Regularization methods under the small ball property

Guillaume Lecué

CNRS, centre de mathématiques appliquées, Ecole Polytechnique.

28 août 2014 - Toulouse

joint work with Shahar Mendelson

Regularization methods in learning theory

data – procedure – aims

Data : $(X_i, Y_i)_{i=1}^N$ iid couples $\sim (X, Y) \in \mathcal{X} \times \mathbb{R}$

data – procedure – aims

Data : $(X_i, Y_i)_{i=1}^N$ iid couples $\sim (X, Y) \in \mathcal{X} \times \mathbb{R}$

Model : F convex class of functions from $\mathcal{X}$ to $\mathbb{R}$

data – procedure – aims

Data : $(X_i, Y_i)_{i=1}^N$ iid couples $\sim (X, Y) \in \mathcal{X} \times \mathbb{R}$

Model : F convex class of functions from $\mathcal{X}$ to $\mathbb{R}$

Oracle : $f^* \in \operatorname{argmin}_{f\in F} \mathbb{E}(Y - f(X))^2$

data – procedure – aims

Data : $(X_i, Y_i)_{i=1}^N$ iid couples $\sim (X, Y) \in \mathcal{X} \times \mathbb{R}$

Model : F convex class of functions from $\mathcal{X}$ to $\mathbb{R}$

Oracle : $f^* \in \mathrm{argmin}_{f \in F} \mathbb{E}(Y - f(X))^2$

Aim : Construct $\hat{f}$ such that $\mathbb{E}(f^*(X) - \hat{f}(X))^2 \le \mathrm{rate}_N(F)$ is small.

data – procedure – aims

Data : $(X_i, Y_i)_{i=1}^N$ iid couples $\sim (X, Y) \in \mathcal{X} \times \mathbb{R}$

Model : F convex class of functions from $\mathcal{X}$ to $\mathbb{R}$

Oracle : $f^* \in \text{argmin}_{f \in F} \mathbb{E}(Y - f(X))^2$

Aim : Construct $\hat{f}$ such that $\mathbb{E}(f^*(X) - \hat{f}(X))^2 \leq \mathrm{rate}_N(F)$ is small.

Problem : F "large" $\Rightarrow \mathrm{rate}_N(F)$ big.

data – procedure – aims

Data : $(X_i, Y_i)_{i=1}^N$ iid couples $\sim (X, Y) \in \mathcal{X} \times \mathbb{R}$

Model : F convex class of functions from $\mathcal{X}$ to $\mathbb{R}$

Oracle : $f^* \in \mathrm{argmin}_{f\in F}\, \mathbb{E}(Y - f(X))^2$

Aim : Construct $\hat{f}$ such that $\mathbb{E}(f^*(X) - \hat{f}(X))^2 \leq \mathrm{rate}_N(F)$ is small.

Problem : F "large" $\Rightarrow \mathrm{rate}_N(F)$ big.

a priori : there exists $\|\cdot\|$ (not necessarily a norm) such that $\|f^*\|$ is small

data – procedure – aims

Data : $(X_i, Y_i)_{i=1}^N$ iid couples $\sim (X, Y) \in \mathcal{X} \times \mathbb{R}$

Model : F convex class of functions from $\mathcal{X}$ to $\mathbb{R}$

Oracle : $f^* \in \text{argmin}_{f \in F} \mathbb{E}(Y - f(X))^2$

Aim : Construct $\hat{f}$ such that $\mathbb{E}(f^*(X) - \hat{f}(X))^2 \leq \mathrm{rate}_N(F)$ is small.

Problem : F "large" $\Rightarrow \mathrm{rate}_N(F)$ big.

a priori : there exists $\|\cdot\|$ (not necessarily a norm) such that $\|f^*\|$ is small

"Truth" : $F^* = \{f \in F : \|f\| \leq \|f^*\|\} \Rightarrow \mathrm{rate}_N(F^*)$

data – procedure – aims

Data : $(X_i, Y_i)_{i=1}^N$ iid couples $\sim (X, Y) \in \mathcal{X} \times \mathbb{R}$

Model : F convex class of functions from $\mathcal{X}$ to $\mathbb{R}$

Oracle : $f^* \in \mathrm{argmin}_{f \in F} \mathbb{E}(Y - f(X))^2$

Aim : Construct $\hat{f}$ such that $\mathbb{E}(f^*(X) - \hat{f}(X))^2 \leq \mathrm{rate}_N(F)$ is small.

Problem : F "large" $\Rightarrow \mathrm{rate}_N(F)$ big.

a priori : there exists $\|\cdot\|$ (not necessarily a norm) such that $\|f^*\|$ is small

"Truth" : $F^* = \{f \in F : \|f\| \leq \|f^*\|\} \Rightarrow \mathrm{rate}_N(F^*)$

RERM : Regularized empirical risk minimization :

$$\hat{f} \in \underset{f \in F}{\mathrm{argmin}} \Big(\frac{1}{N} \sum_{i=1}^N (Y_i - f(X_i))^2 + \lambda \|f\| \Big).$$

data – procedure – aims

Data : $(X_i, Y_i)_{i=1}^N$ iid couples $\sim (X, Y) \in \mathcal{X} \times \mathbb{R}$

Model : F convex class of functions from $\mathcal{X}$ to $\mathbb{R}$

Oracle : $f^* \in \text{argmin}_{f \in F} \mathbb{E}(Y - f(X))^2$

Aim : Construct $\hat{f}$ such that $\mathbb{E}(f^*(X) - \hat{f}(X))^2 \leq \mathrm{rate}_N(F)$ is small.

Problem : F "large" $\Rightarrow \mathrm{rate}_N(F)$ big.

a priori : there exists $\|\cdot\|$ (not necessarily a norm) such that $\|f^*\|$ is small

"Truth" : $F^* = \{f \in F : \|f\| \leq \|f^*\|\} \Rightarrow \mathrm{rate}_N(F^*)$

RERM : Regularized empirical risk minimization :

$$\hat{f} \in \underset{f \in F}{\text{argmin}} \Big(\frac{1}{N} \sum_{i=1}^N (Y_i - f(X_i))^2 + \lambda \|f\| \Big).$$

Aim : $\hat{f} \in F^*$ and $\mathbb{E}(f^*(X) - \hat{f}(X))^2 \leq \mathrm{rate}_N(F^*)$

data – procedure – aims

Data : $(X_i, Y_i)_{i=1}^N$ iid couples $\sim (X, Y) \in \mathcal{X} \times \mathbb{R}$

Model : F convex class of functions from $\mathcal{X}$ to $\mathbb{R}$

Oracle : $f^* \in \mathrm{argmin}_{f \in F} \mathbb{E}(Y - f(X))^2$

Aim : Construct $\hat{f}$ such that $\mathbb{E}(f^*(X) - \hat{f}(X))^2 \leq \mathrm{rate}_N(F)$ is small.

Problem : F "large" $\Rightarrow \mathrm{rate}_N(F)$ big.

a priori : there exists $\|\cdot\|$ (not necessarily a norm) such that $\|f^*\|$ is small

"Truth" : $F^* = \{f \in F : \|f\| \leq \|f^*\|\} \Rightarrow \mathrm{rate}_N(F^*)$

RERM : Regularized empirical risk minimization :

$$\hat{f} \in \underset{f \in F}{\mathrm{argmin}} \Big(\frac{1}{N} \sum_{i=1}^N (Y_i - f(X_i))^2 + \lambda \|f\| \Big).$$

Aim : $\hat{f} \in F^*$ and $\mathbb{E}(f^*(X) - \hat{f}(X))^2 \leq \mathrm{rate}_N(F^*)$

Problem : find the regularization parameter λ and $\mathrm{rate}_N(F^*)$?

data – procedure – aims

Data : $(X_i, Y_i)_{i=1}^N$ iid couples $\sim (X, Y) \in \mathcal{X} \times \mathbb{R}$

Model : F convex class of functions from $\mathcal{X}$ to $\mathbb{R}$

Oracle : $f^* \in \operatorname{argmin}_{f \in F} \mathbb{E}(Y - f(X))^2$

Aim : Construct $\hat{f}$ such that $\mathbb{E}(f^*(X) - \hat{f}(X))^2 \leq \mathrm{rate}_N(F)$ is small.

Problem : F "large" $\Rightarrow \mathrm{rate}_N(F)$ big.

a priori : there exists $\|\cdot\|$ (not necessarily a norm) such that $\|f^*\|$ is small

"Truth" : $F^* = \{f \in F : \|f\| \leq \|f^*\|\} \Rightarrow \mathrm{rate}_N(F^*)$

RERM : Regularized empirical risk minimization :

$$\hat{f} \in \operatorname*{argmin}_{f \in F} \left(\frac{1}{N} \sum_{i=1}^{N} (Y_i - f(X_i))^2 + \lambda \|f\| \right).$$

Aim : $\hat{f} \in F^*$ and $\mathbb{E}(f^*(X) - \hat{f}(X))^2 \leq \mathrm{rate}_N(F^*)$

Problem : find the regularization parameter λ and $\mathrm{rate}_N(F^*)$?

Remark : No Statistical model !

Assumptions on the regularization function $\|\cdot\|$

1. $\|f+g\|, \|f-g\| \leq \eta_1 \big(\|f\| + \|g\|\big)$

Assumptions on the regularization function $\|\cdot\|$

1. $\|f+g\|, \|f-g\| \leq \eta_1 \big(\|f\| + \|g\|\big)$
2. $$\left\{ \begin{array}{ccc} [0,1] & \to & [0,\|f\|] \\ \mu & \mapsto & \|\mu f\| \end{array} \right. \quad \text{is continuous and } \leq \mu\|f\|.$$

Assumptions on the regularization function $\|\cdot\|$

1. $\|f+g\|, \|f-g\| \leq \eta_1\big(\|f\|+\|g\|\big)$
2. $$\left\{ \begin{array}{ccc} [0,1] & \to & [0,\|f\|] \\ \mu & \mapsto & \|\mu f\| \end{array} \right. \text{ is continuous and } \leq \mu\|f\|.$$

Examples :

1) Norms : ℓ_p^d, $w\ell_p^d$, S_p, RKHS-norms, sum of norms

Assumptions on the regularization function $\|\cdot\|$

1. $\|f+g\|, \|f-g\| \le \eta_1 \big(\|f\| + \|g\|\big)$
2. $$\left\{ \begin{array}{ccc} [0,1] & \to & [0,\|f\|] \\ \mu & \mapsto & \|\mu f\| \end{array} \right. \quad \text{is continuous and } \le \mu\|f\|.$$

Examples :

1) Norms : ℓ_p^d, $w\ell_p^d$, S_p, RKHS-norms, sum of norms
2) Quasi-norms : $\ell_p^d, S_p, w\ell_p^d$, $0 < p \le 1$

Assumptions on the regularization function $\|\cdot\|$

1. $\|f+g\|, \|f-g\| \leq \eta_1\left(\|f\|+\|g\|\right)$
2. $$\left\{\begin{array}{ccc} [0,1] & \to & [0,\|f\|] \\ \mu & \mapsto & \|\mu f\| \end{array}\right. \text{ is continuous and } \leq \mu\|f\|.$$

Examples :

1) Norms : ℓ_p^d, $w\ell_p^d$, S_p, RKHS-norms, sum of norms
2) Quasi-norms : $\ell_p^d, S_p, w\ell_p^d$, $0 < p \leq 1$
3) semi-norms : $\|A\cdot\|_1$ (TV)

Assumptions on the regularization function $\|\cdot\|$

1. $\|f+g\|, \|f-g\| \leq \eta_1\big(\|f\|+\|g\|\big)$
2. $$\left\{\begin{array}{ccc} [0,1] & \to & [0,\|f\|] \\ \mu & \mapsto & \|\mu f\| \end{array}\right. \text{ is continuous and } \leq \mu\|f\|.$$

Examples :

1) Norms : ℓ_p^d, $w\ell_p^d$, S_p, RKHS-norms, sum of norms
2) Quasi-norms : $\ell_p^d, S_p, w\ell_p^d$, $0 < p \leq 1$
3) semi-norms : $\|A\cdot\|_1$ (TV)
4) square of norms (elastic-net)

Assumptions on the regularization function $\|\cdot\|$

1. $\|f+g\|, \|f-g\| \leq \eta_1 \big(\|f\| + \|g\|\big)$
2. $$\left\{ \begin{array}{ccc} [0,1] & \to & [0, \|f\|] \\ \mu & \mapsto & \|\mu f\| \end{array} \right. \text{ is continuous and } \leq \mu\|f\|.$$

Examples :

1) Norms : ℓ_p^d, $w\ell_p^d$, S_p, RKHS-norms, sum of norms
2) Quasi-norms : $\ell_p^d, S_p, w\ell_p^d$, $0 < p \leq 1$
3) semi-norms : $\|A\cdot\|_1$ (TV)
4) square of norms (elastic-net)

Does not work for ℓ_0^d and $\mathrm{rank}(\cdot)$.

Two empirical processes – two conditions

quadratic/linear decomposition of the excess loss

$$(y - f)^2 - (y - f^*)^2 = (f - f^*)^2 - 2(y - f^*)(f - f^*).$$

Two empirical processes – two conditions

quadratic/linear decomposition of the excess loss

$$(y-f)^2-(y-f^*)^2=(f-f^*)^2-2(y-f^*)(f-f^*).$$

1. Empirical small ball condition on $F_\rho=\{f\in F:\|f\|\leq\rho\}$:

Two empirical processes – two conditions

quadratic/linear decomposition of the excess loss

$$(y-f)^2-(y-f^*)^2=(f-f^*)^2-2(y-f^*)(f-f^*).$$

1. Empirical small ball condition on $F_\rho=\{f\in F:\|f\|\leq\rho\}$: $\forall f\in F_\rho$

$$\|f-f^*\|_{L_2}\geq s_Q(\rho)\Rightarrow P_N(f-f^*)^2\geq\kappa_0\|f-f^*\|_{L_2}^2$$

where $P_N h=N^{-1}\sum_{i=1}^N h(X_i)$.

Two empirical processes – two conditions

quadratic/linear decomposition of the excess loss

$$(y - f)^2 - (y - f^*)^2 = (f - f^*)^2 - 2(y - f^*)(f - f^*).$$

1. Empirical small ball condition on $F_\rho = \{f \in F : \|f\| \leq \rho\}$: $\forall f \in F_\rho$

$$\|f - f^*\|_{L_2} \geq s_Q(\rho) \Rightarrow P_N(f - f^*)^2 \geq \kappa_0 \|f - f^*\|_{L_2}^2$$

   where $P_N h = N^{-1} \sum_{i=1}^N h(X_i)$.
2. Noise/class interaction on F_ρ : $\forall f \in F_\rho$

$$P_N(Y - f^*)(f - f^*) \leq \kappa_1 \max\left(s_L(\rho)\|f - f^*\|_{L_2}, \|f - f^*\|_{L_2}^2\right)$$

Some hints about s_Q and s_L

Working in model F_ρ (such that $f^* \in F_\rho$).

Some hints about s_Q and s_L

Working in model F_ρ (such that $f^* \in F_\rho$).Consider the Empirical Risk Minimization procedure :

$$\tilde{f} \in \underset{f \in F_\rho}{\operatorname{argmin}} P_N(Y - f)^2$$

Some hints about s_Q and s_L

Working in model F_ρ (such that $f^* \in F_\rho$).Consider the Empirical Risk Minimization procedure :

$$\tilde{f} \in \operatorname*{argmin}_{f \in F_\rho} P_N(Y - f)^2$$

so that $H_N(\tilde{f}) = P_N\big[(Y - \tilde{f})^2 - (Y - f^*)\big] \leq 0.$

Some hints about s_Q and s_L

Working in model F_ρ (such that $f^* \in F_\rho$).Consider the Empirical Risk Minimization procedure :

$$\tilde{f} \in \underset{f \in F_\rho}{\operatorname{argmin}} P_N(Y - f)^2$$

so that $H_N(\tilde{f}) = P_N\left[(Y - \tilde{f})^2 - (Y - f^*)\right] \leq 0.$

$$H_N(f) = P_N(f - f^*)^2 - 2P_N(Y - f^*)(f - f^*).$$

Some hints about s_Q and s_L

Working in model F_ρ (such that $f^* \in F_\rho$).Consider the Empirical Risk Minimization procedure :

$$\tilde{f} \in \operatorname*{argmin}_{f \in F_\rho} P_N(Y - f)^2$$

so that $H_N(\tilde{f}) = P_N\left[(Y - \tilde{f})^2 - (Y - f^*)\right] \leq 0$.

$$H_N(f) = P_N(f - f^*)^2 - 2P_N(Y - f^*)(f - f^*).$$

Under the [small ball] and [noise/class] assumptions :

Some hints about s_Q and s_L

Working in model F_ρ (such that $f^* \in F_\rho$).Consider the Empirical Risk Minimization procedure :

$$\tilde{f} \in \underset{f \in F_\rho}{\operatorname{argmin}} P_N (Y - f)^2$$

so that $H_N(\tilde{f}) = P_N\left[(Y - \tilde{f})^2 - (Y - f^*)\right] \leq 0$.

$$H_N(f) = P_N(f - f^*)^2 - 2P_N(Y - f^*)(f - f^*).$$

Under the [small ball] and [noise/class] assumptions :

$$\text{if } \|f - f^*\|_{L_2} \geq s(\rho) = \max(s_L(\rho), s_Q(\rho)) \text{ then}$$

1. $P_N(f - f^*)^2 > \kappa_0 \|f - f^*\|_{L_2}^2$

Some hints about s_Q and s_L

Working in model F_ρ (such that $f^* \in F_\rho$).Consider the Empirical Risk Minimization procedure :

$$\tilde{f} \in \underset{f \in F_\rho}{\operatorname{argmin}} P_N(Y - f)^2$$

so that $H_N(\tilde{f}) = P_N\left[(Y - \tilde{f})^2 - (Y - f^*)\right] \leq 0.$

$$H_N(f) = P_N(f - f^*)^2 - 2P_N(Y - f^*)(f - f^*).$$

Under the [small ball] and [noise/class] assumptions :

$$\text{if } \|f - f^*\|_{L_2} \geq s(\rho) = \max(s_L(\rho), s_Q(\rho)) \text{ then}$$

1. $P_N(f - f^*)^2 > \kappa_0 \|f - f^*\|_{L_2}^2$
2. $2P_N(Y - f^*)(f - f^*) < 2\kappa_1 \|f - f^*\|_{L_2}^2$

Some hints about s_Q and s_L

Working in model F_ρ (such that $f^* \in F_\rho$).Consider the Empirical Risk Minimization procedure :

$$\tilde{f} \in \operatorname*{argmin}_{f \in F_\rho} P_N(Y - f)^2$$

so that $H_N(\tilde{f}) = P_N\big[(Y - \tilde{f})^2 - (Y - f^*)\big] \leq 0$.

$$H_N(f) = P_N(f - f^*)^2 - 2P_N(Y - f^*)(f - f^*).$$

Under the [small ball] and [noise/class] assumptions :

$$\text{if } \|f - f^*\|_{L_2} \geq s(\rho) = \max(s_L(\rho), s_Q(\rho)) \text{ then}$$

1. $P_N(f - f^*)^2 > \kappa_0 \|f - f^*\|_{L_2}^2$
2. $2P_N(Y - f^*)(f - f^*) < 2\kappa_1 \|f - f^*\|_{L_2}^2$

$\Rightarrow H_N(f) > 0$ (when $2\kappa_1 < \kappa_0$) so (since $H_N(\tilde{f}) \leq 0$),

$$\|\tilde{f} - f^*\|_{L_2} \leq s(\rho).$$

Choice of λ – main result

$$\lambda \gtrsim \sup_{\rho>0} \sup_{f \in F_\rho} \frac{P_N(Y - f^*)(f - f^*)}{\rho}. \quad (1)$$

Choice of λ – main result

$$\lambda \gtrsim \sup_{\rho>0} \sup_{f \in F_\rho} \frac{P_N(Y - f^*)(f - f^*)}{\rho}. \qquad (1)$$

Theorem (L. & Mendelson)

If the following are true :

1. *empirical small ball property in F_{ρ^*} for $\rho^* \sim \|f^*\|$, with level $s_Q(\rho^*)$*

Choice of λ – main result

$$\lambda \gtrsim \sup_{\rho>0} \sup_{f \in F_\rho} \frac{P_N(Y - f^*)(f - f^*)}{\rho}. \tag{1}$$

Theorem (L. & Mendelson)

If the following are true :

1. *empirical small ball property in F_{ρ^*} for $\rho^* \sim \|f^*\|$, with level $s_Q(\rho^*)$*
2. *noise/class interaction of level $s_L(\rho^*)$ in F_{ρ^*}*

Choice of λ – main result

$$\lambda \gtrsim \sup_{\rho>0} \sup_{f \in F_\rho} \frac{P_N(Y-f^*)(f-f^*)}{\rho}. \tag{1}$$

Theorem (L. & Mendelson)

If the following are true :

1. *empirical small ball property in F_{ρ^*} for $\rho^* \sim \|f^*\|$, with level $s_Q(\rho^*)$*
2. *noise/class interaction of level $s_L(\rho^*)$ in F_{ρ^*}*
3. *λ satisfies (1).*

Choice of λ – main result

$$\lambda \gtrsim \sup_{\rho>0} \sup_{f \in F_\rho} \frac{P_N(Y - f^*)(f - f^*)}{\rho}. \qquad (1)$$

Theorem (L. & Mendelson)

If the following are true :

1. *empirical small ball property in F_{ρ^*} for $\rho^* \sim \|f^*\|$, with level $s_Q(\rho^*)$*
2. *noise/class interaction of level $s_L(\rho^*)$ in F_{ρ^*}*
3. *λ satisfies (1).*

Then :

$$\|\hat{f}\| \lesssim \|f^*\|$$

Choice of λ – main result

$$\lambda \gtrsim \sup_{\rho>0} \sup_{f\in F_\rho} \frac{P_N(Y-f^*)(f-f^*)}{\rho}. \tag{1}$$

Theorem (L. & Mendelson)

If the following are true :

1. *empirical small ball property in F_{ρ^*} for $\rho^* \sim \|f^*\|$, with level $s_Q(\rho^*)$*
2. *noise/class interaction of level $s_L(\rho^*)$ in F_{ρ^*}*
3. *λ satisfies (1).*

Then :

$$\|\hat{f}\| \lesssim \|f^*\| \text{ and } \|\hat{f}-f^*\|_{L_2(X)} \lesssim s(\|f^*\|)$$

Choice of λ – main result

$$\lambda \gtrsim \sup_{\rho>0} \sup_{f\in F_\rho} \frac{P_N(Y-f^*)(f-f^*)}{\rho}. \tag{1}$$

Theorem (L. & Mendelson)

If the following are true :

1. *empirical small ball property in F_{ρ^*} for $\rho^* \sim \|f^*\|$, with level $s_Q(\rho^*)$*
2. *noise/class interaction of level $s_L(\rho^*)$ in F_{ρ^*}*
3. *λ satisfies (1).*

Then :

$$\|\hat{f}\| \lesssim \|f^*\| \text{ and } \|\hat{f}-f^*\|_{L_2(X)} \lesssim s(\|f^*\|)$$

where

$$s(\rho) = \max\Big(s_L(\rho), s_Q(\rho), \lambda\rho\Big).$$

Choice of λ – main result

$$\lambda \gtrsim \sup_{\rho>0} \sup_{f\in F_\rho} \frac{P_N(Y-f^*)(f-f^*)}{\rho}. \qquad (1)$$

Theorem (L. & Mendelson)

If the following are true :

1. *empirical small ball property in F_{ρ^*} for $\rho^* \sim \|f^*\|$, with level $s_Q(\rho^*)$*
2. *noise/class interaction of level $s_L(\rho^*)$ in F_{ρ^*}*
3. *λ satisfies (1).*

Then :

$$\|\hat{f}\| \lesssim \|f^*\| \text{ and } \|\hat{f}-f^*\|_{L_2(X)} \lesssim s(\|f^*\|)$$

where

$$s(\rho) = \max\Big(s_L(\rho), s_Q(\rho), \lambda\rho\Big).$$

Rem : For many model, $\max(s_L(\rho), s_Q(\rho))$ is the minimax rate of convergence in F_ρ.

Checking the empirical small ball property

H satisfies an empirical small ball property when $\forall h \in H$,

$$\|h\|_{L_2} \geq s_Q \Rightarrow P_N h^2 \geq \kappa_0 P h^2.$$

Checking the empirical small ball property

H satisfies an empirical small ball property when $\forall h \in H$,

$$\|h\|_{L_2} \geq s_Q \Rightarrow P_N h^2 \geq \kappa_0 P h^2.$$

This holds with probability at least $1 - e^{-N}$, when :

Checking the empirical small ball property

H satisfies an empirical small ball property when $\forall h \in H$,

$$\|h\|_{L_2} \geq s_Q \Rightarrow P_N h^2 \geq \kappa_0 P h^2.$$

This holds with probability at least $1 - e^{-N}$, when :

1. Small ball property[Mendelson, "learning without concentration"] : there exists u_0, β_0 s.t. $\forall h \in H$,

$$P\left[|h(X)| \geq u_0 \|h\|_{L_2(X)}\right] \geq \beta_0.$$

Checking the empirical small ball property

H satisfies an empirical small ball property when $\forall h \in H$,

$$\|h\|_{L_2} \geq s_Q \Rightarrow P_N h^2 \geq \kappa_0 P h^2.$$

This holds with probability at least $1 - e^{-N}$, when :

1. Small ball property[Mendelson, "learning without concentration"] : there exists u_0, β_0 s.t. $\forall h \in H$,

$$P\left[|h(X)| \geq u_0 \|h\|_{L_2(X)}\right] \geq \beta_0.$$

(holds if $\|h\|_{2+\epsilon} \lesssim \|h\|_2$)

Checking the empirical small ball property

H satisfies an empirical small ball property when $\forall h \in H$,

$$\|h\|_{L_2} \geq s_Q \Rightarrow P_N h^2 \geq \kappa_0 P h^2.$$

This holds with probability at least $1 - e^{-N}$, when :

1. Small ball property[Mendelson, "learning without concentration"] : there exists u_0, β_0 s.t. $\forall h \in H$,

$$P\left[|h(X)| \geq u_0 \|h\|_{L_2(X)}\right] \geq \beta_0.$$

   (holds if $\|h\|_{2+\epsilon} \lesssim \|h\|_2$)
2. fixed point equation :

$$\mathbb{E} \sup_{h \in H : \|h\|_{L_2} \leq s_Q} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i h(X_i) \right| \lesssim s_Q.$$

Checking the empirical small ball property

H satisfies an empirical small ball property when $\forall h \in H$,

$$\|h\|_{L_2} \geq s_Q \Rightarrow P_N h^2 \geq \kappa_0 P h^2.$$

This holds with probability at least $1 - e^{-N}$, when :

1. Small ball property[Mendelson, "learning without concentration"] : there exists u_0, β_0 s.t. $\forall h \in H$,

$$P\left[|h(X)| \geq u_0 \|h\|_{L_2(X)}\right] \geq \beta_0.$$

   (holds if $\|h\|_{2+\epsilon} \lesssim \|h\|_2$)

2. fixed point equation :

$$\mathbb{E} \sup_{h \in H: \|h\|_{L_2} \leq s_Q} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i h(X_i) \right| \lesssim s_Q.$$

Rem. : The control on the linear process is more standard.

Learning linear functional by regularization methods

setup – learning linear functional

$F = \{\langle \cdot, t \rangle : t \in \mathcal{H}\}$ where $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is a Hilbert space like $\mathcal{H} \in \{\mathbb{R}^d, \mathbb{R}^{m \times T}, RKHS\}$.

setup – learning linear functional

$F = \{\langle \cdot, t \rangle : t \in \mathcal{H}\}$ where $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is a Hilbert space like $\mathcal{H} \in \{\mathbb{R}^d, \mathbb{R}^{m \times T}, RKHS\}$. We want to estimate

$$t^* \in \underset{t \in \mathcal{H}}{\operatorname{argmin}} \, \mathbb{E}\big(Y - \langle X, t \rangle\big)^2$$

setup – learning linear functional

$F = \{\langle \cdot, t\rangle : t \in \mathcal{H}\}$ where $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is a Hilbert space like $\mathcal{H} \in \{\mathbb{R}^d, \mathbb{R}^{m\times T}, RKHS\}$. We want to estimate

$$t^* \in \underset{t\in\mathcal{H}}{\operatorname{argmin}}\, \mathbb{E}\big(Y - \langle X, t\rangle\big)^2$$

by the RERM

$$\hat{t} \in \underset{t\in\mathcal{H}}{\operatorname{argmin}} \Big(\frac{1}{N}\sum_{i=1}^{N}\big(Y_i - \langle X_i, t\rangle\big)^2 + \lambda\|t\|\Big).$$

setup – learning linear functional

$F = \{\langle \cdot, t\rangle : t \in \mathcal{H}\}$ where $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is a Hilbert space like $\mathcal{H} \in \{\mathbb{R}^d, \mathbb{R}^{m\times T}, RKHS\}$. We want to estimate

$$t^* \in \underset{t\in\mathcal{H}}{\operatorname{argmin}}\, \mathbb{E}\big(Y - \langle X, t\rangle\big)^2$$

by the RERM

$$\hat{t} \in \underset{t\in\mathcal{H}}{\operatorname{argmin}} \left(\frac{1}{N}\sum_{i=1}^{N}\big(Y_i - \langle X_i, t\rangle\big)^2 + \lambda\|t\|\right).$$

Aim :

$$\|\hat{t}\| \lesssim \|t^*\|$$

setup – learning linear functional

$F = \{\langle \cdot, t\rangle : t \in \mathcal{H}\}$ where $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ is a Hilbert space like $\mathcal{H} \in \{\mathbb{R}^d, \mathbb{R}^{m\times T}, RKHS\}$. We want to estimate

$$t^* \in \underset{t\in\mathcal{H}}{\operatorname{argmin}}\, \mathbb{E}\big(Y - \langle X, t\rangle\big)^2$$

by the RERM

$$\hat{t} \in \underset{t\in\mathcal{H}}{\operatorname{argmin}} \Big(\frac{1}{N}\sum_{i=1}^{N}\big(Y_i - \langle X_i, t\rangle\big)^2 + \lambda\|t\|\Big).$$

Aim :

$$\|\hat{t}\| \lesssim \|t^*\| \text{ and } \big(\mathbb{E}\langle X, \hat{t} - t^*\rangle^2\big)^2 \lesssim s(\|t^*\|).$$

General result when the design is sub-gaussian and noise in L_q, $q > 2$

Assume that :

1. X is L-sub-gaussian : $P[|\langle X, t\rangle| \geq Lu\|\langle X, t\rangle\|_{L_2}] \leq 2\exp(-u^2)$,

General result when the design is sub-gaussian and noise in L_q, $q > 2$

Assume that :

1. X is L-sub-gaussian : $P[|\langle X, t\rangle| \geq Lu\|\langle X, t\rangle\|_{L_2}] \leq 2\exp(-u^2)$,
2. "noise" in L_q (q>2), $\sigma_q = \|Y - \langle X, t^*\rangle\|_{L_q} < \infty$.

General result when the design is sub-gaussian and noise in L_q, $q > 2$

Assume that :

1. X is L-sub-gaussian : $P[|\langle X, t\rangle| \geq Lu\|\langle X, t\rangle\|_{L_2}] \leq 2\exp(-u^2)$,
2. "noise" in L_q (q>2), $\sigma_q = \|Y - \langle X, t^*\rangle\|_{L_q} < \infty$.

Denote $B_{\|\cdot\|} = \{t \in \mathcal{H} : \|t\| \leq 1\}$ and $\ell^*(B_{\|\cdot\|}) = \mathbb{E}\sup_{t \in B_{\|\cdot\|}}\langle G, t\rangle$.

General result when the design is sub-gaussian and noise in L_q, $q > 2$

Assume that :

1. X is L-sub-gaussian : $P[|\langle X, t\rangle| \geq Lu\|\langle X, t\rangle\|_{L_2}] \leq 2\exp(-u^2)$,
2. "noise" in L_q (q>2), $\sigma_q = \|Y - \langle X, t^*\rangle\|_{L_q} < \infty$.

Denote $B_{\|\cdot\|} = \{t \in \mathcal{H} : \|t\| \leq 1\}$ and $\ell^*(B_{\|\cdot\|}) = \mathbb{E}\sup_{t\in B_{\|\cdot\|}}\langle G, t\rangle$.
Then, with probability larger than $1 - 2\exp(-N) - (c/u)^q$,

$$\hat{t} \in \underset{t\in\mathcal{H}}{\operatorname{argmin}} \Big(\frac{1}{N}\sum_{i=1}^N \big(Y_i - \langle X_i, t\rangle\big)^2 + u\sigma_q\|t\|\frac{\ell^*(B_{\|\cdot\|})}{\sqrt{N}}\Big)$$

General result when the design is sub-gaussian and noise in L_q, $q > 2$

Assume that :

1. X is L-sub-gaussian : $P[|\langle X, t\rangle| \geq Lu\|\langle X, t\rangle\|_{L_2}] \leq 2\exp(-u^2)$,
2. "noise" in L_q (q>2), $\sigma_q = \|Y - \langle X, t^*\rangle\|_{L_q} < \infty$.

Denote $B_{\|\cdot\|} = \{t \in \mathcal{H} : \|t\| \leq 1\}$ and $\ell^*(B_{\|\cdot\|}) = \mathbb{E}\sup_{t\in B_{\|\cdot\|}}\langle G, t\rangle$.
Then, with probability larger than $1 - 2\exp(-N) - (c/u)^q$,

$$\hat{t} \in \underset{t\in\mathcal{H}}{\operatorname{argmin}}\Big(\frac{1}{N}\sum_{i=1}^N \big(Y_i - \langle X_i, t\rangle\big)^2 + u\sigma_q\|t\|\frac{\ell^*(B_{\|\cdot\|})}{\sqrt{N}}\Big)$$

is such that

$$\|\hat{t}\| \lesssim \|t^*\|$$

General result when the design is sub-gaussian and noise in L_q, $q > 2$

Assume that :

1. X is L-sub-gaussian : $P[|\langle X, t\rangle| \geq Lu\|\langle X, t\rangle\|_{L_2}] \leq 2\exp(-u^2)$,
2. "noise" in L_q (q>2), $\sigma_q = \|Y - \langle X, t^*\rangle\|_{L_q} < \infty$.

Denote $B_{\|\cdot\|} = \{t \in \mathcal{H} : \|t\| \leq 1\}$ and $\ell^*(B_{\|\cdot\|}) = \mathbb{E}\sup_{t\in B_{\|\cdot\|}}\langle G, t\rangle$.
Then, with probability larger than $1 - 2\exp(-N) - (c/u)^q$,

$$\hat{t} \in \underset{t\in\mathcal{H}}{\operatorname{argmin}}\Big(\frac{1}{N}\sum_{i=1}^{N}\big(Y_i - \langle X_i, t\rangle\big)^2 + u\sigma_q\|t\|\frac{\ell^*(B_{\|\cdot\|})}{\sqrt{N}}\Big)$$

is such that

$$\|\hat{t}\| \lesssim \|t^*\| \text{ and } \|\langle X, \hat{t} - t^*\rangle\|_{L_2} \lesssim s(\|t^*\|)$$

where

General result when the design is sub-gaussian and noise in L_q, $q > 2$

Assume that :

1. X is L-sub-gaussian : $P[|\langle X, t\rangle| \geq Lu\|\langle X, t\rangle\|_{L_2}] \leq 2\exp(-u^2)$,
2. "noise" in L_q (q>2), $\sigma_q = \|Y - \langle X, t^*\rangle\|_{L_q} < \infty$.

Denote $B_{\|\cdot\|} = \{t \in \mathcal{H} : \|t\| \leq 1\}$ and $\ell^*(B_{\|\cdot\|}) = \mathbb{E}\sup_{t\in B_{\|\cdot\|}}\langle G, t\rangle$.
Then, with probability larger than $1 - 2\exp(-N) - (c/u)^q$,

$$\hat{t} \in \underset{t\in\mathcal{H}}{\operatorname{argmin}}\left(\frac{1}{N}\sum_{i=1}^{N}\big(Y_i - \langle X_i, t\rangle\big)^2 + u\sigma_q\|t\|\frac{\ell^*(B_{\|\cdot\|})}{\sqrt{N}}\right)$$

is such that

$$\|\hat{t}\| \lesssim \|t^*\| \text{ and } \|\langle X, \hat{t} - t^*\rangle\|_{L_2} \lesssim s(\|t^*\|)$$

where

$$s^2(\rho) = \max\left(u\sigma_q\rho\frac{\ell^*(B_{\|\cdot\|})}{\sqrt{N}}, \rho^2\frac{\ell^*(B_{\|\cdot\|})^2}{N}\right)$$

Examples of Gaussian mean widths in $\mathbb{R}^d$

1. $B_p^d = \{t \in \mathbb{R}^d : \|t\|_p \leq 1\}$, $0 < p \leq \infty$,

$$\ell^*(B_p^d) \sim \begin{cases} \sqrt{\log(ed)} & \text{when } 0 < p < 1 + (\log(ed))^{-1} \\ \sqrt{q} n^{1/q} & \text{when } p \geq 1 + (\log(ed))^{-1} \end{cases}$$

Examples of Gaussian mean widths in $\mathbb{R}^d$

1. $B_p^d = \{t \in \mathbb{R}^d : \|t\|_p \leq 1\}$, $0 < p \leq \infty$,

$$\ell^*(B_p^d) \sim \begin{cases} \sqrt{\log(ed)} & \text{when } 0 < p < 1 + (\log(ed))^{-1} \\ \sqrt{q} n^{1/q} & \text{when } p \geq 1 + (\log(ed))^{-1} \end{cases}$$

for $p = 1$ we recover the rates of estimation of the Lasso in [Koltchinksi]

Examples of Gaussian mean widths in $\mathbb{R}^d$

1. $B_p^d = \{t \in \mathbb{R}^d : \|t\|_p \leq 1\}$, $0 < p \leq \infty$,

$$\ell^*(B_p^d) \sim \begin{cases} \sqrt{\log(ed)} & \text{when } 0 < p < 1 + (\log(ed))^{-1} \\ \sqrt{q} n^{1/q} & \text{when } p \geq 1 + (\log(ed))^{-1} \end{cases}$$

   for $p = 1$ we recover the rates of estimation of the Lasso in [Koltchinksi]

2. $wB_{p\infty}^d = \{t : t_j^* \leq j^{-1/p}\}$, where $t_1^* \geq \cdots \geq t_d^*$,

$$\ell^*(B_{p\infty}) \lesssim \begin{cases} \sqrt{\log(ed)} & \text{when } 0 < p < 1 \\ (\log(ed))^{3/2} & \text{when } p = 1 \end{cases}$$

Examples of Gaussian mean widths in $\mathbb{R}^d$

1. $B_p^d = \{t \in \mathbb{R}^d : \|t\|_p \leq 1\}$, $0 < p \leq \infty$,

$$\ell^*(B_p^d) \sim \begin{cases} \sqrt{\log(ed)} & \text{when } 0 < p < 1 + (\log(ed))^{-1} \\ \sqrt{q}n^{1/q} & \text{when } p \geq 1 + (\log(ed))^{-1} \end{cases}$$

   for $p = 1$ we recover the rates of estimation of the Lasso in [Koltchinksi]

2. $wB_{p\infty}^d = \{t : t_j^* \leq j^{-1/p}\}$, where $t_1^* \geq \cdots \geq t_d^*$,

$$\ell^*(B_{p\infty}) \lesssim \begin{cases} \sqrt{\log(ed)} & \text{when } 0 < p < 1 \\ (\log(ed))^{3/2} & \text{when } p = 1 \end{cases}$$

3. Michelli, Morales and Pontil norms : Θ a convex cone in $(0, \infty)^d$,

$$\Omega(t|\Theta) = \inf_{\theta \in \Theta} \frac{1}{2} \sum_{j=1}^{d} \left(\frac{t_j^2}{\theta_j} + \theta_j \right)$$

   is a norm

Examples of Gaussian mean widths in $\mathbb{R}^d$

1. $B_p^d = \{t \in \mathbb{R}^d : \|t\|_p \leq 1\}$, $0 < p \leq \infty$,

$$\ell^*(B_p^d) \sim \begin{cases} \sqrt{\log(ed)} & \text{when } 0 < p < 1 + (\log(ed))^{-1} \\ \sqrt{q} n^{1/q} & \text{when } p \geq 1 + (\log(ed))^{-1} \end{cases}$$

for $p = 1$ we recover the rates of estimation of the Lasso in [Koltchinksi]

2. $wB_{p\infty}^d = \{t : t_j^* \leq j^{-1/p}\}$, where $t_1^* \geq \cdots \geq t_d^*$,

$$\ell^*(B_{p\infty}) \lesssim \begin{cases} \sqrt{\log(ed)} & \text{when } 0 < p < 1 \\ (\log(ed))^{3/2} & \text{when } p = 1 \end{cases}$$

3. Michelli, Morales and Pontil norms : Θ a convex cone in $(0, \infty)^d$,

$$\Omega(t|\Theta) = \inf_{\theta \in \Theta} \frac{1}{2} \sum_{j=1}^{d} \left(\frac{t_j^2}{\theta_j} + \theta_j \right)$$

is a norm and for $\mathcal{E}$: extreme points of $\Theta \cap S_1^{d-1}$

$$\ell^*(B_{\Omega(\cdot|\Theta)}) \lesssim \|\mathcal{E}\|_\infty \sqrt{\log\left(|\mathcal{E}| \|\mathcal{E}\|_\infty\right)}.$$

Examples of Gaussian mean widths in $\mathbb{R}^{m\times T}$

1. $B_{S_p} = \{A \in \mathbb{R}^{m\times T} : \sum \sigma_j(A)^p \leq 1\}$, $p > 0$,

$$\ell^*(B_{S_p}) \lesssim \begin{cases} \sqrt{m+T} & \text{when } 0 < p < 1 \\ (m \wedge T)^{1-1/p}\sqrt{m+T} & \text{when } p \geq 1 \end{cases}$$

Examples of Gaussian mean widths in $\mathbb{R}^{m\times T}$

1. $B_{S_p} = \{A \in \mathbb{R}^{m\times T} : \sum \sigma_j(A)^p \leq 1\}$, $p > 0$,

$$\ell^*(B_{S_p}) \lesssim \begin{cases} \sqrt{m+T} & \text{when } 0 < p < 1 \\ (m \wedge T)^{1-1/p}\sqrt{m+T} & \text{when } p \geq 1 \end{cases}$$

2. $\|A\|_{max} = \min_{A=UV^\top} \|U\|_{2\to\infty}\|V\|_{2\to\infty}$,

$$\ell^*(B_{max}) \leq \sqrt{mT(m+T)}.$$

[Srebo, Shraibman]

Examples of Gaussian mean widths in $\mathbb{R}^{m\times T}$

1. $B_{S_p} = \{A \in \mathbb{R}^{m\times T} : \sum \sigma_j(A)^p \leq 1\}$, $p > 0$,

$$\ell^*(B_{S_p}) \lesssim \begin{cases} \sqrt{m+T} & \text{when } 0 < p < 1 \\ (m \wedge T)^{1-1/p}\sqrt{m+T} & \text{when } p \geq 1 \end{cases}$$

2. $\|A\|_{max} = \min_{A=UV^\top} \|U\|_{2\to\infty}\|V\|_{2\to\infty}$,

$$\ell^*(B_{max}) \leq \sqrt{mT(m+T)}.$$

   [Srebo, Shraibman]

3. Atomic norm regularization $\|A\|_{\mathcal{A}} = \inf(t > 0 : A \in t\mathrm{conv}(\mathcal{A}))$ for $\mathcal{A} \subset \mathbb{R}^{m\times T}$ (atoms),

$$\ell^*(B_{\|\cdot\|_{\mathcal{A}}}) = \ell^*(\mathcal{A}).$$

   [Chandrasekaran, Recht, Parrilo, Willsky]

Gaussian mean width of the unit ball of a RKHS

$K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ symmetric kernel and the operator $(T_K f) = \int K(\cdot, x) f(x) d\mu(x)$. Denote :

1. $(\lambda_j)_j$ eigenvalues of T_K,

Gaussian mean width of the unit ball of a RKHS

$K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ symmetric kernel and the operator $(T_K f) = \int K(\cdot, x) f(x) d\mu(x)$. Denote :

1. $(\lambda_j)_j$ eigenvalues of T_K,
2. $(\phi_j)_j$ associated eigenfunctions,

Gaussian mean width of the unit ball of a RKHS

$K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ symmetric kernel and the operator $(T_K f) = \int K(\cdot, x) f(x) d\mu(x)$. Denote :

1. $(\lambda_j)_j$ eigenvalues of T_K,
2. $(\phi_j)_j$ associated eigenfunctions,
3. $\|f\|_K = \sum \sqrt{\lambda_i} \langle f, \phi_i \rangle^2$

Gaussian mean width of the unit ball of a RKHS

$K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ symmetric kernel and the operator $(T_K f) = \int K(\cdot, x) f(x) d\mu(x)$. Denote :

1. $(\lambda_j)_j$ eigenvalues of T_K,
2. $(\phi_j)_j$ associated eigenfunctions,
3. $\|f\|_K = \sum \sqrt{\lambda_i} \big\langle f, \phi_i \big\rangle^2$
4. $B_K = \big\{ \sum \sqrt{\lambda_j} \beta_j \phi_j : \|\beta\|_{\ell_2} \leq 1 \big\}$.

Then,

$$\ell^*(B_K) \sim \Big(\sum \lambda_j \Big)^{1/2}.$$

Result for the Lasso under the small ball property

Previous results when X is sub-gaussian.

Result for the Lasso under the small ball property

Previous results when X is sub-gaussian.We can weaken this assumption if we assume a statistical model :

$$Y = \left\langle X, t^* \right\rangle + \zeta$$

where ζ independent of $X \in \mathbb{R}^d$.

Result for the Lasso under the small ball property

Previous results when X is sub-gaussian.We can weaken this assumption if we assume a statistical model :

$$Y = \langle X, t^* \rangle + \zeta$$

where ζ independent of $X \in \mathbb{R}^d$.We assume :

1. a small ball property : there exists u_0, β_0 such that

$$P\left[|\langle X, t \rangle| \geq u_0 \left(\mathbb{E}\langle X, t \rangle^2 \right)^{1/2} \right] \geq \beta_0,$$

Result for the Lasso under the small ball property

Previous results when X is sub-gaussian.We can weaken this assumption if we assume a statistical model :

$$Y = \langle X, t^* \rangle + \zeta$$

where ζ independent of $X \in \mathbb{R}^d$.We assume :

1. a small ball property : there exists u_0, β_0 such that

$$P\left[|\langle X, t \rangle| \geq u_0 \left(\mathbb{E}\langle X, t \rangle^2 \right)^{1/2} \right] \geq \beta_0,$$

2. $X = (x_1, \ldots, x_d)$, $\|x_j\|_{L_p} \leq \kappa \sqrt{p}$ for $1 \leq p \leq \log(d)$,

Result for the Lasso under the small ball property

Previous results when X is sub-gaussian.We can weaken this assumption if we assume a statistical model :

$$Y = \langle X, t^* \rangle + \zeta$$

where ζ independent of $X \in \mathbb{R}^d$.We assume :

1. a small ball property : there exists u_0, β_0 such that

$$P\left[|\langle X, t\rangle| \geq u_0 \left(\mathbb{E}\langle X, t\rangle^2\right)^{1/2}\right] \geq \beta_0,$$

2. $X = (x_1, \dots, x_d)$, $\|x_j\|_{L_p} \leq \kappa\sqrt{p}$ for $1 \leq p \leq \log(d)$,
3. $$\|\zeta\|_{2,1} = \int_0^\infty \sqrt{P[|\zeta| > x]}dx < \infty$$

Result for the Lasso under the small ball property

Then, for the Lasso :

$$\hat{t} \in \underset{t \in \mathbb{R}^d}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} \big(Y_i - \langle X_i, t \rangle \big)^2 + u \|\zeta\|_{2,1} \|t\|_1 \sqrt{\frac{\log d}{N}} \right)$$

Result for the Lasso under the small ball property

Then, for the Lasso :

$$\hat{t} \in \underset{t \in \mathbb{R}^d}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \langle X_i, t \rangle \right)^2 + u \|\zeta\|_{2,1} \|t\|_1 \sqrt{\frac{\log d}{N}} \right)$$

with probability at least $1 - (1/u) - \exp(-N)$,

$$\|\hat{t}\|_1 \lesssim \|t^*\|_1$$

Result for the Lasso under the small ball property

Then, for the Lasso :

$$\hat{t} \in \underset{t \in \mathbb{R}^d}{\operatorname{argmin}} \left(\frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \langle X_i, t \rangle \right)^2 + u \|\zeta\|_{2,1} \|t\|_1 \sqrt{\frac{\log d}{N}} \right)$$

with probability at least $1 - (1/u) - \exp(-N)$,

$$\|\hat{t}\|_1 \lesssim \|t^*\|_1 \text{ and } \|\langle X, \hat{t} - t^* \rangle\|_{L_2} \lesssim s(\|t^*\|_1)$$

where

$$s(\rho) = \max \left(u \|\zeta\|_{2,1} \rho \sqrt{\frac{\log d}{N}}, \rho^2 \frac{\log d}{N} \right).$$

Thanks for your attention