Optimizing Goalie Pulls in Hockey

Laura Roettges (roettges@wisc.edu) — Ryan Moreno (rrmoreno@wisc.edu)

July 2024

1 Problem Overview (Introduction)

In hockey, it is common practice to "pull the goalie" in the last few minutes of a game if a team is losing. Pulling the goalie entails subbing out the goalie for an additional offensive player. This way, the losing team has 6 skaters on the ice (and no goalie) instead of the usual 5 skaters and 1 goalie. The key idea behind this is that winning/losing the game is solely decided by which team ends with more points, not *how many* points they win/lose by. So, if team A is down by 2 points with 1 minute left in the game, there is a very small chance of them catching up by the end of the game if the play remains 5v5. Although pulling the goalie makes it more likely that team A will be scored on (called an "empty net" goal), there is minimal cost associated with getting scored on at this point in the game, because team A was almost certainly going to lose anyways. Pulling the goalie also makes it more likely that team A will score a goal, since they have 6 offensive skaters. Even though there is higher risk of getting scored on, the reward associated with potentially scoring outweighs the risk since scoring could bring team A to a tie or win. In essence, after a goalie is pulled, there is higher variance in the number of goals scored (on both sides). Since team A is about to lose the game, this higher variance is to their benefit.

Our question is, given the score of a hockey game and the number of minutes left, when (if ever) is the optimal time for team A to pull their goalie?

2 Data Definitions and Assumptions

- Inputs
  - A_0: Our team's current score. We will use a stochastic approach for determining this value.
  - B_0: Opponent's current score. We will use a stochastic approach for determining this value.
  - T: number of seconds remaining in the game
- Decision
  - s: how many seconds to wait before pulling the goalie

  - $s = T$ indicates never pulling the goalie
  - we will refer to the period from time 0 (now) until s as "epoch 1" and the period from s until the end of the game as "epoch 2"

- Goal
  - We want to pull the goalie such that we maximize the probability that our team ties or wins the game.

- Simplifications/Assumptions
  - Our team is currently losing
  - It is the third period ($T \leq 1200$)
  - The other team will not pull their goalie
  - Once we pull our goalie, the goalie will not be put back in for the rest of the game.
  - Teams are equal in skill.
  - Each second, a team can score either 0 or 1 goals, according to a Bernoulli random variable.

- Data
  - $p_{b1} = p_{a1} = 5.18 * 10^{-4}$: probability that the opponent scores during a given second in 5v5 play (same as the probability that we score on the opponent during a given second in 5v5 play)
  - $p_{b2} = 8.51 * 10^{-3}$: probability that the opponent scores during a given second in 6v5 play.
  - $p_{a2} = 3.10 * 10^{-3}$: probability that we score during a given second in 6v5 play

  For details on how these probabilities were calculated see the 'Data Calculations' section. Note that p_{b2} and p_{a2} are omputed from NHL data using an exponential survival model (where the end of a period is considered "censoring").

3 Data Calculations

3.1 Pre-computing combinatorics

We precomputed combinatorics $\binom{n}{s}$ with $s \leq 1200$ in order to save some time for executing our model since the probability of success is dependent

3.2 Pre-computing p_{a1} and p_{b1}

To determine the values of p_{b1} and p_{a1} we reviewed goals scored and conceded during playoff games between 2016 and 2024 when there were 5v5 on the ice [League, 2024b] [League, 2024a]. This dataset was not accompanied by exact time ranges for an accurate number of seconds when 5v5 were on the ice, so for simplicity we assume it was 5v5 for the full 3 periods (this will inherently be an underestimate). This resulted in two probabilities:

- Probability of scoring a goal per second with 5v5: $4.98 * 10^{-4}$
- Probability of conceding a goal per second with 5v5: $5.37 * 10^{-4}$

Since we are assuming $p_{b1} = p_{a1}$ we want just a singular probability and are averaging these two values to get $5.18 * 10^{-4}$

3.3 Pre-computing p_{a2}

To determine the probability that we score during a given second in 6v5 play p_{a2} we utilized a set of pre-processed data where the author has already scraped NHL's data to determine goalie pull times and the time the first goal (for or against) the team [Agalea, 2023]. This dataset is from games between 2003-2019 and we used an exponential survival model approach to determine from the data p_{a2}. We utilized the **Kaplan-Meier estimator** to estimate the survival function where survival here is the team's success of scoring a goal after they pull the goalie. To do so we have to identify failures and successes.

- Successes are instances where the team pulls their goalie and subsequently scores, we store the time from when the goalie was pulled to when they scored.
- Failures include:
  - Instances where a team pulled their goalie and subsequently conceded a goal. Here we store the time from when the goalie was pulled to when they conceded a goal.
  - Instances where a team pulled their goalie and no goal was scored by the end of the game. This is censored data, i.e. data where some values are unknown because they are not observed, and the time we store here is from when the goalie was pulled to when the game ended.

Since there was a significant amount of censored data our estimator does not actually achieve values at or below a survival probability of .5 see figure 1 . To determine how many seconds to have a .5 probability of survival we linearly interpolated the data to approach estimate the value at .5 which resulted in an estimate of 322.5 seconds, in other words we would expect our team who pulled a goalie to score by 322.5 seconds after they pulled the goalie. This yields an approximate probability, $p_{a2} = 1/322.5$ or $p_{a2} = 3.10 * 10^{-3}$

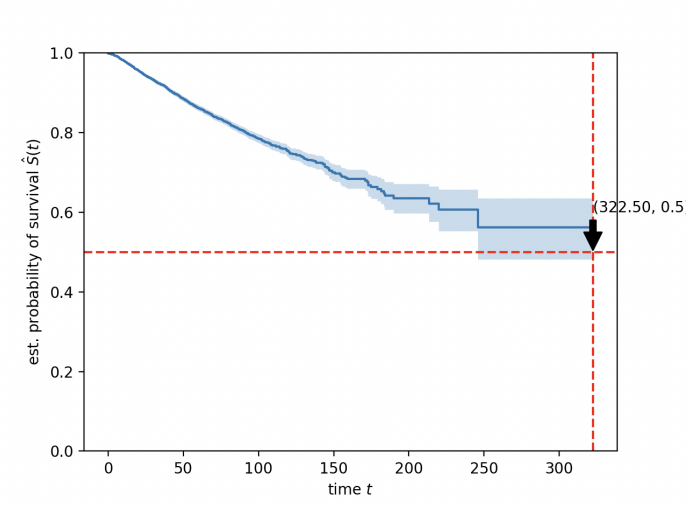

Figure 1: Kaplan Meier Survival Estimation for Team Successfully Scoring After Pulling their Goalie

3.4 Pre-computing p_{b2}

To determine the probability that the opponent scores during a given second in 6v5 play, p_{b2}, we use an equivalent approach as for p_{a2}. We use the Kaplan-Meier estimator for the survival function except now

- Successes are instances a team pulled their goalie and subsequently conceded a goal (aka the team playing down scores).
- Failures include:
  - Instances where a team pulls their goalie and subsequently scores (aka the team playing down 5v6 gets scored on), we store the time from when the goalie was pulled to when they scored.
  - Instances where a team pulled their goalie and no goal was scored by the end of the game, i.e. our censored data

This yielded $p_{b2} = 1/117.491$ or $p_{b2} = 8.51 * 10^{-3}$, see figure 2.

4 Mathematical Model

- Stochastic programming model (mixed integer programming)
  - Stage 1 decision: s - how many seconds until the goalie is pulled
  - Randomness:

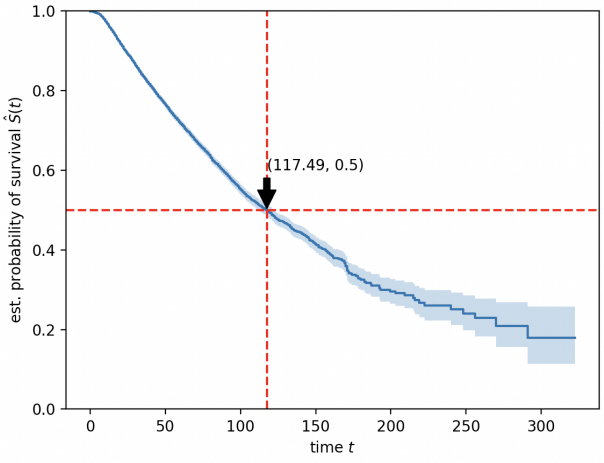

Figure 2: Kaplan Meier Survival Estimation for Oponent Scoring After Goalie Pull

    - $*$ G_{a1}: Number of goals scored by our team during epoch 1 (Binomial RV)
    - $*$ G_{a2}: Number of goals scored by our team during epoch 2 (Binomial RV)
    - $*$ G_{b1}: Number of goals scored by team B during epoch 1 (Binomial RV)
    - $*$ G_{b2}: Number of goals scored by team B during epoch 2 (Binomial RV)
    - $*$ $\mathscr{E}_i = [g_{a1} \in G_{a1}, g_{a2} \in G_{a2}, g_{b1} \in G_{b1}, g_{b2} \in G_{b2}]$: possible event space (all combinations of possibilities for how many goals were scored by each team within each epoch)
    - $*$ $\sim 2 \times 10^{12}$ situations exist for $g_{a1}, g_{a2}, g_{b1}, g_{b2} \in [0 : T]$
  - No Stage 2 recourse decision
  - Decision-dependent probability (endogenous uncertainty)

- **Objective:** $\max(\mathbb{P}[\text{success}]) = \max(\mathbb{P}[\text{our team wins or ties}]) = \max(\mathbb{P}[a_0 + a_1 + a_2 \geq b_0 + b_1 + b_2])$
- $\mathbb{P}[\text{success}] = \sum_{\mathscr{E}_i} \mathbb{P}[\mathscr{E}_i] \mathbb{1}_{\text{success}(\mathscr{E}_i)}$
- $\mathbb{P}[\mathscr{E}_i] = \mathbb{P}[(G_{a1} == g_{a1}) \& (G_{a2} == g_{a2}) \& (G_{b1} == g_{b1}) \& (G_{b2} = g_{b2})]$
  $= \binom{s}{g_{a1}} (p_{a1})^{g_{a1}} (1-p_{a1})^{s-g_{a1}} \cdot \binom{s}{g_{b1}} (p_{b1})^{g_{b1}} (1-p_{b1})^{s-g_{b1}} \cdot \binom{T-s}{g_{a2}} (p_{a2})^{g_{a2}} (1-p_{a2})^{T-s-g_{a2}} \cdot \binom{T-s}{g_{b2}} (p_{b2})^{g_{b2}} (1-p_{b2})^{T-s-g_{b2}}$
- **Objective (simplified)**: $\max(\sum_{\mathscr{E}_i} w_{\mathscr{E}i})$

- **Constraints:**
  - $w_{\mathscr{E}i} \leq \mathbb{P}[\mathscr{E}_i]$ (amount incorporated into the sum can be at most the probability of the event, when indicator variable is 1)
  - want $w_{\mathscr{E}_i} \leq 0$ iff $a_0 + a_1 + a_2 < b_0 + b + 1 + b + 2$
    - let $a = a_0 + a_1 + a_2$ be our final score
    - let $b = b_0 + b_1 + b_2$ be team B's final score
    - let $d = a + 1 - b$. Note that $d \geq 1$ if we tie or win and $d \leq 0$ if we lose
    - let $M = 3599$ be the maximum possible value of d (if their team scored every second of the game and our team never scored)
  - want $w_{\mathscr{E}_i} \leq \max(0, d)$ (will be 0 when indicator variable is 0)
    - $w_{\mathscr{E}_i} \geq 0$
    - $w_{\mathscr{E}_i} \geq d$
    - $w_{\mathscr{E}_i} \leq 0 + M(1 - y)$
    - $w_{\mathscr{E}_i} \leq d + My$ (y is an indicator variable)
    - $y \in \{0, 1\}$ (mixed integer programming!)
  - $s \in \mathbb{Z}^+$ (we can only pull the goalie on the second
  - $0 < s \leq 1200$ (we are in the final period of the game)

5 Discussion

5.1 Decision-dependent uncertainty

Our model is made complicated by the fact that the probability that a given event $\mathscr{E}_i = [g_{a1}, g_{a2}, g_{b1}, g_{b2}]$ occurs is necessarily based on our decision s. In other words, we have *decision-dependent uncertainty* [Hellemo et al., 2018].

One way to handle this decision-dependent uncertainty is to model our choice as selecting between probability distributions. Because we have the constraints $0 < s \leq 1200$ and $s \in \mathbb{Z}^+$, we have a finite number of choices, which means that we are selecting from a finite number of probability distributions; each choice s can be mapped to a discrete probability distribution over our events $\mathscr{E}_i$. These types of "decision-dependent distribution selection" problems are easier to solve than problems with "decision-dependent uncertainty" more generally.

5.2 Complexity

Mixed integer programming (MIP) is NP-hard, so there is no guarantee that we have an efficient method for solving MIP optimization problems. In fact, even mixed integer linear programs are NP-complete. In practice, however, models are able to leverage approximations and take advantage of specific problem's structures in order to solve these problems [Smith, 2024]. Unfortunately, in our

case we have a nonlinear and nonconvex function, so we do not have simple methods to solve this problem.

In order to make solving this problem tractable, we have chosen to model the results of a hockey game such that each time has a binary result (scoring or not scoring) for each second of the game. To solve the problem more efficiently, we may need to broaden these "buckets" of time in which a team can score to 30 seconds. This change would decrease our event space size from $1200^4 \simeq 2 \times 10^{12}$ to only $40^4 \simeq 2 \times 10^6$. In essence, this technique is similar to sampling discrete events from a probability distribution in order to have a smaller event space.

Finally, in order to take advantage of our problem's "distribution selection" flavor, we chose to precompute all values of n choose k necessary for computing the probability distributions. Although we did not precompute the probability distribution for each possible decision, this step saves significant compute time and avoids redundant calcuations.

5.3 Progress

Please review our github repository here: https://github.com/ryan-moreno/CS524-project; the README.md file has concise information about our data processing and model. Note that as a place to start we are first attempting to brute-force the model without using a solver. As indicated above, our problem is complicated by its non-linear and non-convex structure, making it difficult to solve in an efficient manner. We will attempt to use IPopt as our solver to address these concerns. We expect to have the model structured to use IPopt by next week.

References

A. Agalea. Nhl goalie pull optimization: Processed data, 2023. URL https://github.com/agalea91/nhl-goalie-pull-optimization/tree/master/data/processed/csv. Accessed: 2024-07-27.

L. Hellemo, P. I. Barton, and A. Tomasgard. Decision-dependent probabilities in stochastic programs with recourse. *Computational Management Science*, 15(3):369–395, 2018. ISSN 1619-6988. doi: 10.1007/s10287-018-0330-0. URL https://doi.org/10.1007/s10287-018-0330-0.

N. H. League. Team statistics: Goals against by strength, 2024a. URL https://www.nhl.com/stats/teams?aggregate=0&report=goalsagainstbystrength&reportType=game&seasonFrom=20162017&seasonTo=20232024&dateFromSeason&gameType=3&sort=a_teamFullName&page=14&pageSize=100. Accessed: 2024-07-27.

N. H. League. Team statistics: Goals for by strength, 2024b. URL https://www.nhl.com/stats/teams?aggregate=0&report=goalsforbystrength&reportType=game&seasonFrom=20162017&seasonTo=20232024&dateFromSeason&gameType=3&sort=a_teamFullName&page=14&pageSize=100. Accessed: 2024-07-27.

A. Smith. Lecture 1 - problem complexity or, why integer programs are so hard to solve. module 12 - integer programming basics, July 2024.