

A COMPUTING PLAYER RATINGS IN HUMAN DATA

To compute player ratings, we used a regularized logistic outcome model. Specifically, we optimized the loss

$$L(s|\mathcal{D}) = \mathbb{E}_{(i,j) \in \mathcal{D}} [\sigma(s_i - s_j)] + \lambda |s|_2$$

across all pairs of players $(i, j) \in \mathcal{D}$ where player i achieved a "better" outcome than player j in a game. We found this approach led to more plausible scores than Elo (Elo, 1978) or TrueSkill (Herbrich et al., 2007) ratings.

Paquette et al. (2019) took an orthogonal approach to filter poor players from the training data: they only trained on "winning" powers, i.e. those who ended the game with at least 7 SCs. This filtering is sensible for training a policy for play, but is problematic for training policies for search. In a general-sum game, it is crucial for the agent to be able to predict the empirical distribution of actions even for other agents who are destined to lose.

B SEARCH EXAMPLE

The listing below shows the policy generated by one run of our search algorithm for the opening move of Diplomacy when running for 512 iterations and considering 8 possible actions per power.

For each possible action, the listing below shows

- probs : The probability of the action in the final strategy.
- bp_p : The probability of the action in the blueprint strategy.
- avg_u : The average predicted sum-of-squares utility of this action.
- orders : The orders for this action

```
AUSTRIA avg_utility=0.15622
probs    bp_p    avg_u    orders
0.53648  0.13268  0.15697  ('A VIE - TRI', 'F TRI - ALB', 'A BUD - SER')
0.46092  0.52008  0.14439  ('A VIE - GAL', 'F TRI - ALB', 'A BUD - SER')
0.00122  0.03470  0.14861  ('A VIE - TRI', 'F TRI - ALB', 'A BUD - GAL')
0.00077  0.03031  0.11967  ('A VIE - BUD', 'F TRI - ALB', 'A BUD - SER')
0.00039  0.05173  0.11655  ('A VIE - GAL', 'F TRI S A VEN', 'A BUD - SER')
0.00015  0.04237  0.12087  ('A VIE - GAL', 'F TRI H', 'A BUD - SER')
0.00007  0.14803  0.09867  ('A VIE - GAL', 'F TRI - VEN', 'A BUD - SER')
0.00000  0.04009  0.03997  ('A VIE H', 'F TRI H', 'A BUD H')

ENGLAND avg_utility=0.07112
probs    bp_p    avg_u    orders
0.41978  0.20069  0.07151  ('F EDI - NTH', 'F LON - ENG', 'A LVP - YOR')
0.34925  0.29343  0.07161  ('F EDI - NWG', 'F LON - NTH', 'A LVP - YOR')
0.10536  0.06897  0.07282  ('F EDI - NTH', 'F LON - ENG', 'A LVP - WAL')
0.07133  0.36475  0.07381  ('F EDI - NWG', 'F LON - NTH', 'A LVP - EDI')
0.05174  0.01649  0.07202  ('F EDI - NTH', 'F LON - ENG', 'A LVP - EDI')
0.00249  0.00813  0.06560  ('F EDI - NWG', 'F LON - NTH', 'A LVP - WAL')
0.00006  0.00820  0.06878  ('F EDI - NWG', 'F LON - ENG', 'A LVP - EDI')
0.00000  0.03933  0.03118  ('F EDI H', 'F LON H', 'A LVP H')

FRANCE avg_utility=0.21569
probs    bp_p    avg_u    orders
0.92038  0.09075  0.21772  ('F BRE - MAO', 'A PAR - GAS', 'A MAR - BUR')
0.06968  0.42617  0.18878  ('F BRE - MAO', 'A PAR - BUR', 'A MAR S A PAR - BUR')
0.00917  0.07987  0.16941  ('F BRE - MAO', 'A PAR - PIC', 'A MAR - BUR')
0.00049  0.05616  0.16729  ('F BRE - ENG', 'A PAR - BUR', 'A MAR - SPA')
0.00023  0.17040  0.17665  ('F BRE - MAO', 'A PAR - BUR', 'A MAR - SPA')
0.00004  0.04265  0.18629  ('F BRE - MAO', 'A PAR - PIC', 'A MAR - SPA')
0.00001  0.09291  0.15828  ('F BRE - ENG', 'A PAR - BUR', 'A MAR S A PAR - BUR')
0.00000  0.04109  0.06872  ('F BRE H', 'A PAR H', 'A MAR H')

GERMANY avg_utility=0.21252
probs    bp_p    avg_u    orders
0.39050  0.01382  0.21360  ('F KIE - DEN', 'A MUN - TYR', 'A BER - KIE')
0.38959  0.02058  0.21381  ('F KIE - DEN', 'A MUN S A PAR - BUR', 'A BER - KIE')
0.16608  0.01628  0.21739  ('F KIE - DEN', 'A MUN H', 'A BER - KIE')
0.04168  0.21879  0.21350  ('F KIE - DEN', 'A MUN - BUR', 'A BER - KIE')
0.01212  0.47409  0.21287  ('F KIE - DEN', 'A MUN - RUH', 'A BER - KIE')
0.00003  0.05393  0.14238  ('F KIE - HOL', 'A MUN - BUR', 'A BER - KIE')
0.00000  0.16896  0.13748  ('F KIE - HOL', 'A MUN - RUH', 'A BER - KIE')
0.00000  0.03355  0.05917  ('F KIE H', 'A MUN H', 'A BER H')

ITALY avg_utility=0.13444
probs    bp_p    avg_u    orders
0.41740  0.19181  0.13609  ('F NAP - ION', 'A ROM - APU', 'A VEN S F TRI')
0.25931  0.07652  0.12465  ('F NAP - ION', 'A ROM - VEN', 'A VEN - TRI')
```

0.13084	0.29814	0.12831	('F NAP - ION', 'A ROM - VEN', 'A VEN - TYR')
0.09769	0.03761	0.13193	('F NAP - ION', 'A ROM - APU', 'A VEN - TRI')
0.09412	0.16622	0.13539	('F NAP - ION', 'A ROM - APU', 'A VEN H')
0.00034	0.05575	0.11554	('F NAP - ION', 'A ROM - APU', 'A VEN - PIE')
0.00028	0.13228	0.10953	('F NAP - ION', 'A ROM - VEN', 'A VEN - PIE')
0.00000	0.04167	0.05589	('F NAP H', 'A ROM H', 'A VEN H')
RUSSIA avg_utility=0.06623			
probs	bp_p	avg_u	orders
0.64872	0.05988	0.06804	('F STP/SC - FIN', 'A MOS - UKR', 'A WAR - GAL', 'F SEV - BLA')
0.28869	0.07200	0.06801	('F STP/SC - BOT', 'A MOS - STP', 'A WAR - UKR', 'F SEV - BLA')
0.04914	0.67998	0.05929	('F STP/SC - BOT', 'A MOS - UKR', 'A WAR - GAL', 'F SEV - BLA')
0.01133	0.01147	0.05023	('F STP/SC - BOT', 'A MOS - SEV', 'A WAR - UKR', 'F SEV - RUM')
0.00120	0.02509	0.05008	('F STP/SC - BOT', 'A MOS - UKR', 'A WAR - GAL', 'F SEV - RUM')
0.00064	0.09952	0.05883	('F STP/SC - BOT', 'A MOS - STP', 'A WAR - GAL', 'F SEV - BLA')
0.00027	0.01551	0.04404	('F STP/SC - BOT', 'A MOS - SEV', 'A WAR - GAL', 'F SEV - RUM')
0.00000	0.03655	0.02290	('F STP/SC H', 'A MOS H', 'A WAR H', 'F SEV H')
TURKEY avg_utility=0.13543			
probs	bp_p	avg_u	orders
0.82614	0.25313	0.13787	('F ANK - BLA', 'A SMY - ARM', 'A CON - BUL')
0.14130	0.00651	0.12942	('F ANK - BLA', 'A SMY - ANK', 'A CON - BUL')
0.03080	0.61732	0.12760	('F ANK - BLA', 'A SMY - CON', 'A CON - BUL')
0.00074	0.01740	0.11270	('F ANK - CON', 'A SMY - ARM', 'A CON - BUL')
0.00069	0.05901	0.12192	('F ANK - CON', 'A SMY - ANK', 'A CON - BUL')
0.00030	0.00750	0.11557	('F ANK - CON', 'A SMY H', 'A CON - BUL')
0.00001	0.00598	0.10179	('F ANK S F SEV - BLA', 'A SMY - CON', 'A CON - BUL')
0.00001	0.03314	0.04464	('F ANK H', 'A SMY H', 'A CON H')

C RL DETAILS

Each action a in the MDP is a sequence of orders (o_1, \dots, o_t) . The probability of the order a under policy π_θ is defined by an LSTM in auto regressive fashion, i.e., $\pi_\theta(a) = \prod_{i=1}^t (\pi_\theta(o_i | o_1 \dots o_{i-1}))$. To make training more stable, we would like to prevent entropy $H(\pi_\theta) := -E_{a \sim \pi_\theta} \log(\pi_\theta(a))$ from collapsing to zero. The naive way to optimize the entropy of the joint distribution is to use a sum of entropies for each individual order, i.e., $\frac{d}{d\theta} H(\pi_\theta(\bullet)) \approx \sum_{i=1}^t \frac{d}{d\theta} H(\pi_\theta(\bullet | o_1 \dots o_{i-1}))$. However, we found that this does not work well for our case probably because there are strong correlations between orders. Instead we use an unbiased estimate of the joint entropy that is agnostic to the size of the action space and requires only to be able to sample from a model and to adjust probabilities of the samples.

Statement 1. Let $\pi_\theta(\bullet)$ be a probability distribution over a discrete set A , such that $\forall a \in A \pi_\theta(a)$ is a smooth function of a vector of parameters θ . Then

$$\frac{d}{d\theta} (H(\pi_\theta(\bullet))) = -E_{a \sim \pi_\theta} (1 + \log \pi_\theta(a)) \frac{d}{d\theta} \log \pi_\theta(a).$$

Proof. Proof is similar to one for REINFORCE:

$$\begin{aligned}
\frac{d}{d\theta} (H(\pi_\theta(\bullet))) &= \frac{d}{d\theta} (-E_{a \sim \pi_\theta} \log \pi_\theta(a)) \\
&= -\frac{d}{d\theta} \left(\sum_{a \in A} \pi_\theta(a) \log \pi_\theta(a) \right) \\
&= -\sum_{a \in A} \left(\frac{d}{d\theta} \pi_\theta(a) \log \pi_\theta(a) \right) \\
&= -\sum_{a \in A} \left(\pi_\theta(a) \frac{d}{d\theta} \log \pi_\theta(a) + \log \pi_\theta(a) \frac{d}{d\theta} \pi_\theta(a) \right) \\
&= -\sum_{a \in A} \left(\pi_\theta(a) \frac{d}{d\theta} \log \pi_\theta(a) + \log \pi_\theta(a) \pi_\theta(a) \frac{d}{d\theta} \log \pi_\theta(a) \right) \\
&= -\sum_{a \in A} \pi_\theta(a) \left((1 + \log \pi_\theta(a)) \frac{d}{d\theta} \log \pi_\theta(a) \right) \\
&= -E_{a \sim \pi_\theta} \left((1 + \log \pi_\theta(a)) \frac{d}{d\theta} \log \pi_\theta(a) \right).
\end{aligned}$$

□

D SUBGAME EXPLOITABILITY RESULTS

Figure D plots the total exploitability of joint policies computed by ERM in the subgame used for equilibrium search at each phase in 7 simulated Diplomacy games.

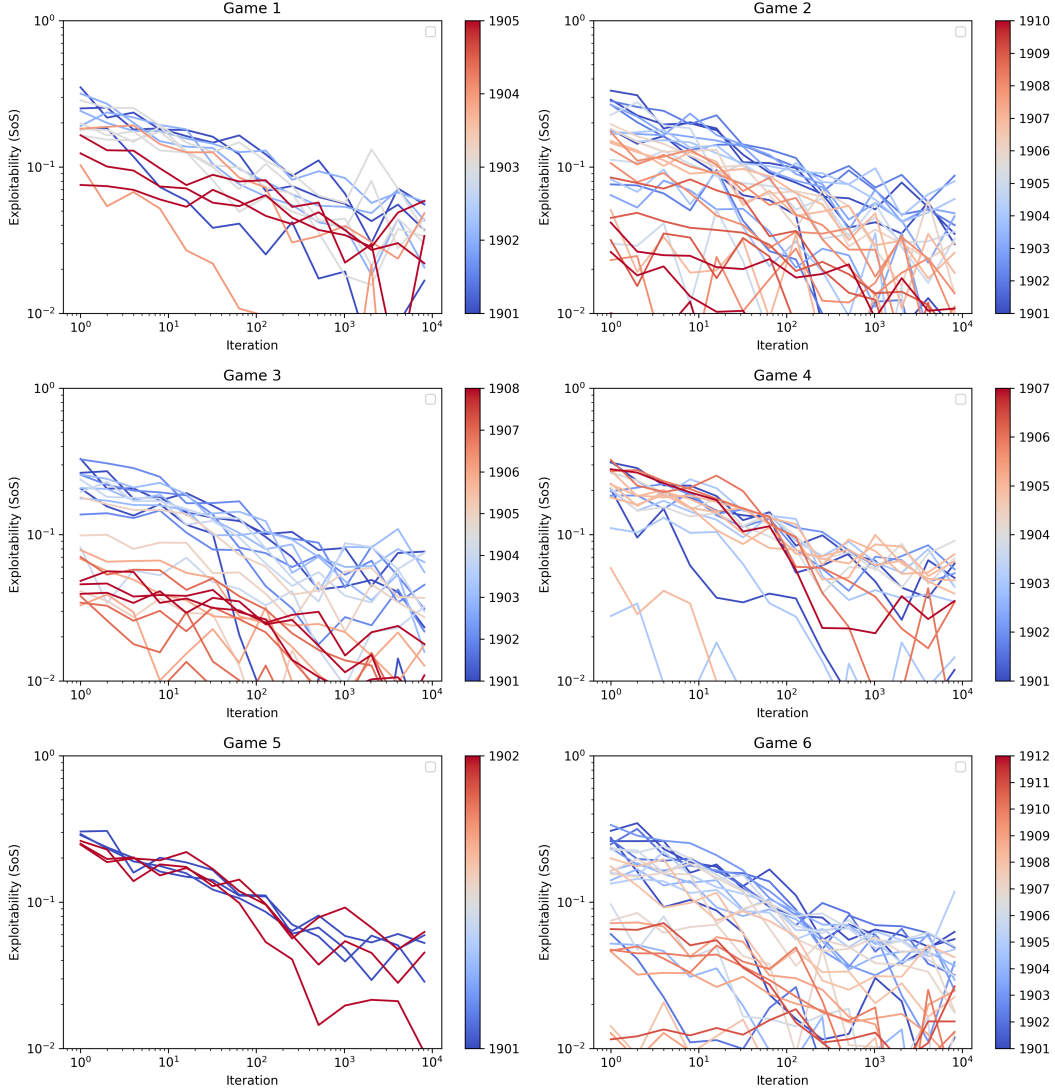


Figure 4: Exploitability as a function of ERM iteration at each phase of 7 simulated games with our search agent (until the game ends or the search agent is eliminated). Aggregate results in Figure 3.

E DETAILS ON EXPERIMENTAL SETUP

Most games on webdiplomacy.net were played with 24-hour turns, though the agent also played in some “live” games with 5-minute turns. Different hyperparameters were used for live games versus non-live games. In non-live games, we typically ran ERM for 2,048 iterations with a rollout length of 3 movement phases, and set M_i equal to 5 times the number of units a player controls. This typically required about 20 minutes to compute. In live games, including games in which one human played against six bots, we typically ran ERM for 256 iterations with a rollout

length of 2 movement phases, and set M_i equal to 3.5 times the number of units a player controls. This typically required about 2 minutes to compute. In all cases, the temperature for the blueprint in rollouts was set to 0.75.

Agent A (1x)	Agent B (6x)	SoS Score	Games
SearchBot	SL DipNet	54.5% \pm 2%	670
SearchBot	RL DipNet	35.6% \pm 2%	699
SearchBot	Blueprint	60.8% \pm 2%	699
SL DipNet	SearchBot	0.2% \pm 0.2%	138
RL DipNet	SearchBot	0.7% \pm 0.3%	700
Blueprint	SearchBot	0.6% \pm 0.3%	140

Table 4: Comparison of average sum-of-squares scores for our agent (SearchBot) in 1v6 games with DipNet agents from Paquette et al. (2019), as well as our own blueprint imitation learning agent. All agents other than SearchBot use a temperature of 0.1.

Power	Score	Human Mean	Games	Wins	Draws	Losses
All Games	25.6% \pm 4.8%	14.3%	50	7	16	27
Normalized By Power	27.0% \pm 5.3%	14.3%	50	7	16	27
Austria	42.4%	11.0%	7	2	2	3
England	29.4%	12.6%	8	1	3	4
France	7.8%	16.9%	9	0	3	6
Germany	41.8%	15.0%	6	1	3	2
Italy	31.7%	11.6%	5	1	2	2
Russia	18.8%	14.8%	8	1	2	5
Turkey	17.1%	18.2%	7	1	1	5

Table 5: Performance of our agent in anonymous games against humans on `webdiplomacy.net`. Average human performance is 14.3%. Score in the case of draws was determined by the rules of the joined game. The \pm shows one standard error. Due to small sample sizes, we do not display a \pm for individual powers. Average human performance for was calculated based on SoS scoring of historical games on `webdiplomacy.net`.

Power	1 Human vs. 6 DipNet	1 Human vs. 6 Blueprint	1 Human vs. 6 SearchBot
All Games	39.1%	22.5%	5.7%
Austria	40%	20%	0%
England	28.4%	20%	0%
France	20%	4.3%	40%
Germany	40%	33%	0%
Italy	60%	20%	0%
Russia	40%	20%	0%
Turkey	45.1%	40%	0%

Table 6: Performance of one expert human playing against six bots under repeated play. A score less than 14.3% suggests the human is unable to exploit the bot. Five games were played for each power for each agent, for a total of 35 games per agent. For each power, the human first played all games against DipNet, then the blueprint model described in Section 3.1, and then finally SearchBot.

The experiments on `webdiplomacy.net` occurred over a three-month timespan, with games commonly taking one to two months to complete (players are typically given 24 hours to act). Freezing research and development over such a period would have been impractical, so our agent was not fixed for the entire time period. Instead, serious bugs were fixed, improvements to the algorithm were made, and the model was updated.

F QUALITATIVE ASSESSMENT OF SEARCHBOT

Qualitatively, we observe that SearchBot performs particularly well in the early and mid game. However, we observe that it sometimes struggles with endgame situations. In particular, when it is clear that one power will win unless the others work together to stop it, SearchBot will sometimes continue to attack its would-be allies. There may be multiple contributing factors to this. One important limitation, which we have verified in some situations, is that the sampled subgame actions may not contain any action that could prevent a loss. This is exacerbated by the fact that players typically control far more units in the endgame and the number of possible actions grows exponentially with the number of units, so the sampled subgame actions contain a smaller fraction of all possible actions. Another possible contributing factor is that the state space near the end of the game is far larger, so there is relatively less data for supervised learning in this part of the game. Finally, another possibility is that because the network only has a dependence on the state of the board and the most recent player actions, it is unable to sufficiently model the future behavior of other players in response to being attacked.

Although the sample size is small, the results suggest that SearchBot performed particularly well with the central powers of Austria, Germany, and Italy. These powers are considered to be the most difficult to play by humans because they are believed to require an awareness of the interactions between all players in the game.