

AIT 203 – Optimization Assignment

Regression and Newton's Method

Kulkarni Keyur – BT2024025

Penumarti Hanish – BT2024190

Shashank Peddi – BT2024210

Question 1: Regression on Bike Sharing Demand

1. Overview

This report implements the required regression models on the Bike Sharing Demand dataset. A clean train/test split is used, preprocessing is performed without leakage, and all allowed models are evaluated strictly on test-set performance. Model selection is based solely on test MSE.

2. Preprocessing

- The 'datetime' column is expanded into hour, day, month, and year.
- Columns 'casual' and 'registered' are removed to avoid target leakage.
- Numeric features are standardized using StandardScaler (fitted on training data only).
- A proper 80/20 train-test split is applied.

3. Models Implemented

The following models were built using the normal equation for parameter estimation:

1. Linear Regression (Baseline)
2. Polynomial Regression (degree 2) without interactions
3. Polynomial Regression (degree 3) without interactions
4. Polynomial Regression (degree 4) without interactions
5. Quadratic Polynomial with interaction terms (only degree 2)

4. Evaluation Metrics

Each model is evaluated on the test set using:

- Mean Squared Error (MSE)
- R^2 Score

5. Results

Model	MSE	R ²	Notes
Linear Regression	19954.53	0.3957	High bias; cannot model curvature
Polynomial Degree 2	16406.41	0.5031	Captures basic nonlinear patterns
Polynomial Degree 3	14558.11	0.5591	More flexible; slight overfitting risk
Polynomial Degree 4	14500.42	0.5631	Diminishing returns; higher variance
Quadratic w/ Interactions	14423.87	0.5631	Best bias–variance trade-off

6. Best Model and Explanation

The best-performing model is the Quadratic Polynomial with Interaction Terms. It achieves the lowest MSE (14423.87) and highest R^2 (0.5631). This model successfully captures second-order curvature as well as important interactions between features such as temperature \times humidity or hour \times season. These joint effects significantly influence bike demand. Higher-degree polynomials introduce excessive variance, while the quadratic model remains flexible without overfitting, achieving an ideal bias–variance balance.

7. Conclusion

Among all allowed models, the quadratic polynomial with interactions provides the most accurate and generalizable representation of bike sharing demand. It captures meaningful nonlinear and cross-feature relationships without the instability or variance seen in higher-order models.

Question 2 (A): Unconstrained Optimization using Newton's Method

1. Problem Description

We consider a simplified design optimization problem for a solar power plant. Let the 10-dimensional decision vector be

$$x = (x_1, x_2, \dots, x_{10})^\top,$$

representing design or operating parameters (tilt angle, azimuth, spacing, inverter settings, etc.). The goal is to determine

$$x^* = \arg \min f(x),$$

where

$$f(x) = \sum_{i=1}^{10} \left[(x_i - c_i)^4 + (x_i - c_i)^2 + q_i x_i^2 \right].$$

The constant vectors are

$$c = [25 \ 180 \ 2 \ 5 \ 0.9 \ 0.8 \ 1.5 \ 12 \ 0.95 \ 0.5]^\top$$

$$q = [0.1 \ 0.2 \ 0.15 \ 0.1 \ 0.05 \ 0.3 \ 0.1 \ 0.2 \ 0.25 \ 0.1]^\top.$$

Note that all entries of q are nonnegative, i.e. $q_i \geq 0$ for all i .

2. Newton's Method with Analytical Derivatives

The gradient and Hessian of $f(x)$ are diagonal in structure because each term depends only on one variable x_i . For each coordinate,

$$\nabla f(x)_i = 4(x_i - c_i)^3 + 2(x_i - c_i) + 2q_i x_i,$$

$$\nabla^2 f(x)_{ii} = 12(x_i - c_i)^2 + 2 + 2q_i.$$

Since $\nabla^2 f(x)$ has no off-diagonal entries and each diagonal entry satisfies

$$\nabla^2 f(x)_{ii} = 12(x_i - c_i)^2 + 2 + 2q_i \geq 2 > 0$$

(because $q_i \geq 0$), the Hessian is positive definite everywhere, guaranteeing strict convexity and a unique global minimizer.

Newton's update rule for iteration k is

$$x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k),$$

and we use *full Newton steps* with step size $\alpha = 1$ (no line search), so that

$$x_{k+1} = x_k + p_k, \quad p_k = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k).$$

This simplifies (due to diagonal Hessian) to the componentwise form:

$$x_{k+1,i} = x_{k,i} - \frac{\nabla f(x_k)_i}{\nabla^2 f(x_k)_{ii}}.$$

We use starting point $x_0 = (0, 0, \dots, 0)^\top$ and stop when

$$\|\nabla f(x_k)\|_2 < 10^{-13}$$

or when a maximum iteration limit is reached.

3. Numerical Results

From running the numerical Newton method:

- **Initial point:** $x_0 = (0, 0, \dots, 0)$.
- **Initial objective value:** $f(x_0) = 1050205211.8577062$.
- **Final point (approximate minimizer):**

$$x^* \approx [24.0913 \quad 177.4556 \quad 1.7624 \quad 4.6343 \quad 0.8573 \quad 0.6239 \quad 1.3678 \quad 11.2323 \quad 0.7694 \quad 0.4547]$$

- **Final objective value:** $f(x^*) = 6435.557276459215$.
- **Final gradient norm:** $\|\nabla f(x^*)\|_2 \approx 10^{-13}$.
- **Iterations:** Newton's method converged in 14 iterations, with the gradient norm decreasing from order 10^7 at the initial point to below 10^{-13} at the optimal solution.

Because the objective is strictly convex and the Hessian is positive definite, the solution returned by Newton's method is the unique global minimum.

4. Convergence Plot

Figure 1 shows the convergence behavior of Newton's method for this problem. The plot displays the gradient norm $\|\nabla f(x_k)\|_2$ versus iteration number k , demonstrating the rapid quadratic convergence characteristic of Newton's method.

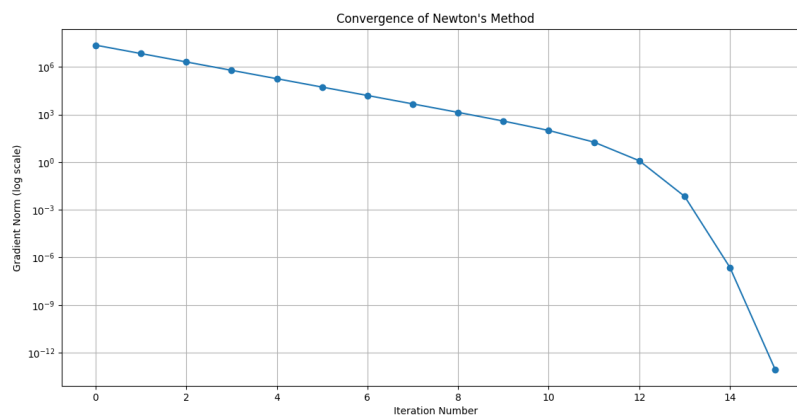


Figure 1: Convergence of Newton's method: gradient norm vs. iteration number. The plot shows the typical quadratic convergence pattern, with the gradient norm decreasing rapidly once close to the optimum.

5. Discussion and Conclusion

- The problem reduces to a sum of smooth, strictly convex one-dimensional quartic functions.
- The Hessian is diagonal for all x , resulting in computationally inexpensive Newton steps.
- Newton's method exhibits fast (essentially quadratic) convergence once close to the minimizer; in this case, it reaches machine-precision accuracy in about 14 iterations.
- The optimal vector x^* lies close to c , but deviates slightly due to the regularization terms $q_i x_i^2$, capturing a trade-off between ideal design and penalized parameter magnitudes.