees

Logistic regression

Linear regression



- **Prediction** problem: predict $y \in \mathbf{R}$ based on covariates $x \in \mathbf{R}^d$
- Random pair $(X, Y) \sim P$ on $\mathbf{R}^d \times \mathbf{R}$, distribution P **unknown**
- **Risk** $R(f) = \mathbf{E}[(f(X) - Y)^2]$ of prediction function $f : \mathbf{R}^d \to \mathbf{R}$
- $\mathcal{F}_{\text{lin}} = \{f_\theta : \theta \in \mathbf{R}^d\}$ with $f_\theta(x) = \langle \theta, x \rangle$ **linear functions**

Linear regression

- **Prediction** problem: predict $y \in \mathbf{R}$ based on covariates $x \in \mathbf{R}^d$
- Random pair $(X, Y) \sim P$ on $\mathbf{R}^d \times \mathbf{R}$, distribution P **unknown**
- **Risk** $R(f) = \mathbf{E}[(f(X) - Y)^2]$ of prediction function $f : \mathbf{R}^d \to \mathbf{R}$
- $\mathcal{F}_{\text{lin}} = \{f_\theta : \theta \in \mathbf{R}^d\}$ with $f_\theta(x) = \langle \theta, x \rangle$ **linear functions**
- Given $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbf{R}^d \times \mathbf{R}$ i.i.d. sample from P, find function $\widehat{f} : \mathbf{R}^d \to \mathbf{R}$ whose **excess risk**

$$\mathcal{E}(\widehat{f}) = R(\widehat{f}) - \inf_{\theta \in \mathbf{R}^d} R(f_\theta)$$

is **small** in expectation/with high probability. *I.e., prediction error $R(\widehat{f})$ of $\widehat{f}$ almost as small as that of best linear function.*

Linear regression: “universal” lower bound

Linear regression with independent noise

Assume $Y = \langle \theta^*, X \rangle + \varepsilon$ with $\varepsilon | X \sim \mathcal{N}(0, 1)$ and $\theta^* \in \mathbf{R}^d$.

Best guarantee uniform over $\mathbf{R}^d$: **minimax risk** (depending on P_X)

$$\mathcal{E}^*(P_X) = \inf_{\widehat{\theta}_n} \sup_{\theta^* \in \mathbf{R}^d} \mathbf{E}_{\theta^*}[R(f_{\widehat{\theta}_n}) - R(f_{\theta^*})] \, .$$

Linear regression with independent noise

Assume $Y = \langle \theta^*, X \rangle + \varepsilon$ with $\varepsilon | X \sim \mathcal{N}(0,1)$ and $\theta^* \in \mathbf{R}^d$.

Best guarantee uniform over $\mathbf{R}^d$: **minimax risk** (depending on P_X)

$$\mathcal{E}^*(P_X) = \inf_{\widehat{\theta}_n} \sup_{\theta^* \in \mathbf{R}^d} \mathbf{E}_{\theta^*}[R(f_{\widehat{\theta}_n}) - R(f_{\theta^*})] \, .$$

Then, **least-squares** $\widehat{\theta}_n^{ls} = \mathrm{argmin}_\theta \, n^{-1} \sum_{i=1}^n (Y_i - \langle \theta, X_i \rangle)^2$ is minimax for every P_X, and (assuming w.l.o.g. that $\mathbf{E} X X^\mathsf{T} = I_d$)

$$\mathcal{E}^*(P_X) = \mathbf{E} \, \mathrm{Tr} \left\{ \left(\sum_{i=1}^n X_i X_i^\mathsf{T} \right)^{-1} \right\} .$$

Gaussian vs. general distributions

Recall the minimax risk for prediction:

$$\mathcal{E}^*(P_X) = \inf_{\widehat{\theta}_n} \sup_{\theta^* \in \mathbf{R}^d} \mathbf{E}_{\theta^*}[R(f_{\widehat{\theta}_n}) - R(f_{\theta^*})] = \mathbf{E}\,\mathrm{Tr}\left\{ \Big(\sum_{i=1}^{n} X_i X_i^{\mathsf{T}} \Big)^{-1} \right\}$$

Using matrix convexity, $\mathcal{E}^*(P_X) \geqslant \mathbf{E}\,\mathrm{Tr}\{(nI_d)^{-1}\} = d/n$.

Gaussian vs. general distributions

Recall the minimax risk for prediction:

$$\mathcal{E}^*(P_X) = \inf_{\widehat{\theta}_n} \sup_{\theta^* \in \mathbf{R}^d} \mathbf{E}_{\theta^*}[R(f_{\widehat{\theta}_n}) - R(f_{\theta^*})] = \mathbf{E}\,\mathrm{Tr}\left\{\Big(\sum_{i=1}^n X_i X_i^{\mathsf{T}}\Big)^{-1}\right\}$$

Using matrix convexity, $\mathcal{E}^*(P_X) \geqslant \mathbf{E}\,\mathrm{Tr}\{(nI_d)^{-1}\} = d/n$.

If $X \sim \mathcal{N}(0, I_d)$ (**Gaussian** case), then (Wishart matrices)

$$\mathcal{E}^*(P_X) = \frac{d}{n-d-1} \quad \Big(\to \frac{\gamma}{1-\gamma} \text{ if } d/n \to \gamma \in (0,1)\Big).$$

Gaussian vs. general distributions

Recall the minimax risk for prediction:

$$\mathcal{E}^*(P_X) = \inf_{\widehat{\theta}_n} \sup_{\theta^* \in \mathbf{R}^d} \mathbf{E}_{\theta^*}[R(f_{\widehat{\theta}_n}) - R(f_{\theta^*})] = \mathbf{E}\,\mathrm{Tr}\left\{\left(\sum_{i=1}^n X_i X_i^\mathsf{T}\right)^{-1}\right\}$$

Using matrix convexity, $\mathcal{E}^*(P_X) \geqslant \mathbf{E}\,\mathrm{Tr}\{(nI_d)^{-1}\} = d/n$.

If $X \sim \mathcal{N}(0, I_d)$ (**Gaussian** case), then (Wishart matrices)

$$\mathcal{E}^*(P_X) = \frac{d}{n-d-1} \quad \left(\to \frac{\gamma}{1-\gamma} \text{ if } d/n \to \gamma \in (0,1)\right).$$

Proposition (M., 2019)

For **every** *distribution P_X on $\mathbf{R}^d$ (with $\mathbf{E}XX^\mathsf{T} = I_d$),*

$$\mathcal{E}^*(P_X) \geqslant \frac{d}{n-d+1} \quad \left(\to \frac{\gamma}{1-\gamma} \textit{ if } d/n \to \gamma \in (0,1)\right).$$

Lower bound in terms of signal strength

Instead of sup over $\theta^* \in \mathbf{R}^d$: **Prior** $\theta^* \sim \mathcal{N}(0, (\eta/d) I_d)$.

$\eta = \mathbf{E}\|\theta^*\|^2 = \mathbf{E}[\langle \theta^*, X \rangle^2]$ **signal-to-noise ratio** (SNR).

Lower bound in terms of signal strength

Instead of sup over $\theta^* \in \mathbf{R}^d$: **Prior** $\theta^* \sim \mathcal{N}(0, (\eta/d) I_d)$.

$\eta = \mathbf{E}\|\theta^*\|^2 = \mathbf{E}[\langle \theta^*, X\rangle^2]$ **signal-to-noise ratio** (SNR).

Theorem ("Marchenko-Pastur" lower bound; M., 2019)

For any distribution P_X with $\mathbf{E}[XX^\mathsf{T}] = I_d$, the Bayes risk writes

$$\mathbf{E}\,\mathrm{Tr}\left\{\Big(\sum_{i=1}^n X_i X_i^\mathsf{T} + \frac{d}{\eta} I_d\Big)^{-1}\right\} \geqslant \frac{d}{n+1}\mathcal{S}_{MP}\Big(\frac{d}{n+1}, \frac{d}{n+1}\eta^{-1}\Big)$$

$$\textit{where} \quad \mathcal{S}_{MP}(\gamma, \lambda) = \frac{-(1-\gamma+\lambda) + \sqrt{(1-\gamma+\lambda)^2 + 4\gamma\lambda}}{2\lambda\gamma}\,.$$

Lower bound in terms of signal strength

Instead of sup over $\theta^* \in \mathbf{R}^d$: **Prior** $\theta^* \sim \mathcal{N}(0, (\eta/d) I_d)$.

$\eta = \mathbf{E}\|\theta^*\|^2 = \mathbf{E}[\langle \theta^*, X \rangle^2]$ **signal-to-noise ratio** (SNR).

Theorem ("Marchenko-Pastur" lower bound; M., 2019)

For any distribution P_X with $\mathbf{E}[XX^\top] = I_d$, the Bayes risk writes

$$\mathbf{E}\,\mathrm{Tr}\left\{ \Big(\sum_{i=1}^n X_i X_i^\top + \frac{d}{\eta} I_d \Big)^{-1} \right\} \geqslant \frac{d}{n+1} \mathcal{S}_{MP}\Big(\frac{d}{n+1}, \frac{d}{n+1} \eta^{-1} \Big)$$

where $$\mathcal{S}_{MP}(\gamma, \lambda) = \frac{-(1-\gamma+\lambda) + \sqrt{(1-\gamma+\lambda)^2 + 4\gamma\lambda}}{2\lambda\gamma}.$$

Matching limit for Gaussian X (**Marchenko-Pastur law**) as $d/n \to \gamma$. General **lower bound** valid for **any** distribution, Gaussian distribution is "asymptotically easiest".

Exact extremality of the spherical distribution?

Question

Is it true that, for any $n > d \geqslant 1$ and $\eta > 0$, the **spherical distribution** P_X (uniform on the sphere of radius $\sqrt{d}$) minimizes the Bayes risk:

$$\mathcal{E}^*(P_X, \eta) = \mathbf{E}\,\mathrm{Tr}\left\{\left(\sum_{i=1}^n X_i X_i^\mathsf{T} + \frac{d}{\eta} I_d\right)^{-1}\right\}$$

among all distributions on $\mathbf{R}^d$ such that $\mathbf{E} XX^\mathsf{T} = I_d$?

True among **spherically invariant** distributions (including the Gaussian), so **asymptotically minimal** as $d/n \to \gamma$.

Related to certain matrix inequalities (would follow from a possible extension of the Golden-Thomson inequality).

Linear regression: distribution-free guarantees

Upper bounds

We considered **lower bounds**, allowing to identify the "best case" (Gaussian covariates). What about **upper bounds**?

Upper bounds

We considered **lower bounds**, allowing to identify the “best case” (Gaussian covariates). What about **upper bounds**?

Under **strong tail assumptions** on $P_{(X,Y)}$ (e.g. “sub-Gaussian” vectors), least squares $\widehat{\theta}_n^{ls} = \mathrm{argmin}_\theta \, n^{-1} \sum_{i=1}^n (Y_i - \langle \theta, X_i \rangle)^2$ has optimal $O(d/n)$ risk with high probability.

Upper bounds

We considered **lower bounds**, allowing to identify the "best case" (Gaussian covariates). What about **upper bounds**?

Under **strong tail assumptions** on $P_{(X,Y)}$ (e.g. "sub-Gaussian" vectors), least squares $\widehat{\theta}_n^{ls} = \mathrm{argmin}_\theta \, n^{-1} \sum_{i=1}^n (Y_i - \langle \theta, X_i \rangle)^2$ has optimal $O(d/n)$ risk with high probability.

A recent line of work on **robust regression** (e.g. Audibert & Catoni '11, Lugosi & Mendelson'19, Oliveira'16, Lecué & Lerasle'20) shows that more sophisticated estimators $\widehat{\theta}_n$ achieve $O(d/n)$ risk under heavy tails, e.g. **moment equivalence** for X (and likewise for errors):

$$\forall \theta \in \mathbf{R}^d, \quad (\mathbf{E}\langle \theta, X \rangle^4)^{1/4} \leqslant \kappa (\mathbf{E}\langle \theta, X \rangle^2)^{1/2} \,.$$

Upper bounds

We considered **lower bounds**, allowing to identify the "best case" (Gaussian covariates). What about **upper bounds**?

Under **strong tail assumptions** on $P_{(X,Y)}$ (e.g. "sub-Gaussian" vectors), least squares $\widehat{\theta}_n^{ls} = \mathrm{argmin}_\theta\, n^{-1} \sum_{i=1}^n (Y_i - \langle \theta, X_i \rangle)^2$ has optimal $O(d/n)$ risk with high probability.

A recent line of work on **robust regression** (e.g. Audibert & Catoni '11, Lugosi & Mendelson'19, Oliveira'16, Lecué & Lerasle'20) shows that more sophisticated estimators $\widehat{\theta}_n$ achieve $O(d/n)$ risk under heavy tails, e.g. **moment equivalence** for X (and likewise for errors):

$$\forall \theta \in \mathbf{R}^d, \quad (\mathbf{E}\langle \theta, X \rangle^4)^{1/4} \leqslant \kappa (\mathbf{E}\langle \theta, X \rangle^2)^{1/2} .$$

Question: Is it possible to **remove any assumption** on X?

Distribution-free regression

Joint distribution $P_{(X,Y)}$ characterized by P_X and $P_{Y|X}$.

We want guarantees that are valid for **any distribution** P_X of X on $\mathbf{R}^d$, and under **minimal assumptions** on $Y|X$.

Distribution-free regression

Joint distribution $P_{(X,Y)}$ characterized by P_X and $P_{Y|X}$.

We want guarantees that are valid for **any distribution** P_X of X on $\mathbf{R}^d$, and under **minimal assumptions** on $Y|X$.

Main assumption (on $P_{Y|X}$)

There exists a constant $m > 0$ such that

$$\sup_{x \in \mathbf{R}^d} \mathbf{E}[Y^2 | X = x] \leqslant m^2.$$

This assumption is **necessary**: no distribution-free guarantee is achievable without it. This holds if Y is **bounded**: $|Y| \leqslant m$ a.s., but also allows for **heavy tails** (only 2 moments).

Limitations of linear predictors

An predictor $\widehat{f_n}$ is **linear** if it consists of a linear function $f_{\widehat{\theta}_n}$.

Remark: this includes least squares, but also most procedures in the literature (including in robust regression).

Limitations of linear predictors

An predictor $\widehat{f}_n$ is **linear** if it consists of a linear function $f_{\widehat{\theta}_n}$.

Remark: this includes least squares, but also most procedures in the literature (including in robust regression).

Proposition

For all $n, d \geqslant 1$ and any linear predictor $\widehat{f}_n$, there exists a distribution P with $|Y| \leqslant 1$ such that

$$\mathbf{E} R(\widehat{f}_n) - \inf_{\theta \in \mathbf{R}^d} R(f_\theta) \gtrsim 1.$$

(Upper bound of 1 trivially achieved by zero function $\widehat{f}_n \equiv 0$.)

No nontrivial distribution-free guarantee for **linear** predictors.

Classical bound for truncated least squares

Truncated least squares: thresholds predictions to $[-1, 1]$

$$\widehat{f}_{\text{trunc}}(x) = \max(-1, \min(1, \langle \widehat{\theta}_n^{ls}, x \rangle)).$$

Nonlinear (due to truncation).

Classical bound for truncated least squares

Truncated least squares: thresholds predictions to $[-1, 1]$

$$\widehat{f}_{\text{trunc}}(x) = \max(-1, \min(1, \langle \widehat{\theta}_n^{ls}, x \rangle)).$$

Nonlinear (due to truncation). Let $f_{\text{reg}}(x) = \mathbf{E}[Y|X = x]$.

Theorem (Györfi, Kohler, Krzyzak, Walk, 2002)

If $\mathbf{E}[Y^2|X] \leqslant 1$*, then truncated least squares satisfies:*

$$\mathbf{E} R(\widehat{f}_{\text{trunc}}) - \inf_{\theta \in \mathbf{R}^d} R(f_\theta) \leqslant c \, \frac{d \log n}{n} + 7 \Big(\inf_{\theta \in \mathbf{R}^d} R(f_\theta) - R(f_{\text{reg}}) \Big)$$

Distribution-free result (**no assumption** on P_X!)

Classical bound for truncated least squares

Truncated least squares: thresholds predictions to $[-1, 1]$

$$\widehat{f}_{\text{trunc}}(x) = \max(-1, \min(1, \langle \widehat{\theta}_n^{ls}, x \rangle)).$$

Nonlinear (due to truncation). Let $f_{\text{reg}}(x) = \mathbf{E}[Y|X = x]$.

Theorem (Györfi, Kohler, Krzyzak, Walk, 2002)

If $\mathbf{E}[Y^2|X] \leqslant 1$, then truncated least squares satisfies:

$$\mathbf{E} R(\widehat{f}_{\text{trunc}}) - \inf_{\theta \in \mathbf{R}^d} R(f_\theta) \leqslant c \, \frac{d \log n}{n} + 7 \Big(\inf_{\theta \in \mathbf{R}^d} R(f_\theta) - R(f_{\text{reg}}) \Big)$$

Distribution-free result (**no assumption** on P_X!)

Approximation term $7(\inf_{\theta \in \mathbf{R}^d} R(f_\theta) - R(f_{\text{reg}}))$, extra $\log n$ factor.

Improved bound in expectation for truncated least squares

Truncated least squares: $\widehat{f}_{\text{trunc}}(x) = \max(-1, \min(1, \langle \widehat{\theta}_n^{ls}, x \rangle))$

Theorem (M., Vaškevičius, Zhivotovskiy, 2021)

If $\mathbf{E}[Y^2|X] \leqslant 1$, *then*

$$\mathbf{E}R(\widehat{f}_{\text{trunc}}) - \inf_{\theta \in \mathbf{R}^d} R(f_\theta) \leqslant \frac{8d}{n+1}.$$

Improved bound in expectation for truncated least squares

Truncated least squares: $\widehat{f}_{\mathrm{trunc}}(x) = \max(-1, \min(1, \langle \widehat{\theta}_n^{ls}, x \rangle))$

Theorem (M., Vaškevičius, Zhivotovskiy, 2021)

If $\mathbf{E}[Y^2|X] \leqslant 1$, *then*

$$\mathbf{E} R(\widehat{f}_{\mathrm{trunc}}) - \inf_{\theta \in \mathbf{R}^d} R(f_\theta) \leqslant \frac{8d}{n+1}.$$

Distribution-free guarantee (as before), $O(d/n)$ rate.

Removes approximation term $7(\inf_{\theta \in \mathbf{R}^d} R(f_\theta) - R(f_{\mathrm{reg}}))$ and extra $\log n$; gives explicit constant $c = 8$. **Simple proof** (leave-one-out argument).

Truncated least squares fails with constant probability

Truncated least squares: $\widehat{f}_{\text{trunc}}(x) = \max(-1, \min(1, \langle \widehat{\theta}_n^{ls}, x\rangle))$, with in-expectation bound $\mathbf{E} R(\widehat{f}_{\text{trunc}}) - \inf_{\theta \in \mathbf{R}^d} R(f_\theta) \leqslant 8d/n$.

Theorem (M., Vaškevičius, Zhivotovskiy, 2021)

For any $n, d \geqslant 1$, there exists a distribution of (X, Y) with $|Y| \leqslant 1$ such that

$$\mathbf{P}\left(R(\widehat{f}_{\text{trunc}}) - \inf_{\theta \in \mathbf{R}^d} R(f_\theta) \geqslant c \right) \geqslant c.$$

With **constant probability**, $\widehat{f}_{\text{trunc}}$ has **trivial/constant** excess risk.

Truncated least squares fails with constant probability

Truncated least squares: $\widehat{f}_{\mathrm{trunc}}(x) = \max(-1, \min(1, \langle \widehat{\theta}_n^{ls}, x \rangle))$, with in-expectation bound $\mathbf{E} R(\widehat{f}_{\mathrm{trunc}}) - \inf_{\theta \in \mathbf{R}^d} R(f_\theta) \leqslant 8d/n$.

Theorem (M., Vaškevičius, Zhivotovskiy, 2021)

For any $n, d \geqslant 1$, there exists a distribution of (X, Y) with $|Y| \leqslant 1$ such that

$$\mathbf{P}\Big(R(\widehat{f}_{\mathrm{trunc}}) - \inf_{\theta \in \mathbf{R}^d} R(f_\theta) \geqslant c \Big) \geqslant c.$$

With **constant probability**, $\widehat{f}_{\mathrm{trunc}}$ has **trivial/constant** excess risk.

Contradiction (?) with d/n bound in expectation?

Truncated least squares fails with constant probability

Truncated least squares: $\widehat{f}_{\text{trunc}}(x) = \max(-1, \min(1, \langle \widehat{\theta}_n^{ls}, x \rangle))$, with in-expectation bound $\mathbf{E} R(\widehat{f}_{\text{trunc}}) - \inf_{\theta \in \mathbf{R}^d} R(f_\theta) \leqslant 8d/n$.

Theorem (M., Vaškevičius, Zhivotovskiy, 2021)

For any $n, d \geqslant 1$, there exists a distribution of (X, Y) with $|Y| \leqslant 1$ such that

$$\mathbf{P}\Big(R(\widehat{f}_{\text{trunc}}) - \inf_{\theta \in \mathbf{R}^d} R(f_\theta) \geqslant c \Big) \geqslant c.$$

With **constant probability**, $\widehat{f}_{\text{trunc}}$ has **trivial/constant** excess risk.

Contradiction (?) with d/n bound in expectation? **No**, since $R(\widehat{f}_{\text{trunc}}) - \inf_{\theta \in \mathbf{R}^d} R(f_\theta)$ can take **negative values** as $\widehat{f}_{\text{trunc}}$ is **nonlinear** (compensates in expectation).

Nearly deviation-optimal estimator

Theorem (M., Vaškevičius, Zhivotovskiy, 2021)

For every $n \geqslant d \geqslant 1$ *and* $\delta \in (0,1)$, *there is a procedure* $\widehat{f}_n$ *such that, for any distribution satisfying* $\mathbf{E}[Y^2|X] \leqslant 1$, *with probability* $1-\delta$,

$$R(\widehat{f}_n) - \inf_{\theta \in \mathbf{R}^d} R(f_\theta) \leqslant c \, \frac{d \log(en/d) + \log(1/\delta)}{n}.$$

Nearly deviation-optimal estimator

Theorem (M., Vaškevičius, Zhivotovskiy, 2021)

For every $n \geqslant d \geqslant 1$ and $\delta \in (0,1)$, there is a procedure $\widehat{f}_n$ such that, for any distribution satisfying $\mathbf{E}[Y^2|X] \leqslant 1$, with probability $1-\delta$,

$$R(\widehat{f}_n) - \inf_{\theta \in \mathbf{R}^d} R(f_\theta) \leqslant c\, \frac{d \log(en/d) + \log(1/\delta)}{n}.$$

Nearly (up to log) **deviation-optimal** procedure, **distribution-free** w.r.t. P_X and only $\mathbf{E}[Y^2|X] \leqslant 1$ (minimal assumption).

Nearly deviation-optimal estimator

Theorem (M., Vaškevičius, Zhivotovskiy, 2021)

For every $n \geqslant d \geqslant 1$ and $\delta \in (0, 1)$, there is a procedure $\widehat{f}_n$ such that, for any distribution satisfying $\mathbf{E}[Y^2|X] \leqslant 1$, with probability $1 - \delta$,

$$R(\widehat{f}_n) - \inf_{\theta \in \mathbf{R}^d} R(f_\theta) \leqslant c \, \frac{d \log(en/d) + \log(1/\delta)}{n}.$$

Nearly (up to log) **deviation-optimal** procedure, **distribution-free** w.r.t. P_X and only $\mathbf{E}[Y^2|X] \leqslant 1$ (minimal assumption).

Explicit, though involved, procedure. Computationally **expensive** (exponential time in dimension d).

Open question: practical optimal estimator?

Question

Is there a procedure $\widehat{f}_n$ **computable in polynomial time** in n and d such that, for any distribution of (X, Y) with $\mathbf{E}[Y^2|X] \leqslant 1$ (or even $|Y| \leqslant 1$ a.s.), with probability $1 - \delta$,

$$R(\widehat{f}_n) - \inf_{\theta \in \mathbf{R}^d} R(f_\theta) \leqslant c \, \frac{d + \log(1/\delta)}{n} \, ?$$

Logistic regression

Logistic regression

Here, **binary target** $y \in \{-1, 1\}$ (instead of $y \in \mathbf{R}$).

Given $x \in \mathbf{R}^d$, assign **conditional probabilities** $p(\pm 1|x)$ on y.

Logistic regression

Here, **binary target** $y \in \{-1, 1\}$ (instead of $y \in \mathbf{R}$).

Given $x \in \mathbf{R}^d$, assign **conditional probabilities** $p(\pm 1|x)$ on y.

Efficient procedures with $\widetilde{O}(d/n)$ excess risk **in expectation** are known (sampling-based Bayesian methods: e.g. Yang'00, Catoni'04, Kakade & Ng'05, or optimization-based "virtual sample" approach in M., Gaïffas'19, see also Jézéquel, Gaillard, Rudi'20).

Logistic regression

Here, **binary target** $y \in \{-1, 1\}$ (instead of $y \in \mathbf{R}$).

Given $x \in \mathbf{R}^d$, assign **conditional probabilities** $p(\pm 1|x)$ on y.

Efficient procedures with $\widetilde{O}(d/n)$ excess risk **in expectation** are known (sampling-based Bayesian methods: e.g. Yang'00, Catoni'04, Kakade & Ng'05, or optimization-based "virtual sample" approach in M., Gaïffas'19, see also Jézéquel, Gaillard, Rudi'20).

However, known procedures with optimal **high-probability** guarantees have computational time **exponential** in dimension d.

Same **open question** as in the linear case: existence of computationally efficient optimal procedures?

Summary

- In high-dimensional linear regression with $d \asymp n$, **Gaussian covariates** are almost/**asymptotically the "easiest"** ones.
- It is possible to obtain $\widetilde{O}(d/n)$ statistical guarantees for linear regression **without any assumption** on the distribution of covariates. However, this requires using **nonlinear predictors**.
- The known procedure is **not practical**/efficiently computable.
- Related results and questions in **logistic regression**.

References

- J. M., "Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices". *Ann. Statist.*, 2022.
- J. M., T. Vaškevičius, N. Zhivotovskiy. "Distribution-free robust linear regression". *Mathematical Statistics and Learning*, 2021.
- J. M., S. Gaïffas. "An improper estimator with optimal excess risk in misspecified density estimation and logistic regression". *Journal of Machine Learning Research*, 2022.

Thank you!