

Τεχνικές Βελτιστοποίησης 3η Εργαστηριακή Άσκηση

Σπύρος Κούγιας, ΑΕΜ: 10124

I. ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ

Στόχος της παρούσας εργασίας είναι η εύρεση μιας αναλυτικής έκφρασης για την άγνωστη συνάρτηση $f(u_1, u_2)$, χρησιμοποιώντας ένα σύνολο δεδομένων εκπαίδευσης και έναν Γενετικό Αλγόριθμο. Το μοντέλο προσέγγισης που επιλέχθηκε είναι ένας γραμμικός συνδυασμός (K σε αριθμό) Γκαουσιανών συναρτήσεων.

Η μαθηματική μορφή του μοντέλου είναι:

$$y(u_1, u_2) = \sum \left[w_k \cdot \exp \left(-\frac{(u_1 - c_{1k})^2}{2\sigma_1^2} - \frac{(u_2 - c_{2k})^2}{2\sigma_2^2} \right) \right]$$

Ζητούμενο ήταν όχι μόνο η ελαχιστοποίηση του σφάλματος, αλλά και η διατήρηση της πολυπλοκότητας του μοντέλου σε χαμηλά επίπεδα, με μέγιστο όριο τους 15 όρους.

II. ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΥΛΟΠΟΙΗΣΗ

Δομή “Χρωμοσώματος”

Κάθε χρωμόσωμα αναπαριστά μια πλήρη λύση και αποτελείται από τα βάρη και τις παραμέτρους των Γκαουσιανών. Για K όρους, το χρωμόσωμα έχει μήκος $5 \cdot K$: $[w_k, c_{1k}, s_{1k}, c_{2k}, s_{2k}]$

Παράμετροι του Αλγορίθμου

Οι βασικές παράμετροι που χρησιμοποιήθηκαν είναι:

- Μέγεθος Πληθυσμού: 100.
- Γενιές: 600 (επαρκείς για σύγκλιση).
- Επιλογή: Tournament Selection με πίεση επιλογής 4, ώστε να προκρίνονται οι καλύτερες λύσεις αλλά να διατηρείται και η ποικιλομορφία.
- Διασταύρωση: Arithmetic Crossover (παιδί \Rightarrow σταθμισμένος μέσος όρος γονέων)
- Μετάλλαξη: Προσθήκη Γκαουσιανού θορύβου με πιθανότητα 15%.

III. ΕΠΙΛΟΓΗ K (ΠΟΛΥΠΛΟΚΟΤΗΤΑ)

Για τον προσδιορισμό της βέλτιστης πολυπλοκότητας, δοκιμάσαμε το μοντέλο για διάφορες τιμές του K (από 3 έως 15) και καταγράψαμε το μέσο τετραγωνικό σφάλμα (MSE) στα δεδομένα validation.

K	MSE	K	MSE
3	0.017739	10	0.007140
4	0.015618	11	0.010741

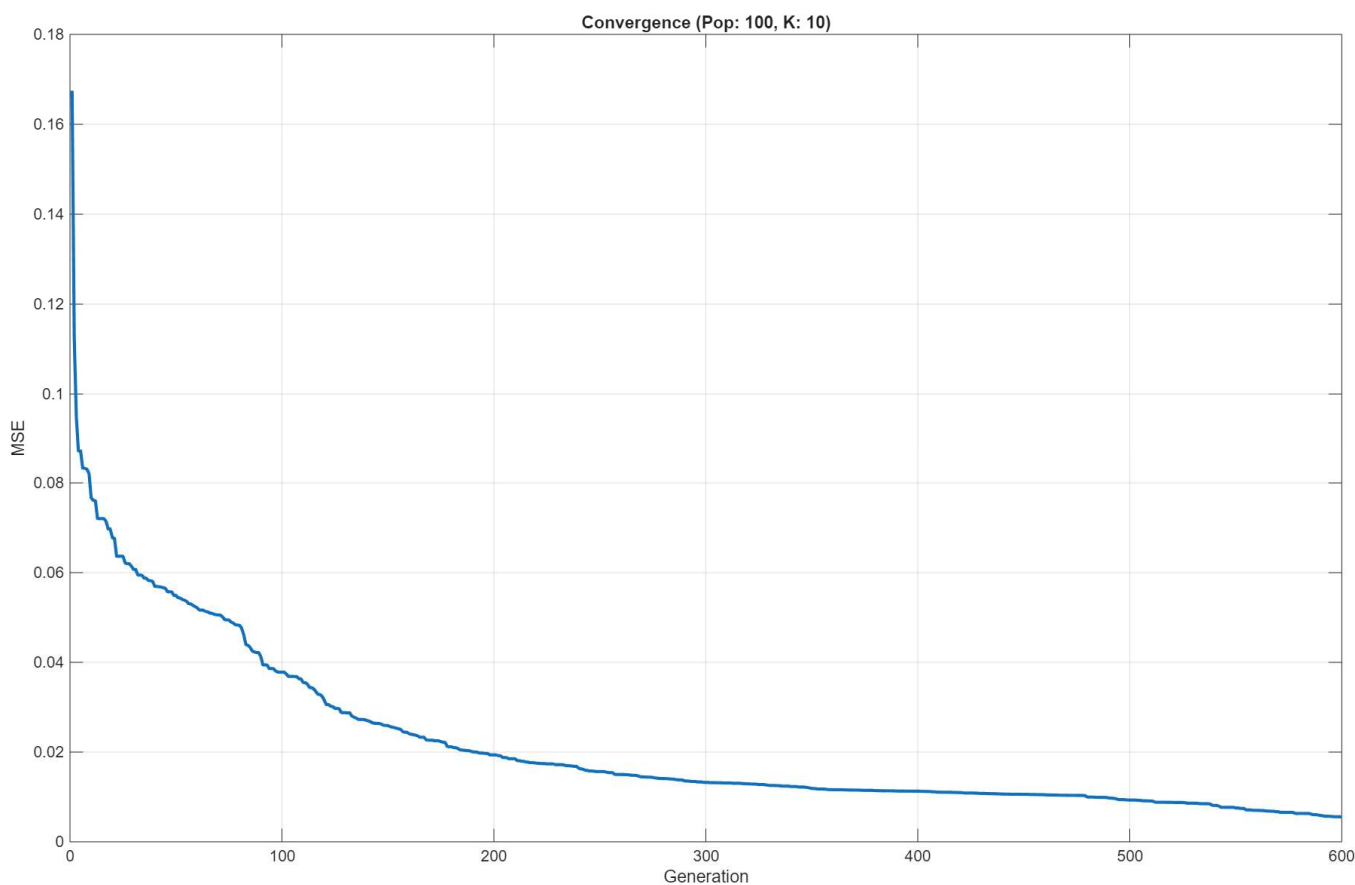
K	MSE	K	MSE
5	0.007594	12	0.020221
6	0.016372	13	0.007694
7	0.008519	14	0.014654
8	0.009638	15	0.014267
9	0.007300		

Όπως φαίνεται και στον πίνακα, $K=10$ έδωσε σημαντική βελτίωση στο σφάλμα. Η περαιτέρω αύξηση των όρων δεν φαίνεται να βελτιώνει το σφάλμα και φυσικά δεν δικαιολογεί την επιπλέον πολυπλοκότητα. Επομένως, επιλέγουμε το $K=10$ ως την βέλτιστη λύση. Αξίζει βέβαια να σημειωθεί ότι το $K = 5$ και $K = 9$ είναι επίσης πολύ καλές επιλογές, μιας και πλησιάζουν το MSE του $K = 10$ και μάλιστα είναι απλότερα.

Τρέχοντας τον αλγόριθμο με $K=10$, λάβαμε τα εξής αποτελέσματα:

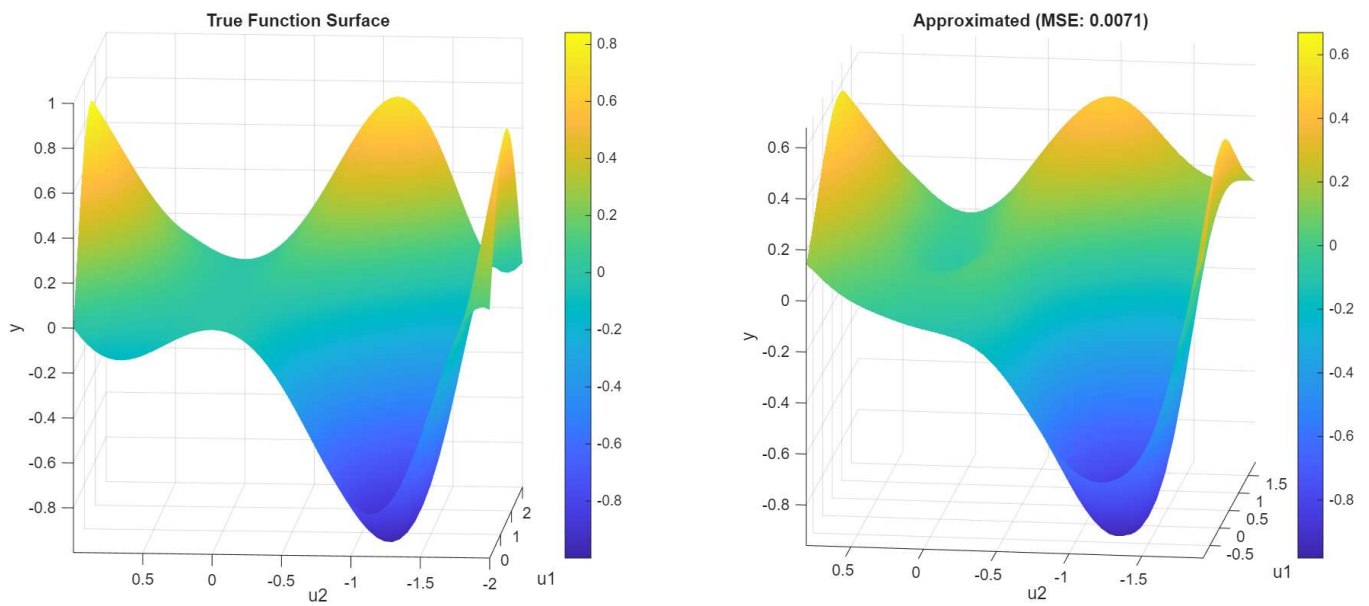
- **Final Training MSE:** 0.005543
- **Validation MSE:** 0.007140

Ακόμα και από $K=10$ παρατηρούμε σημαντική απόσταση μεταξύ validation και training MSE.

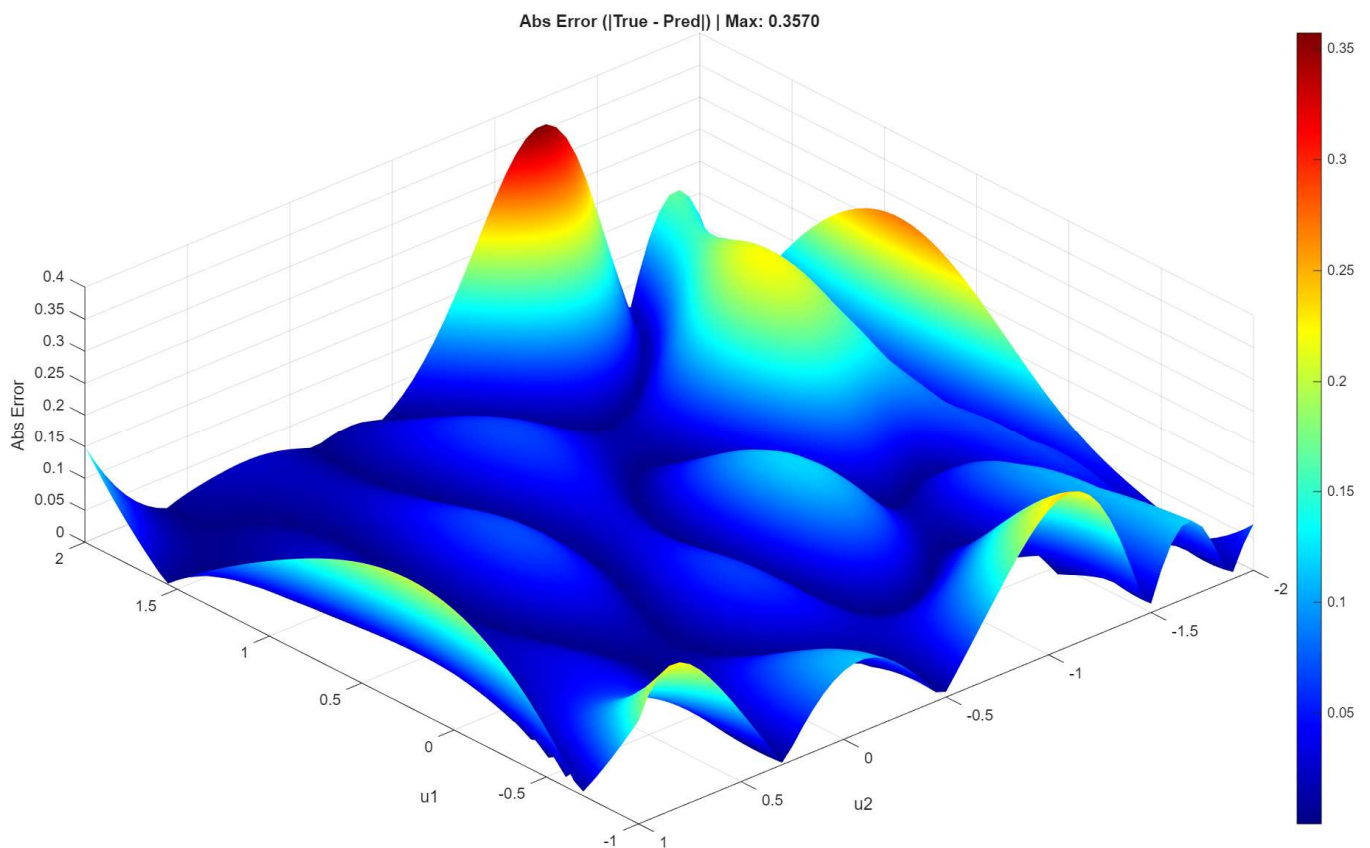


Η εξέλιξη του σφάλματος ανά γενιά φαίνεται στο παραπάνω διάγραμμα, όπου παρατηρούμε πιο έντονη πτώση στις πρώτες ~6 γενιές, έπειτα μια πιο ομαλή πτώση (~100 γενιές) που καταλήγει σε ομαλή σύγκλιση προς τη βέλτιστη λύση.

Στο παρακάτω σχήμα συγκρίνουμε τα πραγματικά δεδομένα (validation) με την έξοδο του μοντέλου μας. Η προσέγγιση φαίνεται ικανοποιητική, καθώς το μοντέλο ακολουθεί πιστά την καμπυλότητα της συνάρτησης στόχου.

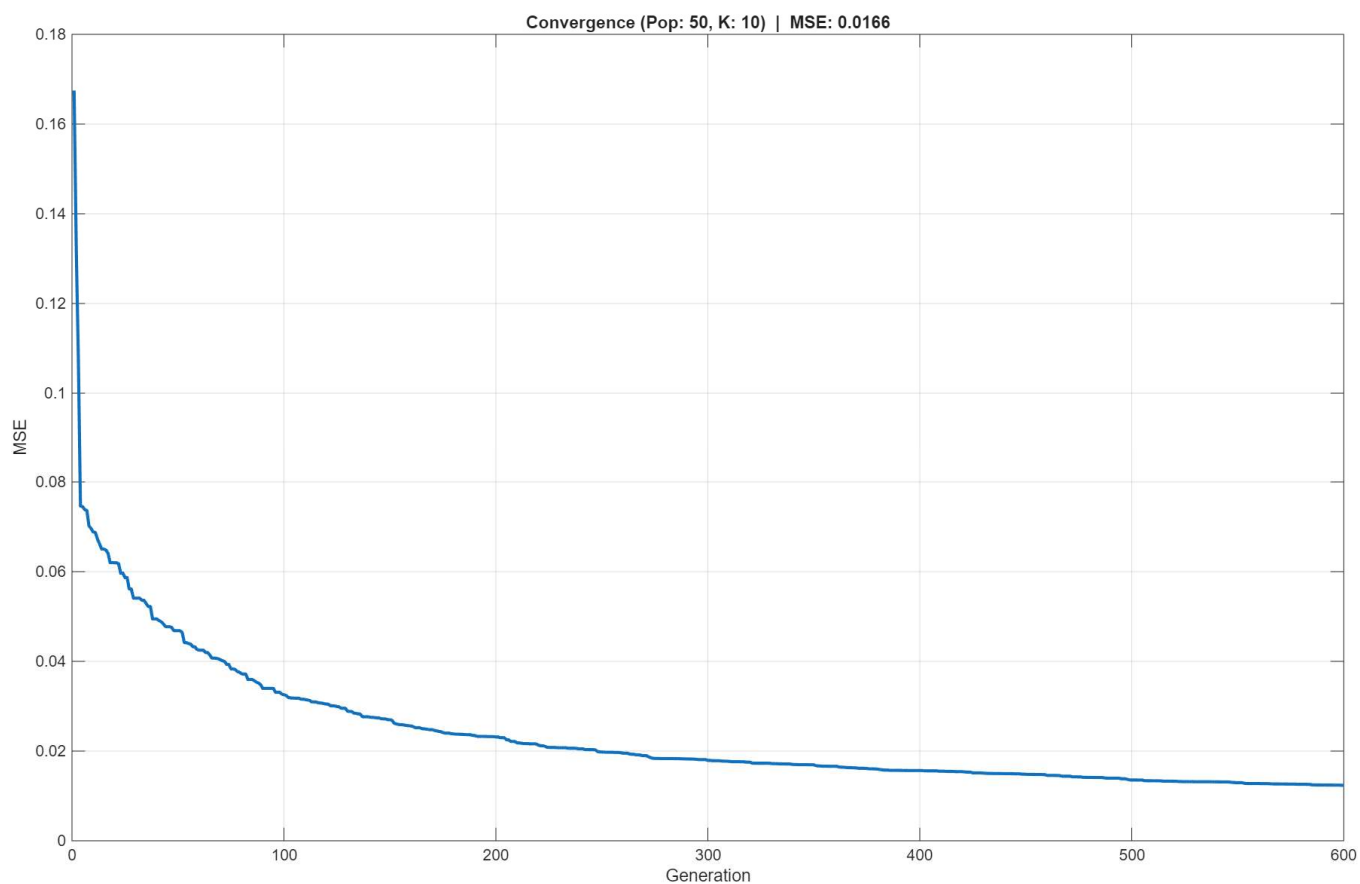


Επιπλέον παρουσιάζουμε και το διάγραμμα του απόλυτου σφάλματος που προδίδει σε ποια τμήματα στο σύνολο που μελετάμε, το μοντέλο μας δυσκολεύεται να αντιπροσωπεύσει την αρχική συνάρτηση.

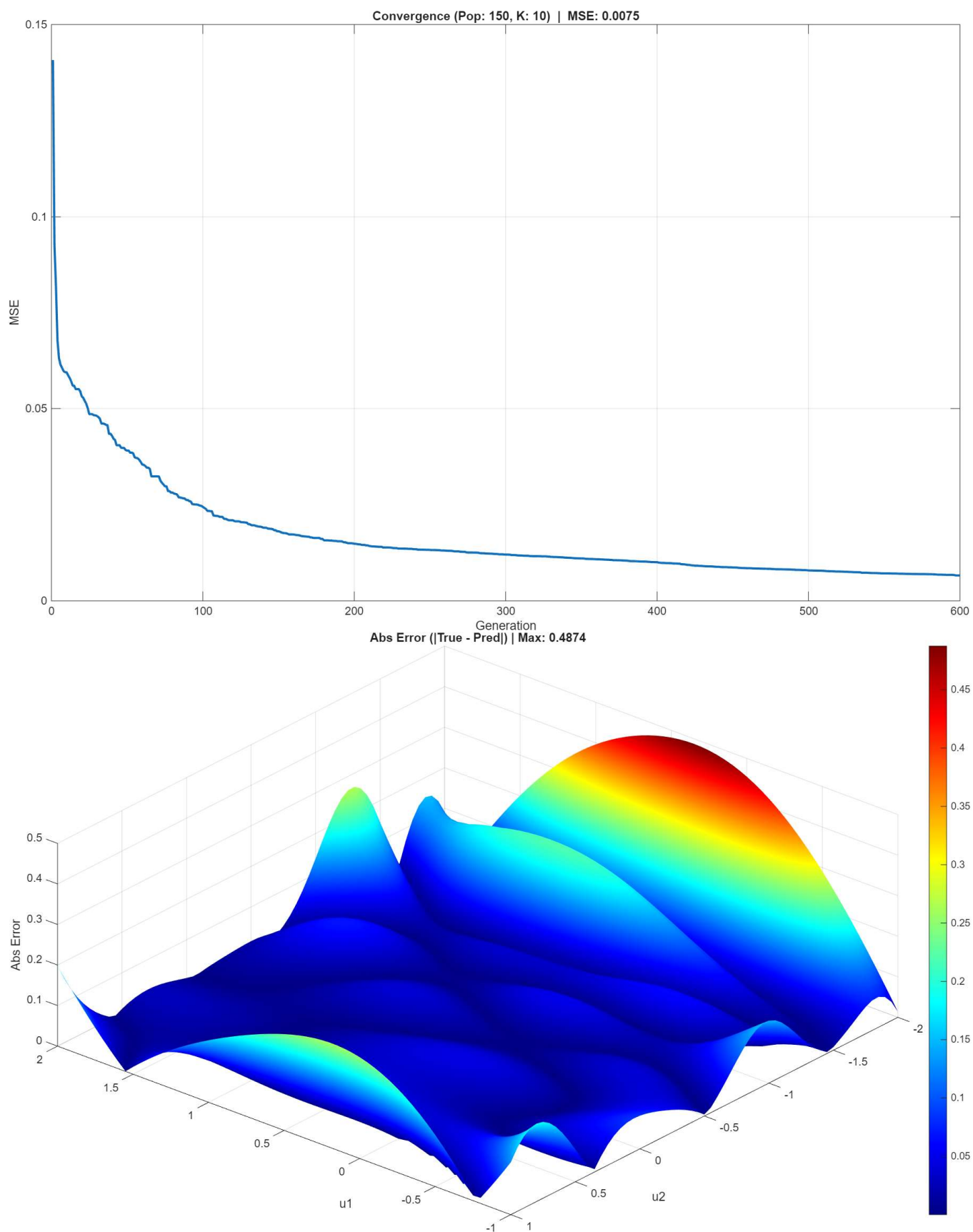


IV. ΕΠΙΡΡΟΗ ΠΛΗΘΥΣΜΟΥ

Θα κάνουμε μερικούς ελέγχους πάνω στην επιρροή του πληθυσμού,



Για πληθυσμό 50 παρατηρούμε, όπως είναι και αναμενόμενο, αύξηση του MSE με το approximated function να έχει παρόμοια μορφή με εκείνη πληθυσμού 100.



Για πληθυσμό 150 το MSE έχει ελαφρώς αυξηθεί και το απόλυτο error έχει αρκετά εντονότερο χαρακτήρα στον άξονα u_1 σε

αντίθεση με το αντίστοιχο διάγραμμα για πληθυσμό 100. Βέβαια η αύξηση του MSE δεν στέκει θεωρητικά και πιθανότατα οφείλεται στην τυχαιότητα του πειράματος, όμως αυτό μας δείχνει ότι η αύξηση του πληθυσμού δεν βελτιώνει σημαντικά το MSE για pop>100 και άρα η αρχική μας επιλογή αρκεί.

V. ΠΡΟΤΕΙΝΟΜΕΝΗ ΑΝΑΛΥΤΙΚΗ ΕΚΦΡΑΣΗ

Η τελική αναλυτική έκφραση που προκύπτει για K=10 και Pop=100 περιγράφεται από τον παρακάτω πίνακα παραμέτρων.

Term	Weight	C1	Sigma1	C2	Sigma2
1	0.6627	1.0607	0.8102	-0.0926	1.3714
2	1.1461	0.0225	0.7567	-1.9608	0.2008
3	-0.3052	0.3849	0.7761	-1.1784	0.9213
4	0.4232	-0.0629	0.5942	-0.2262	0.5857
5	0.4554	0.0645	0.7508	-0.5038	0.3453
6	0.2714	1.8663	0.6661	-0.9766	0.3218
7	0.3316	0.1431	0.9047	0.9979	0.2267
8	-0.5699	1.0380	0.8970	0.0471	0.3840
9	-0.8148	-0.1245	0.8844	-0.9985	0.7059
10	-0.2903	0.1396	0.8152	-1.0653	0.8476

Στον τύπο:

$$y(u_1, u_2) = \sum_{k=1}^{10} \left[w_k \cdot \exp \left(\frac{(u_1 - C_1)^2}{2\sigma_1^2} + \frac{(u_2 - C_2)^2}{2\sigma_2^2} \right) \right]$$

VI. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία ασχοληθήκαμε με την προσέγγιση της άγνωστης συνάρτησης $f(u_1, u_2)$ μέσω ενός Γενετικού Αλγορίθμου, βελτιστοποιώντας τις παραμέτρους ενός γραμμικού συνδυασμού Γκαουσιανών συναρτήσεων. Από τα πειράματα και την ανάλυση των αποτελεσμάτων, προκύπτουν τα ακόλουθα συμπεράσματα:

- Το K = 10 παρουσίασε το μικρότερο σφάλμα ωστόσο το K = 5 μπορεί να θεωρηθεί και αποδοτικότερο εάν δώσουμε περισσότερο βάρος στην πολυπλοκότητα.
- Η ανάλυση ευαισθησίας ως προς τον πληθυσμό έδειξε ότι οι 100 υποψήφιος λύσεις αποτελούν την ιδανική επιλογή. Μικρότεροι πληθυσμοί (50 άτομα) οδήγησαν σε πρόωρη σύγκλιση και υψηλότερο σφάλμα, ενώ η αύξηση του πληθυσμού στα 150 άτομα δεν απέδωσε.
- Η μικρή απόκλιση μεταξύ του σφάλματος εκπαίδευσης και του σφάλματος επαλήθευσης επιβεβαιώνει ότι ο αλγόριθμος απέφυγε την υπερπροσαρμογή.