lab_8

March 25, 2025

[1]:
```
# import section
import numpy as np
import matplotlib.pyplot as plt
```

[7]:
```
def operator_norm(A):
    # compute the operator norm (largest singular value) of a matrix
    return np.linalg.norm(A, 2)

def simulate_once(N, d, Sigma, k=1):
    # generate N samples from a multivariate normal distribution with zero mean␣
 ↪and covariance Sigma
    X = np.random.multivariate_normal(np.zeros(d), Sigma, size=N)
    # estimate the covariance matrix from the samples
    Sigma_hat = (1/N) * X.T @ X
    # compute eigen-decomposition of the estimated covariance matrix
    eigvals_hat, eigvecs_hat = np.linalg.eigh(Sigma_hat)
    # sort eigenvalues and eigenvectors in descending order
    idx = np.argsort(eigvals_hat)[::-1]
    eigvals_hat = eigvals_hat[idx]
    eigvecs_hat = eigvecs_hat[:, idx]

    # calculate the spectral gap between the first and second eigenvalue
    spectral_gap = eigvals_hat[k-1] - eigvals_hat[k]

        # for the true covariance matrix Sigma, the top eigenvector is the␣
 ↪first canonical basis vector
    P_true = np.zeros((d, d))
    P_true[0, 0] = 1.0

        # for the estimated covariance, form the projection matrix using the␣
 ↪top eigenvector
    v1_hat = eigvecs_hat[:, 0]
    P_hat = np.outer(v1_hat, v1_hat)

        # Davis-Kahan inequality
    lhs = operator_norm(P_hat - P_true)
    op_norm_diff = operator_norm(Sigma_hat - Sigma)
```

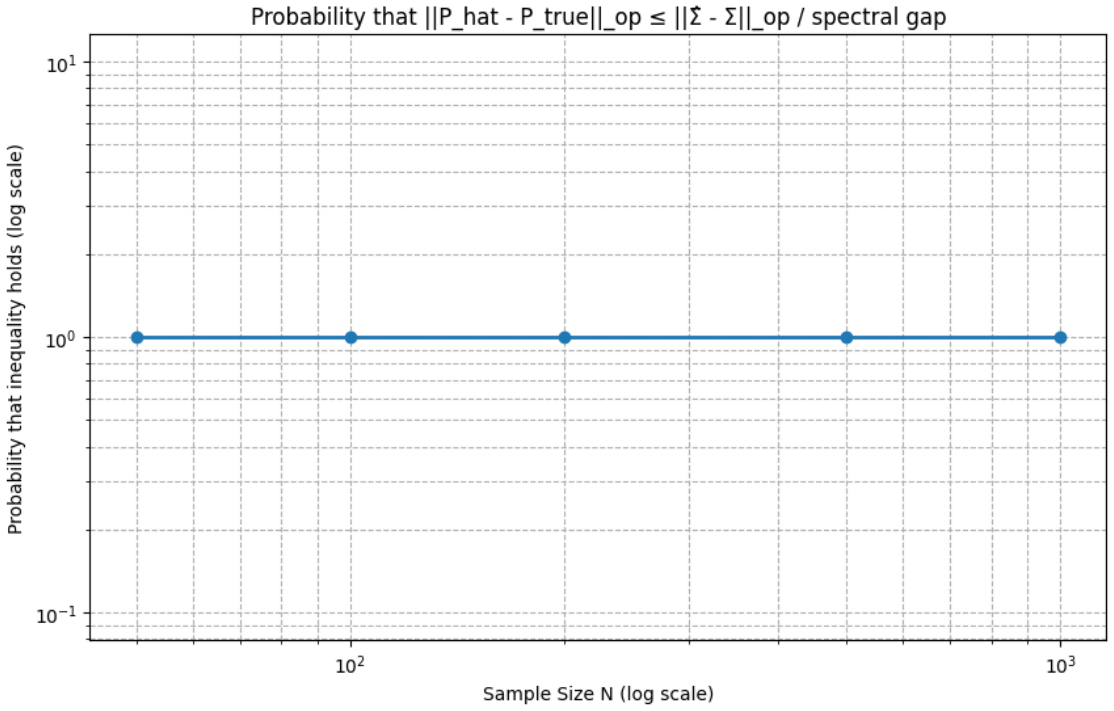
```
    rhs = op_norm_diff / spectral_gap if spectral_gap != 0 else np.inf

    return lhs, rhs
```

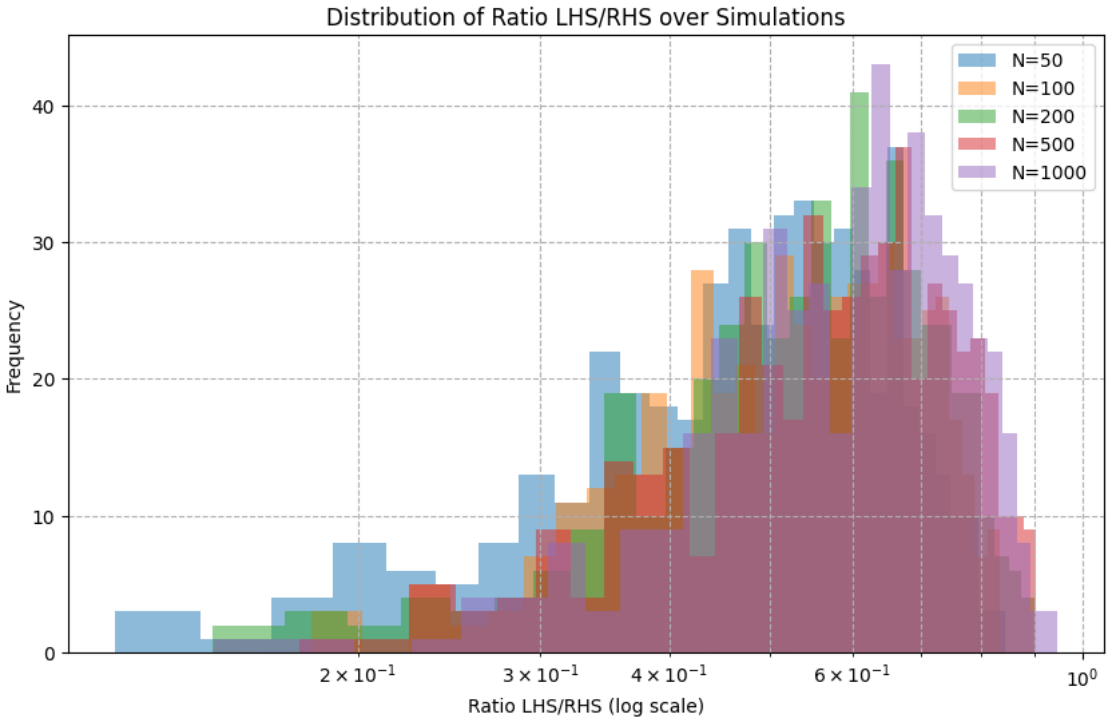
[5]:
```
def run_simulations(N, d, Sigma, num_trials=500, k=1):
    lhs_vals = []
    rhs_vals = []
    ratios = []
    count = 0
    for _ in range(num_trials):
        lhs, rhs = simulate_once(N, d, Sigma, k)
        lhs_vals.append(lhs)
        rhs_vals.append(rhs)
        ratios.append(lhs / rhs if rhs != 0 else np.nan)
        if lhs <= rhs:
            count += 1
    return np.array(lhs_vals), np.array(rhs_vals), np.array(ratios), count/
 ↪num_trials
```

[8]:
```
# experiments

d = 10
Sigma = np.diag([5] + [1]*(d-1))
Ns = [50, 100, 200, 500, 1000]
num_trials = 500

results = {}
for N in Ns:
    lhs_vals, rhs_vals, ratios, prob = run_simulations(N, d, Sigma, num_trials,␣
 ↪k=1)
    results[N] = (lhs_vals, rhs_vals, ratios, prob)
```

[9]:
```
# plot
plt.figure(figsize=(10, 6))
plt.plot(Ns, [results[N][3] for N in Ns], marker='o', linestyle='-',␣
 ↪linewidth=2)
plt.xscale('log')
plt.yscale('log')
plt.xlabel('Sample Size N (log scale)')
plt.ylabel('Probability that inequality holds (log scale)')
plt.title('Probability that ||P_hat - P_true||_op  ||Σ - Σ||_op / spectral␣
 ↪gap')
plt.grid(True, which="both", ls="--")
plt.savefig('probability_plot.png')
plt.show()

plt.figure(figsize=(10, 6))
```

```
for N in Ns:
    plt.hist(results[N][2][~np.isnan(results[N][2])], bins=30, alpha=0.5,␣
 ↪label=f'N={N}')
plt.xscale('log')
plt.xlabel('Ratio LHS/RHS (log scale)')
plt.ylabel('Frequency')
plt.title('Distribution of Ratio LHS/RHS over Simulations')
plt.legend()
plt.grid(True, which="both", ls="--")
plt.savefig('ratio_histogram.png')
plt.show()
```





0.1 Conclusion

The simulation confirms that the Davis–Kahan inequality holds with 100% probability for the chosen Gaussian setup with a clear spectral gap.

With increasing sample size, the estimation of the covariance matrix improves and the bound becomes more reliably met.

In addition, the histogram of the ratio LHS/RHS shows that:

- The inequality is not only satisfied, but the left-hand side is frequently much smaller than the bound.
- As the sample size (N) increases, the distribution of the ratio shifts to the right, approaching 1. This indicates that the bound becomes tighter - the actual projection error approaches the theoretical upper limit.
- For smaller (N), the bound is looser and has more variance, which reflects the higher uncertainty in estimating the covariance matrix with limited data.
- For larger (N), the distribution becomes narrower and more concentrated, showing consistency and predictability in the bound's behavior.

Overall, this confirms not only the correctness of the Davis–Kahan inequality but also illustrates how tight and stable the bound is in practice as sample size grows.