# Sparsity oracle inequality for the Lasso

**Theorem.** Let $\xi_i \sim \mathcal{N}(0,\sigma^2)$ i.i.d. and $\|f_j\| \le 1$, $j=1,\dots,M$. Let $\hat{\theta}^L$ be the Lasso estimator

$$\hat{\theta}^L = \operatorname*{argmin}_{\theta \in \mathbb{R}^M} \left( \|y - f_\theta\|^2 + 2\tau |\theta|_1 \right)$$

with $\tau = A\sigma\sqrt{\frac{\log M}{n}}$, $A = \frac{t}{\delta}$, $t > \sqrt{2}$, $0 < \delta < 1$.
Then, with probability at least $1 - M^{1-t^2/2}$ we have

$$\forall f: \quad \|f_{\hat{\theta}^L} - f\|^2 \le \min_{\theta \in \mathbb{R}^M} \Big[ \|f_\theta - f\|^2 + C \min\Big( \sigma^2 \mu^2(\theta) |\theta|_0 \frac{\log M}{n}, \; |\theta|_1 \sqrt{\frac{\log M}{n}} \Big) \Big]$$

where $C > 0$ depends only on $t$ and $\delta$,

$$\mu(\theta) = \inf \Big\{ \mu > 0 : |\Delta_{J(\theta)}|_1 \le \mu \sqrt{\frac{|\theta|_0}{n}} \, |X\Delta|_2, \; \forall \Delta \in C_\theta \Big\}$$

with

$$C_\theta = \Big\{ \Delta \in \mathbb{R}^M : |\Delta_{J^c(\theta)}|_1 \le \frac{1+\delta}{1-\delta} |\Delta_{J(\theta)}|_1 \Big\}$$

and $0 < \delta < 1$ is an arbitrary fixed number.

Proof of Theorem. Set for brevity $\hat{\theta}^L \stackrel{\triangle}{=} \hat{\theta}$.

$$G(\theta) \stackrel{def}{=} \|y - f_\theta\|^2 + 2\tau|\theta|_1, \quad \boxed{\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^M} G(\theta).}$$

Denote by $(\cdot,\cdot)$ the inner product in $\mathbb{R}^M$,

$$\langle f, g\rangle \stackrel{def}{=} \frac{1}{n}\sum_{i=1}^{n} f(X_i)g(X_i).$$

General fact: for any convex function $G$ we have

Lemma A. $\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^M} G(\theta) \Rightarrow$ there exists

$B \in \partial(G(\hat{\theta}))$ such that

$$(B, \hat{\theta} - \theta) \le 0, \text{ for all } \theta \in \mathbb{R}^M.$$

In our case,

$$\nabla\left(\|y - f_\theta\|^2\right) = \nabla\left(\frac{1}{n}|y - X\theta|_2^2\right) =$$

$$= -\frac{2}{n} X^T(y - X\theta), \text{ and thus}$$

$$\left(\nabla\left(\|y - f_{\hat{\theta}}\|^2\right), \hat{\theta} - \theta\right) = -\frac{2}{n}\left(X^T(y - X\hat{\theta}), \hat{\theta} - \theta\right)$$

$$= -\frac{2}{n}\left(X(\hat{\theta} - \theta), y - X\hat{\theta}\right) = -2\langle f_{\hat{\theta}-\theta}, y - f_{\hat{\theta}}\rangle$$

Applying Lemma A we get that there exists $\hat{V} \in \partial(|\hat{\theta}|_1)$ such that

(1) $$-2\langle f_{\hat\theta-\theta}, y-f_{\hat\theta}\rangle + 2\tau(\hat V, \hat\theta-\theta) \le 0$$

Let $V$ be any element of $\partial(|\theta|_1)$. We use the following result:

---

<u>Lemma B.</u> For any convex function $g(\theta)$ we have
$$(V-V', \theta-\theta') \ge 0, \quad \forall \theta, \theta' \in \mathbb{R}^M,$$
where $V \in \partial g(\theta)$, $V' \in \partial g(\theta')$.

---

From Lemma B and (1) we find

$$-2\langle f_{\hat\theta-\theta}, y-f_{\hat\theta}\rangle + 2\tau(\hat V - V, \hat\theta-\theta) \le$$
$$\le -2\tau(V, \hat\theta-\theta),$$

and if $V \in \partial(|\theta|_1)$,

$$-2\langle f_{\hat\theta-\theta}, y-f_{\hat\theta}\rangle \le -2\tau(V, \hat\theta-\theta).$$

Since $y = f+\xi$, we can rewrite this in the form:

(2) $$2\langle f_{\hat\theta-\theta}, f_{\hat\theta}-f\rangle \le -2\tau(V, \hat\theta-\theta) + 2\langle\xi, f_{\hat\theta-\theta}\rangle$$

<u>for any(!)</u> $V \in \partial(|\theta|_1)$ and <u>any</u>(!) $\theta \in \mathbb{R}^M$.

Note that, by the "cosine theorem",

$$(3)\quad 2\langle f_{\hat\theta-\theta}, f_{\hat\theta}-f\rangle = 2\langle f_{\hat\theta}-f_\theta, f_{\hat\theta}-f\rangle$$

$$= \|f_{\hat\theta}-f_\theta\|^2 + \|f_{\hat\theta}-f\|^2 - \|f_\theta - f\|^2.$$

Let $J = J(\theta)$ be the set of non-zero components of $\theta$. Write

$$V = V_J + V_{J^c}$$

where $V_J \in \mathbb{R}^M$ is the vector with components $V_j\, I(V_j \neq 0)$, $j = 1, \dots, M$, where $V_j$ are the components of $V$, and $J^c = \{1, \dots, M\} \setminus J$ is the complement of $J$. Then

$$(V, \hat\theta-\theta) = (V_J, \hat\theta-\theta) + (V_{J^c}, \hat\theta-\theta) =$$

$$= (V_J, \hat\theta-\theta) + (V_{J^c}, \hat\theta),$$

since the components of $V_{J^c}$ vanish on the support of $\theta$. On the other hand, $V$ is any element of $\partial(|\theta|_1)$, and thus the components of $V$ satisfy

$$\begin{cases} |V_j| \le 1, & j \in J^c \\ V_j = \operatorname{sign}(\theta_j), & j \in J. \end{cases}$$

Choose $V$ such that $V_j = \operatorname{sign}(\hat{\theta}_j)$ for $j \in J^c$. This is possible, since $V_j$ can be any values satisfying $|V_j| \le 1$ for $j \in J^c$.

Then

$$(V, \hat{\theta} - \theta) = (V_J, \hat{\theta} - \theta) + |\hat{\theta}_{J^c}|_1.$$

This and (2) imply

$$(4) \quad 2 \langle f_{\hat{\theta} - \theta}, f_{\hat{\theta}} - f \rangle + 2\tau |\hat{\theta}_{J^c}|_1 \le$$

$$\le 2\tau |\Delta_J|_1 + 2 (H, \underbrace{\hat{\theta} - \theta}_{= \Delta}),$$

where $\Delta = \hat{\theta} - \theta$ and

$$H = \frac{1}{n} X^T \xi = \begin{pmatrix} H_1 \\ \vdots \\ H_M \end{pmatrix}$$

with $H_j = \frac{1}{n} \sum_{i=1}^{n} f_j(x_i) \xi_i$. We have used here the identity $\langle \xi, f_{\hat{\theta} - \theta} \rangle = (H, \hat{\theta} - \theta)$.

First note that if

$$\langle f_{\hat{\theta}-\theta}, f_{\hat{\theta}} - f \rangle \le 0,$$

then, in view of (3) we get

$$\| f_{\hat{\theta}} - f \|^2 \le \| f_\theta - f \|^2,$$

and the result of the theorem follows in a trivial way. So, it is enough to consider the case $\langle f_{\hat{\theta}-\theta}, f_{\hat{\theta}} - f \rangle \ge 0$. But, in this case, in view of (4),

$$\tau \, | \hat{\theta}_{J^c} |_1 \le \tau \, | \Delta_J |_1 + | M |_\infty \, | \Delta |_1 .$$

$\underbrace{\text{Assume}}_{}$ for the moment, that $\boxed{|M|_\infty \le \delta \tau}$ for some $0<\delta<1$. Then, since $|\Delta|_1 = |\Delta_J|_1 + |\Delta_{J^c}|_1$, and $|\hat{\theta}_{J^c}|_1 = |\Delta_{J^c}|_1$, we have:

$$(1-\delta) \, |\Delta_{J^c}|_1 \le (1+\delta) \, |\Delta_J|_1 ,$$

or:

$$|\Delta_{J^c}|_1 \le \frac{1+\delta}{1-\delta} \, |\Delta_J|_1 .$$

In other words, it suffices to consider $\Delta \in C_\theta$, where $C_\theta$ is the following cone:

$$C_\theta = \{ \Delta \in \mathbb{R}^M : |\Delta_{J^c}|_1 \leq c_0 |\Delta_J|_1 \}$$

and $J = J(\theta)$ is the set of non-zero components of $\theta$,

$$c_0 = \frac{1+\delta}{1-\delta}.$$

Next, we return to (4), and bound more accurately the terms on the right-hand side of (4).

We have, using that $|M|_\infty \leq \delta\tau$ and $\Delta \in C_\theta$,

$$2\tau |\Delta_J|_1 + 2(M, \Delta) \leq$$

$$\leq 2\tau |\Delta_J|_1 + 2|M|_\infty |\Delta|_1$$

$$\leq 2\tau |\Delta_J|_1 + 2\delta\tau (|\Delta_J|_1 + |\Delta_{J^c}|_1)$$

$$\leq 2\tau \left(\frac{1+\delta}{1-\delta}\right) |\Delta_J|_1 .$$

This and (4) imply

$$2\langle f_{\hat\theta-\theta}, f_{\hat\theta}-f\rangle \le 2\tau\left(\frac{1+\delta}{1-\delta}\right)|\Delta_J|_1 .$$

Combining this with (3) we get

$$(5)\qquad \|f_{\hat\theta}-f\|^2 \le \|f_\theta - f\|^2 - \|f_{\hat\theta}-f_\theta\|^2 + \\ + 2\tau\left(\frac{1+\delta}{1-\delta}\right)|\Delta_J|_1 .$$

Since $\Delta \in C_\theta$, we get

$$|\Delta_J|_1 \le \mu(\theta)\sqrt{\frac{|\theta|_0}{n}}\,|X\Delta|_2 = \mu(\theta)\sqrt{|\theta|_0}\,\|f_\Delta\| \\ \equiv \mu(\theta)\sqrt{|\theta|_0}\,\|f_{\hat\theta}-f_\theta\| .$$

By the elementary inequality $2ab \le a^2+b^2$,

$$(6)\qquad 2\tau\left(\frac{1+\delta}{1-\delta}\right)|\Delta_J|_1 \le 2\tau\left(\frac{1+\delta}{1-\delta}\right)\mu(\theta)\sqrt{|\theta|_0}\,\|f_{\hat\theta}-f_\theta\| \\ \le \tau^2\left(\frac{1+\delta}{1-\delta}\right)^2\mu^2(\theta)\,|\theta|_0 + \|f_{\hat\theta}-f_\theta\|^2 .$$

Combining (5) and (6) we obtain

(7) $$\boxed{\begin{aligned}&\|f_{\hat\theta}-f\|^2 \le \|f_\theta - f\|^2 + \tau^2\left(\frac{1+\delta}{1-\delta}\right)\mu^2(\theta)|\theta|_0 \\ &\forall\theta,\ \forall f\end{aligned}}$$

Note that this inequality is proved for all $\theta \in \mathbb{R}^M$ and all $f$ under the assumption that

$$\boxed{|\mathbf{H}|_\infty \le \delta\tau.}$$

Let us now show that $|\mathbf{H}|_\infty \le \delta\tau$ holds with probability at least $1 - M^{1-t^2/2}$.

Consider the random event

$$\mathcal{A} = \{\, |\mathbf{H}|_\infty \le \delta\tau \}.$$

The probability of its complement $\mathbb{P}(\mathcal{A}^c)$ is estimated as follows

$$\mathbb{P}(\mathcal{A}^c) = \mathbb{P}\left(\max_{1\le j\le M}\left|\frac{1}{n}\sum_{i=1}^n f_j(X_i)\xi_i\right| > \delta\tau\right)$$

$$\le \sum_{j=1}^M \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n f_j(X_i)\xi_i\right| > \delta\tau\right)$$

Here, for each $j$,

$$\frac{1}{n}\sum_{i=1}^n f_j(X_i)\xi_i \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\|f_j\|^2\right).$$

It follows similarly to Lemma 2 in the handouts that, since $\|f_j\|^2 \le 1$, $\forall j$, we get

$$\mathbb{P}\left(|H|_\infty > \delta\tau\right) = \mathbb{P}\left(|H|_\infty > \sigma t\sqrt{\frac{\log M}{n}}\right)$$

$$\le M^{1-t^2/2}.$$

To finish the proof of the theorem, we prove that, on the same event $\mathcal{A}$,

$$(8) \qquad \boxed{\|f_{\hat\theta} - f\|^2 \le \|f_\theta - f\|^2 + C\,|\theta|_1\sqrt{\frac{\log M}{n}}, \quad \forall\theta, \forall f.}$$

Indeed, since $G(\hat\theta) \le G(\theta)$ for all $\theta \in \mathbb{R}^M$, we get, by a simple algebra,

$$(9) \qquad \|f_{\hat\theta} - f\|^2 \le \|f_\theta - f\|^2 + 2\langle \xi, f_{\hat\theta} - f_\theta\rangle + 2\tau|\theta|_1 - 2\tau|\hat\theta|_1$$

Since $\langle \xi, f_{\hat\theta-\theta}\rangle = (H, \hat\theta - \theta)$ and $|H|_\infty \le \delta\tau$ on $\mathcal{A}$, we find

$$(10) \qquad 2\langle \xi, f_{\hat\theta} - f_\theta\rangle \le 2\delta\tau|\hat\theta - \theta|_1 + 2\tau\left(|\theta|_1 - |\hat\theta|_1\right)$$

$$\le 2\tau(1+\delta)|\theta|_1.$$

Combining (9) and (10) we get (8). Finally, the theorem follows from (7) and (8). ◀

$y = X\theta^* + \xi$

## Б. Модель Гауссовской последовательности

$X$ – ортогональная (ort): $\frac{1}{n}X^TX = E = I_M$ (ед. матрица, $m \geq 2$.

$y = X\theta^* + \xi$, $\xi \sim \mathcal{N}(0, \sigma I_n) \Rightarrow$

$\Rightarrow \frac{1}{n}X^Ty \underset{def}{=} \theta^* + \frac{1}{n}X^T\xi \Leftrightarrow \theta^* = \frac{1}{n}X^T(y - \xi)$

Д $\frac{1}{n}X^Ty = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$; $\frac{1}{n}X^T\xi \overset{def}{=} \mathcal{J} = \begin{pmatrix} \mathcal{J}_1 \\ \vdots \\ \mathcal{J}_m \end{pmatrix} \Rightarrow \theta_j = y_j - \mathcal{J}_j$

$\{y_i, i = 1, \dots, m\}$ – новая выборка.

Очевидно, что $\mathbb{E}\mathcal{J} = \mathbb{E}(\frac{1}{n}X^T\xi) = 0$

Матрица ковариаций: $V(\mathcal{J}) = \frac{1}{n^2}\mathbb{E}(X^T\xi\xi^TX) = \frac{1}{n^2}\sigma^2X^TX \underset{ort}{=} \frac{1}{n}\sigma^2 I$

$\Rightarrow \mathcal{J} \sim \mathcal{N}(0, \frac{\sigma^2}{n}I_m)$

$\varepsilon \overset{def}{=} \frac{\sigma}{\sqrt{n}} \Rightarrow$ → Новая модель – модель Гауссовской последовательности.

$\boxed{y_j = \theta_j^* + \varepsilon\eta_j}$ $\eta_j \sim \mathcal{N}(0, 1)$ i.i.d.

$\theta^*$ – sparse $\Rightarrow |\theta^*|_0 = \sum_{j=1}^{M} I(\theta_j^* \neq 0)$ – мало.

Найти оценку $\tilde{\theta}$ такую, что

$\mathbb{E}\frac{1}{n}|X(\tilde{\theta} - \theta^*)|_2^2 \overset{?}{\leq} C|\theta^*|_0 \cdot \frac{\log M}{n}$

Заметим, что (ort) $\Rightarrow \frac{1}{n}|X\theta|_2^2 = |\theta|_2^2$

$\frac{1}{n}|X\theta|_2^2 = \frac{1}{n}\theta^TX^TX\theta = |\theta|_2^2$

Риск: $\mathbb{E}(\frac{1}{n}|X(\tilde{\theta} - \theta)|_2^2) = \mathbb{E}|\tilde{\theta} - \theta|_2^2$


$\theta^*$

Порог $\tau > 0$.

Если $|y_j| > \tau \Rightarrow$ компонента $j$ не нулевая.

$|y_j| \leq \tau \Rightarrow$ компонента $j$ нулевая.

Если $|y_j| > \tau$, то берем оценку $\tilde{\theta}_j = y_j$

$\tilde{\theta}_j \to \tilde{\theta}_j^H$ ← Hard thresholding

$\tilde{\theta}_j^H = y_j I(|y_j| > \tau)$

См. Лемму 1 и Лемму 2 из этого раздаточного материала.

$$\tilde{\Theta}_j^H = y_j I(|y_j| > \tau)$$

С учетом лемм:

~~$\max\limits_{1\le j\le m} |\varepsilon\eta_j|$~~ $\max\limits_{1\le j\le m} |\varepsilon\eta_j| \le \varepsilon\sqrt{2\log M}$ с вероятностью, близкой к 1 при большом $M$.

Т.е. шум никогда не зашкалит за $\tau = \varepsilon\sqrt{2\log M}$

Если зашкалило, то почти наверняка там $\theta \neq 0$

$\tau = \varepsilon\sqrt{2\log M}$ – универсальный порог.

$\left(\tau = \sigma\sqrt{\frac{2\log M}{n}}\right)$

Основной результат:

~~Th5: Пусть выполнено цело~~

<u>Th5:</u> Рассмотрим модель гауссовской последовательности.

Тогда, если $\tau = \varepsilon\sqrt{2\log M}$

1.) $\mathbb{E}|\hat{\theta}^H - \theta^*|_2^2 \le C\varepsilon^2 |\theta^*|_0 \log M$

2.) Выбор компонент:

Если к тому же $\min\limits_{j:\theta_j^* \neq 0} |\theta_j^*| > 2\tau$

то $J(\hat{\theta}^H) = J(\theta^*)$ – мн-во ненулевых компонент.

с вероятностью $\ge 1 - \frac{1}{\sqrt{\pi \log M}}$

Док-во написано на листочках (кратко:)

1). $\mathbb{E}|\hat{\theta}^H - \theta^*|_2^2 = \sum\limits_{j=1}^M \mathbb{E}(\hat{\theta}_j^H - \theta_j^*)^2$

$\theta_j^* = 0 \Rightarrow \begin{cases} |\hat{\theta}_j^H - \theta_j^*| = |\hat{\theta}_j^H| = |\varepsilon\eta_j| I(\varepsilon|\eta_j| > \tau) \Rightarrow \varepsilon|\eta_j| I(|\eta_j| > \sqrt{2\log M}). \\ y_j = \varepsilon\eta_j \end{cases}$

$\theta_j^* \neq 0 \quad |\hat{\theta}_j^H - \theta_j^*| = |y_j - \theta_j^* - y_j I(|y_j| \le \tau)| \le \varepsilon|\eta_j| + \tau$

$$\sum_{j=1}^M \mathbb{E}|\hat{\theta}_j^H + \theta_j^*|^2 = \underbrace{\sum_{j:\theta_j^*=0} ( )}_{\text{много}} + \underbrace{\sum_{j:\theta_j^*\neq 0} ( )}_{\text{мало}} \le \begin{cases} M\,\mathbb{E}\left[\varepsilon^2\eta_1^2 I(|\eta_1| > \sqrt{2\log M})\right] + \\ + |\theta^*|_0\, \mathbb{E}\left[(\varepsilon|\eta_j| + \tau)^2\right] \end{cases} \xRightarrow{\text{лемма 2}} \text{см. стр. 3.}$$

$$\le \underbrace{M\varepsilon^2\mathbb{E}[\eta^2 I(|\eta|>\sqrt{2\log M})]}_{\text{согласно лемме 1, 2 } \le C\left(\frac{1}{\sqrt{\log M}}+\sqrt{\log M}\right)\cdot\frac{1}{M}} + \mathbb{E}[(\varepsilon|\eta|+\tau)^2]\cdot|\theta^*|_0$$

$$\le \mathbb{E}[2\varepsilon^2\eta^2+2\tau^2] = O(\varepsilon^2+\tau^2) = O(\varepsilon^2\log M)$$

$$= \underbrace{|\theta^*|_0\,\varepsilon^2\log M}_{\text{доминирует}} + \varepsilon^2\left(\frac{1}{\sqrt{\log M}}+\sqrt{\log M}\right)$$

$\log M$ в оценке — плата за незнание ненулевых компонент.

Оптимальная (лучшая) оценка:

$$\mathbb{E}|\hat\theta^H-\theta^*|_2^2 \le C\varepsilon^2|\theta^*|_0\log\frac{M}{|\theta^*|_0}.$$

Д-во части 2.

Мн-во $A=\{|\underbrace{y_j-\theta_j^*}_{\varepsilon\eta_j}|<\tau,\ \forall i=1,\dots,M\} = \{|\eta_j|<\sqrt{2\log M},\ \forall j\}$

Лемма 2: $P(A)\ge 1-\frac{1}{\sqrt{\pi\log M}}$

Док-м: $J(\hat\theta^H)\subseteq J(\theta^*)$ на $A$: $\quad$ $\lrcorner\ j\in J(\hat\theta^H)\Rightarrow\hat\theta_j^H\neq0$ с вер. 1.

$J(\hat\theta^H)\supseteq J(\theta^*)$

$\Leftrightarrow |y_j|>\tau \Leftrightarrow |\theta_j^*+\varepsilon\eta_j|>\tau\Rightarrow$

$\Rightarrow |\theta_j^*|>\tau-|\varepsilon\eta_j|>0\Rightarrow\theta_j^*\neq0\Rightarrow j\in J(\theta^*)$

Обратно: $J(\theta^*)\subseteq J(\hat\theta^H)$

$j\in J(\theta^*)\Rightarrow|\theta_j^*|>2\tau$

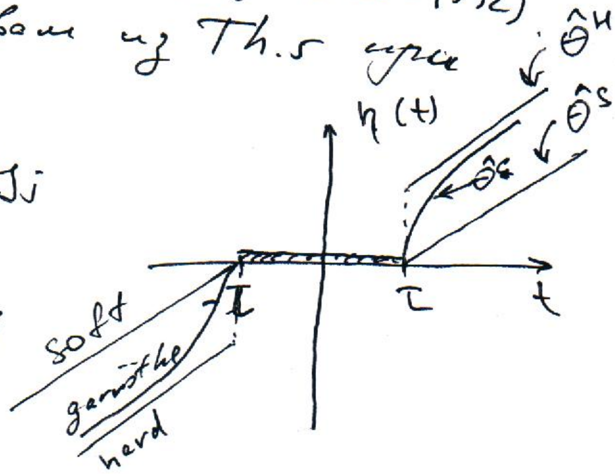
$|y_j|=|\theta_j^*+\varepsilon\eta_j|>2\tau-|\varepsilon\eta_j|\ge2\tau-\tau=\tau\Rightarrow\hat\theta_j^H=y_j$

$\Rightarrow\hat\theta_j^H\neq0$ с вер. 1 $\Rightarrow j\in J(\hat\theta^H)$

Все это было введено Donoho & Johnstone (~1990–1992)

Другие оценки (похожие по свойствам из Th.5 при таких же ограничениях):

- Soft thresholding: $\hat\theta_j^S=\left(1-\frac{\tau}{|y_j|}\right)_+ y_j$
- Non negative garrothe: $\hat\theta_j^G=\left(1-\frac{\tau^2}{y_j^2}\right)_+ y_j$




Все эти оценки нелинейны.

$\hat\theta_j=\eta(y_j)$

Предложение 3: (Вариационная формулировка hard и soft tr.)

$$\hat\theta^H=\arg\min_{\theta\in\mathbb{R}^M}\left[\sum_{j=1}^M(y_j-\theta_j)^2+\tau^2|\theta|_0\right]$$

$$\hat\theta^S=\arg\min_{\theta\in\mathbb{R}^M}\left[\sum_{j=1}^M(y_j-\theta_j)^2+2\tau|\theta|_1\right]$$

Идея док-ва: $\sum_{j=1}^M\{(y_j-\theta_j)^2+2\tau|\theta_j|\}\to\min_\theta$ можно min по координатно.

Возврат к регрессии при условии ортогональности 4. ← пока!

$$\sum_{j=1}^{M} (y_j - \theta_j)^2 =$$ Вспомним, что $\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \frac{1}{n} X^T y$.

$$= \left|\frac{1}{n} X^T y - \theta\right|_2^2 = |\theta|_2^2 - \frac{2}{n}\theta^T X^T y + \underbrace{\frac{1}{n^2} y^T X X^T y}_{\text{от } \theta \text{ не зависит}} \overset{\text{Ort}}{=} \frac{1}{n}|X\theta|_2^2 - \frac{2}{n}(X\theta)^T y + \text{const} + \frac{1}{n}|y|_2^2 - \frac{1}{n}|y|_2^2 =$$

$$= \underbrace{\frac{1}{n}(X\theta - y)_2^2}_{RSS(\theta)} + \text{const}.$$

$$\hat{\theta}^{H} = \arg\min_{\theta}\left(RSS(\theta) + \tau^2|\theta|_0\right)$$
$$\hat{\theta}^{S} = \arg\min_{\theta}\left(RSS(\theta) + 2\tau|\theta|_1\right)$$
} годится для регрессии.

## 6. Методы BiC и Lasso.

$y = X\theta^* + \xi$ — парам.

$y = f + \xi$ — не парам. модель.

BiC: Bayesian Informat. Criterion.

$f_1, \dots, f_M \to X$.

$$\hat{\theta}^{BiC} = \arg\min_{\theta\in\mathbb{R}^M}\left[\cancel{RSS} \; \frac{1}{n}|y - X\theta|_2^2 + \overbrace{\tau^2|\theta|_0}^{pen}\right]$$

$$\hat{\theta}^{Lasso} = \arg\min_{\theta\in\mathbb{R}^m}\left[\frac{1}{n}|y - X\theta|_2^2 + \overbrace{2\tau|\theta|_1}^{pen}\right].$$

Замечание: 1. Классический выбор $\tau = \sigma\sqrt{\frac{A\log M}{n}}$, $A > 2$.

2. Методы со штрафом (пенализация)
- BiC штрафует за большие значения $|\theta|_0$ → интуитивно то, что надо
- Lasso штрафует за большие значения $|\theta|_1$ → ?

3. Численный аспект: Lasso – min выпуклой ф-ции (полином. вр. счета). BiC не считается за полиномиальное время. $|\theta|_0$ – не выпуклая ф-ция. комбинаторная минимизация. (перебор).

$$\hat{\theta}^{BiC} = \arg\min_{\theta\in\mathbb{R}^m}\{RSS(\theta) + \tau^2|\theta|_0\} = \min_{0\le m\le M}\left[\min_{\theta:|\theta|_1 = m} RSS(\theta) + \tau^2 m\right]$$

Сложность $2^M$ — NP-hard.

4. Выбор переменных для Lasso. Исходя из вида оценки всегда есть координаты, которые обнуляются. При некоторых условиях: $J(\hat{\theta}^L) = J(\theta^*)$ с $P \sim 1$.

Для BiC вопрос открытый.