3.2 ст1

а) Как было показано на лекции,

$$\hat{\theta} = (X^T X + \lambda I_d)^{-1} X^T Y$$

Тогда $MSE_{\hat{\theta}}(\theta) = \mathbb{E}[(\hat{\theta}-\theta)^T(\hat{\theta}-\theta)] =$

$= \mathbb{E}[\hat{\theta}^T\hat{\theta} - 2\hat{\theta}^T\theta + \theta^T\theta] =$

$= \mathbb{E}[Y^T X \underbrace{(X^T X + \lambda I_d)^{-1}}_{W_\lambda} (X^T X + \lambda I_d)^{-1} X^T Y -$

$- 2((X^T X + \lambda I_d)^{-1} X^T Y)^T \theta + \theta^T\theta] =$

$= \mathbb{E}[Y^T X W_\lambda^2 X^T Y - 2\theta^T \cdot W_\lambda X^T Y + \theta^T\theta]$

$= \mathbb{E}[tr(X W_\lambda^2 X^T \cdot Y Y^T) - 2\theta^T W_\lambda X^T Y + \theta^T\theta] =$

$= tr(X W_\lambda^2 X^T \cdot \mathbb{E}[Y Y^T]) - 2\theta^T W_\lambda X^T \mathbb{E}[Y] + \theta^T\theta$

$= \{$ в предположении, что модель гауссовская $\}$ $=$

~~$= tr(X W_\lambda^2 X^T \sigma^2) - 2\theta^T W_\lambda X^T \cdot X\theta + \theta^T\theta$~~

~~$= tr(X W_\lambda^2 X^T \cdot (Cov Y + (\mathbb{E}Y)^T \cdot \mathbb{E}Y)) - 2\theta^T W_\lambda X^T X\theta + \theta^T\theta = tr(X W_\lambda X^T \cdot \sigma^2 + (X\theta)^T \cdot X\theta) - 2\theta^T W_\lambda X^T X\theta + \theta^T\theta$~~

$= \{$ а также т.к. $Cov Y = \mathbb{E}[(Y - \mathbb{E}Y)(Y - \mathbb{E}Y)^T] =$

$= \mathbb{E}[Y Y^T - 2Y(\mathbb{E}Y)^T + \mathbb{E}Y(\mathbb{E}Y)^T] \}$ $=$

$= tr(X W_\lambda^2 X^T \cdot (Cov Y + 2\mathbb{E}Y(\mathbb{E}Y)^T - \mathbb{E}Y(\mathbb{E}Y)^T)) -$

$- 2\theta^T W_\lambda X^T \mathbb{E}Y + \theta^T\theta = tr(X W_\lambda^2 X^T \cdot (\sigma^2 I_n +$

$+ X\theta \cdot (X\theta)^T)) - 2\theta^T W_\lambda X^T \cdot X\theta + \theta^T\theta$

Проверка: при $\lambda = 0$ ~~имеем~~ $W_\lambda = (X^TX)^{-1} \Rightarrow$

$\Rightarrow MSE_{\hat{\theta}}(\theta) = tr(X(X^TX)^{-1}X^T \cdot (\sigma^2 I_n + X\theta(X\theta)^T))$

$- 2\theta^T\theta + \theta^T\theta = tr(X(X^TX)^{-1}X^T\sigma^2 + X\cancel{(X^TX)^{-1}X^TX}\theta \cdot$

$\cdot \theta^T X^T) - \theta^T\theta = tr(X((X^TX)^{-1}\sigma^2 + \cancel{I}\theta \cdot \theta^T)X^T)$

$- \theta^T\theta$

б) $\langle \hat{y}, e \rangle = \langle \hat{y}, y - \hat{y} \rangle = (X \cdot \underbrace{(X^TX + \lambda I_d)^{-1}}_{W_\lambda} X^T y)^T \cdot y -$

$- (X(X^TX + \lambda I_d)^{-1}X^Ty)^T \cdot (-//-) =$

$= y^T X W_\lambda X^T y - y^T X W_\lambda X^T \cdot (X W_\lambda X^T y) =$

$= \underset{\neq 0}{y^T}\, \underset{\neq 0}{X}\, \underset{\neq 0}{W_\lambda} \underbrace{(I_d - X^TXW_\lambda)}_{\neq 0} \underset{\neq 0}{X^T} \underset{\neq 0}{y} \neq 0$, т.к.

3.2 а) 2

Сначала одномерный случай:

$$Pr_{\lambda x^2}(p) = \underset{x}{\operatorname{argmin}} \Big(\underbrace{\tfrac{1}{2}(x-p)^2 + \lambda x^2}_{g(x,p)} \Big)$$

$\partial g(x,p) = x - p + 2\lambda x$

Т.к. $x_0 = Pr_{\lambda x^2}(p) \Leftrightarrow$ ~~$0 \in p - x_0$~~ $0 \in \partial g(x_0, p)$,

имеем: $0 \in \partial g(x_0, p) \Leftrightarrow 0 = x_0 - p + 2\lambda x_0 \Leftrightarrow$

$\Leftrightarrow x_0 = \frac{p}{1+2\lambda}$ — всегда $\exists$-т, т.к. $\lambda \geq 0$.

$\Rightarrow Pr_{\lambda x^2}(p) = \frac{p}{1+2\lambda}$.

По утв-ю 5 с лекции: если θ - реш-е

$F(\theta)+\lambda R(\theta) \to \min_\theta$, то $\theta = Pr_{\lambda R}(\theta - \nabla F(\theta))$

В предположении, что есть сх-ть ~~ит~~

итер. метода, имеем: $\theta_{k+1} = \dfrac{\theta_k - \nabla F(\theta_k)}{1+2\lambda} =$

$= \dfrac{\theta_k - \nabla \|y - X\overline{\theta_k}\|_2^2}{1+2\lambda}$, где $\overline{\theta_k} = \begin{pmatrix} 0 \\ \vdots \\ \theta_k \\ 0 \\ \vdots \end{pmatrix}$ ← k-я позиция

$\theta_{k+1} = \dfrac{\theta_k + 2y^T \circ [X]_k - 2\theta_k}{1+2\lambda} = \dfrac{2y^T[X]_k - \theta_k}{1+2\lambda}$ (k-й столбец)

Многом. случай: $Pr_{\lambda\|x\|_2^2}(p) = \arg\min_x \Big(\frac{1}{2} \cdot \|x-p\|_2^2 + \lambda\|x\|_2^2\Big) = \arg\min_x \Big(\frac{1}{2}\sum_{i=1}^d (x_i - p_i)^2 + \lambda \sum_{i=1}^d x_i^2\Big) = \big(\dots, Pr_{\lambda x_i^2}(p_i), \dots\big)^T$

Св-ва λ: 1) отличие от лассо-регрессии в том, что итерации "плавные", т.к. нет резкого зануления θ_k; а ещё сх-ть более быстрая

2) Сх-ть геометрическая, т.к. в знаменателе каждый раз $(1+2\lambda)$ стоит $\Rightarrow$ при $\lambda \to \infty$ очень быстро $\theta_k \to 0$, а при $\lambda = 0$ получается градиентный спуск для $F(\theta)$.