On Par Golf

Leveraging Gaussian Process Regression for Perfect Golf

Abstract

Golf is fundamentally a problem of sequential decision-making under uncertainty, where a surplus of strategic options often obscure the optimal path to the hole. This project presents a data-driven framework that quantifies player-specific optimal strategies by leveraging Gaussian Process Regression (GPR). Moving beyond traditional biomechanical analysis, we model the golf course as a continuous risk surface, learning the "scoring potential" of every location based on course geometry, historical data, and the player's shot tendencies. The primary output of this framework is a tangible, actionable strategy: a specific optimal club and aimpoint recommendation for any coordinate on a given hole. By simulating thousands of shot outcomes against our risk surfaces, the model solves for the decision rule that best satisfies the player's objective. Whether the goal is minimising the expected strokes to hole out or maximising the probability of securing a birdie, our framework is capable of providing a mathematically grounded roadmap, replacing intuition with calculated risk management in order to optimise performance.

December 13, 2025

1. Introduction and Research Question / Background & significance

Golf is a deceptively simple sport: get the ball into the hole in the fewest number of strokes possible. Upon closer inspection, however, this objective is marred by complexity. Over the course of a round, a player faces 18 holes made up of differing lengths and topographies, riddled with obstacles: water hazards, sand traps, and out-of-bounds limits with associated penalties. Consequently, no player will ever face the exact same shot twice.

Traditionally, golf performance has been viewed through a biomechanical lens. Striking a ball smaller than 2 inches in diameter with a club head moving at 135 mph toward a distant target is no small feat. Consequently, the status quo is an obsession with technical perfection: the pursuit of optimising the biomechanical motion of the swing to maximise energy transfer and consistency. This fixation is evidenced by the proliferation of AI-driven swing analysis tools such as Sportsbox.AI, GolfFix, and Sparrow Golf.

Said focus on mechanics is not statistically unfounded. In 2014, Mark Broadie provided statistical evidence that driving distance is a key factor and strong predictor of success on the PGA Tour[1] - often more important than accuracy or putting (which had long been hypothesised to be the most important part of the golf game) (Broadie 2014). PGA Tour professionals who drive their ball 20 yards further than their peers typically save about 0.75 strokes per round, translating to a substantial 3 strokes over a four-day tournament. The only way to hit the ball further is to maximise "smash factor", which implies optimising technique.

While mechanical questions are valid, they give rise to critical counter-questions: is there a way to optimise the tools we already have? How can we improve our scores without necessarily "getting better" at swinging a club?

Ideally, the tour professional's swing is mechanically optimised, yet scoring divergences persist between the top 50 players, consistent winners, and the rest. For the recreational amateur, while technique has a large margin for improvement, meaningful mechanical changes require time, patience, and financial investment. Both professionals and amateurs can benefit from taking a different approach.

Golf is not solely about physical execution; it is a problem of complex sequential decision-making under uncertainty. There are a million sequences of different shots that can get the ball into the bottom of the hole. A player faces the course with up to 14 clubs and the possibility of hitting dozens of shot types. Such flexibility creates a "paradox of choice"; a psychological hurdle where an abundance of options obscures the best path forward (Schwartz 2015) and where decision theory and statistics become relevant. By treating every shot as a mathematical calculation of risk versus reward, we can determine the optimal path to the hole.

To this end, our paper proposes a data-driven framework allowing an individual player to quantify their optimal strategy on a given golf course by leveraging both venue- and player-specific data. We approach the decision theory scenario through a Bayesian lens, using prior performance data to help inform future strategic choices. By implementing Gaussian Process Regression (GPR), we construct continuous predictive surfaces that evaluate the "scoring potential" of all possible landing zones on a golf course. This model allows us to mathematically define the player's objective through different "loss functions" - whether minimising the expected number of strokes to hole out for consistent play, or maximising the raw probability of a birdie when the need to exhibit a more aggressive play style arises. Ultimately, our framework replaces guesswork with calculation, identifying the specific club and aimpoint combination that statistically optimises a player's outcome from any location on the course.

The remainder of this paper is organised as follows: Section 2 places this work within the broader context of sports analytics and reviews existing literature. Section 3 describes the data we used and Section 4 explores the modeling mechanisms employed. Section 5 presents the results of our simulations, and Section 6 discusses the limitations of the current framework alongside potential avenues for future expansion. Appendix A presents the data generating mechanism we used to supplement our existing data and Appendix B includes a glossary of relevant terms.

[1]PGA Tour: Professional Golf Association, US-based organisation running the main professional golf circuit.

2. Background and Significance

The shift toward objective strategy in sports is not unprecedented. This project draws direct inspiration from SABRmetrics (Albert 2010): the empirical analysis of baseball that uses statistics to evaluate player performance and inform objective, data-driven decisions. Coined by Bill James, this non-traditional statistical investigation famously allowed the Oakland Athletics to remain competitive against teams with much larger budgets by using data to identify and acquire undervalued players (Lewis 2003). Here, we apply the same philosophy to golf: replacing subjective intuition with statistical decision-making.

Historically, golf analysis relied heavily on total score and basic counting stats (fairways hit, putts per round), which often failed to capture the true quality of a shot. This landscape changed with the introduction of Strokes Gained, SG (Broadie 2014). The SG metric redefined performance evaluation by establishing baselines for the expected number of strokes to hole out from any distance and lie on the course. For the first time, players had access to interpretable metrics that could isolate specific strengths and weaknesses in their game relative to the field.

Strokes Gained paved the way for applied course strategy systems, most notably DECADE (Fawcett 2024). DECADE is a proprietary course-management system built largely on generalised SG estimates, allowing players to make informed decisions grounded in statistical insight rather than emotion. The efficacy of such systems is well-documented; for instance, Will Zalatoris rapidly ascended from a World Amateur Golf Ranking (WAGR) of 3,000+ to the top 3 in under three months after adopting the DECADE pedagogy (Fawcett 2024).

The application of statistics to golf strategy is still an evolving field. Previous work has primarily focussed on three distinct methodologies: Markov Decision Processes (MDPs), neural network classification and reinforcement learning. While these studies provide a solid foundation for objective decision making, they each exhibit limitations that this project seeks to address.

Stauffer and Guillot 2025 present a study on strategy optimisation in Stroke Play for professional golfers. Their main goal was to demonstrate that a data-driven MDPs could be solved exactly for large-scale, real-world instances without relying on heuristics or Q-learning approximations. That is, they could solve the golf strategy problem perfectly for every point on the course without implementing approximations.

They inferred personalised “TrackMan profiles” for PGA Tour players using historical ShotLink data and constructed a 2D simulation model of a course in `C++`. By defining states as pixels and actions as triplets of (state, targeted distance, direction), they applied a value iteration algorithm to solve the stochastic shortest path (SSP) problem exactly. The model successfully demonstrated computational feasibility, allowing for the automated construction of strategic booklets outlining optimal shots from any position. The study was limited by a number of simplifying assumptions: it treated greens as homogeneous flat surfaces, assumed that in historical data players were always aiming at the pin, and always assumed risk-neutral decision-making, often leading to simulated strategies that were riskier than those observed in professional play.

Khazaeli and Javadpour 2024 focussed on implementing a probabilistic classification model to predict approach shot outcomes incorporating environmental factors using multi-level neural networks. They combined over 700 shots from NCAA Division I golfers, and augmented them via Monte Carlo simulation. They then trained a model on 27 features, including distance, terrain, and atmospheric conditions such as wind strength and lie. They validated the hypothesis that incorporating environmental features leads to higher predictive accuracy than solely relying on distance (which still remained the dominant attribute).

Despite its predictive power, this study was limited in scope, focussing solely on approach shots and relying heavily on synthetic data augmentation. Furthermore, it functioned primarily as a classification tool used to predict outcome zones rather than providing overall strategy optimisation.

Sugawara, Kawamura, and Suzuki 2013 explored skill level and strategy through reinforcement learning. They applied Q-learning within MDP frameworks to minimise the expected score for a hole. They defined golfer skill via shot variance and utilised conditional probability distributions to account for penalties from trees and hazards. By comparing strategies for skilled versus unskilled agents on a simulated Augusta National, they provided the key insight that optimal strategy is highly dependent on skill level. Their simulation, however, lacked realism; the putting model was oversimplified, ignoring crucial information such as green slopes and calculating expected putts solely based on distance. The failure to model crucial topographical meant that their strategies were optimised for two-dimensional versions of golf, rather than the three-dimensional reality of the game.

Current literature largely relies on generalised estimates, rigid assumptions regarding course geometry, or limited subsets of the game (e.g., only modelling approach). This project addresses these gaps through a novel methodological framework. As an individual sport, golf requires player-specific solutions. We move past generalised estimates in commercial systems by integrating course-specific data and individual player tendencies into our model. Unlike previous studies that generally optimised for a single metric (lowest score), this framework implements variable loss functions. This allows for strategy optimisation based on different definitions of success, such as minimising expected strokes (conservative) versus maximising birdie probability (aggressive). Uniquely, our project leverages Gaussian Process Regression (GPR). Previously unexplored in this domain, we shift the analysis to the Bayesian paradigm, allowing us to inherently model uncertainty via posterior variance, providing a robust probabilistic measure of reliability that frequentist approaches lack.

3. Data

In order to understand the data inputs that we used when modelling the optimal golf strategy, it is valuable to distinguish between the ideal dataset (which would drive a fully deployed version of our tool) and the proxy dataset which we actually used in this proof-of-concept. Our model is agnostic to its data source; it requires three specific layers of information to work: course geometry for lie classification, location-specific outcome data, and player shot distribution data.

3.1 The Ideal Data

In a fully realized application, such as one using the PGA Tour's ShotLink system and a user's shot history, our model would consume the following:

1. **Course Geometry for Lie Classification:** A high resolution, bird's-eye view vector map of the hole (where the tee is at $(0, 0)$). This allows our model to embed course geometry into the model by instantly classifying any coordinate (x_1, x_2) as a specific lie type (i.e. Fairway, Rough, Deep Rough or Bunker), triggering the appropriate modelling and decision logic.

2. **Location-Specific Outcome Data:** A large repository of historical shots for a hole with a given position. For any recorded shot from a given starting coordinate, (x_1^0, x_2^0) which can be thought of as the origin in the Cartesian plane, we would observe the discrete outcomes of how many shots it took players to hole out to the corresponding hole. The numbers of shots would encode the underlying "difficulty" of a specific spot: locations with the same Euclidean distance to the pin where the average strokes to hole out are higher than others are inherently "harder" (see Figure 1).

3. **Player Shot Distribution Data** A library of "stock shots" labeled by the player. Whether recorded on a launch monitor (e.g., Trackman, GCQuad) or manually logged, the dataset would consist of the landing coordinates (x_1^l, x_2^l) generated by hitting shots from (x_1^0, x_2^0) aiming down a set target line (where the lateral offset is zero) for every repeatable shot in a player's repertoire[2].

[2] In addition to the 14 clubs they usually carry in a bag, competitors usually identify different kinds of shots they can hit with their clubs. These shot-types tend to impact the ball's flight affecting factors such as the ball's lateral curve, apex height, or total carry distance. While most players recognise and play standard full swing shots, a player may choose to name additional shots in

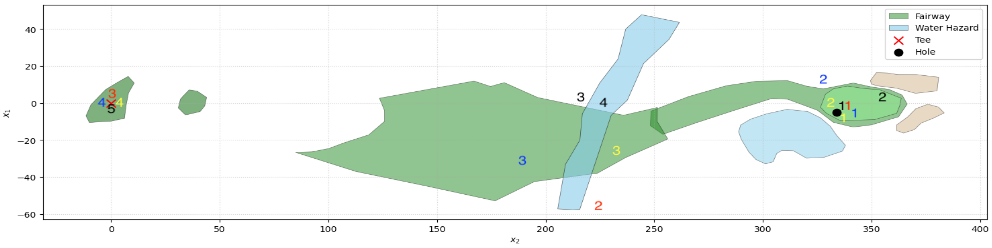


Figure 1. Visualisation of Location-Specific Outcome Data. The figure illustrates the recorded paths of different players (distinguished by colour) on the digitised hole. The numeric labels at each coordinate indicate the observed number of strokes they needed to hole out from that specific location. The discrete observations with coordinates serve as the raw input for encoding the underlying difficulty of every position on the course.

Ideally, we would have labelled data from different starting lies such as rough or fairway bunkers, accounting for the increased dispersion and variance that results from hitting off of different surfaces. More concretely, the player's shot library would contain observations encoding the tuple $\left((x_1^0, x_2^0),\ (x_1^l, x_2^l),\ Lie\right)$ for every specific labelled $Shot$ type. From data in said format, we derive enough information to fully characterise the conditional landing distribution $P\left((x_1^l, x_2^l) \mid Shot,\ (x_1^0, x_2^0),\ Lie\right)$.

3.2 Data Implemented in this Study

To demonstrate the viability of our framework, we developed and implemented a robust proxy dataset (defined in Appendix A) that provided an appropriate substitute for the three core data layers required by the model. We successfully digitised a real golf course using satellite imagery and vector mapping tools, achieving the fidelity of the ideal data for our course geometry. In absence of proprietary PGA Tour ShotLink data, we built a data-generating mechanism to emulate location-specific strokes-to-hole-out outcomes, supplemented by empirical estimates. Finally, we established a representative LPGA[3] Tour player profile using publicly available launch monitor statistics to model the player's shot distribution data.

3.2.1 Course Geometry

We digitised hole 9 of Mountain Meadows Golf Course in California, a par 3 chosen for its rich interaction of obstacles such as bunkers and water hazards near the green. The course layout was scraped using OpenStreetMap (Bennett 2010), and refined in QGIS (QGIS Development Team 2024) to create accurate vector polygons compatible with the `Shapely` (Gillies et al. 2024) library in Python. To test chained-decision-making (par 4 strategy), we manually generated a hypothetical extension of hole 9 adding a secondary water hazard to increase strategic complexity.

3.2.2 Emulating Location-Specific Outcome Data

- **On the green:** We generated a dense sample of synthetic putting outcomes for our digitised green taking into consideration slope and distance in a three dimensional setting (see Appendix A.2). The data generated simulates the "ideal" data described in subsection 3.1, providing a count of the strokes to hole out for hundreds of points on the putting surface. The proxy data is displayed in Figure 2.
- **Off/Around the Green:** Generating realistic outcomes around the green, while having the potential

distinct ways (e.g., a player may name a specific shot an L-shot, while another may name it a 3/4 swing). Trackman allows the player to record and save shots under different names, which are then modelled in such a way that the output of our framework includes recommendations implementing such labelled shots.

[3] LPGA: Ladies Professional Golf Association

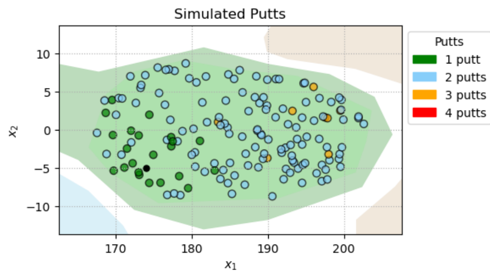


Figure 2. Synthetic putting dispersion pattern on our green. The figure shows the modelled distribution of putt outcomes relative to a fixed pin location (black dot). The position of each point represents a starting position of a sequence of putts, and the colours indicate the number of strokes it took to hole the ball from said position: green for one-putt, blue for two-putt, orange for three-putt, and red for four-putt results.

for being extremely informative for our model, was deemed infeasible (due to the lack of available estimates, and the variance of outcomes and approaches that different players may take when tackling short range chips or bunker shots). We therefore interpolated average "strokes to hole out" as a function of distance values from different lies from Mark Broadie's empirical estimates (derived from the PGA Tour Shot Link data from 2003-2012) and published in Every Shot Counts (Broadie 2014). These estimates can also be found in Appendix A.3 and allow us to assign a baseline expected value to any lie around the green without needing a generated point cloud for that specific coordinate.

It is worth noting that while we utilise LPGA Tour data for full-swing dispersion (see Section 3.2.3), we rely on PGA Tour baselines for short game outcomes due to the availability of their empirical estimates. We posit that the relative difficulty of short-game scenarios (the penalty of bunker shot or a short sided chip) remains comparable across professional tours. Moreover, as this framework is data-agnostic, the presence of mixed-source data does not undermine the validity of the optimisation logic.

- **Long Game:** We opted not to use granular historical outcome data for the long game, as such data is inherently confounded by variations in player skill and suboptimal strategic choices. Relying on raw stroke averages would pollute the model with historical inefficiencies rather than revealing the theoretical optimal path. Consequently, we strictly define the value of long-game locations recursively based on the potential of the subsequent approach shot, as outlined in Section 4.4.1.

3.2.3 Player Shot Distribution Data

We utilised a proxy shot distribution data set based on publicly available Trackman launch monitor statistics for a representative "average" LPGA tour player. The library encompassed 13 standard clubs from 60° wedge to a Driver (ranging approximately 66 to 250 yards), plus several partial wedges for shorter distances. For each of the 20 shot-types recorded in this proxy player's repertoire, the dataset contains 30 recorded landing coordinates, totalling 600 observations. We assume all of these shots are hit off the fairway (or a tee in the case of a Driver). These coordinates were generated from shots hit along a single target line (zero lateral offset), providing the necessary dispersion characteristics to model the player's skill profile, and can be viewed in Figure 3.

4. Modelling and Methods

4.1 Our Mathematical Framework: Decision Theory

Ben Hogan, who was widely regarded as one of the greatest ball strikers in golf history, famously stated that even on a great day, he would hit at most three truly great shots in a round (Jenkins 1998). The inherent

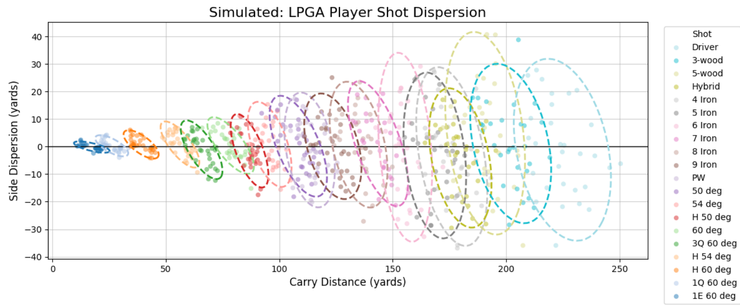


Figure 3. Stochastic Modelling of Player Dispersion. Visualisation of the proxy LPGA player's shot library. The covariance ellipses represent the 95% confidence region for landing coordinates across varying club types. Note the heteroscedasticity of the data, where dispersion (Σ_s) increases non-linearly with distance, a critical factor for the GPR model's risk assessment.

imperfection in performance means golf fundamentally poses a problem of decision-making under uncertainty.

In statistical decision theory, we usually seek a decision rule δ that minimises a specific loss function $\mathcal{L}$. In the context of a golf shot, our decision rule is a tuple $\delta = (s, \theta)$, where s represents the discrete choice of club/shot type and $\theta \in \mathbb{R}$ represents the continuous choice of aimpoint's lateral offset from the pin (where negative values denote aiming left of the pin, and positive values denote aiming right). The outcome of this decision, which is the actual landing coordinate of the ball, is denoted by the random vector $\mathbf{X} \in \mathbb{R}^2$. Given that human performance is imperfect and that there are ever-changing variables, the landing is stochastic, drawn from a probability density function conditioned on the player's decision: $\mathbf{X} \sim f(\mathbf{x}|s, \theta)$. We denote the realised outcome of this variable as $\mathbf{x}^l = (x_1^l, x_2^l)$.

The objective of our model is to find the decision δ minimising the expected cost of a shot. The loss function, $\mathcal{L}(\mathbf{X})$, measures the cost (or penalty, in strokes) incurred when the ball lands at coordinate $\mathbf{X}$. In a deterministic scenario where the outcome $\mathbf{X}$ was certain, we would simply minimise the known loss $\mathcal{L}(\mathbf{X})$. However, we are operating in an imperfect and therefore stochastic paradigm, where the outcome is governed by the probability density function $f(\mathbf{x}|\delta)$. The nature of golf requires minimising the risk, $R(\delta)$, defined as the expected loss over the player's shot outcome distribution:

$$R(\delta) = \mathbb{E}_{\mathbf{X}|\delta}[\mathcal{L}(\mathbf{X})] = \int_{\mathbb{R}^2} \mathcal{L}(\mathbf{x}) f(\mathbf{x}|\delta) d\mathbf{x}$$

The choice of loss function, $\mathcal{L}$, critically shapes the optimal strategy. In standard statistical modeling, we often select symmetric loss functions. For example, the symmetric Squared Error Loss, $\mathcal{L}^*(P, \hat{P}) = (P - \hat{P})^2$, used to derive the mean, assumes that an error of magnitude ϵ is equally detrimental regardless of direction.

However, golf, like many real-world scenarios often presents asymmetric loss landscapes. Consider civil engineers designing bridges: building a bridge a foot too long is a minor inefficiency and cost when compared to the catastrophic failure of building it a foot too short. Golf exhibits similar non-linearities. A shot missing a couple yards to the left might rest safely in a spot fractionally less optimal than the "perfect" spot, whereas one missing a couple yards right could plunge into a water hazard, incurring a fixed penalty stroke and a positional disadvantage.

While we can conceptually define these loss scenarios, the resulting risk function $R(\delta)$ remains an elusive unknown surface over the two dimensional domain of a golf course. We don't have an analytical formula that outputs the expected loss for any given coordinate $\mathbf{x}^l$: we only have sparse, noisy observations of shots and scores. To perform any sort of optimisation, we need to first learn the associated risk surface for a specific hole (no two holes are exactly alike!). The definition of such a surface requires a statistical method capable of taking in scattered data to predict a continuous landscape of expected strokes to hole out (ESHO), a task for which we can leverage Gaussian Process Regression (GPR).

In this project, we define two distinct risk landscapes that our GPR models can learn, corresponding to different player objectives.

1. *Minimising Expected Strokes to Hole Out.* Here, our loss function, $\mathcal{L}_\mu$, is the count of strokes to hole out from any given landing position. $\mathbf{x} \in \mathbb{R}^2$, and outputs an integer count of strokes.

   The function is defined as: $\mathcal{L}_\mu : \mathbb{R}^2 \to \mathbb{Z}^+$

   $$\mathcal{L}_\mu(\mathbf{x}^l) = \text{Strokes to hole out from } \mathbf{x}^l$$

   The associated risk, R_μ, is the expectation of the aforementioned loss given a specific decision δ. The risk function domain is consistent with the decision rule: $R_\mu : \mathcal{S} \times \mathbb{R} \to \mathbb{R}$, where, $\mathcal{S}$ is the discrete set of shot types a player chooses from.

   $$R_\mu(\delta) = \mathbb{E}[\mathcal{L}_\mu(\mathbf{X})|\delta]$$

   By minimising $R_\mu(\delta)$ we identify decisions where the average outcome (Expected Strokes to Hole Out) is the lowest. We naturally account for asymmetric hazards because the penalised observations fundamentally skew the local mean upwards. The actual optimisation mechanism is explored in Section 4.3 and 4.4.

2. *Maximising Birdie Probability.*

   Here, we employ a utility-based loss function, where "success" and, by complementation, loss, is binary. The loss is defined based on whether the expected score from the landing spot $\mathbf{x}$ fails to achieve the birdie target.

   The loss function $\mathcal{L}_B$, maps the landing position $\mathbf{x}$ to a binary outcome, $\mathcal{L}_B : \mathbb{R}^2 \to \{0, 1\}$.

   $$\mathcal{L}_B(\mathbf{x}) = \begin{cases} 0 & \text{if score from } \mathbf{x} < \text{Par Score, (i.e. Birdie or better)} \\ 1 & \text{if score from } \mathbf{x} \geq \text{Par Score (i.e. Par or worse)} \end{cases}$$

   The associated risk, R_B, is the probability of failure (not making a birdie or better), $R_B : (\mathcal{S} \times \mathbb{R}) \to [0, 1]$. Since the loss is binary, the expectation is equivalent to the probability of the loss occurring:

   $$R_B(\delta) = \mathbb{E}[\mathcal{L}_B(\mathbf{X})|\delta] = \mathbb{P}\,(\text{Strokes to Hole Out from } \mathbf{X} \geq \text{Par Score}|\delta)$$

   The minimisation of this value reflects the "must-score" scenarios where the magnitude of a miss is irrelevant (a par is as useless as a triple bogey).

   The actual optimisation mechanism is explored in Section 4.5.

4.2 Our Modelling Engine: Gaussian Process Regression

4.2.1 Gaussian Process Regression to Learn Risk in Golf

In order to minimise the risk functions defined above, helping us find the optimal decision rules, we have to first estimate their outcomes. We treat the unknown risk surfaces (expected strokes or birdie probability) corresponding to different holes as latent functions $f(\mathbf{x})$ to be learned from the observed data. An ideal mechanism to estimate the risk surface is Gaussian Process Regression.

Gaussian Process Regression is a tool that works intuitively with this project because golf courses are inherently spatial environments where nearby locations (on the same kind of surface) tend to have similar strategic values. If a player is 100 yards from the hole in the fairway, moving a couple yards on the same surface is unlikely to drastically change their expected score. GPR's smoothness assumptions align with the green, allowing us to interpolate values across unobserved regions based on spatial proximity.

Furthermore, golf scoring relationships are complex and non linear. A putt from 10 feet is not necessarily twice as hard as one from 5 feet, nor do two different 100-yard shots share the same difficulty if one is from deep rough and the other is off the fairway. Unlike rigid parametric models (such as linear regression) which might force specific functional forms, GPR is non-parametric. It lets the data speak for itself, "learning" arbitrary functional relationships and adapting to the true complexity of the three dimensional course surface.

4.2.2 What are Gaussian Processes and what is Gaussian Process Regression?

Gaussian Processes are distributions of functions where any finite subset of points follows a multivariate Gaussian distribution. As expected, they are fully specified by their mean function $m(\mathbf{x}) = \mathbb{E}(f(\mathbf{x}))$ and covariance kernel $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

In our context, $\mathbf{x}$ would represent the spatial coordinates on the hole ($\mathbf{x} = (x_1, x_2)$), and $f(\mathbf{x})$ would represent the latent scoring potential of any given position (i.e., f encodes the number of strokes to hole out from a given position).

4.2.3 Covariance Kernels

The heart of Gaussian Processes are the kernel functions, $k(\mathbf{x}, \mathbf{x}')$ which encode our prior assumptions about the structure of the data. To capture the aforementioned spatial coherence of a golf course, we use the Radial Basis Function (RBF) kernel:

$$k(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(-\sigma^2 \frac{||\mathbf{x} - \mathbf{x}'||^2}{2l^2}\right)$$

Where the hyperparameters dictate the surface's behaviour and flexibility. The length-scale, l, controls the flexibility of our functions controlling smoothness by dictating how quickly scoring difficulty changes over distance. Smaller values of l drive correlations to 0, allowing functions to exhibit "wilder" or more jagged behaviour, while also relaxing the width of confidence bounds. The signal variance, σ^2, on the other hand, controls the vertical amplitude of the function we are trying to estimate.

4.2.4 The Posterior: From Prior to Prediction

We model our observed data $\mathbf{y}$ (i.e., actual strokes taken from a spot) as the latent function with Gaussian noise: $y = f(\mathbf{X}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$.

Given a set of n observed training points $\mathbf{X}$ and outcomes $\mathbf{y}$, and a new test location $\mathbf{x}_*$ (a new spot we want to evaluate), the joint distribution of the observed data and the unobserved latent function value f_* is given by:

$$\begin{bmatrix} \mathbf{y} \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} k(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I & k(\mathbf{X}, \mathbf{x}_*) \\ k(\mathbf{x}_*, \mathbf{X}) & k(\mathbf{x}_*, \mathbf{x}_*) \end{bmatrix}\right)$$

By conditioning on the observed $\mathbf{y}$, we derive the posterior predictive distribution for the risk at our new location $\mathbf{x}_*$:

$$f_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_* \sim \mathcal{N}\left(\bar{f}_*, \mathrm{Var}(f_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)\right)$$

Where the posterior mean $\bar{f}_*$, at a specific location, provides our point estimate for optimal strategy described in Section 4.3 (predicted ESHO or Birdie Probability) and the posterior variance quantifies our uncertainty.

4.3 Modelling an Optimal Shot. Par 3s & Minimisation of Strokes.

With our modelling mechanism in place, we can now simulate the decision-making process for a specific shot. We begin with par 3 holes where players usually try to land on the green in a singular shot expecting to hole out in an additional two putts. Since we are modelling decisions with different shots and aimpoints available, we only need to consider the first shot to the green. Once on the green, the player rarely uses clubs other than their putter; the slope informs their aimpoint decision, not our model.

We model the problem as a decision taken from a fixed starting coordinate (x_1^0, x_2^0) which we treat as the origin $(0, 0)$ in the shot's local coordinate system. Our optimisation follows a "green-to-tee" philosophy: by understanding the value of every landing spot on the green, we can inform the optimal decision from the tee (the choice that precedes it). Such an implementation is viable because the problem exhibits optimal substructure, where the ultimate solution to a problem depends on the optimal solutions to its smaller subproblems. Said characteristic is best known as Bellman's Principle of Optimality (Rosu 2002).

4.3.1 Step 1: Fitting GPR to the Green

Our strategy builds on the GPR model fitted to the green data leveraging the `Python` library, `GPyTorch` (Gardner et al. 2021), evaluated at the posterior mean. The fitted Gaussian defines a continuous Expected Strokes to Hole Out (ESHO) surface $f_G : \Omega_{green} \subset \mathbb{R}^2 \to \mathbb{R}$, which outputs the expected number of strokes to hole out (ESHO) from any coordinate on the putting surface. In other words, f_G directly implements our loss function $\mathcal{L}_\mu$ for the putting region.

Because f_G is fit to the unique sample of data specific to hole 9's green and particular pin position, f_G intrinsically accounts for the non-linear difficulties (slope, tiering) that a simple distance-based function would miss.

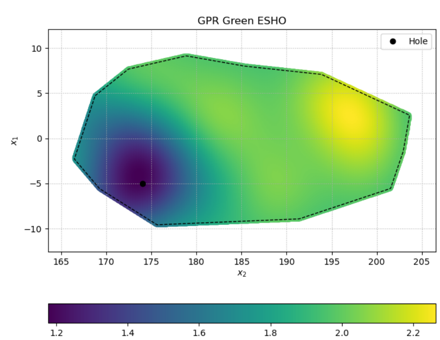


Figure 4. The Continuous Risk Surface (f_G). Posterior mean predictions of the Gaussian Process Regression model fitted to the green's synthetic putting data. The colour gradient represents the Expected Strokes to Hole Out (ESHO), effectively interpolating the difficulty of unobserved locations based on spatial proximity and topographical slope.

When a simulated shot from the fairway or tee lands on the green, we compute its "cost" by evaluating its scoring potential from its resting position.

If a shot lands on the green, the cost is retrieved directly by evaluating $f_G(\mathbf{x})$.

Alternatively, if the shot misses the green, we employ an interpolation function $f_A : \mathbb{R}^+ \times C \to \mathbb{R}^+$, where $C = \{F, R, S\}$ represents the set of lie conditions (fairway, rough or sand trap). f_A is a function of lie and distance from the hole that outputs the expected number of strokes to hole out. The function is derived from Broadie's empirical estimates for each lie (See Appendix A.3).

Finally, if a shot lands in a water hazard, we simulate a "drop" at the nearest point of relief (drawing a direct line from the ball's starting point, and simulating that the ball landed at the first intersection with land), incurring a penalty stroke and recursively evaluating the cost from the new position.

Formally, we define a global "Around the Green" cost function $AG(\mathbf{x})$, representing the expected strokes to hole out from any location $\mathbf{x} = (x_1, x_2)$ near or on the green:

$$AG(\mathbf{x}) = \begin{cases} f_G(\mathbf{x}^l) & \text{if } \mathbf{x}^l \in \Omega_{green} \\ f_A\left(\mathbf{x}^l, \text{Lie}(\mathbf{x}^l)\right) & \text{if } \mathbf{x}^l \in \Omega_{fairway} \cup \Omega_{rough} \cup \Omega_{sand} \\ 1 + AG(\mathbf{x}^{drop}) & \text{if } \mathbf{x}^l \in \Omega_{water} \end{cases}$$

4.3.2 Step 2: Simulating Player Outcomes

In order to evaluate the success of a strategy in approaching the green, we must be able to model and simulate the stochastic nature of a player's ball hitting ability. We achieve this by sampling from a normal distribution fitted to the player shot distribution data described in 3.1, and then transforming the outcomes based on the chosen aimpoint.

1. *Fitting Bivariate Gaussians.* First, we analyse the player's historical shot data to create statistical profiles for each of the available shots s. This is done by empirically estimating the two-dimensional centre of data $\hat{\mu}_\mathbf{s}$ and the covariance matrix $\hat{\mathbf{\Sigma}}_\mathbf{s}$ for the landing coordinates of shots aimed directly at the pin (zero lateral offset) hit with shot s (assuming the player is hitting off the fairway). This step "learns" a player's unique dispersion shape, capturing how their distance control and directional accuracy may vary with different shots, allowing us to generate more data as needed in the following step.

2. *Monte Carlo Simulation.* We then employ a Monte Carlo simulation to generate a synthetic dataset of potential outcomes. For a selected shot s, we draw N random samples from the fitted multivariate normal distribution.
$$\mathbf{Z}^s = \{\mathbf{z}^{s1}, \mathbf{z}^{s2}, \ldots, \mathbf{z}^{sN}\} \text{ where } \mathbf{z}^{si} \sim \mathcal{N}(\hat{\mu}_s, \hat{\Sigma}_s)$$
This creates a sample of shot outcomes statistically mirroring the player's tendencies, ensuring we have sufficient statistical volume to calculate reliable risk estimates even with sparse historical data.

3. *Changing Aimpoints.* Finally, in order to evaluate different strategic aimpoints, we apply a linear transformation to our simulated shots. If a player aims at an angle θ relative to the pin, we rotate the entire set of simulated outcomes $\mathbf{Z}^s$ using a standard rotation matrix $\mathbf{M}(\theta)$:
$$\mathbf{M}(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$
$$\mathbf{z}^{si\theta} = \mathbf{M}(\theta)\mathbf{z}^{si}$$
Giving us $\mathbf{Z}^{s\theta} = \{\mathbf{z}^{s1\theta}, \mathbf{z}^{s2\theta}, \ldots, \mathbf{z}^{sN\theta}\}$ and allowing us to evaluate the player's dispersion patterns at different aimpoints relative to the pin, ultimately evaluating the expected cost of rotated samples across the course's risk surface.

4.3.3 Step 3: Minimising Expected Strokes

We proceed to the optimisation step, which returns the optimal strategy. For a given strategy, defined by a labelled shot type s and aimpoint θ, we have a set of N simulated landing positions $\mathbf{Z}^{s\theta} = \{\mathbf{z}^{s1\theta}, \mathbf{z}^{s2\theta}, \ldots, \mathbf{z}^{sN\theta}\}$.

For each simulated shot i (which has a defined s and θ), we evaluate the outcome by querying our $AG(\mathbf{x})$ at the landing coordinate $\mathbf{z}^{si\theta} = \mathbf{M}(\theta)\mathbf{z}^{si}$. The "cost" of the shot is the sum of the stroke taken to reach the first point plus the expected strokes required to hole out.

$$\text{Cost}^{si\theta} = 1 + AG(\mathbf{z}^{si\theta})$$

We then aggregate these individual outcomes to compute the empirical risk (average expected strokes) for the specific strategic configuration:

$$R_\mu(\delta) = R_\mu(s, \theta) = \frac{1}{N}\sum_{i=1}^{N}(1 + AG(\mathbf{z}^{si\theta}))$$

This computation is repeated iteratively across all feasible shots in the player's bag and a discretised range of aimpoints (e.g., $\theta \in [-20°, +20°]$ in $1°$ increments). The optimal strategy for shot on a specific hole is the pair (s^*, θ^*) which minimises the risk surface:

$$\delta_{opt} = \arg\min_{s,\theta} R_\mu(s, \theta) = (s^*, \theta^*)$$

We convert the optimal angle θ^* to yards (left or right of the pin) by using basic trigonometric conversions which make the output actionable for the player. Our three-step process transforms abstract data into concrete recommendations on which the player can then operate.

4.4 Modelling Chained Decisions. Par 4s & Recursive Optimisation.

Extending on the innovative use of GPR modelling on par 3s, we apply GPR to longer holes requiring chained decision-making. Our optimisation strategy for par-4 holes continues the recursive "green-to-tee" philosophy.

The key insight is that we believe par 4s are par 3s with a variable teeboxes when approaching the green (that is, we need to consider different positions we may be presented with, when approaching the green). To optimise the actual tee shot, we first understand the strategic value of every possible intermediate landing zone in the fairway, rough or sand. We achieve this through a two-stage process: pre-computing the optimal approach from all possible positions and then fitting a GPR surface to those optimal values.

4.4.1 Step 1: The "Strategic Blanket" (Pre-computing Approaches)

We begin by evaluating a dense grid of potential second-shot locations on and near the fairway, in our case ranging from 50 to 350 yards from the green. We define a discrete grid of points as $\mathcal{G} = \{\mathbf{g}^1, \mathbf{g}^2, \ldots, \mathbf{g}^M\}$, where each point is $\mathbf{g}^j = (x_1^j, x_2^j)$; we run the complete simulation algorithm described in Section 4.3.3.

1. Identify the lie at $\mathbf{g}^j$ - fairway, rough, or bunker – to select the appropriate shot distribution to sample from the player shot distribution dataset. If $\mathbf{g}^j$ is on water, simulate dropping with the same logic described in the previous section.

2. Simulate shots for all s and θ as described in Section 4.3.2 from the temporary "origin", $\mathbf{g}^j$.

3. Identify the strategy (s^*, θ^*) that minimises expected strokes said location, $\mathbf{g}^j$.

4. Report the minimum risk value (ESHO) $R_\mu(s^*, \theta^*)^j$, to use the values in the next phase.

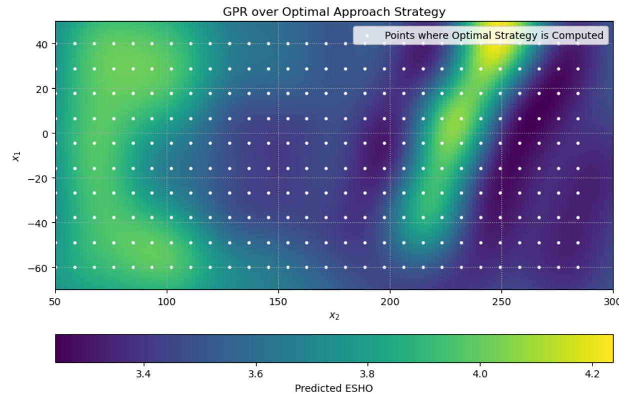


Figure 5. The "Strategic Blanket" ($f_{approach}$). A visualisation of the recursive optimisation step. The heatmap does not represent the difficulty of the hole, but rather the optimal expected score achievable from any given grid point $\mathbf{g}^j$ (50–350 yards), assuming optimal play from that point forward. The "valleys" (dark blue) represent ideal landing zones for a tee shot.

This process yields a minimal Expected Strokes to Hole Out (ESHO) value for every point on the grid (this can be seen in Figure 6 (b)). We then fit a *second* Gaussian Process model to these ESHO values associated to each $\mathbf{g}^j$, creating a new continuous surface $f_{approach} : \mathbb{R}^2 \to \mathbb{R}^+$ which we name a "strategic blanket".

Evaluated at a coordinate $\mathbf{x}^l$ on the golf course, $f_{approach}$ function does not return a value representing the raw difficulty of the hole, but rather represents the best possible expected score a player can achieve from 50 - 350 yards from the green, assuming they play optimally from that point forward.

4.4.2 Step 2: Optimising the Tee Shot

With $f_{approach}(\mathbf{x})$ acting as our main evaluation function (replacing $AG(\mathbf{x})$), the par 4 tee shot optimisation becomes mathematically equivalent to the par 3 problem.

We simulate tee shots for long-range clubs (Driver, Woods, and Long Irons) using the player's dispersion parameters $\hat{\mu}_s, \hat{\Sigma}_s$, now assuming to be hitting off the tee. For each simulated landing point $\mathbf{z}^{si\theta}$ (derived from the same Monte Carlo process described in Section 4.3.2), the cost is no longer derived from the green model, but from our strategic blanket:

$$\text{Cost}^{si\theta} = 1 + f_{approach}(\mathbf{z}^{si\theta})$$

Our optimisation seeks the shot type and aimpoint that lands the ball in the "valleys" of $f_{approach}$ – areas where the subsequent shot is easiest and therefore the expectation of the optimal choice is lowest - while avoiding the "peaks" (hazards or difficult angles).

We aggregate these costs once again to find the risk of the tee shot:

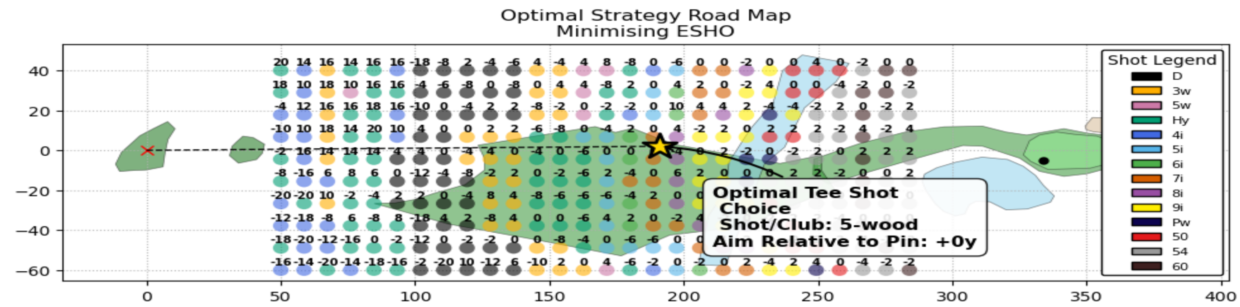
$$R_{\mu_{tee}}(\delta) = R_{\mu_{tee}}(s, \theta) = \frac{1}{N} \sum_{i=1}^{N} (1 + f_{approach}(\mathbf{z}^{si\theta}))$$

As such, the optimal tee strategy is defined as:

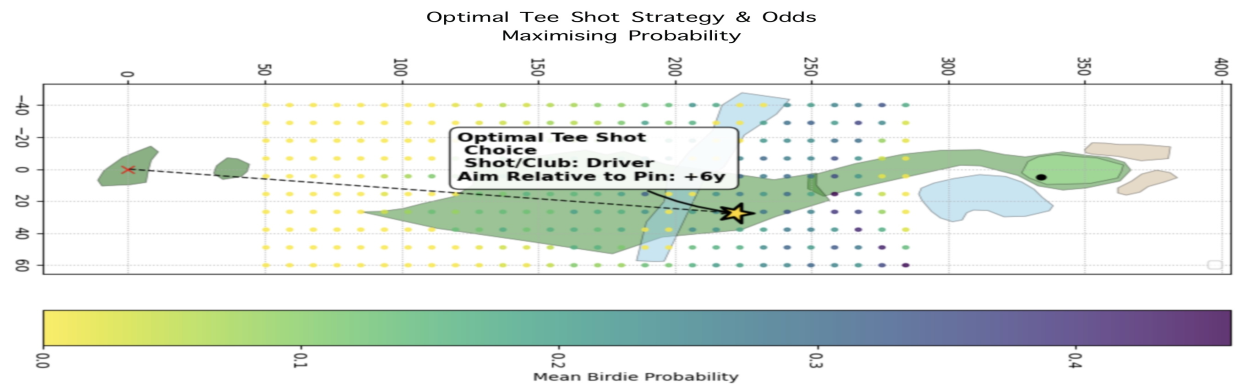
$$\delta_{opt} = \arg\min_{s,\theta} R_{\mu_{tee}}(s, \theta) = (s^*_{tee}, \theta^*_{tee})$$

Our recursive framework allows the model to make nuanced trade-offs, potentially suggesting more conservative options such as hitting the shorter distance club not just to ensure landing on the fairway, but to avoid bringing other trouble into play (water or bunkers).

After assessing these values at the corresponding coordinate, the golfer would do best to choose the provided s^* and θ^* for their shot. In our example (Figure 6 (a)) the decision is 5-wood aimed at $+0y$ (with no offset to the pin).



(a) Conservative Strategy (R_μ)



(b) Aggressive Strategy (R_B)

Figure 6. Strategic Divergence based on Loss Function. (a) When minimising Expected Strokes to Hole Out (R_μ), the model identifies a risk-averse strategy: a 5-wood lay-up short of the water hazard, prioritising variance reduction. The dotted grid acts as a recursive roadmap for the subsequent approach shot: the fill colour of each dot indicates the optimal club selection (as indicated by the legend), while the text label specifies the optimal aimpoint offset in yards (positive values aim right of the pin, negative aim left). (b) When maximising Birdie Probability (R_B), the model shifts to a high-variance strategy: a Driver aimed over the hazard, accepting the penalty risk in exchange for a higher probability of a short approach shot. While a similar club roadmap as (a) exists for this objective, the heatmap here visualises the computed *likelihood* of securing a birdie from each coordinate, illustrating the high-reward zones that drive this aggressive strategy.

4.5 Modelling Alternative Loss Functions. Maximisation of Birdie Probabilities.

Up until now, our optimisation strategy has focused on minimising the expected number of strokes to hole out. This approach naturally produces conservative strategies – decisions that prioritise avoiding high-penalty disasters (such as hitting the ball into water hazards or Out of Bounds) over chasing rare, high-reward outcomes. In statistical terms, this conservative play minimises the variance of the score distribution to protect the mean.

However, golfers are often faced with high-tension scenarios where consistency becomes inconsequential. If a player needs a birdie to make the cut, force a playoff, or win a tournament, a par may be functionally equivalent to a double bogey. In these moments, aggressive strategies are required. Here, the player willingly accepts a higher expected score (and higher variance) in exchange for increasing the probability mass in the lower tail of the distribution (i.e., in the case of a par 4, the chance of holing it in 3 or fewer strokes).

To model the change in objective, we alter our loss function by converting the data used to train the GPR.

4.5.1 Step 1: The Probability Surface

We begin by returning to the raw data generated on the green (as described in Section 3.2.2). In the par 3 and par 4 expectation minimisation models described earlier, each data point collected on a green was a count of strokes (e.g., 1, 2, or 3) from location (x_1, x_2). For our new objective, we perform a binary conversion on the training data outcomes $\mathbf{y}$.

$$y_{binary} = \begin{cases} 1, & \text{if strokes} \leq 1 \text{ (holed)} \\ 0, & \text{if strokes} > 1 \text{ (missed)} \end{cases}$$

Why do we care about "one-putts"? A typical hole's par is set such that a typical golfer will achieve a par by taking at most two putts to hole out once on the green. Therefore, the most secure way to score a birdie is to take at most one putt to hole the ball. The likelihood of one-putting is what we are modelling here.

We then fit a new GPR model to the binary outcomes. Although the inputs are discrete, the GPR learns a continuous latent function $f_{birdie} : \mathbb{R}^2 \to [0, 1]$ representing the conditional probability of one-putting from any location on the green. This creates a "Birdie Probability Surface", which decays exponentially: being 3 feet away from the cup offers massive reward, whereas being 15 feet away might offer near-zero value in "birdie necessary" circumstances.

4.5.2 Step 2: Recursive Maximisation for Birdie Likelihood Maximisation

With the birdie probability surface defined, we re-run the original simulation mechanism described in Section 4.3.3 with a modified evaluation logic:

1. **Approach Simulation:** For all simulated approach shots $\mathbf{z}^{si\theta}$ (the rotated samples from the Monte Carlo step), we query the new surface:

$$\mathbb{P}(\text{Birdie} \mid \mathbf{z}^{si\theta}) = \begin{cases} f_{birdie}(\mathbf{z}^{si\theta}) & \text{if } \mathbf{z}^{si\theta} \in \Omega_{green} \\ 0 & \text{else} \end{cases}$$

   Note that we make the simplifying assumption that shots missing the green ($\mathbf{z}^{si\theta} \notin \Omega_{green}$) have negligible probability of being holed for birdie.

2. **Risk Calculation:** The risk we seek to minimise is the probability of failure (not achieving a birdie or better). We therefore aggregate the probabilities across all N samples:

$$R_B(\delta) = R_B(s, \theta) = 1 - \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}(\text{Birdie} \mid \mathbf{z}^{si\theta})$$

3. **Optimal Aggressive Strategy on Par 3s:** Finally, we identify the strategy that minimises the probability of failure (thereby maximising birdie probability) by iterating over all feasible shots and aim angles:

$$\delta_{opt} = \arg\min_{s,\theta} R_B(s, \theta) = (s^*, \theta^*)$$

   This yields the specific club s and aimpoint θ^* required maximise the odds of acquiring a birdie, ignoring the potential higher variance of the shot.

4.5.3 Step 3: Aggressive Tee Shots

For par 4s, we repeat the "strategic blanket" process described in Section 4.4, but with a fundamental shift in the evaluation metric.

Instead of fitting a GPR to the ESHO from every fairway location, we now map the highest birdie probability achievable from every grid point $\mathbf{g}^j$ on or around the fairway.

Like the par 3 birdie optimisation mechanism, this creates a new surface, $f_{approach_b} : \mathbb{R}^2 \to [0, 1]$, where the "peaks" represent fairway locations that offer the best statistical chance of converting a birdie on the subsequent shot.

When optimising the tee shot, the model aggregates these probabilities to find the strategy that maximises the potential for a birdie. We simulate the tee shots as before, but the cost function is inverted to represent the probability of success:

$$R_{B_{tee}}(s, \theta) = 1 - \frac{1}{N} \sum_{i=1}^{N} f_{approach_b}(\mathbf{z}^{si\theta})$$

Consequently, the optimal aggressive tee strategy is identified by minimising this risk (failure) function:

$$\delta_{opt} = \arg\min_{s,\theta} R_{B_{tee}}(s, \theta) = (s^*_{tee}, \theta^*_{tee})$$

After assessing these values at the tee, the golfer would do best to choose the provided s^* and θ^* for their first shot. In our example (Figure 6 (b)) the decision is Driver aimed $+6y$ (aimed 6 yards right). The output of the optimisation for birdie can often be drastically different from optimisation for ESHO. Which in some ways, is exactly what we want to see – different optimal decisions for different objectives.

5. Results

The primary objective of this study was to establish a computationally grounded methodology for converting individual player performance data into actionable strategic recommendations by employing Gaussian Process Regression. Ultimately, we wanted to demonstrate that we could generate actionable strategic recommendations based on varying player objectives. Our simulation results confirm that our model effectively learns course-specific risk surfaces and adapts its recommendations accordingly.

5.1 Risk Surface Generation

The GPR model successfully learns continuous risk surfaces from discrete, noisy data points. For the par 3 case study (hole 9, Mountain Meadows), the model generated a smooth Expected Strokes to Hole Out (ESHO) surface that accurately reflected the penalty structure of the hole. The surface naturally exhibited higher "risk" values near the water hazard and bunkers, and lower values in the wider part of the green, effectively interpolating difficulty in unobserved regions. Our model subsequently gave us a very reasonable recommendation.

5.2 Strategic Divergence: Conservative vs. Aggressive

The most valuable finding is the clear divergence in optimal strategy when switching between loss functions (minimising expected strokes versus maximising birdie probability), visualised in Figure 6.

- **Minimising Expected Strokes (Conservative):** When minimising $R_\mu(\delta)$, the model recommends strategies that prioritise variance reduction. On the par 4 simulation, the "Strategic Blanket" identified that the optimal tee shot was not necessarily the driver. Instead, our model suggested hitting a 5-wood (a safe choice off the tee) aimed at the widest part of the fairway, sacrificing distance to avoid the high-penalty zone (the water penalty crossing the hole) that the driver's wider dispersion would bring into play.

- **Maximising Birdie Probability (Aggressive):** Conversely, when minimising $R_B(\delta)$, the model shifted to a high-variance, high-reward approach. The model recommends a more aggressive approach by hitting the driver over the water and accepting a non-zero probability of a penalty in exchange for substantially shorter approach shot, which would intuitively have a smaller dispersion.

In summary, the framework works as intended: it does not simply output a single "best" shot, but effectively quantifies the trade-offs between consistency and scoring potential, providing players with a roadmap that they can implement regardless of the outcome of their first shot.

6. Discussion

Our project serves as a proof-of-concept for the derivation of mathematically reasonable golf strategies under a Bayesian paradigm. By treating the course as a continuous probabilistic surface, we accept the variability and uncertainty behind a shot, directly modelling it. As outlined earlier, our approach works; however, several limitations must be addressed before the framework can be deployed widely.

Perhaps the most important of our limitations is the simplification of our ball flight physics. Our model currently assumes that the ball stops at its carry distance, when in reality shots usually have roll-out. How much a ball rolls out is complicated to compute as it depends on landing angle, spin rate, and ground firmness. On hard fairways or greens a ball may roll-out 20-40 yards, potentially bringing hazards into play that we deemed unreachable in our statistical analysis based on carry distance alone. Furthermore, we do not model the vertical trajectory (apex height) or shot shape (draw/fade), meaning the model cannot currently account for obstacles that require clearance, such as tall trees.

The current iteration of our project also assumes a static environment. We ignore meteorological conditions such as wind speed, direction, temperature, and air density, all of which drastically affect carry distances. While these variables can be parametrised, implementing variables that can change quickly during a tournament presents a formidable challenge. Additionally, our model assumes a constant green speed and ignores elevation changes from tee to green, which are critical for accurate club selection.

While our green putting model is robust, the "around-the-green" data used in our analysis relies on sparse interpolated averages from historical PGA Tour baselines rather than player-specific data. This assumes the player possesses "average" professional scrambling ability, which may mask individual weaknesses (e.g., a player who is statistically poor from sand bunkers should be advised to avoid them more aggressively than our model currently suggests).

Our project also presents unique opportunities. A distinct advantage of GPR is the generation of posterior variance estimates, which quantify the model's uncertainty. Future iterations should explicitly incorporate these interval estimates into the strategy recommendation engine. By distinguishing between high-confidence predictions (where data is dense) and regions of high uncertainty (where data is sparse), future models could differentiate between strategies that are genuinely "safe" and those that merely appear optimal because the model lacks sufficient data on local hazards.

Similarly, we would like to extend our framework to model par 5 holes which introduce the consideration of three chained shots in a row. Additionally, testing the model on diverse course styles ranging from links-courses (firm course with high roll and few hazards) to target golf (courses with many obstacles that force the player to fly the ball past them) would validate the robustness of our proposed framework.

Currently, the model utilizes a proxy "average LPGA" profile. Future work involves ingesting real shot data from distinct populations (e.g., elite male amateurs vs. seniors vs. collegiate athletes) to observe how optimal risk surfaces change based on power and consistency disparities.

Ultimately, this project lays the mathematical groundwork for a "digital caddie" evolving with the player by turning every shot recorded into a smarter decision for the future.

A. Data Generation Methodology

In the absence of granular, shot-level tournament data (such as PGA Tour ShotLink data) and player-specific data, we developed a robust synthetic data generation pipeline to construct realistic two-dimensional scoring distributions by combining course digitisation, empirical statistical interpolation, and probabilistic simulation.

We constructed a proxy player profile representing an "average" LPGA Tour player using publicly available launch monitor data (Trackman 2024).

A.1 Course Geometry and Lie Classification

To create a spatially accurate environment for our model, we digitised the layout of hole 9 at Mountain Meadows Golf Course in Pomona, California.

1. **Vector Mapping:** The course layout was initially scraped using OpenStreetMap (Bennett 2010) and imported into QGIS (QGIS Development Team 2024) for refinement.

2. **Polygon Classification:** Features were exported as WKT (Well-Known Text) polygons, allowing for the discrete classification of any coordinate (x_1, x_2) into specific lie types (Fairway, Rough, Bunker, Green, or Water) using the `Shapely` library in Python (Gillies et al. 2024). This geometric classification is the foundational layer for all subsequent scoring evaluations.

A.2 Green Surface Data Generation

To model putting difficulty beyond simple distance estimates, we constructed a multivariate dataset that accounted stroke differences across the complex topography of a green. This process involved a multi-stage pipeline designed to transform univariate averages into spatially aware discrete outcomes, creating data that would be more representative of that collected at an actual tournament.

1. **Surface Topographical Modelling.** We defined a continuous height function $z(x_1, x_2)$ to simulate a realistic putting surface featuring a general tilt, undulating bumps, and a mid-green tier. By defining a height function, we were then able to calculate slope vectors and break severity at any point on the green.

2. **Baseline Interpolation with Noise.** We first interpolated baseline average putts values from Mark Broadie's empirical estimates in Table 1. To mimic real-world heteroscedasticity (where variance increases with distance) and data sparsity at long range, we generated noisy synthetic samples centred at Broadie's empirical means.
$$y_{sample} \sim \mathcal{N}(f_{Broadie}(d), \sigma_d^2)$$
We then fit an intermediate GPR model to the synthetic samples, creating a smooth, continuous interpolator capable of generating posterior predictive distributions of ESHO for any distance d. The Gaussian Process also captured both the mean trend and the increasing uncertainty at longer ranges.

3. **Difficulty Multipliers** To account for the non-linear difficulty introduced by the green's shape, we computed difficulty multipliers based on the local topography. We evaluated the gradient of the surface at four critical points along the putt's trajectory (50%, 70%, 85%, and 95% of the distance to the hole) to capture the relevant slope and break.

   - **Fall Line Slope:** Penalising downhill putts (where the ball picks up speed) more heavily than uphill ones.

Table 1. PGA Tour Putting Statistics by Distance. The "Average putts" column represents the PGA Tour putting benchmark used in strokes gained computations. The values are based on an analysis of nearly four million putts on the PGA Tour from 2003 to 2012, as detailed in (Broadie 2014).

Distance (feet)	2	3	4	5	6	7	8	9	10
One-putt probability	99%	96%	88%	77%	66%	58%	50%	45%	40%
Three-putt probability	0.0%	0.1%	0.3%	0.4%	0.4%	0.5%	0.6%	0.7%	0.7%
Average putts	1.01	1.04	1.13	1.23	1.34	1.42	1.50	1.56	1.61

Distance (feet)	15	20	30	40	50	60	90
One-putt probability	23%	15%	7%	4%	3%	2%	1%
Three-putt probability	1.3%	2.2%	5.0%	10.0%	17.0%	23.0%	41.0%
Average putts	1.78	1.87	1.98	2.06	2.14	2.21	2.40

- **Tier Difference:** Adding a strong penalty if the ball and hole are located on different tiers.
- **Side Slope:** Calculating the perpendicular slope at different points of the putt's path to account for large amounts of break.

The specific multipliers (provided in Table 2) were estimated heuristically based on domain expertise, as empirical data for slope-specific putting difficulty is not publicly available. The multipliers adjust the baseline ESHO (μ_{base}) to form an adjusted mean μ_{adj}:

$$\mu_{adj} = \mu_{base} \times M_{tier} \times M_{side} \times M_{slope}$$

Table 2. Topographical Difficulty Multipliers. Multipliers are applied to the baseline expected strokes to hole out (ESHO) to account for green complexity. *Note: These values are heuristic estimates derived from domain knowledge.*

Factor	Condition	Multiplier Range
Tier Difference	No Tier (Diff < 0.15)	1.00
(Height diff. between ball & hole)	Level 1 (Diff 0.15 – 0.45)	1.05 – 1.07
	Level 2 (Diff 0.45 – 0.70)	1.08 – 1.12
	Level 3 (Diff 0.70 – 1.00)	1.12 – 1.17
	Level 4 (Diff 1.00 – 2.00)	1.18 – 1.25
	Level 5 (Diff > 2.00)	1.40
Side Slope	Negligible (< 0.5%)	1.00
(Max perpendicular slope)	Mild (0.5 – 1.0%)	1.00 – 1.03
	Moderate (1.0 – 2.0%)	1.04 – 1.09
	Severe (2.0 – 3.0%)	1.10 – 1.20
	Extreme (3.0 – 4.0%)	1.20 – 1.25
	Unplayable (> 4.0%)	1.35
Uphill / Downhill	Flat	1.00
(Max parallel slope)	Uphill (Mild to Steep)	1.03 – 1.10
	Downhill (Mild)	1.01 – 1.07
	Downhill (Steep)	1.10 – 1.45
	Downhill (Extreme > 3%)	1.45 – 1.60

4. **Generating Discrete Outcomes**

   To generate the final discrete stroke counts used for modeling we executed a probabilistic simulation process that transforms continuous difficulty estimates into discrete scoring distributions:

(a) **Sampling and Evaluation:** We selected random (x_1, x_2) coordinates on the digitized green surface. For each point, we computed the Euclidean distance to the hole and queried the intermediate GPR (from Step 2, Appendix A.2) to obtain the posterior distribution of ESHO at that distance. We then applied the topographical multipliers (from Step 3, Appendix A.2) to this distribution, shifting the mean to reflect the specific difficulty of the putt's slope and break.

(b) **Base Probability Retrieval:** Simultaneously, we retrieved the baseline probabilities p_1, p_2, p_{3+} for holing out in 1, 2, or 3+ strokes from Broadie's empirical Table 1, based solely on the Euclidean distance to the hole.

(c) **Probability Mass Redistribution:** We utilised the shifted ESHO from the GPR to adjust the baseline probabilities. Rather than simply rounding a continuous mean, we redistributed probability mass between discrete outcomes. For example, in locations with high difficulty multipliers (e.g., steep downhill slopes or cross-tier putts), probability mass was shifted from p_1 to p_2 and to p_{3+}, decreasing the likelihood of a one-putt while increasing the risk of a three-putt. This shift was calibrated to ensure the resulting expected value matched our adjusted ESHO (μ_{adj}) while preserving the internal logic of discrete golf scoring.

(d) **Sampling and Optimization:** We sampled final discrete scores using a discrete inverse transformation on the adjusted probability vectors. Crucially, our framework allowed us to directly extract the probability of a one-putt, $P(\text{Score} = 1 \mid x, y)$, from any location, and to fit the GPR in Section 4.4 to data that emulated the ideal data we would collect in a tournament described in Section 3.1.

A.3 Short Game Data Generation

For shots not on the green, we employ a direct interpolation method. Since slope is less critical for these shots than the lie condition, we estimate the expected strokes to hole out (ESHO) by interpolating directly from historical PGA Tour averages based on the distance to the hole and the specific lie type (Fairway, Rough, or Sand). Table 3 displays the baseline values used for the interpolation, once again derived from Broadie's work (Broadie 2014).

Table 3. Average Shots to Hole Out by Lie and Distance. Values represent the expected number of strokes to hole out from specific distances and lies, derived from PGA Tour ShotLink data (2003-2012) (Broadie 2014).

Distance (yards)	**Fairway (avg shots)**	**Rough (avg shots)**	**Sand (avg shots)**
20	2.40	2.59	2.53
40	2.60	2.78	2.82
60	2.70	2.91	3.15
80	2.75	2.96	3.24

A.3.1 Simulating Player-Specific Data

We constructed a proxy player profile representing an "average" right-handed LPGA Tour player using a simulation based on publicly available carry distance averages (Trackman 2024).

1. **Base Averages:** We defined average carry distances for 20 distinct shot types, ranging from a Driver (223 yards) to a 60° wedge (78 yards) based off of official data; also including partial wedge shots to bridge yardage gaps for the shorter distances (e.g., a 3/4 60-degree wedge).

2. **Dispersion Scaling:** To model the natural increase in variance with shot length, we scaled the standard deviations for lateral error (σ_{side}) and distance error (σ_{dist}) linearly with carry distance, setting $\sigma_{side} \approx 0.08 \times$ Avg. Carry and $\sigma_{dist} \approx 0.04 \times$ Avg. Carry.

3. **Directional Bias:** We introduced a correlation between lateral miss and carry distance to mimic real ball flight physics. Shots missing left (simulating pulls) were modelled with a positive distance bias (+5 yards on average), while shots missing right (simulating fades or pushes) were modelled with a negative distance bias (-2 yards on average), this adjustment was heuristically based on domain expertise, as empirical data for these distributions is not publicly available.

4. **Sampling:** For each club, we generated $N = 30$ synthetic shot outcomes by sampling from the resulting bivariate Gaussian distributions, creating the dataset used to fit the player's statistical profiles.

B. Glossary of Terms

- **Lie:** The ground condition or surface where a ball rests. Common lies include:
  - *Fairway*: Closely mown grass offering the standard hitting surface.
  - *Rough*: Longer grass bordering fairway lies, thicker and more unpredictable.
  - *Sand/bunker*: Obstacle or hazard filled with sand.
  - *Green*: Short grass area around the hole used for putting.
- **Club:** Instrument used to hit golf balls. Clubs vary in loft and length, affecting distance and shot shape. Examples used in this paper include: 5-iron, Wood or Driver.
- **Tee/Tee box:** Designated starting area for each hole.
- **Pin/Flag:** The flagstick marking the hole's location on the green.
- **Hazard:** Course obstacles like water or sand bunkers that penalise errant shots.
- **Drop:** Placing the ball back in play after it lands in a water hazard, typically incurring a penalty stroke.
- **Par:** The expected number of strokes a skilled golfer should take to complete a hole. A par-3 hole expects a golfer to finish in 3 shots.
- **Birdie:** The score that is one stroke under par on a hole.
- **Bogey:** The score that is one stroke over par on a hole.
- **Stroke:** A single swing attempt at the ball.
- **Stock shot:** A player's standard, repeatable "full swing" given a club.
- **Approach shot:** A shot intended to reach the green, or "approach" the pin.
- **Carry distance:** The distance a ball travels through the air before it first touches the ground.
- **Dispersion:** The spread of a player's shots around their intended target, variable depending on golf club difficulty and individual golfer's tendencies.
- **Aimpoint:** Directional target chosen by the player, defining the centre of their shot dispersion.
- **Short game:** Shots played close to the green (chipping, pitching, putting).
- **Break:** The curve a putt takes due to green slope.
- **Trackman:** A radar launch monitor system that captures detailed shot data, including launch angle, ball speed, spin, and dispersion.
- **ShotLink:** The PGA Tour's proprietary real-time scoring system that captures 3-dimensional coordinates for every shot taken during a tournament.
- **Strokes Gained (SG):** A statistical metric that compares a player's performance against a baseline (typically the PGA Tour average) to quantify the quality of a shot.
- **Smash Factor:** The ratio of ball speed to clubhead speed, used as a measure of energy transfer efficiency.
- **WAGR:** World Amateur Golf Ranking.

References

Albert, Jim (2010). "Sabermetrics: The past, the present, and the future". In: *Mathematics and sports* 43, p. 15.

Bennett, Jonathan (2010). *OpenStreetMap*. Packt Publishing Ltd.

Broadie, Mark (Mar. 2014). URL: https://business.columbia.edu/faculty/research/every-shot-counts-using-revolutionary-strokes-gained-approach-improve-your-golf.

Fawcett, Scott (2024). *DECADE Golf Course Management System*. URL: https://decade.golf/.

Gardner, Jacob R. et al. (2021). *GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration*. arXiv: 1809.11165 [cs.LG]. URL: https://arxiv.org/abs/1809.11165.

Gillies, Sean et al. (2024). *Shapely: manipulation and analysis of geometric objects*. https://shapely.readthedocs.io/. Version 2.x.

Jenkins, Dan (Aug. 1998). *Three perfect shots makes Hogan's day*. URL: https://www.irishtimes.com/sport/three-perfect-shots-makes-hogan-s-day-1.179937.

Khazaeli, Mehdi and Leili Javadpour (Sept. 2024). "Golf Club Selection with AI-Based Game Planning". en. In: *Entropy* 26.9, p. 800. ISSN: 1099-4300. DOI: 10.3390/e26090800. URL: https://www.mdpi.com/1099-4300/26/9/800 (visited on 12/08/2025).

Lewis, Michael (2003). *Moneyball: The Art of Winning an Unfair Game*. W. W. Norton Company.

QGIS Development Team (2024). *QGIS Geographic Information System*. https://qgis.org/. Open Source Geospatial Foundation Project.

Rosu, Ioanid (2002). *The bellman principle of optimality*.

Schwartz, Barry (2015). "The paradox of choice". In: *Positive psychology in practice: Promoting human flourishing in work, health, education, and everyday life*, pp. 121–138.

Stauffer, Gautier and Matthieu Guillot (Aug. 2025). *Golf Strategy Optimization and the "Drive for show, putt for dough" adage*. en. arXiv:2309.00485 [math]. DOI: 10.48550/arXiv.2309.00485. URL: http://arxiv.org/abs/2309.00485 (visited on 12/08/2025).

Sugawara, Shogo, Hidenori Kawamura, and Keiji Suzuki (July 2013). "Skill-based simulation model for optimizing strategy in golf". en. In: *2013 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*. Wollongong, NSW: IEEE, pp. 1591–1596. ISBN: 978-1-4673-5320-5 978-1-4673-5319-9. DOI: 10.1109/AIM.2013.6584323. URL: http://ieeexplore.ieee.org/document/6584323/ (visited on 12/08/2025).

Trackman (May 2024). *Introducing Updated Tour Averages*. https://www.trackman.com/blog/golf/introducing-updated-tour-averages. Accessed: 2025-07-10.

Generative AI Statement

During the preparation of this work, we employed generative AI tools to expedite the software development process. ChatGPT (models GPT-4o and GPT-4.1) was used to identify necessary software libraries, outline implementation steps for the coding framework, and assist with debugging. Google Gemini was used to assist in refining some sentence structure and improving the conciseness of some prose. Grammarly was employed for standard grammatical review. These tools served strictly as technical aids for code implementation and language refinement; the conceptual modelling, mathematical logic, and strategic analysis remain as original intellectual work.