HW3P1: Proof of in-sample probabilistic risk bound

linear model:

$$y = X\theta^* + \varepsilon, \text{ where } X \in \mathbb{R}^{n \times d}, \ \theta^* \in \mathbb{R}^d, \text{ and } \varepsilon \in SG_n(\sigma)$$

In-sample risk:

$$\frac{1}{n} \mathbb{E} \| X\hat{\theta} - X\theta^* \|^2, \text{ where } \hat{\theta} \text{ is the estimator, } \theta^* \text{ is the true param}$$

Probabilistic bound for in sample risk:

$$\mathbb{P}\left[\frac{1}{n} \| X\hat{\theta} - X\theta^* \|^2 \geq c\sigma^2 \left(\frac{r + \log(1/\delta)}{n} \right) \right]$$

where $c > 0$, $r = \text{rank}(X^T X)$

- Proof -

1. optimality of $\hat{\theta}$ (since $\hat{\theta}$ minimizes $\|y - X\hat{\theta}\|^2$)

$$\|y - X\hat{\theta}\|^2 \leq \|y - X\theta^*\|^2 = \|\varepsilon\|^2$$

($= \|\varepsilon\|^2$ by the definition of linear model)

2. expand the squared norm ($\|a+b\|^2 = \|a\|^2 + \|b\|^2 + 2a^T b$)

$$\|y - X\hat{\theta}\|^2 = \|X\theta^* + \varepsilon - X\hat{\theta}\|^2 = \|X(\theta^* - \hat{\theta})\|^2 + \|\varepsilon\|^2 + 2\varepsilon^T X(\theta^* - \hat{\theta})$$

($= \|X\theta^* + \varepsilon - X\hat{\theta}\|^2$ by the definition of linear model)

3. substitute back into inequality and simplify

$$\|X(\theta^* - \hat{\theta})\|^2 \leq 2\varepsilon^T X(\theta^* - \hat{\theta})$$

4. normalize the inequality

$$\|X(\theta^* - \hat{\theta})\| \leq 2\varepsilon^T \frac{X(\theta^* - \hat{\theta})}{\|X(\theta^* - \hat{\theta})\|}$$

(unit vector that depends on ε)

5. introduce supremum over unit vectors - to replace random component with worst-case non-random component

$$\|X(\theta^* - \hat{\theta})\| \leq 2 \sup_{v \in S^{n-1}} \varepsilon^T v$$

(S^{n-1}: all unit vectors v in the unit sphere)

6. reduce dimensionality using rank (replace d with r)

$X(\theta^* - \hat{\theta}) = \Phi w$, where $\Phi \in \mathbb{R}^{n \times r}$ and $w \in \mathbb{R}^r$

$\tilde{\varepsilon} = \Phi^T \varepsilon$ ($\tilde{\varepsilon}$ is the projection of ε onto the column space of X)

7. rewrite the inequality on lower dimensions

$$\|X(\theta^* - \hat{\theta})\| \leq 2\, \hat{\varepsilon}^T \frac{w}{\|w\|} = 2 \sup_{v \in S^{r-1}} \hat{\varepsilon}^T v$$

since $\hat{\Phi}^T \hat{\Phi} = I$, $\hat{\varepsilon}^T \frac{w}{\|w\|}$ simplifies to an inner product over unit vectors.

8. supremum over unit sphere

$$2 \sup_{v \in S^{r-1}} \hat{\varepsilon}^T v = 2\, \|\hat{\varepsilon}\|$$

↑ the maximum inner product occurs when v is in the direction of $\hat{\varepsilon}$.

$$\therefore \quad \|X(\theta^* - \hat{\theta})\| \leq 2\|\hat{\varepsilon}\| \leq 2\sqrt{r}\, |\hat{\varepsilon}_i| \quad \text{for } \{\hat{\varepsilon}_i\}_i^r$$

9. From the expected bound for maxima over a sequence of r.v.

$$\mathbb{E}\left[\max_{i=1\dots r} |\hat{\varepsilon}_i|\right] \leq \sigma\sqrt{2\log(2r)} \qquad \text{(from previous chapter)}$$

$$2\sqrt{r}\, \mathbb{E}\left[\max_{i=1\dots r} |\hat{\varepsilon}_i|\right] \leq 2\sigma\sqrt{2r\log(2r)}$$

$$\therefore \quad \mathbb{E}\left[\tfrac{1}{n}\|X(\theta^* - \hat{\theta})\|^2\right] \leq \frac{r}{n} 8\sigma^2 \log(2r)$$

↑ proportional to $\frac{r}{n}$

HW3 P2: prove the given representation of the Lasso solution under an orthogonal design.

$$\hat{\beta}_{Lasso,j} = \begin{cases} \hat{\beta}_{LS,j} + \lambda/2 & \text{if } \hat{\beta}_{Lasso,j} < 0 \\ \hat{\beta}_{LS,j} - \lambda/2 & \text{if } \hat{\beta}_{Lasso,j} > 0 \\ 0 & \text{otherwise} \end{cases}$$

1. Lasso optimization problem / loss function

$$L(\beta) = \min_{\beta \in \mathbb{R}^d} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1, \text{ where } X^T X = I_d \text{ under the orthogonal design assumption}$$

2. expand the squared error term ↙ cancel bc of orthogonality

$$\|y - X\beta\|_2^2 = y^T y - 2\beta^T X^T y + \beta^T X^T X \beta = y^T y - 2\beta^T X^T y + \beta^T \beta$$

3. The loss function simplifies to

$$L(\beta) = y^T y - 2\beta^T X^T y + \beta^T \beta + \lambda \|\beta\|_1 \quad \leftarrow y^T y \text{ is constant w.r.t. } \beta$$

$$L(\beta) = \sum_{j=1}^{d} \left(-2\beta_j X_j^T y + \beta_j^2 + \lambda|\beta_j|\right)$$

4. In the orthogonal case, each β_j can be optimized independently

$$L(\beta_j) = -2\beta_j X_j^T y + \beta_j^2 + \lambda|\beta_j|$$

5. write this for positive and negative β_j, set derivative w.r.t β_j to zero and solve.

I: $\beta_j > 0 \Rightarrow |\beta_j| = \beta_j$:

$$L(\beta_j) = -2\beta_j X_j^T y + \beta_j^2 + \lambda\beta_j$$

$$\frac{d}{d\beta} L(\beta_j) = -2X_j^T y + 2\beta_j + \lambda = 0$$

$$\beta_j = X_j^T y - \lambda/2$$

II: $\beta_j < 0 \Rightarrow |\beta_j| = -\beta_j$

$$L(\beta_j) = -2\beta_j X_j^T y + \beta_j^2 - \lambda\beta_j$$

$$\frac{d}{d\beta} L(\beta_j) = -2X_j^T y + 2\beta_j - \lambda = 0$$

$$\beta_j = X_j^T y + \lambda/2$$

6. relate to the least squares (LS) solution:

$$\hat{\beta}_{LS,j} = X_j^T y$$

$$\therefore \hat{\beta}_{Lasso,j} = \begin{cases} \hat{\beta}_{LS,j} + \lambda/2 & \text{if } \hat{\beta}_{Lasso,j} < 0 \\ \hat{\beta}_{LS,j} - \lambda/2 & \text{if } \hat{\beta}_{Lasso,j} > 0 \\ 0 & \text{otherwise} \end{cases}$$

HW3P3: Weak rate for Lasso

To show that:

$$\mathbb{P}(\|X^T\varepsilon\|_\infty > t) \le 2de^{-t^2/2n\sigma^2}, \text{ where } \varepsilon \sim SG(\sigma^2) \in \mathbb{R}^n \text{ and } X \in \mathbb{R}^{n\times d}$$

$\max_{1\le j\le d} |X_j^T\varepsilon|$, where X_j is the j^{th} column of X

By maximum inequalities from module 2:

$$\mathbb{P}\left[\max_{1\le j\le d} |X_j^T\varepsilon| > t\right] = \mathbb{P}\left[\bigcup_{j=1}^d \{|X_j^T\varepsilon| > t\}\right]$$

$$\le \sum_{j=1}^d \mathbb{P}\left[|X_j^T\varepsilon| > t\right]$$

Since $\{\varepsilon_i\}_{i=1}^n$ is a sequence of independent sub-Gaussian r.v., then $x_j^T \varepsilon$ is $SG(\sigma^2 \|x_j\|_2^2)$. Assuming column normalization, $\|x_j\|_2^2 \leq n$, then by the sub-Gaussian property:

$$\mathbb{P}[|x_j^T \varepsilon| > t] \leq 2e^{-t^2/2n\sigma^2}$$

Summing over all $j = 1 \dots d$ columns:

$$\mathbb{P}[\|X^T \varepsilon\|_\infty > t] \leq 2d\, e^{-t^2/2n\sigma^2}$$

By setting $t = n\lambda$, the bound becomes:

$$\mathbb{P}[\|X^T \varepsilon\|_\infty > n\lambda] \leq 2d\, e^{-n\lambda^2/2\sigma^2}$$

Gabriel Riegner: 06 DEC 2024

HW3 P4:

$X \sim$ Multivariate Gaussian with five different settings of n, d, s.

Show the following MSE upper bound is valid:

$\frac{1}{n}||X\hat{\theta}(\lambda) - X\theta^*||_2^2 \leq 32\lambda^2 s$

where $\lambda \approx \frac{s \cdot \log(d)}{n}$.

Assume:

1. $Y = X\theta^* + \epsilon$
2. column normalization
3. $||\theta^*||_0 \leq s$, where $s << n$ (sparsity assumption)
4. $X \in \mathbb{R}^{n \times d}$ needs to satisfy incoherence condition (s) or "sparsity"

Lasso estimator:

$\hat{\theta}(\lambda) = \text{argmin}_{\theta \in \mathbb{R}^d}\{\frac{1}{n}||Y - X\theta||_2^2 + 2\lambda||\theta||_1\}$

In [1]:
```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import Lasso

# setting
ns = np.arange(1000, 5001, 1000)
ss = [20, 30, 40, 50, 60]  # s << n

rng = np.random.default_rng(0)
results = {f"s={s}": {"mse": [], "upper_bound": []} for s in ss}

for s in ss:
    for n in ns:
        d = n  # d = n
        lambda_ = (s * np.log(d)) / n

        # X
        X = rng.standard_normal((n, d))
        X /= np.linalg.norm(X, axis=0)  # column normalization

        # sparse theta*
        theta_star = np.zeros(d)
        nonzeros = rng.choice(d, s, replace=False)
        theta_star[nonzeros] = rng.standard_normal(s)

        # y
        y = X @ theta_star + rng.standard_normal(n)

        # lasso regression
        lasso = Lasso(alpha=2 * lambda_ / n, fit_intercept=False)
        lasso.fit(X, y)
        theta_hat = lasso.coef_

        # empirical MSE
        mse_ = (1 / n) * np.linalg.norm(X @ theta_hat - X @ theta_star) ** 2

        # theoretical upper bound
        upper_bound = 32 * (lambda_**2) * s

        # store results
        results[f"s={s}"]["mse"].append(mse_)
        results[f"s={s}"]["upper_bound"].append(upper_bound)
```

In [2]:
```
# plot results
fig, axs = plt.subplots(5, 1, figsize=(10, 10), sharex=True, layout="constrained")

for ax, (label, values) in zip(axs, results.items()):
    ax.plot(ns, values["upper_bound"], c="k", linestyle="--", label="Upper Bound")
    ax.scatter(ns, values["mse"], c="k", alpha=0.5, label="Empirical")
    ax.set_ylabel("MSE")
    ax.set_title(label)
    ax.legend()
    ax.grid(True)

axs[-1].set_xlabel("n = d")
```

Out[2]:
```
Text(0.5, 0, 'n = d')
```