Outlier-robust additive matrix decomposition and robust matrix completion

Philip Thompson

FGV EMAp

COLMEA, PUC-Rio
November, 2023
TRACE (LINEAR) REGRESSION: Given an iid sample \( \{(y_i^\circ, X_i)\} \) from a random label-feature pair \((y, X) \in \mathbb{R} \times \mathbb{R}^{d_1 \times d_2} \), estimate the parameter

\[
\Theta^* \in \arg\min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} \mathbb{E} \left( y - \text{tr}(X^\top \Theta) \right)^2.
\]

Equivalently,

\[
y = \text{tr}(X^\top \Theta^*) + \xi,
\]

for some \( \xi \) satisfying \( \mathbb{E}[\xi X] = 0 \). We may also assume that \((X, \xi)\) is centered.
Some notation

**Definition**

Define the **design operator** $\mathcal{X} : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^n$ with coordinates

$$V \mapsto \mathcal{X}_i(V) := \text{tr}(X_i^\top V).$$
Some notation

Definition
Define the **design operator** $\mathcal{X} : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^n$ with coordinates

$$V \mapsto \mathcal{X}_i(V) := \text{tr}(X_i^\top V).$$

Notation:

- Let $y := (y_i)_{i \in [n]}$ and $\xi := (\xi_i)_{i \in [n]}$ for which $y_i = \text{tr}(X_i^\top \Theta^*) + \xi_i$.
- $\langle A, B \rangle := \text{tr}(A^\top B)$.
- $A^{(n)} := A/\sqrt{n}$.
- $\mathbb{I}^{d_1 \times d_2} := \mathbb{R}^{d_1 \times d_2}$. 
Least-squares estimator

It is well known that, under “subgaussian distributions”, the *Empirical Risk Minimization* (ERM) methodology justifies why the *least-squares estimator*

\[
\hat{\Theta} \in \arg\min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \text{tr}(X_i^\top \Theta) \right)^2
\]

is the “optimal” choice among other possible estimators.
Least-squares estimator

It is well known that, under “subgaussian distributions”, the Empirical Risk Minimization (ERM) methodology justifies why the least-squares estimator

$$\hat{\Theta} \in \arg\min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \text{tr}(X_i^\top \Theta) \right)^2$$

is the “optimal” choice among other possible estimators.

In fact, without further assumptions on $\Theta^*$, nothing is lost by seeing it simply as a vector (linear regression).
High-dimensional statistics

- In case $p \gg n$, further assume $\Theta^*$ belongs to a parsimonious class $\mathcal{F}$. 

Compressive sensing: $\theta^* \in \mathbb{R}^p$, $\|\theta^*\|_0 \leq s$ and $\xi = 0$.

Sparse linear regression or noisy compressive sensing: $\theta^* \in \mathbb{R}^p$, $\|\theta^*\|_0 \leq s$.

Trace regression: $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$, $\text{rank}(\Theta^*) \leq r$.

Additive matrix decomposition: $\Theta^* = B^* + \Gamma^*$, $\text{rank}(B^*) \leq r$, $\|\Gamma^*\|_0 \leq s$.

Identity designs: $X = I_n$.

Robust PCA.

Multi-task learning: $y_i = x_i^\top V \mathcal{X}(V) + \xi_i$, where the label is a vector of $d_2$ "tasks" and feature is a $d_1$-dimensional vector.
High-dimensional statistics

- In case $p \gg n$, further assume $\Theta^*$ belongs to a parsimonious class $\mathcal{F}$.

- *Compressive sensing*: $\theta^* \in \mathbb{R}^p$, $\|\theta^*\|_0 \leq s$ and $\xi = 0$. 

- *Sparse linear regression* or *noisy compressive sensing*: $\theta^* \in \mathbb{R}^p$, $\|\theta^*\|_0 \leq s$.

- *Trace regression*: $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$, $\text{rank}(\Theta^*) \leq r$.

- *Additive matrix decomposition*: $\Theta^* = B^* + \Gamma^*$, $\text{rank}(B^*) \leq r$, $\|\Gamma^*\|_0 \leq s$.

- *Identity designs*: $X = I_n$

- *Robust PCA*.

- *Multi-task learning*: $y_i = x_i^\top \bar{V} \bar{X}^i(V) + \xi_i$, where the label is a vector of $d_2$ "tasks" and feature is a $d_1$-dimensional vector.
In case $p \gg n$, further assume $\Theta^*$ belongs to a parsimonious class $\mathcal{F}$.

- **Compressive sensing:** $\theta^* \in \mathbb{R}^p$, $\|\theta^*\|_0 \leq s$ and $\xi = 0$.
- **Sparse linear regression or noisy compressive sensing:** $\theta^* \in \mathbb{R}^p$, $\|\theta^*\|_0 \leq s$. 

- **Trace regression:** $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$, rank($\Theta^*$) $\leq r$.
- **Additive matrix decomposition:** $\Theta^* = B^* + \Gamma^*$, rank($B^*$) $\leq r$, $\|\Gamma^*\|_0 \leq s$.
- **Identity designs:** $X = I_n$
- **Robust PCA**.
- **Multi-task learning:** $y_i = x_i^\top V(X_i(V)) + \xi_i$, where the label is a vector of $d_2$ “tasks” and feature is a $d_1$-dimensional vector.
In case $p \gg n$, further assume $\Theta^*$ belongs to a parsimonious class $\mathcal{F}$. 

- **Compressive sensing**: $\theta^* \in \mathbb{R}^p$, $\|\theta^*\|_0 \leq s$ and $\xi = 0$.  
- **Sparse linear regression or noisy compressive sensing**: $\theta^* \in \mathbb{R}^p$, $\|\theta^*\|_0 \leq s$.  
- **Trace regression**: $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$, $\text{rank}(\Theta^*) \leq r$. 

...
High-dimensional statistics

- In case $p \gg n$, further assume $\Theta^*$ belongs to a parsimonious class $\mathcal{F}$.

- **Compressive sensing**: $\theta^* \in \mathbb{R}^p$, $\|\theta^*\|_0 \leq s$ and $\xi = 0$.

- **Sparse linear regression or noisy compressive sensing**: $\theta^* \in \mathbb{R}^p$, $\|\theta^*\|_0 \leq s$.

- **Trace regression**: $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$, $\text{rank}(\Theta^*) \leq r$.

- **Additive matrix decomposition**: $\Theta^* = B^* + \Gamma^*$, $\text{rank}(B^*) \leq r$, $\|\Gamma^*\|_0 \leq s$.
  - Identity designs: $X = I_n$
  - Robust PCA.
  - Multi-task learning: $y_i = x_i^\top V + \xi_i$, where the label is a vector of $\underbrace{x_i(V)}_{d_2 \text{ "tasks"}}$ and feature is a $d_1$-dimensional vector.

- (...)
Penalized least-squares estimator

Under “subgaussian distributions and proper class $\mathcal{F}$”, a penalized least-squares estimator

$$
\hat{\Theta} \in \arg\min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \operatorname{tr}(X_i^\top \Theta) \right)^2 + \lambda \mathcal{R}(\Theta)
$$

is “optimal” for properly chosen penalization hyper-parameter $\lambda > 0$ and regularization norm $\mathcal{R}$. 
Penalized least-squares estimator

**Figure:** Chapter 7 on Additive Matrix Decomposition
In this work we are concerned about ...

1. **Additive decomposition in trace regression**: \( \Theta^* = B^* + \Gamma^* \)
   where \( \text{rank}(B^*) \leq r \) and \( \|\Gamma^*\|_0 \leq s \).
In this work we are concerned about ...

1. **Additive decomposition in trace regression**: $\Theta^* = B^* + \Gamma^*$ where $\text{rank}(B^*) \leq r$ and $\|\Gamma^*\|_0 \leq s$.

2. **Label contamination** (Robust Statistics/Learning): the observed label sample $\{y_i\}_{i=1}^n$ differs from the iid sample $\{y_i^\diamond\}_{i=1}^n$ by $o$ arbitrary outliers. The “sample contamination fraction” is

   $$\epsilon := \frac{o}{n}.$$  

How much $\epsilon$ affects the estimation?
In this work we are concerned about ...

1. **Additive decomposition in trace regression**: \( \Theta^* = B^* + \Gamma^* \) where \( \text{rank}(B^*) \leq r \) and \( \| \Gamma^* \|_0 \leq s \).

2. **Label contamination** (Robust Statistics/Learning): the observed label sample \( \{ y_i \}_{i=1}^n \) differs from the iid sample \( \{ y_i^\circ \}_{i=1}^n \) by \( o \) arbitrary outliers. The “sample contamination fraction” is

\[
\epsilon := \frac{o}{n}.
\]

How much \( \epsilon \) affects the estimation?

3. **Feature-dependent noise**: “Nothing is assumed” beyond marginal sub-gaussianity of \( (X, \xi) \) and \( \mathbb{E}[\xi X] = 0 \).

   - For instance, \( \xi \) can be centered non-symmetric and have zero mass around the origin.
In this work we are concerned about ...

1. **Additive decomposition in trace regression**: \( \Theta^* = B^* + \Gamma^* \) where \( \text{rank}(B^*) \leq r \) and \( \| \Gamma^* \|_0 \leq s \).

2. **Label contamination** (Robust Statistics/Learning): the observed label sample \( \{y_i\}_{i=1}^n \) differs from the iid sample \( \{y_i^o\}_{i=1}^n \) by \( o \) arbitrary outliers. The “sample contamination fraction” is

\[
\epsilon := \frac{o}{n}.
\]

How much \( \epsilon \) affects the estimation?

3. **Feature-dependent noise**: “Nothing is assumed” beyond marginal sub-gaussianity of \( (X, \xi) \) and \( \mathbb{E}[\xi X] = 0 \).

   - For instance, \( \xi \) can be centered non-symmetric and have zero mass around the origin.

4. Dependence on **failure probability** \( \delta \): optimal rate and adaptivity.
Let us consider trace-regression assuming

- no matrix decomposition: \( \Theta^* = B^* \) with \( \text{rank}(B^*) \leq r \).
- no label contamination.
- \( \xi \) is independent of \( \{X_i\}_{i=1}^n \). E.g. the Gaussian model.
Let us consider trace-regression assuming

- no matrix decomposition: $\Theta^* = B^*$ with $\text{rank}(B^*) \leq r$.
- no label contamination.
- $\xi$ is independent of $\{X_i\}_{i=1}^n$. E.g. the Gaussian model.

In this case,

$$\hat{B} \in \min_{B \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^n (y_i - \langle X_i, B \rangle)^2 + \lambda \mathcal{R}(B),$$

with $\mathcal{R} = \| \cdot \|_N$ the nuclear norm.
The classical “LASSO proof”
for M-estimation with decomposable regularizers

The previous problem is nonsmooth convex so it is equivalent to its first condition:

\[ \exists \mathbf{V} \in \partial \mathcal{R}(\hat{\mathbf{B}}), \forall \mathbf{B} \in \mathbb{R}^p, \]
\[ \sum_{i \in [n]} \left[ y_i^{(n)} - \mathcal{X}_i^{(n)}(\hat{\mathbf{B}}) \right] \langle \mathcal{X}_i^{(n)}, \hat{\mathbf{B}} - \mathbf{B} \rangle \geq \lambda \langle \mathbf{V}, \hat{\mathbf{B}} - \mathbf{B} \rangle. \]
The classical “LASSO proof”
for M-estimation with decomposable regularizers

Using that
\[ y^{(n)} = \mathcal{X}^{(n)}(B^*) + \xi^{(n)}, \]
and defining \( \Delta B^* := \hat{B} - B^* \) one gets
\[ \| \mathcal{X}^{(n)}(\Delta B^*) \|_2^2 \leq \langle \xi^{(n)}, \mathcal{X}^{(n)}(\Delta B^*) \rangle - \lambda \langle V, \Delta B^* \rangle. \]
The classical “LASSO proof”
for M-estimation with decomposable regularizers

Using that
\[ y^{(n)} = \mathcal{X}^{(n)}(B^*) + \xi^{(n)}, \]
and defining \( \Delta_{B^*} := \hat{B} - B^* \) one gets
\[
\| \mathcal{X}^{(n)}(\Delta_{B^*}) \|_2^2 \leq \langle \xi^{(n)}, \mathcal{X}^{(n)}(\Delta_{B^*}) \rangle - \lambda \langle V, \Delta_{B^*} \rangle.
\]

Lemma
\[
-\langle V, \Delta_{B^*} \rangle \leq \mathcal{R}(B^*) - \mathcal{R}(\hat{B}).
\]
The classical “LASSO proof”
for M-estimation with decomposable regularizers

We obtain the “recursion”

\[ \| \mathcal{X}^{(n)}(\Delta_{B^*}) \|_2^2 \leq \langle \xi^{(n)}, \mathcal{X}^{(n)}(\Delta_{B^*}) \rangle - \lambda(\mathcal{R}(B_*) - \mathcal{R}(\hat{B})). \]
We obtain the “recursion”

$$\| \mathcal{X}^{(n)}(\Lambda_{B^*}) \|^2 \leq \langle \xi^{(n)}, \mathcal{X}^{(n)}(\Lambda_{B^*}) \rangle - \lambda (\mathcal{R}(B^*) - \mathcal{R}(\hat{B})).$$

We will need to bound two random processes:

- **upper bound**: the *multiplier process*

  $$V \mapsto \langle \xi^{(n)}, \mathcal{X}^{(n)}(V) \rangle = \frac{1}{n} \sum_{i=1}^{n} \xi_i \langle X_i, V \rangle.$$

- **lower bound**: the *quadratic process*

  $$V \mapsto \| \mathcal{X}^{(n)}(V) \|^2 = \frac{1}{n} \sum_{i=1}^{n} \langle X_i, V \rangle^2.$$
Upper bound: “classical approach”
for M-estimation with decomposable regularizers

The first design property we need is a suitable upper bound on the multiplier process. By the dual-norm inequality,

$$\langle \xi^{(n)}, \mathcal{X}^{(n)}(\Delta_{B^*}) \rangle \leq \left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i X_i \right\|_{\text{op}} \mathcal{R}(\Delta_{B^*}).$$

Lemma (MP)
Assuming "\((\xi, X)\) are independent and marginally 1-subgaussian and \(X\) is isotropic", for any \(\delta \in (0, 1]\), with probability \(\geq 1 - \delta\),

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i X_i \right\|_{\text{op}} \leq C \sqrt{(d_1 + d_2) + \log(1/\delta)n}.$$
Upper bound: “classical approach”
for M-estimation with decomposable regularizers

The first design property we need is a suitable upper bound on the multiplier process. By the dual-norm inequality,

\[ \langle \xi^{(n)}, \mathcal{X}^{(n)}(\Delta_B^*) \rangle \leq \left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i X_i \right\|_{op} \mathcal{R}(\Delta_B^*). \]

Lemma (MP)
Assuming “\((\xi, X)\) are independent and marginally 1-subgaussian and \(X\) is isotropic”, for any \(\delta \in (0, 1)\), with probability \(\geq 1 - \delta\),

\[ \left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i X_i \right\|_{op} \leq C \sqrt{\frac{(d_1 + d_2) + \log(1/\delta)}{n}}. \]
Upper bound: “classical approach”

for M-estimation with decomposable regularizers

The first design property we need is a suitable upper bound on the multiplier process. By the dual-norm inequality,

\[ \langle \xi^{(n)}, \mathcal{X}^{(n)}(\Delta B^*) \rangle \leq \left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i X_i \right\|_{\text{op}} \mathcal{R}(\Delta B^*). \]

Lemma (MP)

Assuming “(\(\xi, X\)) are independent and marginally 1-subgaussian and \(X\) is isotropic”, for any \(\delta \in (0, 1)\), with probability \(\geq 1 - \delta\),

\[ \left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i X_i \right\|_{\text{op}} \leq C \sqrt{\frac{(d_1 + d_2) + \log(1/\delta)}{n}}. \]

NOTE: actually, independence is not required.
Hence, asking the tuning to be

\[ \lambda \geq 2C \sqrt{\frac{(d_1 + d_2) + \log(1/\delta)}{n}}, \]

we get

\[ \| \mathcal{X}^{(n)}(\Delta_{B^*}) \|_2^2 \leq \frac{\lambda}{2} \mathcal{R}(\Delta_{B^*}) + \lambda(\mathcal{R}(B^*) - \mathcal{R}(\hat{B})). \]
Upper bound: “classical approach”
for M-estimation with decomposable regularizers

Definition (Decomposable norms)
A norm $\mathcal{R}$ over $\mathbb{R}^p$ is said to be decomposable if, for all $\mathbf{B} \in \mathbb{R}^p$, there exists linear map $\mathbf{V} \mapsto \mathcal{P}_{\mathbf{B}}^\perp(\mathbf{V})$ such that, for all $\mathbf{V} \in \mathbb{R}^p$, defining

$\mathcal{P}_{\mathbf{B}}(\mathbf{V}) := \mathbf{V} - \mathcal{P}_{\mathbf{B}}^\perp(\mathbf{V}),$

- $\mathcal{P}_{\mathbf{B}}(\mathbf{B}) = 0,$
- $\langle \mathcal{P}_{\mathbf{B}}(\mathbf{V}), \mathcal{P}_{\mathbf{B}}^\perp(\mathbf{V}) \rangle = 0,$
- $\mathcal{R}(\mathbf{V}) = \mathcal{R}(\mathcal{P}_{\mathbf{B}}(\mathbf{V})) + \mathcal{R}(\mathcal{P}_{\mathbf{B}}^\perp(\mathbf{V})).$
Upper bound: “classical approach”
for M-estimation with decomposable regularizers

Example (Nuclear norm)

Let $B \in \mathbb{R}^p$ with rank $r := \text{rank}(B)$, singular values $\{\sigma_j\}_{j \in [r]}$ and singular vector decomposition

$$B = \sum_{j \in [r]} \sigma_j u_j v_j^\top.$$ 

Here $\{u_j\}_{j \in [r]}$ are the left singular vectors spanning the subspace $U$ and $\{v_j\}_{j \in [r]}$ are the right singular vectors spanning the subspace $V$. The pair $(U, V)$ is sometimes referred as the low-rank support of $B$. Given subspace $S \subset \mathbb{R}^\ell$ let $P_{S^\perp}$ denote the matrix defining the orthogonal projection onto $S^\perp$. Then, the map

$$V \mapsto P_B^\perp(V) := P_{U^\perp} V P_{V^\perp}^\top$$

satisfy the decomposability condition for the nuclear norm $\| \cdot \|_N$. 
Upper bound: “classical approach”
for M-estimation with decomposable regularizers

Lemma
Let $\mathcal{R}$ be a decomposable norm. Let $B, \hat{B} \in \mathbb{R}^p$ and $V := \hat{B} - B$. Then

$$
\frac{1}{2} \mathcal{R}(V) + \mathcal{R}(B) - \mathcal{R}(\hat{B}) \leq \frac{3}{2} \mathcal{R}(P_B(V)) - \frac{1}{2} \mathcal{R}(P_B^\perp(V)).
$$
Upper bound: “classical approach”
for M-estimation with decomposable regularizers

Lemma
Let $\mathcal{R}$ be a decomposable norm. Let $\mathbf{B}, \hat{\mathbf{B}} \in \mathbb{R}^p$ and $\mathbf{V} := \hat{\mathbf{B}} - \mathbf{B}$. Then

$$
\frac{1}{2} \mathcal{R}(\mathbf{V}) + \mathcal{R}(\mathbf{B}) - \mathcal{R}(\hat{\mathbf{B}}) \leq \frac{3}{2} \mathcal{R}(\mathcal{P}_B(\mathbf{V})) - \frac{1}{2} \mathcal{R}(\mathcal{P}_B^\perp(\mathbf{V})).
$$

In conclusion, for $\mathcal{R} = \| \cdot \|_N$, we get

$$
\| \hat{\mathbf{x}}^{(n)}(\Delta_{\mathbf{B}^*}) \|_2^2 \leq \frac{3\lambda}{2} \mathcal{R}(\mathcal{P}_{\mathbf{B}^*}(\Delta_{\mathbf{B}^*})) - \frac{\lambda}{2} \mathcal{R}(\mathcal{P}_{\mathbf{B}^*}^\perp(\Delta_{\mathbf{B}^*})).
$$
Upper bound: “classical approach”
for M-estimation with decomposable regularizers

In particular, the estimation error $\Delta_{B^*} = \hat{B} - B^*$ belongs to the dimension-reduction cone

$$C_{B^*}(3) := \left\{ V : \mathcal{R}(\mathcal{P}_{B^*}^\perp(V)) \leq 3 \mathcal{R}(\mathcal{P}_{B^*}(V)) \right\}.$$
Upper bound: “classical approach”
for M-estimation with decomposable regularizers

In particular, the estimation error $\Delta_{B^*} = \hat{B} - B^*$ belongs to the dimension-reduction cone

$$
C_{B^*}(3) := \left\{ V : \mathcal{R}(\mathcal{P}_{B^*}(V)) \leq 3\mathcal{R}(\mathcal{P}_{B^*}(V)) \right\}.
$$

This is good because for $V \in C_{B^*}(3)$,

$$
\mathcal{R}(V) \leq 4\mathcal{R}(\mathcal{P}_{B^*}(V)) \leq 4\sqrt{r\|V\|_F}.
$$
Lower bound
for M-estimation with decomposable regularizers

The second design property we need is to ask for strong-convexity restricted to the dimension cone. Turns out that this a consequence of

**Definition (RSC)**

$X$ satisfies RSC$_\mathcal{R}(a_1, a_2)$ if for all $V \in \mathbb{R}^p$,

$$
\|X^{(n)}(V)\|_2 \geq a_1 \|V\|_\Pi - a_2 \mathcal{R}(V).
$$

Here:

$$
\|V\|_\Pi^2 := \mathbb{E}[\langle X, V \rangle^2].
$$

In case, $X$ is isotropic, $\|V\|_\Pi = \|V\|_F$. 
Lemma (RSC)

“Assume $X$ is 1-subgaussian and isotropic”. Suppose that $n \gtrsim 1 + \log(1/\delta)$. Then, with probability $\geq 1 - \delta$, $\text{RSC}_R(a_1, a_2)$ holds with constants $a_1 \in (0, 1)$ and

$$a_2 \lesssim \sqrt{\frac{r}{n}}.$$
Lower bound
for M-estimation with decomposable regularizers

**Lemma (RSC)**

“Assume $\mathbf{X}$ is 1-subgaussian and isotropic”. Suppose that $n \gtrsim 1 + \log(1/\delta)$. Then, with probability $\geq 1 - \delta$, RSC$\mathcal{R}(a_1, a_2)$ holds with constants $a_1 \in (0, 1)$ and

$$a_2 \lesssim \sqrt{\frac{r}{n}}.$$

In conclusion, if

$$n \gtrsim r \lor (1 + \log(1/\delta)),$$

then restricted strong convexity holds:

$$\inf_{\mathbf{v} \in \mathcal{C}_{B^*}(3)} \frac{\|\mathcal{X}^{(n)}\mathbf{v}\|_2}{\|\mathbf{v}\|_2} \geq \frac{a_1}{2}. $$
Conclusion of “LASSO proof”
for M-estimation with decomposable regularizers

Recall:

\[ \Delta_{B^*} \in C_{B^*}(3). \]
\[ \| \mathcal{X}^{(n)}(\Delta_{B^*}) \|_2^2 \leq \frac{3\lambda}{2} \mathcal{R}(\mathcal{P}_{B^*}(\Delta_{B^*})) - \frac{\lambda}{2} \mathcal{R}(\mathcal{P}_{B^*}^\perp(\Delta_{B^*})). \]

In conclusion,

\[ \frac{a_1}{2} \| \Delta_{B^*} \|_2^2 \leq \| \mathcal{X}^{(n)}(\Delta_{B^*}) \|_2^2 \leq \frac{3\lambda}{2} \mathcal{R}(\mathcal{P}_{B^*}(\Delta_{B^*})) \leq \frac{3\lambda}{2} \sqrt{r} \| \Delta_{B^*} \|_2, \]

so, “on the event that both design properties hold”

with probability at least \( \geq 1 - \delta \),

\[ \| \Delta_{B^*} \|_2 \leq \frac{3\lambda \sqrt{r}}{a_1} \leq \sqrt{\frac{r(d_1 + d_2)}{n}} + \sqrt{\frac{r \log(1/\delta)}{n}}. \]
Failure probability $\delta$

Some observations:

▶ One can show that

$$\sqrt{\frac{r(d_1 + d_2)}{n}}$$

is the **optimal rate in average**.
Failure probability $\delta$

Some observations:

- One can show that
  \[ \sqrt{\frac{r(d_1 + d_2)}{n}} \]
  is the **optimal rate in average**.

- The penalization is $r$-adaptive. Still, the approach using the dual norm inequality leads to the $\delta$-**dependent tuning**
  \[ \lambda \approx \sqrt{\frac{(d_1 + d_2) + \log(1/\delta)}{n}}. \]
Failure probability $\delta$

Some observations:

- One can show that
  \[ \sqrt{\frac{r(d_1 + d_2)}{n}} \]
  is the **optimal rate in average**.

- The penalization is $r$-adaptive. Still, the approach using the dual norm inequality leads to the $\delta$-dependent tuning
  \[ \lambda \approx \sqrt{\frac{(d_1 + d_2) + \log(1/\delta)}{n}}. \]

- What is the **optimal rate in probability**? Under a reasonable regime, is $\delta$-dependent tuning necessary?
Failure probability $\delta$

- Bellec-Guillaume-Tsybakov (2018) was the first to answer these questions in the case of a sparse parameter. The changes for the case of a low-rank parameter are trivial.
Failure probability $\delta$

- Bellec-Guillaume-Tsybakov (2018) was the first to answer these questions in the case of a sparse parameter. The changes for the case of a low-rank parameter are trivial.

- From their bounds, the optimal rate in probability is

$$\sqrt{\frac{r(d_1 + d_2)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}},$$

showing the previous \textbf{multiplicative term} $r \log(1/\delta)$ is suboptimal.
Failure probability $\delta$

- Bellec-Guillaume-Tsybakov (2018) was the first to answer these questions in the case of a sparse parameter. The changes for the case of a low-rank parameter are trivial.

- From their bounds, the optimal rate in probability is

$$\sqrt{\frac{r(d_1 + d_2)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}},$$

showing the previous multiplicative term $r \log(1/\delta)$ is suboptimal.

- This is achieved with an $\delta$-adapted tuning

$$\lambda \asymp \sqrt{\frac{(d_1 + d_2)}{n}}.$$
Definition (MP)

We will say $(\mathcal{X}, \xi)$ satisfies $\text{MP}_R(f_1, f_2)$ if for all $V \in \mathbb{R}^p$,

$$\left| \langle \xi^{(n)}, \mathcal{X}^{(n)}(V) \rangle \right| \leq f_1 \|V\|_\Pi + f_2 R(V).$$

Observations:

- Via dual-norm inequality, one proves MP with

\[ f_1 = 0, \quad f_2 \approx \sqrt{\frac{d_1 + d_2 + \log(1/\delta)}{n}}. \]

Observations:

- Via dual-norm inequality, one proves \( MP \) with

\[
    f_1 = 0, \quad f_2 \propto \sqrt{\frac{d_1 + d_2 + \log(1/\delta)}{n}}.
\]

- Bellec et al (2018) proved \( MP \) with \textbf{non-null} \( f_1 \), namely,

\[
    f_1 \propto \sqrt{\frac{1 + \log(1/\delta)}{n}}, \quad f_2 \propto \sqrt{\frac{d_1 + d_2}{n}}.
\]

Observations:

- Via dual-norm inequality, one proves MP with
  \[ f_1 = 0, \quad f_2 \approx \sqrt{\frac{d_1 + d_2 + \log(1/\delta)}{n}}. \]

- Bellec et al (2018) proved MP with non-null \( f_1 \), namely,
  \[ f_1 \approx \sqrt{\frac{1 + \log(1/\delta)}{n}}, \quad f_2 \approx \sqrt{\frac{d_1 + d_2}{n}}. \]

- This is the technical reason for the previous improvement in terms of the failure probability \( \delta \).
A caveat in Bellec et al (2018)

Their approach in proving $\mathcal{MP}$ with non-null constant $f_1$ works in case $(\xi, X)$ is independent:

Define the random norm

$$\hat{T}(V) := \frac{\|X^{(n)}(V)\|_2}{L} \vee \|V\|_N$$

with $L := \sqrt{n/\log(1/\delta)}/\sigma$. In case $X$ is fixed, they bound the multiplier process concentrating the linear process

$$\sup_{V \in B_{\hat{T}}} \langle \xi^{(n)}, V \rangle = \frac{1}{\sqrt{n}} \sum_{i \in [n]} \xi_i V_i.$$  

— see Proposition 9.2 in Bellec et al (2018). The proof of this elegant result follows from a simple application of a tail symmetrization-comparison argument and the gaussian concentration inequality. Peeling is not necessary — as homogeneity of norms suffices.
A sharp Multiplier Process Inequality

The *multiplier process* over functions $f \in F$ is defined as

$$M(f) := \frac{1}{n} \sum_{i \in [n]} (\xi_i f(X_i) - \mathbb{E}[\xi f(X)]) .$$

“Assume iid sample, $\xi$ is subgaussian and $F \ni f \mapsto f(X)$ is subgaussian”.

**Theorem (Multiplier process)**

*There exists universal constant $c > 0$, such that for all $f_0 \in F$, $n \geq 1$, $u \geq 1$ and $v \geq 1$, with probability at least $1 - ce^{-u/4} - ce^{-nv}$,

$$\sup_{f \in F} |M(f) - M(f_0)| \lesssim (\sqrt{v} + 1) \|\xi\|_{\psi_2} \frac{\gamma_2(F)}{\sqrt{n}}$$

$$+ \left( \sqrt{\frac{2u}{n}} + \frac{u}{n} + \sqrt{\frac{uv}{n}} \right) \|\xi\|_{\psi_2} \bar{\Delta}(F).$$
A sharp Multiplier Process Inequality

“Nothing is assumed” beyond marginal subgaussianity of \((\xi, X)\).

The “sharpness” lies in the fact that the confidence parameter \(u > 0\) does not appear in the effective dimension term.

The proof follows from generic chaining bounds, pioneered by Talagrand (for the empirical process). Precisely, we follow a method from Dirksen (2015) showing sharp bounds for the quadratic process.

NOTE: Mendelson (2014) already obtains impressive MP inequalities that holds for heavy-tailed processes. Particularizing for the subgaussian class, the confidence parameter still multiplies the “effective dimension”.
Corollary

We can prove MP with

\[ f_1 \asymp \sqrt{\frac{1 + \log(1/\delta)}{n}}, \quad f_2 \asymp \sqrt{\frac{d_1 + d_2}{n}}, \]

assuming only marginal subgaussianity of \((\xi, X)\) and \(\mathbb{E}[\xi X] = 0\).

Without independence assumption, obtain optimal rate in probability with \(\delta\)-adapted tuning.
Additive decomposition

- It is now well understood that it is hopeless to “separate two low-rank and sparse components” without imposing a identifiability condition.
Additive decomposition

- It is now well understood that it is hopeless to “separate two low-rank and sparse components” without imposing an identifiability condition.
- Since Candès et al (2011), multiple works have studied matrix decomposition or matrix completion under the notion of incoherence.
- Since Wainwright and collaborators, e.g. Agarwal et al (2012), other works use the notion of low-spikeness

\[ \|B^*\|_\infty \leq \frac{a}{\sqrt{d_1d_2}}. \]
Additive decomposition (no contamination)

- One of the issues with additive decomposition is to ensure restricted strong convexity. Agarwal et al (2012) proposes a suitable notion of RSC to ensure convergence of a penalized least-squares estimator.
Additive decomposition (no contamination)

▶ One of the issues with additive decomposition is to ensure restricted strong convexity. Agarwal et al (2012) proposes a suitable notion of RSC to ensure convergence of a penalized least-squares estimator.

▶ In the concrete applications considered, showing this property holds with high probability is easy.

▶ For instance, in multi-task learning, the design components are $\mathcal{X}_i(B) := x_i^\top B$. When $n \gtrsim d_1$, standard concentration inequalities imply $\frac{1}{n} \sum_{i=1}^n (x_i^\top b)^2 \geq c\|b\|_1^2$ for all $b \in \mathbb{R}^{d_1}$ with high probability for some absolute constant $c \in (0, 1)$. Thus, $\frac{1}{n} \|\mathcal{X}(B)\|_F^2 \geq c\|B\|_1^2$ for all $B \in \mathbb{R}^p$.\[IMPORTANT: in trace-regression, the design operator is singular when $n \leq d_1 d_2$, so ensuring restricted strong convexity under additive decomposition is harder.\]
Additive decomposition (no contamination)

- One of the issues with additive decomposition is to ensure restricted strong convexity. Agarwal et al (2012) proposes a suitable notion of RSC to ensure convergence of a penalized least-squares estimator.

- In the concrete applications considered, showing this property holds with high probability is easy.

- For instance, in multi-task learning, the design components are $\mathcal{X}_i(B) := x_i^\top B$. When $n \gtrsim d_1$, standard concentration inequalities imply
  \[ \frac{1}{n} \sum_{i=1}^n (x_i^\top b)^2 \geq c \|b\|_1^2 \]
  for all $b \in \mathbb{R}^{d_1}$ with high probability for some absolute constant $c \in (0, 1)$. Thus,
  \[ \frac{1}{n} \| \mathcal{X}(B) \|_F^2 \geq c \|B\|_1^2 \]
  for all $B \in \mathbb{R}^p$.

- **IMPORTANT:** in trace-regression, the design operator is singular when $n \leq d_1 d_2$, so ensuring restricted strong convexity under additive decomposition is harder.
Label contamination in sparse regression

In sparse linear regression with label contamination,

\[ y = X(b^*) + \sqrt{n}\theta^* + \xi, \]

with an arbitrary \( o \)-sparse vector \( \theta^* \).

Prior work has considered the estimator

\[
[\hat{b}, \hat{\theta}] \in \arg\min_{b \in \mathbb{R}^p, \theta \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle x_i, b \rangle - \sqrt{n}\theta)^2 + \lambda \|b\|_1 + \tau \|\theta\|_1.
\]

This is in fact penalized Huber regression:

\[
\hat{b} \in \arg\min_{b \in \mathbb{R}^p} \tau^2 \sum_{i=1}^{n} \Phi \left( \frac{y_i - \langle x_i, b \rangle}{\tau \sqrt{n}} \right) + \lambda \|b\|_1.
\]
Label contamination in sparse regression

- Dalalyan-Thompson (2019): identified and proved a design property (IP) enabling to show near-optimality (up to \( \log n \) and \( \log(1/\epsilon) \) factors) of sparse Huber regression.
  - Chevet’s inequality.

- Chinot (2019): removed both logs and allows heavy-tailed noise.
  - Unlike Dalalyan-Thompson (2019), allows feature-dependent noise and optimal rate in \( \delta \).
  - BUT for the oblivious model and assuming knowledge of \((s, o)\) and convexity constant.
  - Mild additional conditions: symmetric noise with positivity mass around the origin.

- PROOF METHOD: “localization + sparsity equation” (Lecué-Mendelson) instead of “M-estimation with decomposable regularizers”.

- Both works use \( \delta \)-non-adapted estimators.
Label contamination in sparse regression

- Dalalyan-Thompson (2019): identified and proved a design property (IP) enabling to show near-optimality (up to $\log n$ and $\log(1/\epsilon)$ factors) of sparse Huber regression.
  - Chevet’s inequality.

- Chinot (2019): removed both logs and allows heavy-tailed noise.
  - Unlike Dalalyan-Thompson (2019), allows feature-dependent noise and optimal rate in $\delta$.
  - BUT for the oblivious model and assuming knowledge of $(s, o)$ and convexity constant.
  - Mild additional conditions: symmetric noise with positivity mass around the origin.

- PROOF METHOD: “localization + sparsity equation” (Lecué-Mendelson) instead of “M-estimation with decomposable regularizers”.
Label contamination in sparse regression

- Dalalyan-Thompson (2019): identified and proved a design property (IP) enabling to show near-optimality (up to $\log n$ and $\log(1/\epsilon)$ factors) of sparse Huber regression.
  - Chevet’s inequality.
- Chinot (2019): removed both logs and allows heavy-tailed noise.
  - Unlike Dalalyan-Thompson (2019), allows feature-dependent noise and optimal rate in $\delta$.
  - BUT for the oblivious model and assuming knowledge of $(s, o)$ and convexity constant.
  - Mild additional conditions: symmetric noise with positivity mass around the origin.
  - PROOF METHOD: “localization + sparsity equation” (Lecué-Mendelson) instead of “M-estimation with decomposable regularizers”.
- Both works use $\delta$-non-adapted estimators.
Our estimators

**Definition (Slope norm)**

Given nonincreasing positive sequence $\mathbf{\omega} := \{\omega_i\}_{i \in [n]}$, the Slope norm at a point $\mathbf{u} \in \mathbb{R}^n$ is defined by

$$\|\mathbf{u}\|^\# := \sum_{i \in [n]} \omega_i u_i^\#,$$

where $u_1^\# \geq \ldots \geq u_n^\#$ denotes the nonincreasing rearrangement of the absolute coordinates of $\mathbf{u}$. Unless otherwise stated, $\mathbf{\omega} \in \mathbb{R}^n$ will be the sequence with coordinates $\omega_i = \sqrt{\log(An/i)}$ for some $A \geq 2$.

Our estimators

Definition (Sorted Huber-type losses)

Define the functions $\rho_1(u) := \|u\|_2$ and $\rho_2(u) := \frac{1}{2}\|u\|_2^2$ over $\mathbb{R}^n$. For $q \in \{1, 2\}$ and $\tau > 0$, let $\rho_{\tau q, \omega} : \mathbb{R}^n \rightarrow \mathbb{R}_+$ be the infimal convolution of $\rho_q$ and $\tau \| \cdot \|_\#$, i.e.,

$$\rho_{\tau q, \omega}(u) := \min_{z \in \mathbb{R}^n} \rho_q(u - z) + \tau \|z\|_\#.$$ 

Finally, define the loss

$$L_{\tau q, \omega}(B) := \rho_{\tau q, \omega}(y - \bar{x}(B)/\sqrt{n}).$$

When $\omega_1 = \ldots = \omega_n = 1$,

$$L_{\tau q, 2}(B) = \tau^2 \sum_{i=1}^n \Phi \left( \frac{y_i - \bar{x}_i(B)}{\tau \sqrt{n}} \right),$$

where $\Phi : \mathbb{R} \rightarrow \mathbb{R}$ is the Huber’s function.
Our estimators

In case of **matrix decomposition**,

$$[\hat{B}, \hat{\Gamma}] \in \arg\min_{[B, \Gamma] \in (\mathbb{R}^p)^2} \mathcal{L}_{\tau \omega, 2}(B + \Gamma) + \lambda \mathcal{R}(B) + \chi \mathcal{S}(\Gamma)$$

s.t. \( \|B\|_\infty \leq a \),

or equivalently,

$$\min_{[B, \Gamma, \theta] \in (\mathbb{R}^p)^2 \times \mathbb{R}^n} \frac{1}{2n} \sum_{i=1}^{n} \left( y_i - \langle \langle X_i, B + \Gamma \rangle \rangle + \sqrt{n} \theta_i \right)^2 + \Psi(B, \Gamma, \theta)$$

s.t. \( \|B\|_\infty \leq a \),

where

$$\Psi(B, \Gamma, \theta) := \lambda \mathcal{R}(B) + \chi \mathcal{S}(\Gamma) + \tau \|\theta\|_\#.$$

Optimization: alternated proximal gradient between \( \ell_\infty \)-constrained/\( \ell_1 \)-norm (\( \ell_\infty \)-constrained soft-thresholding) and nuclear-norm (soft-thresholded SVD).
Our estimators

Desconsidering matrix decomposition, we consider, for \( q \in \{1, 2\} \), estimators of the form

\[
\hat{B} \in \arg\min_{B \in \mathbb{R}^p} \mathcal{L}_{\tau \omega, q}(B) + \lambda \mathcal{R}(B).
\]

Equivalently, for \( q = 2 \),

\[
\min_{[B, \theta] \in \mathbb{R}^p \times \mathbb{R}^n} \frac{1}{2n} \sum_{i=1}^n \left( y_i - \langle X_i, B \rangle + \sqrt{n} \theta_i \right)^2 + \lambda \mathcal{R}(B) + \tau \| \theta \|_\#,
\]

and, for \( q = 1 \),

\[
\min_{[B, \theta] \in \mathbb{R}^p \times \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^n \left( y_i - \langle X_i, B \rangle + \sqrt{n} \theta_i \right)^2 \right\}^{\frac{1}{2}} + \lambda \mathcal{R}(B) + \tau \| \theta \|_\#.
\]

When \( \theta \equiv 0 \) and \( \mathcal{R} = \| \cdot \|_1 \), the above estimator corresponds to the square-root lasso estimator.
Figure: Huber vs “Sorted” Huber losses in sparse regression: $\sqrt{\text{MSE}}$ versus $\epsilon$. 
We will need two design properties to upper bound the “perturbed multiplier process”

\[
[V, W, u] \mapsto \langle \xi^{(n)} - u, \mathcal{X}^{(n)}(V + W) \rangle = \frac{1}{n} \sum_{i=1}^{n} \xi_i \langle X_i, V + W \rangle - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_i \langle X_i, V + W \rangle.
\]

Moreover, we wish to avoid mere use of dual-norm inequalities: optimality in $\delta$. 
Upper bound
with additive decomposition, contamination & feature-dependent noise

Definition

\( (\mathcal{X}, \xi) \) satisfies \( MP_{\mathcal{R}, \mathcal{S}, \mathcal{Q}}(f_1, f_2, f_3, f_4) \) if for all \( [V, W, u] \in (\mathbb{R}^p)^2 \times \mathbb{R}^n \),

\[
|\langle \xi^{(n)}, \mathcal{X}^{(n)}(V + W) + \sqrt{n}u \rangle| \leq f_1\|[V, W, u]\|_\Pi + f_2\mathcal{R}(V) + f_3\mathcal{S}(W) + f_4\mathcal{Q}(u).
\]

\( \mathcal{X} \) satisfies \( IP_{\mathcal{R}, \mathcal{S}, \mathcal{Q}}(b_1, b_2, b_3, b_4) \) if for all \( [V, W, u] \in (\mathbb{R}^p)^2 \times \mathbb{R}^n \),

\[
|\langle u, \mathcal{X}^{(n)}(V + W) \rangle| \leq b_1\|[V, W]\|_\Pi \|u\|_2 + b_2\mathcal{R}(V)\|u\|_2 + b_3\mathcal{S}(W)\|u\|_2 + b_4\|[V, W]\|_\Pi \mathcal{Q}(u).
\]
Lower bound
with additive decomposition, contamination & feature-dependent noise

With additive decomposition and label contamination, “standard restricted convexity is not enough”. We need a lower bound in terms of a “augmented restricted convexity” for the augmented design

\[ [V, W, u] \mapsto \mathcal{M}^{(n)}(V, W, u) := \mathcal{X}^{(n)}(V + W) + u. \]

A third design property we will need is

**Definition (ARSC)**

\( \mathcal{X} \) satisfies ARSC\(_{R,S,Q}(d_1, d_2, d_3, d_4) \) if for all

\[ [V, W, u] \in (\mathbb{R}^p)^2 \times \mathbb{R}^n, \]

\[ \left\{ \left\| \mathcal{X}^{(n)}(V + W) + \sqrt{n}u \right\|_2^2 - 2\langle \mathbf{V}, \mathbf{W} \rangle \Pi \right\}^{\frac{1}{2}}_{+} \geq d_1\| [V, W, u] \|_{\Pi} - d_2 R(V) - d_3 S(W) - d_4 Q(u). \]
Lower bound
with additive decomposition, contamination & feature-dependent noise

We can show that ARSC holds if both IP and a fourth design property holds:

**Definition (PP)**

\( \mathcal{X} \) satisfies \( \text{PP}_{R,S}(c_1, c_2, c_3, c_4) \) if for all \([V, W] \in (\mathbb{R}^p)^2\),

\[
|\langle V, W \rangle_n - \langle V, W \rangle_{\Pi}| \leq c_1 \| V \|_{\Pi} \| W \|_{\Pi}
+ c_2 R(V) \| W \|_{\Pi} + c_3 \| V \|_{\Pi} S(W)
+ c_4 R(V) S(W).
\]
A sharp Product Process Inequality

The product process is defined as

$$A(f, g) := \frac{1}{n} \sum_{i \in [n]} \left\{ f(X_i)g(X_i) - \mathbb{E}f(X_i)g(X_i) \right\},$$

over two distinct classes $F$ and $G$ of measurable functions. When $F = G$, the correspondent process is often termed the quadratic process.

“Assume iid sample, and the maps $F \ni f \mapsto f(X)$ and $G \ni g \mapsto G(X)$ are subgaussian”.

Theorem (Product process)

Let $F, G$ be subclasses of $L_{\psi_2}$. There exist universal constants $c, C > 0$, such that for all $n \geq 1$ and $u \geq 1$, with probability at least $1 - e^{-u}$,

$$\sup_{(f, g) \in F \times G} |A(f, g)| \leq C \left[ \frac{\gamma_2(F)\gamma_2(G)}{n} + \Delta(F) \frac{\gamma_2(G)}{\sqrt{n}} + \Delta(G) \frac{\gamma_2(F)}{\sqrt{n}} \right] + c \sup_{(f, g) \in F \times G} \|fg - Pf\|_{\psi_1} \left( \sqrt{\frac{u}{n}} + \frac{u}{n} \right).$$
Results with additive matrix decomposition and label contamination

Theorem \((q = 2)\)

Assume that the covariate is 1-subgaussian with identity covariance and the noise is 1-subgaussian (for simplicity). Assume the low-spikeness condition

\[
\|B^*\|_\infty \leq \frac{a^*}{\sqrt{n}}.
\]

Assume that \(\epsilon \leq c < 0.5\). Assume \(B^* \in \mathbb{R}^{d_1 \times d_2}\) has rank \(r\) and \(\Gamma^*\) is \(s\)-sparse. Take tuning \(a := \frac{a^*}{\sqrt{n}}\) and

\[
\lambda \asymp \sqrt{(d_1 + d_2)/n}, \quad \chi \asymp \sqrt{\frac{\log(d_1 d_2)}{n} + \frac{a^*}{\sqrt{n}}}, \quad \tau \asymp \frac{1}{\sqrt{n}}.
\]

For any failure probability \(\delta \in (0, 1)\) assume that

\[
n \geq Cr(d_1 + d_2) + Cs \log(d_1 d_2) + C(a^*)^2 s,
\]

\[
\delta \geq \exp(-c_0 n).
\]
Results with additive matrix decomposition

Theorem ($q = 2$)

Then with probability at least $1 - \delta$, the square root of MSE is

$$\sqrt{r(d_1 + d_2)/n} + \sqrt{s \log p/n} + a^* \sqrt{s/n} + \sqrt{\log(1/\delta)/n} + \epsilon \log(1/\epsilon).$$

OBSERVATIONS:

- Rate is optimal up to log factors $\log(d_1d_2)$ and $\log(1/\epsilon)$.
- Optimal rate in $\delta$, $\delta$-uniformity and $\delta$-adaptivity.
- For simplicity, I do not present the formal rate in case of miss-specification: it holds as long as there is $[\mathbf{B}, \Gamma]$ such that

$$\frac{1}{n} \| \mathcal{X}(\mathbf{B} + \Gamma) - \mathbf{f} \|_2^2 \lesssim \sigma^2.$$

- The “machinery” required for additive matrix decomposition and label contamination imply as “corollary” the optimal rates for the particular cases of sparse linear regression and low-rank
Results with additive matrix decomposition and label contamination

(a) Fixed sparsity.

(b) Fixed rank.

**Figure:** Trace regression with additive matrix decomposition with no label contamination: plot of MSE.
Results for sparse or low-rank and label contamination

Figure: Sparse or low-rank linear regression with label contamination: plot of $\sqrt{\text{MSE}}$. 

(a) Fixed sparsity.

(b) Fixed rank.
σ-adaptive results (with no matrix decomposition)

**Theorem (q = 1)**

Assume that the covariate is 1-subgaussian (for simplicity) and the noise is σ-subgaussian. Assume that \( \epsilon \leq c < 0.5 \). Assume \( B^* \in \mathbb{R}^{d_1 \times d_2} \) is s-sparse. Take tuning

\[
\lambda \asymp \sqrt{\frac{\log p}{n}}, \quad \tau \asymp \frac{1}{\sqrt{n}}.
\]

Take \( \mathcal{R} := \| \cdot \|_1 \). For any failure probability \( \delta \in (0, 1) \) assume that

\[
n \geq Cs \log p,
\]

\[
\delta \geq \exp \left( -c_0 \frac{n}{\sigma^2} \right).
\]
\(\sigma\)-adaptive results (with no matrix decomposition)

**Theorem \((q = 1)\)**

*Then with probability at least \(1 - \delta\), the square root of MSE is*

\[
\sqrt{s \log p / n} + \sqrt{\log(1/\delta)/n} + \epsilon \log(1/\epsilon).
\]

**OBSERVATIONS:**

- Taking \(\mathcal{R} := \| \cdot \|_{\#}\), the Slope norm in \(\mathbb{R}^p\), we can obtain the optimal rate \(\sqrt{s \log(p/s)/n}\).
- Analogous bound when \(\text{rank}(B^*) \leq r\) and \(\mathcal{R} := \| \cdot \|_N\).
- Rate is optimal up to log factor \(\log(1/\epsilon)\).
- Optimal rate in \(\delta\), \(\delta\)-uniformity and \(\delta\)-adaptivity.
- **Miss-specification:** it holds as long as there is \([B, \Gamma]\) such that

\[
\frac{1}{n} \| \mathcal{X}(B) - f \|_2^2 \lesssim MSE^2(n, d_{\text{eff}}, \epsilon, \delta).
\]

- “Machinery” and proofs: adaptations of additive matrix decomposition and label contamination.
Comments on the proof

- The traditional proof for LASSO does not work.
- New design properties to handle, jointly, additive decomposition, label contamination and feature-dependent noise.
  - PP -> Product Process Inequality
  - IP -> Chevet’s Inequality
  - MP -> Multiplier Process Inequality
  - Combination of them to handle sharp dependence on regularization norms for low-rank, sparse and corruption structures.
“Possible” generalizations of this “theory” with contaminated scalar labels:

- The approach is using convex penalization.
- Several recent work has studied the optimization/statistical landscape of gradient descent methods using Burer-Monteiro matrix factorization.
- Benefit: under suitable initialization, the computation is faster and the MSE seems to improve in constants when compared to convex relaxations.
- “General idea”: non-convexity but with “hidden convexity”.
- Uses the incoherence condition.
- Is there a corresponding analysis for robust trace regression with additive decomposition?
Inference?

Outlier-robust *Generalized Linear Models* (with Robert Basset):
- Logistic-regression and general exponential families. This may include more complicated models based on this family of distributions?
- Single-index models?

Outlier-robust non-parametric least-squares with *Reproducing kernel Hilbert spaces* with decomposable regularizers.

Online trace regression?

Tensor regression?

Times Series?
THANK YOU!