International Journal of Information Technology & Decision Making


# ANALYSIS OF HYPER-PARAMETERS FOR ALPHAZERO-LIKE DEEP REINFORCEMENT LEARNING

FIRST AUTHOR

*University Department, University Name, Address*
*City, State ZIP/Zone, Country*

Received Day Month Year
Revised Day Month Year

The landmark achievements of AlphaGo Zero have created great research interest into self-play in reinforcement learning. In self-play, Monte Carlo Tree Search is used to train a deep neural network, which is then used itself in tree searches. The training is governed by many hyper-parameters. There has been surprisingly little research on design choices for hyper-parameter values and loss functions, presumably because of the prohibitive computational cost to explore the parameter space. In this paper, we investigate 12 hyper-parameters in an AlphaZero-like self-play algorithm and evaluate how these parameters contribute to training. We study them on small games, to achieve meaningful exploration with moderate computational effort. The experimental results show that training is highly sensitive to hyper-parameter choices. Through multi-objective analysis, we identify 4 important hyper-parameters to further assess. To start, we find surprising results where too much training can sometimes lead to lower performance. Our main result is that the number of self-play iterations subsumes MCTS-search simulations, game episodes, and training epochs. The intuition is that these three increase together as self-play iterations increase and that increasing them individually is suboptimal. As a consequence of our experiments, we provide recommendations on setting hyper-parameter values in self-play. The outer loop of self-play iterations should be emphasized, in favor of the inner loop. This means hyper-parameters for the inner loop, should be set to lower values. A secondary result of our experiments concerns the choice of optimization goals, for which we also provide recommendations.

*Keywords*: AlphaZero; parameter sweep; parameter evaluation; loss function.

## 1. Introduction

The AlphaGo series of papers [1,2,3] have sparked much interest of researchers and the general public alike into deep reinforcement learning. Despite the success of AlphaGo and related methods in Go and other application areas [4,5], there are unexplored and unsolved puzzles in the design and parameterization of the algorithms. Different hyper-parameter settings can lead to very different results. However, hyper-parameter design-space sweeps are computationally very expensive, and in the original publications, we can only find limited information of how to set the values of some important parameters and why. Also, there are few works on how to set the hyper-parameters for these algorithms, and more insight into the

hyper-parameter interactions is necessary. In our work, we study the most general framework algorithm in the aforementioned AlphaGo series by using a lightweight re-implementation of AlphaZero: AlphaZeroGeneral [6].

In order to optimize hyper-parameters, it is important to understand their function and interactions in an algorithm. A single iteration in the AlphaZeroGeneral framework consists of three stages: self-play, neural network training and arena comparison. In these stages, we explore 12 hyper-parameters (see section 4.1) in AlphaZeroGeneral. Furthermore, we observe 2 objectives (see section 4.2): training loss and time cost in each single run. A sweep of the hyper-parameter space is computationally demanding. In order to provide a meaningful analysis we use small board sizes of typical combinatorial games. This sweep provides an overview of the hyper-parameter contributions and provides a basis for further analysis. Based on these results, we choose 4 interesting parameters to further evaluate in depth.

As performance measure, we use the Elo rating that can be computed during training time of the self-play system, as a running relative Elo, and computed separately, in a dedicated tournament between different trained players.

Our contributions can be summarized as follows:

- We find (1) that in general higher values of all hyper-parameters lead to higher playing strength, but (2) that within a limited budget, a higher number of outer iterations is more promising than higher numbers of inner iterations: these are subsumed by outer iterations.
- We evaluate 4 alternative loss functions for 3 games and 2 board sizes, and find that the best setting depends on the game and is usually not the sum of policy and value loss. However, the sum may be a good default compromise if no further information about the game is present.

The paper is structured as follows. We first give an overview of the most relevant literature, before describing the considered test games in Sect. 3. Then we describe the AlphaZero-like self-play algorithm in Sect. 4. After setting up experiments, we present the results in Sect. 6. Finally, we conclude our paper and discuss the promising future work.

## 2. Related work

Hyper-parameter tuning by optimization is very important for many practical algorithms. In reinforcement learning, for instance, the $\epsilon$-greedy strategy of classical Q-learning is used to balance exploration and exploitation. Different $\epsilon$ values lead to different learning performance [7]. Another well known example of hyper-parameter tuning is the parameter $C_p$ in Monte Carlo Tree Search (MCTS) [8]. There are many works on tuning $C_p$ for different kinds of tasks. These provide insight on setting its value for MCTS in order to balance exploration and exploitation [9]. In deep reinforcement learning, the effect of the many neural network parameters are a black-box that precludes understanding, although the strong decision accuracy of

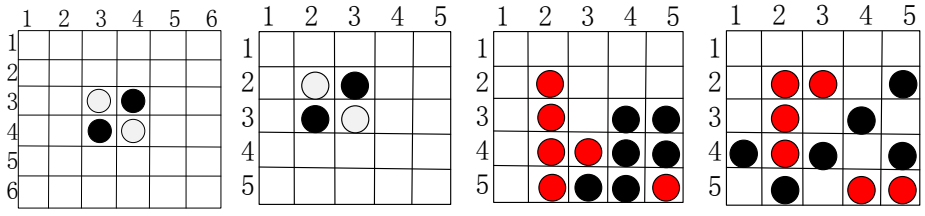
deep learning is undeniable [10], as the results in Go (and many other applications) have shown [11]. After AlphaGo [1], the role of self-play became more and more important. Earlier works on self-play in reinforcement learning are studied [12,13,14,15] and an overview is provided [16].

On hyper-parameters and loss-functions for AlphaZero-like systems there are a few studies: Chen et al. [17] tuned some parameters (in particular MCTS-related parameters in self-play game playing) in AlphaGo with Bayesian optimization, which leads to abandoning the fast rollout in AlphaGo Zero and AlphaZero. Mandai et al. [18] studied policy and value network optimization as a multi-task learning problem [19]. Matsuzaki compares MCTS with evaluation functions of different quality, and finds different results in Othello [20] than AlphaGo's PUCT. Moreover, Matsuzaki et al. [21] showed that the value function has more importance than the policy function in the PUCT algorithm for Othello. In our study, we extend this work and look more deeply into the relationship between value and policy functions in games.

Our experiments are also performed using AlphaZeroGeneral [6] which learns from from tabula rasa on several smaller games, namely 5×5 and 6×6 Othello [22], 5×5 and 6×6 Connect Four [23] and 5×5 and 6×6 Gobang [24]. The smaller size of these games allows us to do more experiments, and they also provide us largely uncharted territory where we hope to find effects that cannot be seen in Go or Chess.[a]

## 3. Test Games

In our hyper-parameter sweep experiments, we use Othello with a 6×6 board size, see Fig. 1(a). In the alternative loss function experiments, we use the games Othello, Connect Four and Gobang, each with 5×5 and 6×6 board sizes.



(a) 6×6 Othello (b) 5×5 Othello (c) 5×5 Connect Four (d) 5×5 Gobang

Fig. 1. Starting positions for Othello, examples for Connect Four and Gobang

Othello is a two-player game. Players take turns placing their own color pieces. Any opponent's color pieces that are in a straight line and bounded by the piece just placed and another piece of the current player's are flipped to the current player's

[a]This part of the work is published at the IEEE SSCI 2019 conference [25], and is included here for completeness.

color. While the last legal position is filled, the player who has most pieces wins the game. Fig. 1(a/b) show the start configurations for Othello. Connect Four is a two-player connection game. Players take turns dropping their own pieces from the top into a vertically suspended grid. The pieces fall straight down and occupy the lowest position within the column. The player who first forms a horizontal, vertical, or diagonal line of four pieces wins the game. Fig. 1(c) is a game termination example for 5×5 Connect Four where the red player wins the game. Gobang is another connection games that traditionally is played with Go pieces (black and white stones) on a Go board. Players alternate turns, placing a stone of their color on an empty position. The winner is the first player to form an unbroken chain of 4 stones horizontally, vertically, or diagonally. Fig. 1(d) is a game termination example for 5×5 Gobang where the black player wins the game. These games are usually employed as test cases in game playing [26,27,28,29,30,31].

## 4. AlphaZero-like Self-play

### 4.1. *The base algorithm*

Following the works by Silver et al. [2,3] the fundamental structure of AlphaZero-like Self-play is an iteration over three different stages (see Algorithm 4.1).

**Algorithm 4.1** AlphaZero-like Self-play Algorithm

**function** AlphaZeroGeneral
  Initialize $f_\theta$ with random weights; Initialize retrain buffer $D$ with capacity $N$
  **for** iteration=1, ..., $I$ **do**
    **for** episode=1,..., $E$ **do** ▷ stage 1
      **for** t=1, ..., $T\prime$, ..., $T$ **do**
        Get an enhanced best move prediction $\pi_t$ by performing MCTS based on $f_\theta(s_t)$
        Before step $T\prime$, select random action $a_t$ based on probability $\pi_t$, else select action $a_t = \arg\max_a(\pi_t)$
        Store example ($s_t$, $\pi_t$, $z_t$) in $D$
        Set $s_t$=excuteAction($s_t$, $a_t$)
      Label reward $z_t$ ($t \in [1, T]$) as $z_T$ in examples
    Randomly sample minibatch of examples ($s_j$, $\pi_j$, $z_j$) from $D$ ▷ stage 2
    $f_{\theta\prime} \leftarrow$ Train $f_\theta$ by performing optimizer to minimize Eq. 1 based on sampled examples
    Set $f_\theta = f_{\theta\prime}$ if $f_{\theta\prime}$ is better than $f_\theta$ ▷ stage 3
  **return** $f_\theta$;

The first stage is a **self-play** tournament. The computer player performs several games against itself in order to generate data for further training. In each step of a game (episode), the player runs MCTS to obtain, for each move, an enhanced policy $\pi$ based on the probability **p** provided by the neural network $f_\theta$. We now introduce the hyper-parameters, and their abbreviation that we use in this paper. In MCTS, hyper-parameter $C_p$ is used to balance exploration and exploitation of game tree search, and we abbreviate it to $c$. Hyper-parameter $m$ is the number of times to run down from the root for building the game tree, where the parameterized network

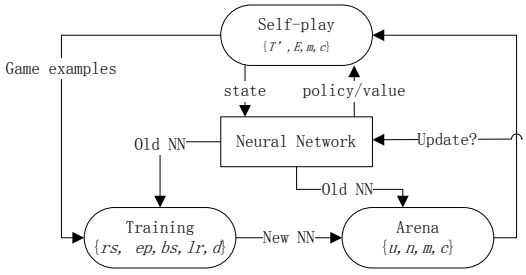



Fig. 2. A diagram of the schema of AlphaZero-like Self-play Algorithm over 3 stages with corresponding hyper-parameters. Hyper-parameter $I$ controls the loop over these 3 stages.

$f_\theta$ provides the value ($v$) of the states for MCTS. For actual (self-)play, from $T'$ steps on, the player always chooses the best move according to $\pi$. Before that, the player always chooses a random move based on the probability distribution of $\pi$. After finishing the games, the new examples are normalized as a form of $(s_t, \pi_t, z_t)$ and stored in $D$.

The second stage consists of **neural network training**, using data from the self-play tournament. Training lasts for several epochs. In each epoch ($ep$), training examples are divided into several small batches [32] according to the specific batch size ($bs$). The neural network is trained to minimize [33] the value of the *loss function* which (see Eq. 1) sums up the mean-squared error between predicted outcome and real outcome and the cross-entropy losses between $\mathbf{p}$ and $\pi$ with a learning rate ($lr$) and dropout ($d$). Dropout is used as probability to randomly ignore some nodes of the hidden layer in order to avoid overfitting [34].

The last stage is **arena comparison**, in which the newly trained neural network model ($f_{\theta\prime}$) is run against the previous neural network model ($f_\theta$). The better model is adopted for the next iteration. In order to achieve this, $f_{\theta\prime}$ and $f_\theta$ play against each other for $n$ games. If $f_{\theta\prime}$ wins more than a fraction of $u$ games, it is replacing the previous best $f_\theta$. Otherwise, $f_{\theta\prime}$ is rejected and $f_\theta$ is kept as current best model. Compared with AlphaGo Zero, AlphaZero does not entail the arena comparison stage anymore. However, we keep this stage for making sure that we can safely recognize improvements.

Furthermore, we intuitively present a diagram to describe the Algorithm 4.1 with necessary components for 3 stages and corresponding hyper-parameters in Fig. 2:

### 4.2. *Loss function*

The **training loss function** consists of $l_{\mathbf{p}}$ and $l_v$. The neural network $f_\theta$ is parameterized by $\theta$. $f_\theta$ takes the game board state $s$ as input, and provides the value $v_\theta \in [-1, 1]$ of $s$ and a policy probability distribution vector $\mathbf{p}$ over all legal actions as outputs. $\mathbf{p}_\theta$ is the policy provided by $f_\theta$ to guide MCTS for playing games. After performing MCTS, we obtain an improvement estimate for policy $\pi$. Training aims

at making $\mathbf{p}$ more similar to $\pi$. This can be achieved by minimizing the cross entropy of both distributions. Therefore, $l_{\mathbf{p}}$ is defined as $-\pi^\top \log \mathbf{p}$. The other aim is to minimize the difference between the output value ($v_\theta(s_t)$) of the state $s$ according to $f_\theta$ and the real outcome ($z_t \in \{-1, 1\}$) of the game. Therefore, $l_v$ is defined as the mean squared error $(v - z)^2$. Summarizing, the total loss function of AlphaZero is defined in Eq. 1.

$$l_+ = -\pi^\top \log \mathbf{p} + (v - z)^2 \tag{1}$$

Note that in AlphaZero's loss function, there is an extra regularization term to guarantee the training stability of the neural network. In order to pay more attention to two evaluation function components, instead, we apply standard measures to avoid overfitting such as the **dropout** mechanism.

### 4.3. *Bayesian Elo system*

The **Elo rating function** has been developed as a method for calculating the relative skill levels of players in games. Usually, in zero-sum games, there are two players, A and B. If their Elo ratings are $R_A$ and $R_B$, respectively, then the expectation that player A wins the next game is $E_A = \frac{1}{1+10^{(R_B - R_A)/400}}$. If the real outcome of the next game is $S_A$, then the updated Elo of player A can be calculated from its original Elo by $R_A = R_A + K(S_A - E_A)$, where K is the factor of the maximum possible adjustment per game. In practice, K should be bigger for weaker players but smaller for stronger players. Following AlphaZero [3], in our design, we adopt the Bayesian Elo system [35] to show the improvement curve of the learning player during self-play. We also employ this method to assess the playing strength of the final models.

### 4.4. *Time cost function*

Because of the high computational cost of self-play reinforcement learning, the running time of self-play is of great importance. We have created a **time cost function** to predict the running time, based on the algorithmic structure in Algorithm 4.1. According to Algorithm 4.1, the whole training process consists of several iterations with three steps as introduced in Section 4.1. Please refer to the algorithm and to Eq. 2. In $i$th iteration ($1 \leq i \leq I$), if we assume that in $j$th episode ($1 \leq j \leq E$), for $k$th game step (the size of $k$ mainly depends on the game complexity), the time cost of $l$th MCTS ($1 \leq l \leq m$) simulation is $t^{(i)}_{jkl}$, and assume that for $p$th epoch ($1 \leq p \leq ep$), the time cost of pulling $q$th batch ($1 \leq q \leq trainingExampleList.size/bs$)[b] through the neural network is $t^{(i)}_{pq}$, and assume that in $w$th arena comparison ($1 \leq w \leq n$), for $x$th game step, the time cost of $y$th MCTS simulation ($1 \leq y \leq m$) is $t^{(i)}_{xyw}$. The time cost of the whole

[b]the size of *trainingExampleList* is also relative to the game complexity

training process is summarized in Eq. 2.

$$\sum_{i}(\overbrace{\sum_{j}\sum_{k}\sum_{l}t_{jkl}^{(i)}}^{self-play}+\overbrace{\sum_{p}\sum_{q}t_{pq}^{(i)}}^{training}+\overbrace{\sum_{x}\sum_{y}\sum_{w}t_{xyw}^{(i)}}^{arena\ comparison}) \tag{2}$$

Please refer to Table 1 for an overview of the hyper-parameters. From Algorithm 4.1 and Eq. 2, we can see that the hyper-parameters, such as $I$, $E$, $m$, $ep$, $bs$, $rs$, $n$ etc., influence training time. In addition, $t_{jkl}^{(i)}$ and $t_{xyw}^{(i)}$ are simulation costs that rely on hardware capacity and game complexity. $t_{uv}^{(i)}$ also relies on the structure of the neural network. In our experiments, all neural network models share the same structure, which consists of 4 convolutional layers and 2 fully connected layers.

## 5. Experimental Setup

Our experiments are run on a machine with 128GB RAM, 3TB local storage, 20-core Intel Xeon E5-2650v3 CPUs (2.30GHz, 40 threads), 2 NVIDIA Titanium GPUs (each with 12GB memory) and 6 NVIDIA GTX 980 Ti GPUs (each with 6GB memory). In order to keep using the same GPUs, we deploy each run of experiments on the NVIDIA GTX 980 Ti GPU. Each run of experiments takes 2 to 3 days.

### 5.1. *Hyper-Parameter sweep*

We sweep the 12 hyper-parameters by configuring 3 different values (minimum value, default value and maximum value) to find the most promising parameter values. In each single run of training, we play 6×6 Othello [22] based on Algorithm 4.1 and change the value of one hyper-parameter, keeping the other hyper-parameters at default values (see Table 1).

Table 1. Default Hyper-Parameter Setting

| - | Description | Minimum | Default | Maximum |
|---|---|---|---|---|
| $I$ | number of iteration | 50 | 100 | 150 |
| $E$ | number of episode | 10 | 50 | 100 |
| $T$' | step threshold | 10 | 15 | 20 |
| $m$ | MCTS simulation times | 25 | 100 | 200 |
| $c$ | weight in UCT | 0.5 | 1.0 | 2.0 |
| $rs$ | number of retrain iteration | 1 | 20 | 40 |
| $ep$ | number of epoch | 5 | 10 | 15 |
| $bs$ | batch size | 32 | 64 | 96 |
| $lr$ | learning rate | 0.001 | 0.005 | 0.01 |
| $d$ | dropout probability | 0.2 | 0.3 | 0.4 |
| $n$ | number of comparison games | 20 | 40 | 100 |
| $u$ | update threshold | 0.5 | 0.6 | 0.7 |

### 5.2. *Hyper-Parameters correlation evaluation*

Based on the above experiments, we further explore the correlation of interesting hyper-parameters (i.e. $I$, $E$, $m$ and $ep$) in terms of their best final player's playing strength and overall training time. We set values for these 4 hyper-parameters as Table 2, and other parameters values are set to the default values in Table 1. In addition, for (and only for) this part of experiments, the stage 3 of Algorithm 4.1 is cut off. Instead, for every iteration, the trained model $f_{\theta\prime}$ is accepted as the current best model $f_\theta$ automatically, which is also adopted by AlphaZero and saves a lot of time.

Table 2. Correlation Evaluation Hyper-Parameter Setting

| - | Description | Minimum | Middle | Maximum |
|---|---|---|---|---|
| $I$ | number of iteration | 25 | 50 | 75 |
| $E$ | number of episode | 10 | 20 | 30 |
| $m$ | MCTS simulation times | 25 | 50 | 75 |
| $ep$ | number of epoch | 5 | 10 | 15 |

Note that due to computation resource limitations, for hyper-parameter sweep experiments on 6×6 Othello, we only perform single run experiments. This may cause noise, but still provides valuable insights on the importance of hyper-parameters under the AlphaZero-like self-play framework.

### 5.3. *Alternative loss function evaluation*

As we want to assess the effect of different loss functions, we employ a weighted sum loss function based on (3):

$$l_\lambda = \lambda(-\pi^\top \log \mathbf{p}) + (1-\lambda)(v-z)^2 \tag{3}$$

where $\lambda$ is a weight parameter. This provides some flexibility to gradually change the nature of the function. In our experiments, we first set $\lambda$=0 and $\lambda$=1 in order to assess $l_\mathbf{p}$ or $l_v$ independently. Then we use Eq. 1 as training loss function. Furthermore, we note from the theory of multi-attribute utility functions in multi-criteria optimization [36] that a sum tends to prefer extreme solutions, whereas a product prefers a more balanced solution. We employ a product combination loss function as follows:

$$l_\times = -\pi^\top \log \mathbf{p} \times (v-z)^2 \tag{4}$$

For all loss function experiments, each setting is run 8 times to get statistically significant results (we show error bars) using the hyper-parameters of Table 1 with their default values. However, in order to allow longer training, we enhance the

iteration number to 200 in the smaller games (5×5 Othello, 5×5 Connect Four and 5×5 Gobang).

The loss function in question is used to guide each training process, with the expectation that smaller loss means a stronger model. However, in practice, we have found that this is not always the case and another measure is needed to check. Following DeepMind's work, we use Bayesian Elo ratings [35] to describe the playing strength of the model in every iteration. In addition, for each game, we use all best players trained from the four different targets ($l_{\mathbf{p}}$, $l_v$, $l_+$, $l_\times$) and **8 repetitions**[c] plus a random player to play 20 round-robin rounds. We calculate the Elo ratings of the 33 players as the real playing strength of a player, rather than the one measured while training.

## 6. Experimental Results

In order to better understand the training process, first, we depict training loss evolution for default settings in Fig. 3.

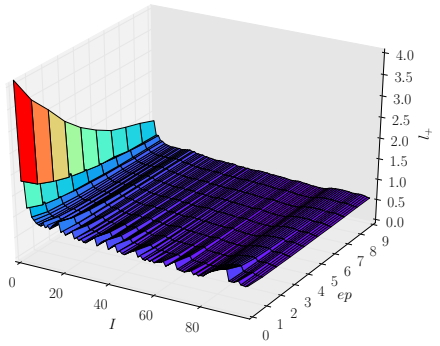



Fig. 3. Single run training loss over iterations I and epochs ep

We plot the training loss of each epoch in every iteration and see that (1) in each iteration, loss decreases along with increasing epochs, and that (2) loss also decreases with increasing iterations up to a relatively stable level.

### 6.1. *Hyper-Parameter sweep results*

***I:*** In order to find a good value for $I$ (iterations), we train 3 different models to play 6×6 Othello by setting $I$ at minimum, default and maximum value respectively. We keep the other hyper-parameters at their default values. Fig. 4(a) shows that

[c]In alternative loss function evaluation experiments, multiple runs for each setting are employed to avoid bias

training loss decreases to a relatively stable level. However, after iteration 120, the training loss unexpectedly increases to the same level as for iteration 100 and further decreases. This surprising behavior could be caused by a too high learning rate, an improper update threshold, or overfitting. This is an unexpected result since in theory more iterations lead to better performance.

***E:*** Since more episodes mean more training examples, it can be expected that more training examples lead to more accurate results. However, collecting more training examples also needs more resources. This shows again that hyper-parameter optimization is necessary to find a reasonable value of for $E$. In Fig. 4(b), for $E$=100, the training loss curve is almost the same as the 2 other curves for a long time before eventually going down.

***T':*** The step threshold controls when to choose a random action or the one suggested by MCTS. This parameter controls exploration in self-play, to prevent deterministic policies from generating training examples. Small $T'$ results in more deterministic policies, large $T'$ in policies more different from the model. In Fig. 4(c), we see that $T'$=10 is a good value.

***m:*** In theory, more MCTS simulations $m$ should provide better policies. However, higher $m$ requires more time to get such a policy. Fig. 4(d) shows that a value for 200 MCTS simulations achieves the best performance in the 70th iteration, then has a drop, to reach a similar level as 100 simulations in iteration 100.

***c:*** This hyper-parameter $C_p$ is used to balance the exploration and exploitation during tree search. It is often set at 1.0. However, in Fig. 4(e), our experimental results show that more exploitation ($c$=0.5) can provide smaller training loss.

***rs:*** In order to reduce overfitting, it is important to retrain models using previous training examples. Finding a good retrain length of historical training examples is necessary to reduce training time. In Fig. 4(f), we see that using training examples from the most recent single previous iteration achieves the smallest training loss. This is an unexpecrted result, suggesting that overfitting is prevented by other means and that the time saving works out best overall.

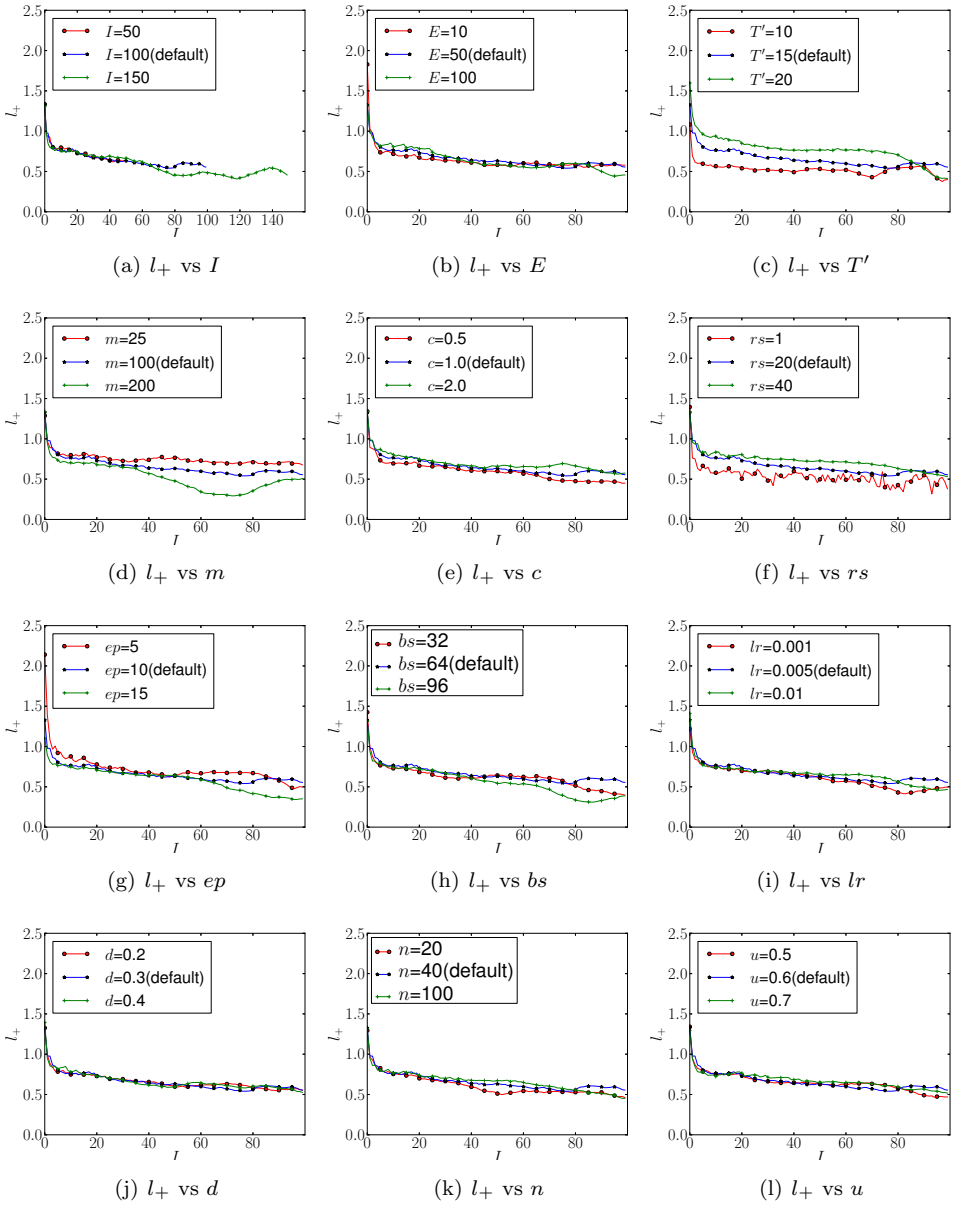
***ep:*** The training loss of different $ep$ is shown in Fig. 4(g). For $ep$=15 the training loss is the lowest. This result shows that along with the increase of epoch, the training loss decreases, which is as expected.

***bs:*** a smaller batch size $bs$ increases the number of batches, leading to higher time cost. However, smaller $bs$ means less training examples in each batch, which may cause more fluctuation (larger variance) of training loss. Fig. 4(h) shows that $bs$=96 achieves the smallest training loss in iteration 85.

***lr:*** In order to avoid skipping over optima, a small learning rate is generally suggested. However, a smaller learning rate learns (accepts) new knowledge slowly. In Fig. 4(i), $lr$=0.001 achieves the lowest training loss around iteration 80.

***d:*** Dropout is a popular method to prevent overfitting. Srivastava et al. claim that dropping out 20% of the input units and 50% of the hidden units is often found to be good [34]. In Fig. 4(j), however, we can not see a significant difference.

***n:*** The number of games in the arena comparison is a key factor of time cost. A




(a) $l_+$ vs $I$ (b) $l_+$ vs $E$ (c) $l_+$ vs $T'$

(d) $l_+$ vs $m$ (e) $l_+$ vs $c$ (f) $l_+$ vs $rs$

(g) $l_+$ vs $ep$ (h) $l_+$ vs $bs$ (i) $l_+$ vs $lr$

(j) $l_+$ vs $d$ (k) $l_+$ vs $n$ (l) $l_+$ vs $u$

Fig. 4. Training loss for different parameter settings over iterations.

small value may miss accepting good new models and too large a value is a waste of time. Our experimental results in Fig. 4(k) show that there is no significant difference. A combination with $u$ can be used to determine the acceptance or rejection of a newly learnt model. In order to reduce time cost, a small $n$ combined with a

large $u$ may be a good choice.

***u:*** This hyper-parameter is the update threshold. Normally, in two-player games, player A is better than player B if it wins more than 50% games. A higher threshold avoids fluctuations. However, if we set it too high, it becomes too difficult to accept better models. Fig. 4(l) shows that $u$=0.7 is too high, 0.5 and 0.6 are acceptable.

Table 3. Time Cost (hr) of Different Parameter Setting

| Parameter | Minimum | Default | Maximum | Type |
|---|---|---|---|---|
| *I* | **23.8** | 44.0 | 60.3 | time-sensitive |
| *E* | **17.4** | 44.0 | 87.7 | time-sensitive |
| *T'* | 41.6 | 44.0 | 40.4 | time-friendly |
| *m* | **26.0** | 44.0 | 64.8 | time-sensitive |
| *c* | 50.7 | 44.0 | 49.1 | time-friendly |
| *rs* | **26.5** | 44.0 | 50.7 | time-sensitive |
| *ep* | **43.4** | 44.0 | 55.7 | time-sensitive |
| *bs* | 47.7 | 44.0 | **37.7** | time-sensitive |
| *lr* | 47.8 | 44.0 | 40.3 | time-friendly |
| *d* | 51.9 | 44.0 | 51.4 | time-friendly |
| *n* | **33.5** | 44.0 | 57.4 | time-sensitive |
| *u* | 39.7 | 44.0 | 40.4 | time-friendly |

To investigate the impact on running time, we present the effect of different values for each hyper-parameter in Table 3. We see that for *I*, *E*, *m*, *rs*, *n*, smaller values lead to quicker training, which is as expected. For *bs*, larger values result in quicker training. The other hyper-parameters are indifferent, changing their values will not lead to significant changes in training time. Therefore, tuning these hyper-parameters shall reduce training time or achieve better quality in the same time.

Table 4. Importance Summary in Different Objectives

| Parameter | Default Value | Loss | Time Cost |
|---|---|---|---|
| *I* | 100 | 100 | **50** |
| *E* | 50 | **10** | **10** |
| *T'* | 15 | **10** | similar |
| *m* | 100 | **200** | **25** |
| *c* | 1.0 | **0.5** | similar |
| *rs* | 20 | **1** | **1** |
| *ep* | 10 | **15** | **5** |
| *bs* | 64 | **96** | **96** |
| *lr* | 0.005 | **0.001** | similar |
| *d* | 0.3 | 0.3 | similar |
| *n* | 40 | insignificant | **20** |
| *u* | 0.6 | insignificant | similar |

Based on the aforementioned results and analysis, we summarize the impor-

tance by evaluating contributions of each parameter to training loss and time cost, respectively, in Table 4 (best values in bold font). For training loss, different values of $n$ and $u$ do not result in a significant difference. Modifying time-indifferent hyper-parameters does not much change training time, whereas larger value of time-sensitive hyper-parameters lead to higher time cost.

### 6.2. *Hyper-Parameter correlation evaluation results*

In this part, we investigate the correlation between promising hyper-parameters in terms of time cost and playing strength. There are $3^4 = 81$ final best players trained based on 3 different values of 4 hyper-parameters ($I$, $E$, $m$ and $ep$) plus a random player (i.e. 82 in total). Any 2 of these 82 players play with each other. Therefore, there are 82×81/2=3321 pairs, and for each of these, 10 games are played.

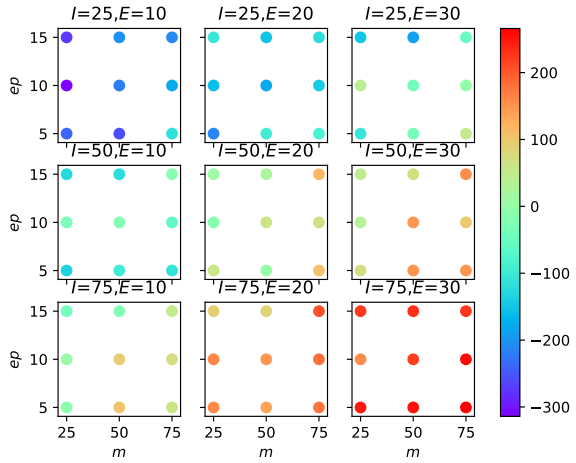



Fig. 5. Elo ratings of the final best players of the full tournament (3 parameters, 1 target value)

In each sub-figures of Fig. 5, all models are trained from the same value of $I$ and $E$, according to the different values in x-axis and y-axis, we find that, generally, larger $m$ and larger $ep$ lead to higher Elo ratings. However, in the last sub-figure, we can clearly notice that the Elo rating of $ep$=10 is higher than that of $ep$=15 for $m$=75, which shows that sometimes more training can not improve the playing strength but decreases the training performance. We suspect that this is caused by overfitting. Looking at the sub-figures, the results also show that more (outer) training iterations can significantly improve the playing strength, also more training examples in each iteration (bigger $E$) helps. These outer iterations are clearly more important than optimizing the inner hyper-parameters of $m$ and $ep$. Note that higher values for the outer hyper-parameters imply more MCTS simulations and

more training epochs, but not vice versa. This is an important insight regarding tuning hyper-parameters for self-play.

According to (Eq. 2) and Table. 4, we know that smaller values of time sensitive hyper-parameters result in quicker training. However, some time sensitive hyper-parameters influence the training of better models. Therefore, we analyze training time versus Elo rating of the hyper-parameters, to achieve the best training performance for a fixed time budget.

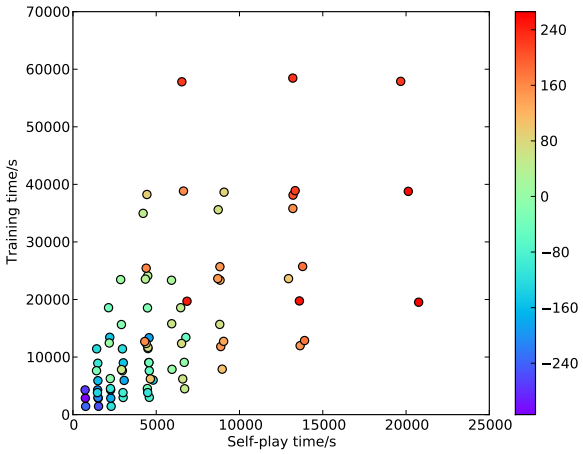


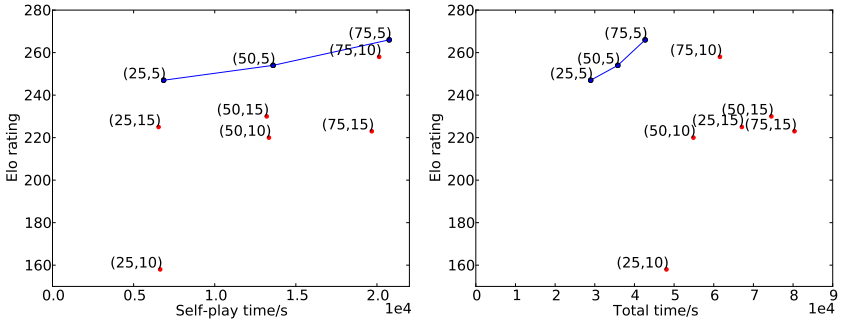

Fig. 6. Elo ratings of the final best players with different time cost of Self-play and neural network training (same base data as in Fig. 5)

In order to find a way to assess the relationship between time cost and Elo ratings, we categorize the time cost into two parts, one part is the self-play (stage 1 in Algorithm 4.1, iterations and episodes) time cost, the other is the training part (stage 2 in Algorithm 4.1, training epochs). In general, spending more time in training and in self-play gives higher Elo. In self-play time cost, there is also an other interesting variable, searching time cost, which is influenced by the value of $m$.

In Fig. 6 we also find high Elo points closer to the origin, confirming that high Elo combinations of low self-play time and low training time exist, as was indicated above, by choosing low epoch $ep$ and simulation $m$ values, since the outer iterations already imply adequate training and simulation.

In order to further analyze the influence of self-play and training on time, we present in Fig. 7(a) the full-tournament Elo ratings of the lower right panel in Fig. 5. The blue line indicates the Pareto front of these combinations. We find that low epoch values achieves the highest Elo in a high iteration training session: more outer

self-play iterations implies more training epochs, and the data generated is more diverse such that training reaches more efficient stable state (no overfitting).




(a) Self-play time vs Elo (b) Total time vs Elo

Fig. 7. Elo ratings of final best players to self-play, training and total time cost while $I$=75 and $E$=30.The values of tuple ($m$, $ep$) are given in the figures for every data point. In long total training, for $m$, larger values cost more time and generally improve the playing strength. For $ep$, more training within one iteration does not show improvement for Elo ratings. The lines indicate the Pareto fronts of Elo rating vs. time.

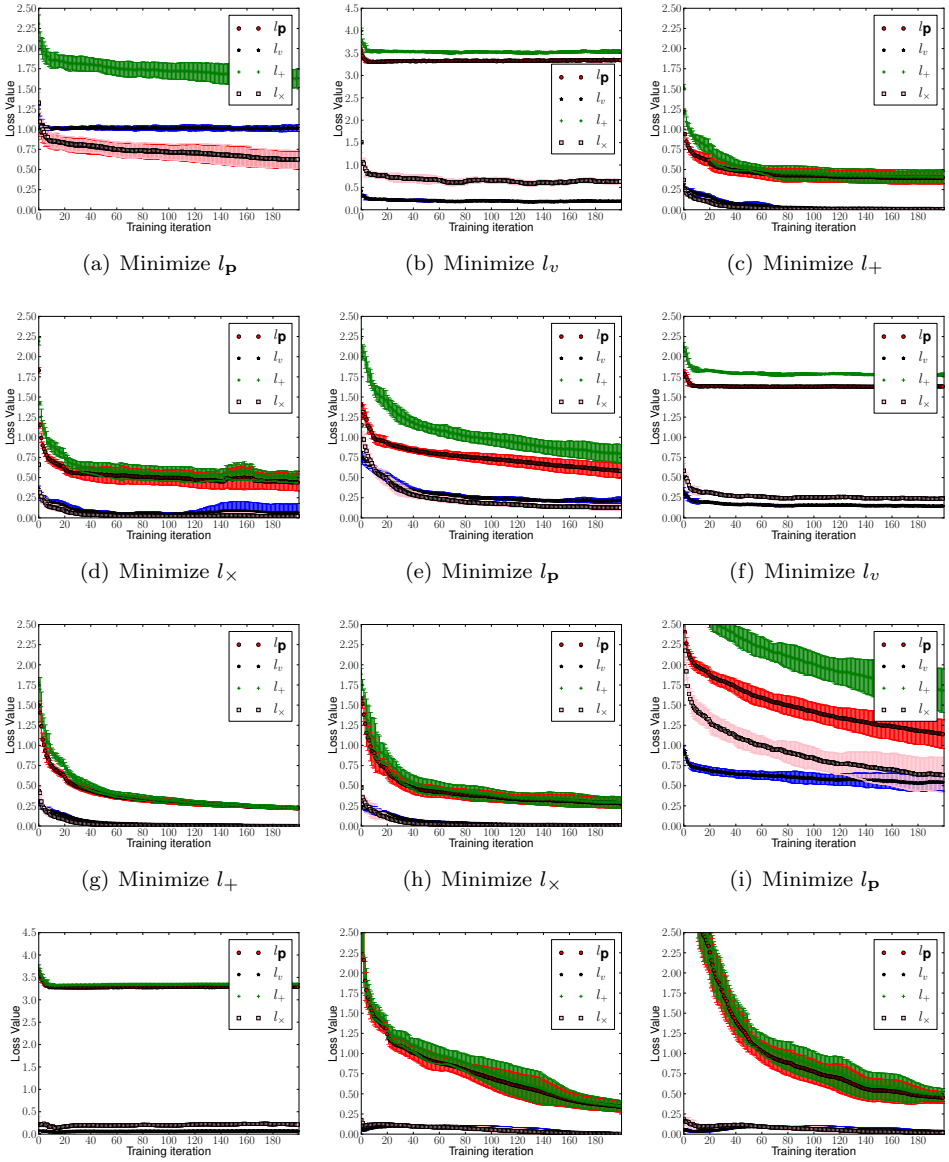
### 6.3. *Alternative loss function results*

In the following, we present the results of different loss functions. We have measured individual value loss, individual policy loss, the sum of the two, and the product of the two, for the three games. We report training loss, the training Elo rating and the tournament Elo rating of the final best players. Error bars indicate standard deviation of 8 runs.

#### 6.3.1. *Training loss*

We first show the training losses in every iteration with one minimization task per diagram, hence we need four of these per game. In these graphs we see what minimizing for a specific target actually means for the other loss types.

For 5×5 Othello, from Fig. 8(a), we find that when minimizing $l_{\mathbf{p}}$ only, the loss decreases significantly to about 0.6 at the end of each training, whereas $l_v$ stagnates at 1.0 after 10 iterations. Minimizing only $l_v$ (Fig. 8(b)) brings it down from 0.5 to 0.2, but $l_{\mathbf{p}}$ remains stable at a high level. In Fig. 8(c), we see that when the $l_+$ is minimized, both losses are reduced significantly. The $l_{\mathbf{p}}$ decreases from about 1.2 to 0.5, $l_v$ surprisingly decreases to 0. Fig. 8(d), it is similar to Fig. 8(c), while the $l_\times$ is minimized, the $l_{\mathbf{p}}$ and $l_v$ are also reduced. The $l_{\mathbf{p}}$ decreases to 0.5, the $l_v$ also surprisingly decreases to about 0. Figures for 6×6 Othello are not shown since they are very similar to 5×5 (for the same reason we do not show loss pictures for 6×6 Connect Four and 6×6 Gobang).




(a) Minimize $l_{\mathbf{p}}$ (b) Minimize $l_v$ (c) Minimize $l_+$

(d) Minimize $l_\times$ (e) Minimize $l_{\mathbf{p}}$ (f) Minimize $l_v$

(g) Minimize $l_+$ (h) Minimize $l_\times$ (i) Minimize $l_{\mathbf{p}}$

(j) Minimize $l_v$ (k) Minimize $l_+$ (l) Minimize $l_\times$

Fig. 8. Training losses for minimizing different targets in 5×5 Othello (Fig. 8(a) to Fig. 8(d)), 5×5 Connect Four (Fig. 8(e) to Fig. 8(h)) and 5×5 Gobang (Fig. 8(i) to Fig. 8(l)) averaged over 8 runs. All measured losses are shown, but only one of these is minimized for (The caption of each subfigure indicates the minimized target). Note the different scaling for subfigure (b) and (j). In most cases, the target that is minimized for is also the lowest.

For 5×5 Connect Four (see Fig. 8(e)), we find that when only minimizing $l_{\mathbf{p}}$, it significantly reduces from 1.4 to about 0.6, whereas $l_v$ is minimized much quicker from 1.0 to about 0.2, where it is almost stationary. Minimizing $l_v$ (Fig. 8(f)) leads to some reduction from more than 0.5 to about 0.15, but $l_{\mathbf{p}}$ is not moving much after an initial slight decrease to about 1.6. For minimizing the $l_+$ (Fig. 8(g)) and the $l_\times$ (Fig. 8(h)), the behavior of $l_{\mathbf{p}}$ and $l_v$ is very similar, they both decrease steadily, until $l_v$ surprisingly reaches 0. Of course the $l_+$ and the $l_\times$ arrive at different values, but in terms of both $l_{\mathbf{p}}$ and $l_v$ they are not different.

For 5×5 Gobang game, we find that, in Fig. 8, when only minimizing $l_{\mathbf{p}}$, $l_{\mathbf{p}}$ value decreases from around 2.5 to about 1.25 while the $l_v$ value reduces from 1.0 to 0.5 (see Fig. 8(i)). When minimizing $l_v$, $l_v$ value quickly reduces to a very level which is lower than 0.1 (see Fig. 8(j)). Minimizing $l_+$ and $l_\times$ both lead to stationary low $l_v$ values from the beginning of training which is different from Othello and Connect Four.

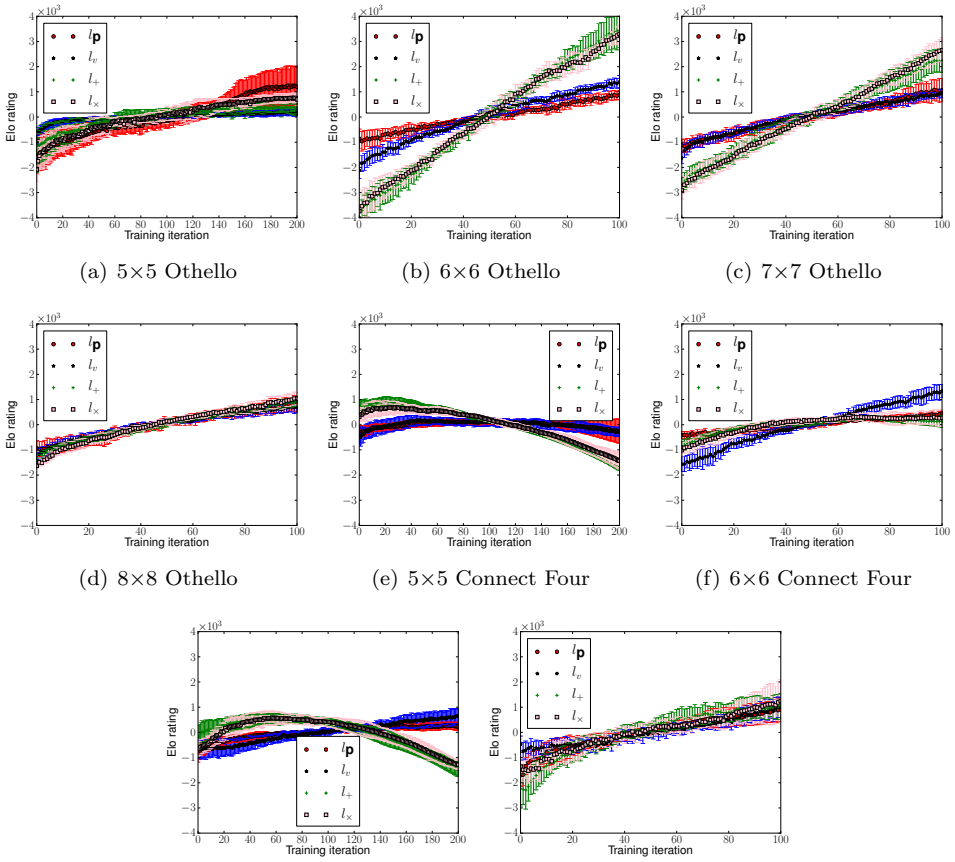
#### 6.3.2. *Training Elo rating*

Following the AlphaGo papers, we also investigate the training Elo rating of every iteration during training. Instead of showing results from single runs, we provide means and variances for 8 runs for each target, categorized by different games in Fig. 9.

From Fig. 9(a) (small 5×5 Othello) we see that for all minimization tasks, Elo values steadily improve, while they raise fastest for $l_{\mathbf{p}}$. In Fig. 9(b), we find that for 6×6 Othello version, Elo values also always improve, but much faster for the $l_+$ and $l_\times$ target, compared to the single loss targets, also for 7×7 Othello (Fig. 9(c)). But for 8×8 Othello, the improving trend becomes gentle (Fig. 9(d)).

Fig. 9(e) and Fig. 9(f) show the Elo rate progression for training players with the four different targets on the small and larger Connect Four setting. This looks a bit different from the Othello results, as we find stagnation (for 6×6 Connect Four) as well as even degeneration (for 5×5 Connect Four), which can be also found in Fig. 9(g) and Fig. 9(h) for Gobang. The latter actually means that for decreasing loss in the training phase, we achieve decreasing Elo rates, such that the players get weaker and not stronger. In the larger Connect Four (and Gobang) setting, we still have a clear improvement, especially if we minimize for $l_v$. Minimizing for $l_{\mathbf{p}}$ leads to stagnation quickly, or at least a very slow improvement.

Overall, we display the Elo progression obtained from the different minimization targets for one game together. However, one must be aware that their numbers are not directly comparable due to the high self-play bias (as they stem from players who have never seen each other). Nevertheless, the trends are important, and it is especially interesting to see if Elo values correlate with the progression of losses. Based on the experimental results, we can conclude that the training Elo rating is certainly good for assessing if training actually works, whereas the losses alone do not always show that. We may even experience contradicting outcomes as stagnating




(a) 5×5 Othello (b) 6×6 Othello (c) 7×7 Othello

(d) 8×8 Othello (e) 5×5 Connect Four (f) 6×6 Connect Four

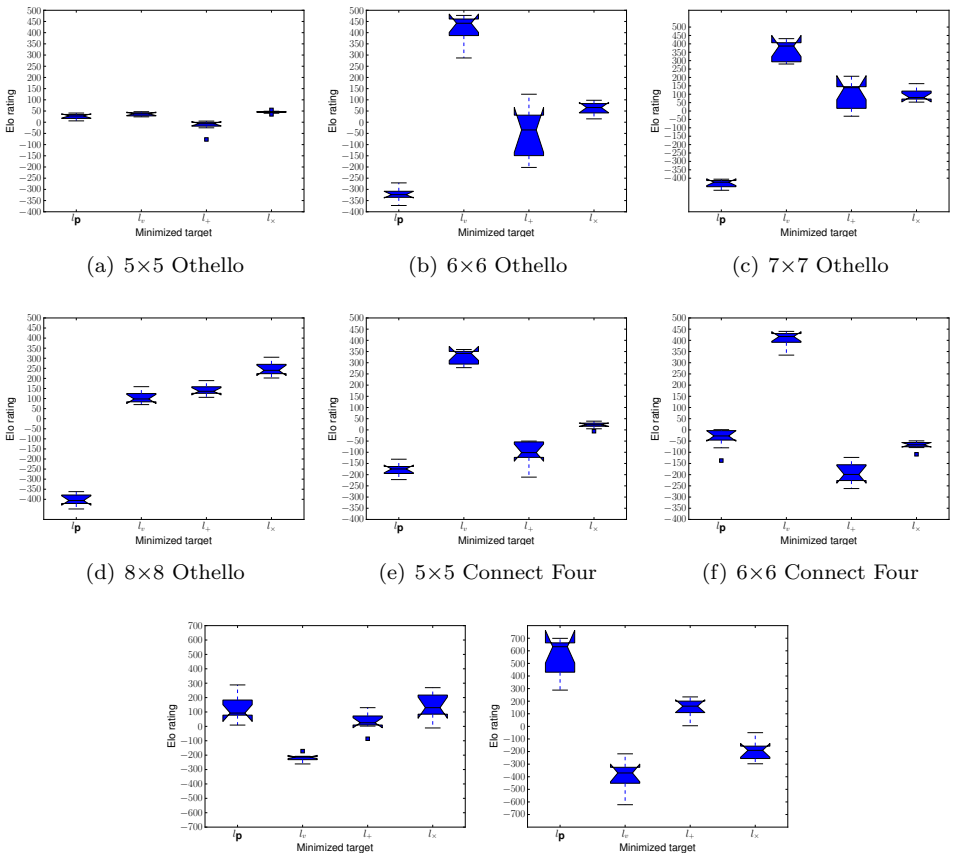
(g) 5×5 Gobang (h) 6×6 Gobang

Fig. 9. The whole history Elo rating at each iteration during training for different games, aggregated from 8 runs. The training Elo for $l_+$ and $l_\times$ in panel b and c for example shows inconsistent results

losses and rising Elo ratings (for the big Othello setting and $l_v$) or completely counterintuitive results as for the small Connect Four setting where Elo ratings and losses are partly anti-correlated. We have experimental evidence for the fact that training losses and Elo ratings are by no means exchangeable as they can provide very different impressions of what is actually happening.

### 6.3.3. *The final best player tournament Elo rating*

In order to measure which target can achieve better playing strength, we let all final models trained from 8 runs and 4 targets plus a random player pit against each other for 20 times in a full round robin tournament. This enables a direct comparison of the final outcomes of the different training processes with different targets. It is




(a) 5×5 Othello (b) 6×6 Othello (c) 7×7 Othello

(d) 8×8 Othello (e) 5×5 Connect Four (f) 6×6 Connect Four

(g) 5×5 Gobang (h) 6×6 Gobang

Fig. 10. Round-robin tournament of all final models from minimizing different targets. For each game 8 final models from 4 different targets plus a random player (i.e. 33 in total). In panel (a) the difference is small. In panel b, c, and d, the Elo rating of $l_v$ minimized players clearly dominates. However, in panel (f), the Elo rating of $l_{\mathbf{p}}$ minimized players clearly achieve the best performance.

thus more informative than the training Elo due to the self-play bias, but provides no information during the self-play training process. In principle, it is possible to do this also during the training at certain iterations, but this is computationally very expensive.

The results are presented in Fig. 10. and show that minimizing $l_v$ achieves the highest Elo rating with small variance for 6×6 Othello, 7×7 Othello, 5×5 Connect Four and 6×6 Connect Four. For 5×5 Othello, with 200 training iterations, the difference between the results is small. We therefore presume that minimizing $l_v$ is the best choice for the games we focus on. This is surprising because we expected the $l_+$ to perform best as documented in the literature. However, this may apply

to smaller games only. Since we see that (1) for 7×7 Othello, $l_v$ is still the best, but comparing with 8×8 Othello, the relative Elo rating of $l_v$ reduced. And (2) for 8×8 Othello (in Fig 10(d)), $l_v$ continues to reduce and $l_\times$ becomes the best. More clearly, based on the Fig. 10(a) to Fig. 10(d), an independent comparison among different board sizes for Othello is shown in Fig 11. Note that 5×5 Othello already seems to be a border case where overfitting levels out all differences.

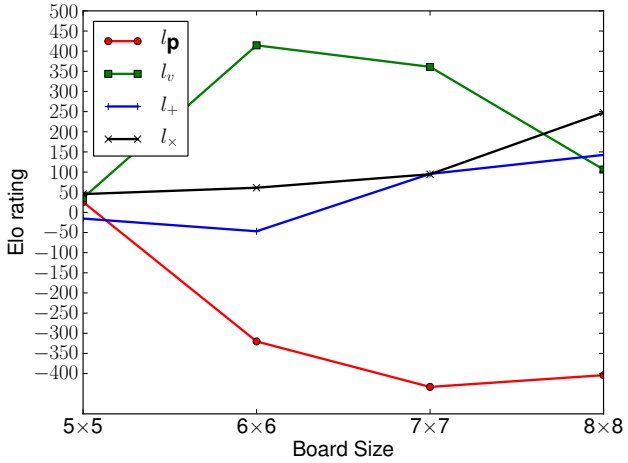



Fig. 11. Average Elo ratings of the final best players with different board size for Othello (same base data as in Fig. 10(a) to Fig. 10(d)). The data points are connected as lines to show trends.

In conclusion, we find that minimizing $l_v$ only is an alternative to the $l_+$ target for certain small cases. And for larger cases, $l_\times$ could be the other alternative. We also report exceptions, especially in relation to the Elo rating as calculated during training. The relation between Elo and loss during training is sometimes inconsistent (5×5 Connect Four training shows Elo decreasing while the losses are actually minimized) due to training bias. And for Gobang game, only minimizing $l_\mathbf{p}$ is the best alternative. A combination achieves lowest loss, but $l_v$ achieves the highest training Elo. If we minimize product loss $l_\times$, this can result in higher Elo rating for certain games. More research into training bias is needed.

## 7. Conclusion

AlphaGo has taken reinforcement learning by storm. The performance of the novel approach to self-play is stunning, yet the computational demands are high, prohibiting the wider applicability of this method. Little is known about the impact of the values of the many hyper-parameters on the speed and quality of learning. In this work, we analyze important hyper-parameters and combinations of loss-functions.

We gain more insight and find recommendations for faster and better self-play. We have used small games to allow us to perform a thorough sweep using a large number of hyper-parameters, within a reasonable computational budget. We sweep 12 parameters in AlphaZeroGeneral [6] and analyze loss and time cost for 6×6 Othello, and select the 4 most promising parameters for further optimization.

We evaluate the interaction between these four time-related hyper-parameters more thoroughly. Thereby we find that: i) generally, higher values lead to higher playing strength; ii) within a limited budget, a higher number of the outer self-play iterations is more promising than higher numbers of the inner training epochs, search simulations, and game episodes. At first, this is a surprising result since conventional wisdom tells us that deep learning networks should be trained well, and MCTS needs many play-out simulations to find good training targets.

In AlphaZero-like self-play, the outer-iterations subsume the inner training and search. Performing more outer iterations automatically implies that more inner training and search are performed. The training and search improvements carry over from one self-play iteration to the next, and long self-play sessions with many iterations can get by with surprisingly few inner training epochs and MCTS simulations. The sample efficiency of self-play is higher than the simple composition of the constituent elements would predict. Also, the implied high number of training epochs may cause overfitting, to be reduced by small values for epochs.

Much research in self-play uses the default loss function (sum of value and policy loss). More research is needed into the relative importance of value function and policy function. We evaluate four alternative loss functions for three games and two board sizes, and find that the best setting depends on the game and is usually not the sum of policy and value loss, but simply the value loss. However, the sum may be a good compromise.

Finally, the experiments show that care must be taken in computing Elo ratings. Computing Elo based on game-play results during training typically gives biased results that differ greatly from tournaments between multiple opponents. Moreover, we pointed out that the Elo calculation based on the final best models tournament should be used.

For future work, more insight into training bias is needed. Also, automatic optimization frameworks can be explored [38,39]. Also, reproducibility studies should be performed to see how our results carry over to larger games (like Go), computational load permitting. Given that Chen et al. [17] tuned some MCTS-related parameters (like exploration and exploitation balancing which we also adopt as parameter $c$) in AlphaGo with Bayesian optimization, resulting in Elo improvements, which evidenced our findings in self-play. However, Chen et al. [17] did not directly study the parameters in neural network training, we believe our work provide insightful analysis for future work on larger games.

## References

1. Silver D, Huang A, Maddison C J, et al: Mastering the game of Go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016)
2. Silver D, Schrittwieser J, Simonyan K, et al: Mastering the game of go without human knowledge. Nature **550**(7676), 354–359 (2017)
3. Silver D, Hubert T, Schrittwieser J, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science, 2018, 362(6419): 1140-1144.
4. Tao J, Wu L, Hu X: Principle Analysis on AlphaGo and Perspective in Military Application of Artificial Intelligence. Journal of Command and Control **2**(2), 114–120 (2016)
5. Zhang Z: When doctors meet with AlphaGo: potential application of machine learning to clinical medicine. Annals of translational medicine **4**(6), (2016)
6. N. Surag, https://github.com/suragnair/alpha-zero-general, 2018.
7. Wang H, Emmerich M, Plaat A. Monte Carlo Q-learning for General Game Playing. arXiv preprint arXiv:1802.05944 (2018)
8. Browne C B, Powley E, Whitehouse D, et al: A survey of monte carlo tree search methods. IEEE Transactions on Computational Intelligence and AI in games **4**(1), 1–43 (2012)
9. B Ruijl, J Vermaseren, A Plaat, J Herik: Combining Simulated Annealing and Monte Carlo Tree Search for Expression Simplification. In: Béatrice Duval, H. Jaap van den Herik, Stéphane Loiseau, Joaquim Filipe. Proceedings of the 6th International Conference on Agents and Artificial Intelligence 2014, vol. 1, pp. 724–731. SciTePress, Setúbal, Portugal (2014)
10. Schmidhuber J: Deep learning in neural networks: An overview. Neural networks **61** 85–117 (2015)
11. Clark C, Storkey A. Training deep convolutional neural networks to play go. International Conference on Machine Learning. pp. 1766–1774 (2015)
12. Tesauro, Gerald. Temporal difference learning and TD-Gammon. Communications of the ACM **38**(3) 58–68 (1995).
13. Heinz E A: New self-play results in computer chess. International Conference on Computers and Games. Springer, Berlin, Heidelberg. pp. 262–276 (2000)
14. Wiering M A: Self-Play and Using an Expert to Learn to Play Backgammon with Temporal Difference Learning. Journal of Intelligent Learning Systems and Applications **2**(2), 57–68 (2010)
15. Van Der Ree M, Wiering M: Reinforcement learning in the game of Othello: Learning against a fixed opponent and learning from self-play. In Adaptive Dynamic Programming And Reinforcement Learning. pp. 108–115 (2013)
16. Aske Plaat, Learning to Play: Reinforcement Learning and Games, Leiden, 2020, forthcoming.
17. Chen, Yutian, Aja Huang, Ziyu Wang, Ioannis Antonoglou, Julian Schrittwieser, David Silver, and Nando de Freitas. "Bayesian optimization in alphago." arXiv preprint arXiv:1812.06855 (2018).
18. Mandai Y, Kaneko T. Alternative Multitask Training for Evaluation Functions in Game of Go. 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI). IEEE, 2018: 132-135.
19. Caruana R. Multitask learning. Machine learning, 1997, 28(1): 41-75.
20. Matsuzaki K, Kitamura N. Do evaluation functions really improve Monte-Carlo tree search?. ICGA Journal, 2018 (Preprint): 1-11
21. Matsuzaki K. Empirical Analysis of PUCT Algorithm with Evaluation Functions of Different Quality. 2018 Conference on Technologies and Applications of Artificial In-

telligence (TAAI). IEEE, 2018: 142-147.
22. Iwata S, Kasai T. The Othello game on an n×n board is PSPACE-complete. Theoretical Computer Science. **123**(2), 329–340 (1994)
23. Allis V. A knowledge-based approach of Connect-Four-the game is solved: White wins. 1988.
24. Reisch, S. Gobang ist PSPACE-vollständig. Acta Informatica 13, 59–66 (1980).
25. Wang H, Emmerich M, Preuss M and Plaat A. Alternative Loss Functions in AlphaZero-like Self-play. 2019 IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China, 155–162 (2019).
26. Buro M. The Othello match of the year: Takeshi Murakami vs. Logistello. ICGA Journal, 1997, 20(3): 189-193.
27. Chong S Y, Tan M K, White J D. Observing the evolution of neural networks learning to play the game of Othello. IEEE Transactions on Evolutionary Computation, 2005, 9(3): 240-251.
28. Thill M, Bagheri S, Koch P, et al. Temporal difference learning with eligibility traces for the game connect four. 2014 IEEE Conference on Computational Intelligence and Games. IEEE, 2014: 1-8.
29. Zhang M L, Wu J, Li F Z. Design of evaluation-function for computer Gobang game system. Journal of Computer Applications, 2012, 7: 051.
30. Banerjee B, Stone P. General Game Learning Using Knowledge Transfer. IJCAI. 2007: 672-677.
31. Wang H., Emmerich M., Plaat A. (2019) Assessing the Potential of Classical Q-learning in General Game Playing. In: Atzmueller M., Duivesteijn W. (eds) Artificial Intelligence. BNAIC 2018. Communications in Computer and Information Science, vol 1021. Springer, Cham.
32. Ioffe S, Szegedy C: Batch normalization: accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37. pp. 448–456 (2015)
33. Kingma D P, Ba J: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
34. Srivastava N, Hinton G, Krizhevsky A, et al: Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research. **15**(1), 1929–1958 (2014)
35. Coulom R. Whole-history rating: A Bayesian rating system for players of time-varying strength. International Conference on Computers and Games. Springer, Berlin, Heidelberg, 113–124, 2008
36. Emmerich M T M, Deutz A H. A tutorial on multiobjective optimization: fundamentals and evolutionary methods. Natural computing, 2018, 17(3): 585-609.
37. Wang H, Emmerich M, Preuss M, Plaat A. Analysis of Hyper-Parameters for Small Games: Iterations or Epochs in Self-Play?. arXiv preprint arXiv:2003.05988 (2020)
38. Birattari M, Stützle T, Paquete L, et al. A racing algorithm for configuring metaheuristics. Proceedings of the 4th Annual Conference on Genetic and Evolutionary Computation. Morgan Kaufmann Publishers Inc. 11-18 (2002)
39. Hutter F, Hoos H H, Leyton-Brown K: Sequential model-based optimization for general algorithm configuration. International Conference on Learning and Intelligent Optimization. Springer, Berlin, Heidelberg, pp. 507–523 (2011)