Distribution-free robust linear regression

Jaouad Mourtada (CREST, ENSAE)

Joint work with:

Tomas Vaškevičius (University of Oxford) and Nikita Zhivotovskiy (ETH Zürich)

Séminaire Palaisien
March 2nd, 2021

Contents

Setting

Statistical learning (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 $f : \mathbf{R}^d \to \mathbf{R}$
- $\mathcal{F}_{\mathrm{lin}} = \{x \mapsto \langle w, x \rangle : w \in \mathbf{R}^d\}$ class of **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**

$$\mathcal{E}(\widehat{f}) = R(\widehat{f}) - \inf_{f \in \mathcal{F}_{\mathrm{lin}}} R(f)$$

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

Basic facts

Let $f_w : x \mapsto \langle w, x \rangle$, and $\mathcal{F}_{\text{lin}} = \{f_w : w \in \mathbf{R}^d\}$.

Assuming $\mathbf{E}Y^2 < \infty$, $\mathbf{E}\|X\|^2 < \infty$, the risk minimizer is f_{w^*}, with

$$w^* = \Sigma^{-1}\mathbf{E}[YX], \quad \text{where} \quad \Sigma = \mathbf{E}XX^\mathsf{T}.$$

Excess risk of a linear function f_w is

$$\begin{aligned}\mathcal{E}(f_w) = R(f_w) - R(f_{w^*}) &= \mathbf{E}(f_w(X) - f_{w^*}(X))^2 \\ &= \|\Sigma^{1/2}(w - w^*)\|^2.\end{aligned}$$

Least squares estimator

Population risk is $R(f) = \mathbf{E}(f(X) - Y)^2$. Define empirical risk by

$$\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2$$

Minimized in $\mathcal{F}_{\text{lin}}$ by least squares/emp. risk minimizer $\widehat{f}_{\text{erm}}$:

$$\widehat{f}_{\text{erm}} = \operatorname*{argmin}_{f \in \mathcal{F}_{\text{lin}}} \widehat{R}_n(f) = f_{\widehat{w}_{\text{erm}}}, \quad \text{where} \quad \widehat{w}_{\text{erm}} = \widehat{\Sigma}_n^{-1} \cdot \frac{1}{n}\sum_{i=1}^{n} Y_i X_i,$$

with $\widehat{\Sigma}_n := \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i X_i^{\mathsf{T}}$ the empirical covariance matrix

Overview of existing results

Performance of the least squares estimator

$w^* = \text{argmin}_{w \in \mathbf{R}^d} R(f_w)$ best parameter, error $\xi = Y - \langle w^*, X \rangle$

Excess risk of the least squares estimator $\widehat{f}_{\text{erm}}$ is

$$R(\widehat{f}_{\text{erm}}) - R(f_{w^*}) = \left\| \Sigma^{1/2} \widehat{\Sigma}_n^{-1} \Sigma^{1/2} \cdot \frac{1}{n} \sum_{i=1}^{n} \xi_i \Sigma^{-1/2} X_i \right\|^2$$

$$\leqslant \underbrace{\lambda_{\min}(\Sigma^{-1/2} \widehat{\Sigma}_n \Sigma^{-1/2})^{-2}}_{\text{matrix fluctuations/random design}} \cdot \underbrace{\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i \Sigma^{-1/2} X_i \right\|^2}_{\text{"noise"}}$$

Least squares under boundedness or light tails

Boundedness assumption: $\|\Sigma^{-1/2}X\| \leqslant C\sqrt{d}$ a.s.

Or **sub-Gaussian** tail: $\mathbf{P}(|\langle w, X\rangle| \geqslant t\|\Sigma^{1/2}w\|) \leqslant 2\exp(-t^2/\kappa^2)$

Strong/restrictive assumptions on X, imply (two-sided) **matrix concentration**: $\frac{1}{2}\Sigma \preccurlyeq \widehat{\Sigma}_n \preccurlyeq 2\Sigma$ for $n \gtrsim d$.

If errors are also light-tailed (sub-Gaussian), then least squares achieves the optimal bound

$$R(\widehat{f}_{\mathrm{erm}}) - \inf_{f \in \mathcal{F}_{\mathrm{lin}}} R(f) \lesssim \frac{d}{n}.$$

Intuition: empirical risk is close to population risk over $\mathcal{F}_{\mathrm{lin}}$

Some references: Caponnetto, De Vito, 2007; Catoni, 2004; Hsu et al., 2014

Linear regression under weaker tail assumptions

Weakened assumptions: finite **moment equivalence** for X:

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

(Oliveira, 2016). Related "small-ball" assumption (Koltchinskii & Mendelson, 2015; Lecué & Mendelson, 2016). **Weaker** assumption on X, implies (one-sided) **lower isometry** $\widehat{\Sigma}_n \succcurlyeq \frac{1}{2}\Sigma$.

- If error is light-tailed, least squares has $O(d/n)$ **excess risk**.

Intuition: functions with large excess risk have large empirical risk.

- If error ξ is heavy-tailed, least squares $\widehat{f}_{\mathrm{erm}}$ is **suboptimal**, but some **robust estimators** achieve $O(d/n)$ bound (Audibert & Catoni 2010, Lugosi & Mendelson 2019, Catoni 2016)

Assumptions on the distribution of covariates

- Strong assumptions on X, e.g., **subgaussian**

$$\forall w \in \mathbf{R}^d, \quad \mathbf{P}(|\langle w, X\rangle| \geqslant t\|\Sigma^{1/2}w\|) \leqslant 2\exp(-t^2/\kappa^2)$$

- Weaker **moment equivalence** conditions:

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

  Still **non-trivial** restriction. In some simple cases, κ depends on d, leading to suboptimal bounds.

- Can we **remove any assumption** on the distribution of X?

Distribution-free setting

"Distribution-free" setting

Joint distribution $P = P_{(X,Y)}$ of (X, Y) is characterized by:

- Distribution P_X of X, probability distribution on $\mathbf{R}^d$
- Conditional distribution $P_{Y|X} = (P_{Y|X=x})_{x \in \mathbf{R}^d}$ (family of distributions on $\mathbf{R}$ indexed by $x \in \mathbf{R}^d$).
  Remark: Risk $R(f)$ is minimized (among all functions) by the regression function

$$f_{\text{reg}}(x) = \mathbf{E}[Y|X = x] \, .$$

A guarantee is **distribution-free** if it holds for all distributions P_X.

1. Is it possible to obtain **distribution-free guarantees**?
2. If so, what are the **minimal conditions** on $P_{Y|X}$?

Minimal assumption on the conditional distribution

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 condition holds if Y is **bounded**: $|Y| \leqslant m$ a.s.

But **much weaker**: compatible with heavy tails of Y, only **(conditional) second moment** bound.

(For instance, one may have $\mathbf{E} Y^{2+\varepsilon} = +\infty$ for any $\varepsilon > 0$.)

Minimal assumption to obtain P_X-free guarantees (lower bounds)

Limitations of proper estimators

A procedure $\widehat{f}_n$ is called **proper** (or: **linear**) if it always returns a linear function $\widehat{f}_n \in \mathcal{F}_{\text{lin}}$.

Remark: includes least squares $\widehat{f}_{\text{erm}}$, but also most procedures in the literature (including in robust regression).

Proposition (Shamir, 2015)

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

$$\mathbf{E} R(\widehat{f}_n) - \inf_{f \in \mathcal{F}_{\text{lin}}} R(f) \gtrsim 1.$$

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

No nontrivial distribution-free guarantee for **proper** procedures

Classical bound for truncated least squares

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

$$\widehat{f}_{\text{trunc}}(x) = \max(-m, \min(m, \langle \widehat{w}_{\text{erm}}, x \rangle)).$$

Improper/nonlinear (due to truncation).

Theorem (Györfi et. al, 2002)

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

$$\mathbf{E} R(\widehat{f}_{\text{trunc}}) - \inf_{f \in \mathcal{F}_{\text{lin}}} R(f) \leqslant c \, \frac{m^2 d \log n}{n} + 7\Big(\inf_{f \in \mathcal{F}_{\text{lin}}} R(f) - R(f_{\text{reg}}) \Big)$$

Distribution-free result (no assumption on P_X!)

Approximation term $7(\inf_{f \in \mathcal{F}_{\text{lin}}} R(f) - R(f_{\text{reg}}))$

Main results

Improved bound in expectation for truncated least squares

Truncated least squares: $\widehat{f}_{\mathrm{trunc}}(x) = \max(-m, \min(m, \langle \widehat{w}_{\mathrm{erm}}, x \rangle))$

Theorem (M., Vaškevičius, Zhivotovskiy, 2021)

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

$$\mathbf{E} R(\widehat{f}_{\mathrm{trunc}}) - \inf_{f \in \mathcal{F}_{\mathrm{lin}}} R(f) \leqslant \frac{8m^2 d}{n+1}.$$

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

Removes approximation term $7(\inf_{f \in \mathcal{F}_{\mathrm{lin}}} R(f) - R(f_{\mathrm{reg}}))$ from previous bound (and extra $\log n$; gives explicit constant $c = 8$).

Simpler proof (leave-one-out argument)!

Similar bound for another procedure (Forster & Warmuth, 2002)

In-expectation vs. high-probability guarantees

Previous results (for e.g. truncated least squares) **in expectation**:

$$\mathbf{E} R(\widehat{f}_n) - \inf_{f \in \mathcal{F}_{\mathrm{lin}}} R(f) \lesssim \frac{m^2 d}{n}.$$

What about **high-probability** guarantees? Given **confidence** parameter δ, bound of the form

$$\mathbf{P}\Big(R(\widehat{f}_n) - \inf_{f \in \mathcal{F}_{\mathrm{lin}}} R(f) \geqslant \varepsilon(n, d, \delta) \Big) \leqslant \delta.$$

Under assumption $\mathbf{E}[Y^2 | X] \leqslant m^2$, **ideal accuracy** (lower bounds):

$$\varepsilon(n, d, \delta) \asymp \frac{m^2 \big(d + \log(1/\delta)\big)}{n}.$$

("Exponential" bound)

Truncated least squares fails with constant probability

Truncated least squares: $\widehat{f}_{\mathrm{trunc}}(x) = \max(-m, \min(m, \langle \widehat{w}_{\mathrm{erm}}, x \rangle))$, with in-expectation bound $\mathbf{E} R(\widehat{f}_{\mathrm{trunc}}) - \inf_{f \in \mathcal{F}_{\mathrm{lin}}} R(f) \lesssim m^2 d/n$.

Theorem (M., Vaškevičius, Zhivotovskiy, 2021)

For any $n, d \geqslant 1$, there exists a distribution P of (X, Y) with $|Y| \leqslant m$ such that (same lower bound for Forster-Warmuth)

$$\mathbf{P}\left(R(\widehat{f}_{\mathrm{trunc}}) - \inf_{f \in \mathcal{F}_{\mathrm{lin}}} R(f) \geqslant c\, m^2 \right) \geqslant c.$$

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

Contradiction (?) with $m^2 d/n$ bound in expectation?

No, since $R(\widehat{f}_{\mathrm{trunc}}) - \inf_{f \in \mathcal{F}_{\mathrm{lin}}} R(f)$ can take **negative values** as $\widehat{f}_{\mathrm{trunc}}$ is **improper/nonlinear** (compensates in expectation).

Deviation-optimal estimator

Theorem (M., Vaškevičius, Zhivotovskiy, 2021)

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

$$R(\widehat{f}_n) - \inf_{f \in \mathcal{F}_{\mathrm{lin}}} R(f) \leqslant c \, \frac{m^2 \big(d \log(n/d) + \log(1/\delta)\big)}{n}.$$

Deviation-optimal procedure, **distribution-free** w.r.t. P_X and only $\mathbf{E}[Y^2|X] \leqslant m^2$ (robustness to heavy tails).

Depends on confidence δ (unavoidable).

Explicit, though involved, procedure. Computationally **expensive**.

Some ideas behind the procedure

Two difficulties: **no assumption on X**, and **heavy-tailed Y**.

- First step: truncate linear functions to m, class $\mathcal{F}_{\mathrm{trunc}}$. Only **reduces risk**, gives **bounded functions**, but **non-convex** class!
- Second step: form some random/empirical finite discretization of the class $\mathcal{F}_{\mathrm{trunc}}$. Needed for technical reasons (heavy tails).
- Third step: use ideas from **model aggregation** theory (Star-type algorithm, Audibert 2008) to handle **non-convexity** of the class.
- Fourth step: extend above from bounded to heavy-tailed setting through **robust mean estimators** and min-max procedures. (Audibert, Catoni 2010; Lugosi, Mendelson 2019; Lecué, Lerasle 2020)

Note: the resulting procedure is **hard to compute** for large d!

Conclusion

Distribution-free linear regression, **no restriction** on P_X; minimal assumption (on $Y|X$) $\mathbf{E}[Y^2|X] \leqslant m^2$

No **proper/linear** procedure (least squares or robust alternatives) gives any useful bound in this distribution-free setting

Truncated least squares achieves $m^2 d/n$ excess risk in expectation (improving 'classical' bound)...

...but fails (m^2 risk) with **constant probability**.

Robust procedure **optimal with high probability** (extends to nonlinear VC-subgraph classes).

Future directions: Practical procedure? Adapting to m?

Thank you!