Регуляризация (гребневая регрессия)

Штраф за увеличение нормы вектора весов $\|\alpha\|$:

$$Q_\tau(\alpha) = \|F\alpha - y\|^2 + \frac{\tau}{2}\|\alpha\|^2,$$

где τ — неотрицательный *параметр регуляризации*.

Модифицированное МНК-решение (τI_n — «гребень»):

$$\alpha_\tau^* = (F^\intercal F + \tau I_n)^{-1} F^\intercal y.$$

Преимущество сингулярного разложения:
можно подбирать параметр τ, вычислив SVD только один раз.

Вектор регуляризованного МНК-решения α^*_τ
и МНК-аппроксимация целевого вектора $F\alpha^*_\tau$:

$$\alpha^*_\tau = U(D^2 + \tau I_n)^{-1} D V^\intercal y = \sum_{j=1}^{n} \frac{\sqrt{\lambda_j}}{\lambda_j + \tau} u_j (v_j^\intercal y);$$

$$F\alpha^*_\tau = VDU^\intercal \alpha^*_\tau = V \,\mathrm{diag}\left(\frac{\lambda_j}{\lambda_j + \tau}\right) V^\intercal y = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \tau} v_j (v_j^\intercal y);$$

$$\|\alpha^*_\tau\|^2 = \|(D^2 + \tau I_n)^{-1} D V^\intercal y\|^2 = \sum_{j=1}^{n} \frac{\lambda_j}{(\lambda_j + \tau)^2} (v_j^\intercal y)^2.$$

$F\alpha^*_\tau \neq F\alpha^*$, но зато решение становится гораздо устойчивее.

Выбор параметра регуляризации τ

Контрольная выборка: $X^k = (x'_i, y'_i)_{i=1}^k$;

$$\underset{k\times n}{F'} = \begin{pmatrix} f_1(x'_1) & \dots & f_n(x'_1) \\ \dots & \dots & \dots \\ f_1(x'_k) & \dots & f_n(x'_k) \end{pmatrix}, \quad \underset{k\times 1}{y'} = \begin{pmatrix} y'_1 \\ \dots \\ y'_k \end{pmatrix}.$$

Вычисление функционала Q на контрольных данных T раз потребует $O(kn^2 + knT)$ операций:

$$Q(\tau) = \|F'\alpha^*_\tau - y'\|^2 = \Big\| \underbrace{F'U}_{k\times n} \operatorname{diag}\Big(\frac{\sqrt{\lambda_j}}{\lambda_j+\tau}\Big) \underbrace{V^\intercal y}_{n\times 1} - y' \Big\|^2.$$

Зависимость $Q(\tau)$ обычно имеет характерный минимум.

Сжатие (shrinkage) или *сокращение весов* (weight decay):

$$\|\alpha^*_\tau\|^2 = \sum_{j=1}^{n} \frac{\lambda_j}{(\lambda_j + \tau)^2}(v_j^\top y)^2 \quad < \quad \|\alpha^*\|^2 = \sum_{j=1}^{n} \frac{1}{\lambda_j}(v_j^\top y)^2.$$

Почему говорят о *сокращении эффективной размерности*?

Роль размерности играет след проекционной матрицы:

$$\operatorname{tr} F(F^\top F)^{-1}F^\top = \operatorname{tr}(F^\top F)^{-1}F^\top F = \operatorname{tr} I_n = n.$$

При использовании регуляризации:

$$\operatorname{tr} F(F^\top F + \tau I_n)^{-1}F^\top = \operatorname{tr}\operatorname{diag}\left(\frac{\lambda_j}{\lambda_j + \tau}\right) = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \tau} < n.$$