Black-Box Optimization Capstone Project

Week-by-Week ML Strategy and Results Journal

Complete 12-Week Analysis

Week 1: Initial Exploration

F1 (2D): Started at center [0.10, 0.10] giving output 0.0. This function appeared completely flat in the low region. No signal to work with yet. Pure exploration strategy.

F2 (2D): Also tested center [0.10, 0.10] giving 0.0892. Seemed like a simple 2D function, didn't realize yet how stochastic it would be. No modeling yet.

F3 (3D): Center point [0.80, 0.80, 0.80] giving -0.1055. Trying to minimize (get closer to zero). Started in mid-range to explore from there. Manual exploration.

F4 (4D): Perfect center [0.5, 0.5, 0.5, 0.5] giving -3.986. Quadratic bowl shape was immediately obvious from this single point. Mathematical intuition guided this.

F5 (4D): Center [0.30, 0.30, 0.30, 0.30] giving 136.85. Low output, didn't know yet this was a completely different regime. Random sampling approach.

F6 (5D): Started high [0.75, 0.75, 0.75, 0.75, 0.75] giving -1.521. Lower is better for this function, so starting high gave me room to explore downward.

F7 (6D): Started at maximum [1.0, 1.0, 1.0, 1.0, 1.0, 1.0] giving 0.000034. Terrible result, clearly corners weren't good here.

F8 (8D): Center [0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1] giving 9.542. Decent starting point for 8 dimensions.

Key Learning: Initial exploration showed F4 was a simple bowl, F1/F7 had sparse signals, and F5 was hiding something. No ML yet.

Week 2: Random Perturbations

F1: [0.12, 0.08] giving 0.0. Still zero, no signal found yet.

F2: [0.12, 0.08] giving 0.0705. Slight perturbation from W1, slightly lower output.

F3: [0.95, 0.95, 0.95] giving -0.0919. Moved toward upper boundary, improved toward zero.

F4: [0.3, 0.3, 0.3, 0.3] giving -4.306. Moved away from center, performance degraded. Bowl confirmed.

F5: [0.28, 0.32, 0.30, 0.29] giving 137.29. Tiny perturbation, same output. Still in low regime.

F6: [0.3, 0.3, 0.3, 0.3, 0.3] giving -1.139. Moved lower, improvement.

F7: [0.2, 0.2, 0.2, 0.2, 0.2, 0.2] giving 0.408. Massive improvement from W1. Mid-range better than corners.

F8: [0.12, 0.09, 0.11, 0.10, 0.08, 0.13, 0.11, 0.09] giving 9.554. Slight random perturbation.

Key Learning: F7 showed biggest sensitivity. F4 bowl structure confirmed. Still manual exploration.

Week 3: Introduction to ML

F1: [0.21, 0.11] giving 0.0. Started fitting Gaussian Process with RBF kernel but with 3 zeros, GP predicted flat zeros. Useless.

F2: [0.21, 0.11] giving 0.0295. GP with RBF kernel showed slight gradient toward lower values but huge uncertainty.

F3: [0.98, 0.99, 0.87] giving -0.0856. GP with Matérn kernel (nu=2.5) suggested upper boundary exploration. First useful model prediction.

F4: [0.44, 0.29, 0.35, 1.25] giving -30.129. GP showed bowl shape but I tested boundary anyway. Terrible.

F5: [0.345, 0.265, 0.374, 0.204] giving 131.78. GP with RBF showed flat landscape, no gradient.

F6: [0.49, 0.02, 0.45, 0.40, 0.32] giving -1.123. Random Forest feature importance suggested x1, x2 matter most.

F7: [0.21, 0.19, 0.21, 0.19, 0.17, 0.19] giving 0.347. GP uncertainty huge in 6D with 3 points.

F8: [0.29, 0.25, 0.02, 0.29, 0.14, 0.22, 0.25, 0.30] giving 9.548. Used Latin Hypercube sampling theory.

Key Learning: First week using ML. GP useful for F3, useless for high-D functions with too little data.

Week 4: Model Refinement

F1: [0.14, 0.14] giving 0.0. GP predicting zeros everywhere. Model right but unhelpful.

F2: [0.14, 0.14] giving 0.0150. GP with RBF showed weak gradient. Used Expected Improvement acquisition.

F3: [0.949, 0.966, 0.808] giving -0.0786. GP gradient analysis with Matérn kernel and ARD (Automatic Relevance Determination). Best result, became reference.

F4: [0.51, 0.60, 0.57, 0.01] giving -12.492. GP showed symmetric bowl but wanted to verify boundary.

F5: [0.197, 0.320, 0.300, 0.290] giving 140.74. Tried Neural Network (2-layer MLP) to find non-linearities GP missed. Both models agreed: stuck in local minimum.

F6: [0.69, 0.001, 0.04, 0.001, 0.001] giving -2.067. Random Forest identified x1, x2 as important. Direction was wrong.

F7: [0.08, 0.32, 0.15, 0.28, 0.41, 0.27] giving 0.568. GP with Matérn suggested mid-range. Used UCB acquisition with beta=2.0.

F8: [1.0, 0.001, 1.0, 0.001, 0.001, 1.0, 1.0, 0.001] giving 4.180. Extreme corner test. GP useless in 8D with 4 points.

Key Learning: GP with ARD useful for identifying important dimensions. F3 and F7 responding well to models.

Week 5: The F5 Breakthrough

F1: [0.08, 0.08] giving 0.0. GP confidently predicted zeros everywhere after 5 zeros. Model accurate but function has sparse signal.

F2: [0.111, 0.100] giving 0.1300. GP suggested this region with Expected Improvement. First clear GP success.

F3: [1.01, 1.01, 0.82] giving -1.1543. Pushed past boundary despite GP warning. Constraint violation.

F4: [0.66, 0.30, 0.30, 0.36] giving -7.262. Still exploring asymmetry despite GP showing bowl.

F5: [0.99, 0.90, 0.98, 0.93] giving 5549.45. BREAKTHROUGH. Random Forest feature importance showed x1, x3 critical. Pushed toward corners. GP trained on low-regime (137±5) completely failed to predict this 40x jump. Models can't predict regime shifts. This changed everything.

F6: [0.26, 0.18, 0.50, 0.48, 0.41] giving -1.092. Used GP gradient to push x1/x2 lower. Model was right.

F7: [0.05, 0.50, 0.25, 0.20, 0.15, 0.85] giving 0.836. Pushed x6 high based on GP extrapolation. GP uncertainty high, should have been cautious.

F8: [0.05, 0.25, 0.25, 0.25, 0.25, 0.25, 0.05, 0.05] giving 9.643. Kernel density estimation of successful points.

Key Learning: F5 has two regimes. GP failed at predicting regime shift. Random Forest feature importance more useful for regime discovery. This function is now 99% of total score.

Week 6: Model Retraining

F1: [0.45, 0.45] giving 0.0128. Finally got non-zero! GP trained on zeros suggested diagonal exploration.

F2: [0.11, 0.10] giving 0.0468. Near W5 input, much worse. GP predicted similar performance but was wrong. First sign of stochasticity.

F3: [0.928, 0.832, 0.004] giving -0.1161. Tried low x3 despite GP suggesting high. Ignored model, made worse.

F4: [0.2, 0.2, 0.95, 0.4] giving -19.009. Still testing asymmetry despite GP bowl model.

F5: [0.985, 0.905, 0.975, 0.925] giving 5398.58. Tried x2=0.905 using GP gradient on high-regime data. Dropped 151 points. Retrained GP on high-regime only, kernel length scales changed dramatically.

F6: [0.1, 0.1, 0.7, 0.7, 0.6] giving -1.231. Expected Improvement acquisition didn't improve.

F7: [0.06, 0.48, 0.25, 0.20, 0.40, 0.75] giving 1.435. GP posterior mean surface showed peak around 0.75. Major improvement. GP was right.

F8: [0.18, 0.15, 0.20, 0.15, 0.25, 0.15, 0.15, 0.18] giving 9.676. K-means clustering on successful past points.

Key Learning: Retrained GP on F5 high-regime only. x2 optimum might be lower than 0.90. F7 x6 around 0.75 per GP.

Week 7: GP Refinement & Peaks

F1: [0.48, 0.48] giving 0.000008. GP gradient extrapolation, performance collapsed. GP overconfident outside training data.

F2: [0.111, 0.100] giving -0.0246. Same input as W5, different output. Added WhiteKernel for noise modeling. Didn't help.

F3: [0.99, 0.99, 0.99] giving -0.4273. Ignored GP with ARD showing x3 should be lower.

F4: [0.65, 0.65, 0.65, 0.65] giving -15.158. GP showed center optimal. Finally decided to trust GP.

F5: [1.0, 0.853, 1.0, 0.977] giving 6158.08. Major refinement using GP gradient on high-regime. GP with Matérn and ARD showed x1, x3 matter most (short length scales). New best by 610 points. Model worked perfectly.

F6: [0.15, 0.15, 0.50, 0.50, 0.70] giving -1.5517. GP gradient suggested this but was wrong.

F7: [0.038, 0.462, 0.239, 0.171, 0.378, 0.734] giving 1.478289. GP posterior mean optimization. Found peak at 0.734. Matérn nu=2.5 kernel worked perfectly.

F8: [0.177, 0.194, 0.170, 0.194, 0.294, 0.143, 0.109, 0.208] giving 9.692075. Bayesian Optimization with Expected Improvement. Model-guided optimization successful.

Key Learning: F5 optimum found using GP with ARD. F7 peaked at 0.734 per GP. F8 GP working well. ARD length scales show which variables are critical.

Week 8: Catastrophic Failure

F1: [0.405, 0.428] giving 0.462531. GP posterior mean showed peak around [0.40, 0.43]. Found it. GP successful after recalibrating.

F2: [0.111, 0.100] giving -0.0246. Third different output. Tried Bayesian Ridge Regression. Still useless. No model handles pure noise.

F3: [0.944, 0.965, 0.807] giving -0.088593. GP correctly predicted this would be similar to W4.

F4: [0.498, 0.502, 0.500, 0.500] giving -3.9857. GP quadratic bowl model identified exact minimum. Model completely correct.

F5: [0.855, 0.852, 1.000, 0.979] giving 4415.99. DISASTER. Changed x1 from 1.0 to 0.855. GP with ARD had short length scale for x1 meaning very sensitive. Output crashed by 1742 points. x1=1.0 is critical. Model told me x1 was critical (short length scale), I misinterpreted it.

F6: [0.258, 0.178, 0.501, 0.482, 0.412] giving -1.064064. GP gradient descent worked. New best.

F7: [0.039, 0.463, 0.240, 0.172, 0.379, 0.742] giving 1.463533. GP uncertainty suggested 0.73-0.75 similar. Slight decline. Peak at 0.734 confirmed.

F8: [0.179, 0.196, 0.172, 0.196, 0.292, 0.145, 0.111, 0.210] giving 9.691885. GP posterior variance to find low-uncertainty region.

Key Learning: x1=1.0 non-negotiable for F5. ARD length scales show sensitivity - short means don't touch it. F1 peaked per GP. F4 at optimum.

Week 9: Model-Guided Recovery

F1: [0.404, 0.425] giving 0.457432. GP posterior mean micro-adjustment, slight decline. GP uncertainty very low, suggesting we're at optimum.

F2: [0.100, 0.103] giving 0.033044. Tried ensemble model (GP + Random Forest). Still terrible. No model works on noise.

F3: [0.954, 0.967, 0.808] giving -0.086019. GP predicted -0.084, got -0.086. Pretty accurate. Using GP carefully now.

F4: [0.489, 0.491, 0.485, 0.486] giving -4.430389. GP test if bowl is perfectly quadratic. Made worse. GP symmetric bowl was correct.

F5: [1.0, 0.831, 1.0, 0.988] giving 6117.763. Reverted x1 to 1.0, tried x2=0.831, x4=0.988 per GP gradient. Got 6118. GP predicted 6150±80, pretty close. Model working but not finding improvements.

F6: [0.245, 0.162, 0.507, 0.482, 0.418] giving -1.068886. GP fine-tuning. GP predicted -1.065. Very accurate.

F7: [0.022, 0.448, 0.257, 0.153, 0.397, 0.694] giving 1.431577. Tried x6=0.694 to verify GP surface. Dropped as predicted. Peak at 0.734.

F8: [0.206, 0.195, 0.196, 0.188, 0.324, 0.148, 0.082, 0.213] giving 9.680707. Bayesian Optimization. Converging, GP uncertainty very low.

Key Learning: GP models becoming very accurate for F3, F5, F7, F8. Starting to suspect x4=1.0 boundary worth testing - GP shows high uncertainty there because never sampled.

Week 10: Duplicate Failure

All Functions: Exact duplicates of W9. Wasted query. No duplicate checking implemented.

Noise Detection: F2 output changed (0.033 to 0.006) - stochastic confirmed. F6 changed (-1.069 to -1.163) - moderate noise. F3 slight change - possible noise. F1, F4, F5, F7, F8 exact same outputs - deterministic confirmed.

Model Updates: Updated all GP models with appropriate noise kernels. F2 high noise, F6 moderate noise, F3 slight noise, others zero noise.

Key Learning: Noise characterization complete. Implemented hash-based duplicate detection for next week. Wasted entire week but got valuable noise data for GP modeling.

Week 11: Empirical Over Models

F1: [0.410, 0.433] giving 0.525860. GP showed plateau but empirical pattern (W8 to W11 both improved stepping up) trusted. Improved 13.7%. Peak extends higher than GP predicted.

F2: [0.111, 0.101] giving 0.068939. Fourth different output on same input. Abandoned all modeling.

F3: [0.949, 0.966, 0.830] giving -0.061424. GP gradient with Matérn predicted -0.058, got -0.061. Close. Improved 21.9%. GP gradient correct.

F4: [0.501, 0.499, 0.501, 0.499] giving -3.985737. GP quadratic model predicted same as W8. Basically identical.

F5: [1.0, 0.853, 1.0, 1.0] giving 6526.44. BREAKTHROUGH. Pure empirical gap-filling - tested 0.93, 0.925, 0.977, 0.979, 0.988 but never 1.0. GP predicted 6288±109 (trained on high-regime only). Got 6526, within 2 sigma. Gained 368 points. x4=1.0 unlocked new regime. Sometimes empirical gaps beat model optimization.

F6: [0.200, 0.162, 0.507, 0.482, 0.418] giving -1.185411. GP gradient suggested lower. Made worse. GP wrong, empirical data right.

F7: [0.022, 0.448, 0.240, 0.153, 0.397, 0.760] giving 1.315567. Tested x6=0.760 gap. GP suggested 0.800+. Dropped 11%. Empirical pattern (0.694 to 0.734 to 0.742 to 0.750 to 0.850) showed peak at 0.734. Empirical data beats GP extrapolation.

F8: [0.207, 0.196, 0.197, 0.189, 0.325, 0.149, 0.083, 0.214] giving 9.681707. GP posterior mean. Converged. GP confident, low uncertainty.

Key Learning: x4=1.0 massive breakthrough (+368). Empirical gap-filling (testing obvious untested values) can beat sophisticated models. GP good for smooth interpolation, bad at boundaries and regime shifts.

Week 12: Regime-Shift Testing

F1: [0.420, 0.443] - Testing if peak continues upward. GP shows slight positive gradient but high uncertainty. Following empirical trend (W8 to W11 both improved) rather than uncertain GP. Conservative step 0.010.

F2: [0.111, 0.100] - Using W5 historical best (0.130). Completely abandoned all modeling. No GP, no ensemble, nothing works on noise.

F3: [0.949, 0.966, 0.850] - Pushing x3 from 0.830 to 0.850. W11 confirmed GP gradient correct (Matérn with ARD). GP predicts -0.048±0.015. Model-guided with confidence.

F4: [0.498, 0.502, 0.500, 0.500] - Using W8 near-center. GP quadratic bowl shows within 0.001 of optimum. Already at optimum per model.

F5: [1.0, 0.920, 1.0, 1.0] - THE BIG BET. Testing if x2=0.920 stacks with x4=1.0. Regime-shift hypothesis. Retrained GP on high-regime only (dropped W1-W4 low-regime to avoid kernel collapse). GP with Matérn predicts 7144±163 but uncertainty very high - extrapolating beyond x2=0.905 (previous max). Historical pattern with x4=0.93-0.977 showed x2=0.905 failed, but maybe x4=1.0 changed landscape. Testing if regime boundaries shift when variables hit boundaries. Model-informed hypothesis.

F6: [0.245, 0.162, 0.507, 0.482, 0.418] - Reverting to W8 best. W11 proved GP gradient wrong. Using empirical best over model.

F7: [0.038, 0.462, 0.239, 0.171, 0.378, 0.734] - Reverting to W7 peak. W11 confirmed 0.734 optimal. GP suggested higher, empirical data showed peak. Trusting data over model.

F8: [0.206, 0.195, 0.196, 0.188, 0.324, 0.148, 0.082, 0.213] - Using W9 best. GP shows converged around 9.68 with very low variance. No further optimization possible per model.

Strategy: 90% on F5 regime-shift hypothesis (does x4=1.0 enable higher x2?), F1 empirical trend, F3 GP gradient. 10% safety nets. Hybrid: using GP for smooth functions (F3, F8) but trusting empirical patterns for boundaries (F5, F7) and trends (F1). Floor is 6526. GP predicts expected value ~6620 with high uncertainty. Model-informed but empirically-grounded.

ML Techniques Used: Gaussian Processes with RBF and Matérn kernels, ARD for feature importance, Expected Improvement and UCB acquisition, Bayesian Optimization, Random Forest, noise kernel modeling, high-regime-only training, empirical gap analysis, ensemble averaging. Models work for interpolation and smooth gradients, empirical pattern recognition better for boundaries and regime shifts.