

Tile Pattern KL-Divergence for Analysing and Evolving Game Levels

Simon M. Lucas

Queen Mary University of London, UK
simon.lucas@qmul.ac.uk

Vanessa Volz

Queen Mary University of London, UK
v.volz@qmul.ac.uk

ABSTRACT

This paper provides a detailed investigation of using the Kullback-Leibler (KL) Divergence as a way to compare and analyse game-levels, and hence to use the measure as the objective function of an evolutionary algorithm to evolve new levels. We describe the benefits of its asymmetry for level analysis and demonstrate how (not surprisingly) the quality of the results depends on the features used. Here we use tile-patterns of various sizes as features.

When using the measure for evolution-based level generation, we demonstrate that the choice of variation operator is critical in order to provide an efficient search process, and introduce a novel convolutional mutation operator to facilitate this. We compare the results with alternative generators, including evolving in the latent space of generative adversarial networks, and Wave Function Collapse. The results clearly show the proposed method to provide competitive performance, providing reasonable quality results with very fast training and reasonably fast generation.

CCS CONCEPTS

• **Mathematics of computing** → Probabilistic representations;
• **Applied computing** → Computer games; • **Theory of computation** → Evolutionary algorithms;

KEYWORDS

Kullback-Leibler Divergence, Procedural Content Generation, Latent Vector Evolution, Wave Function Collapse

ACM Reference Format:

Simon M. Lucas and Vanessa Volz. 2019. Tile Pattern KL-Divergence for Analysing and Evolving Game Levels. In *Genetic and Evolutionary Computation Conference (GECCO '19)*, July 13–17, 2019, Prague, Czech Republic. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3321707.3321781>

1 INTRODUCTION

Procedural Content Generation (PCG) uses algorithms that to generate content, specifically in the context of games and digital entertainment. The type of content varies widely, including levels, landscapes, narratives and weapons.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

GECCO '19, July 13–17, 2019, Prague, Czech Republic

© 2019 Copyright held by the owner/author(s). Publication rights licensed to the Association for Computing Machinery.

ACM ISBN 978-1-4503-6111-8/19/07...\$15.00

<https://doi.org/10.1145/3321707.3321781>

A popular subset of PCG techniques is based on searching a space of potential content in order to identify high quality examples. Search-based PCG approaches are especially useful if it is difficult to find an approach that only generates content that satisfies certain constraints, such as e.g. solvability of generated puzzles. In order to apply a search-based approach, three components are needed [12]:

- A searchable space where each point represents a specific instance of generated content;
- A search algorithm;
- A fitness function to guide the search.

However, specifying a suitable fitness function is a difficult problem due to the subjective nature of preferences in games and game design. The functions used in successful PCG approaches are usually very specific to the content they are designed to evaluate (see [2] for an example). The ideal measure should correlate with human preferences in some way, and be fast to compute in order to provide efficient analysis. Any such measure also has the potential to be used as the fitness function for an evolutionary algorithm and hence to be used to generate levels.

A common usecase of search-based PCG is the generation of novel content that is still similar in various aspects to pre-existing, manually designed examples. This usecase is a large part of the motivation behind PCGML approaches (PCG via Machine Learning) [9], which revolves around the identification of patterns in data sets. Level generation is a typical example of this usecase. Here, a fitness measures generally express similarity to existing content.

In this paper, we propose to use the Kullback-Leibler Divergence (D_{KL}) between generated content and training samples as a measure for similarity in PCGML. The D_{KL} is an asymmetric measure, which enables us to measure and control the degree of novelty in the generated content. We demonstrate how this can be used to gain detailed insights into the behaviour of a given PCG method using Mario level generation as an illustrative example.

After a survey of related work on search-based PCG in section 2, we formally introduce the D_{KL} measure and its adaptation in the context of Mario levels in section 3. We show that the D_{KL} correlates at least in some way with human intuition of level similarity in section 3.2, thus validating the measure for the purposes of this paper. Following that, in section 4 we present an efficient level generation method called ETPKLDiv (Evolution using Tile-Pattern KLDiv) based around D_{KL} , which successfully creates Mario levels in the style of existing ones. We compare the levels with others generated from Wave Function Collapse (WFC, see section 2.2) and Latent Vector Evolution (MarioGAN, see section 2.3), two state-of-the-art PCG methods. The evaluation of the results can be found in section 4.5. Finally, we summarise our results and give an outlook on future work in section 5.

We find that ETPKLDiv offers competitive performance in terms of speed of training, speed of generation and generalisation from small samples. Solution quality is harder to assess but the method is able to evolve reasonable looking levels that often satisfy the objective function better snippets of the training level.

2 SEARCH-BASED PROCEDURAL CONTENT GENERATION

Although there has been significant work on many aspects of PCG, the viral success of Wave Function Collapse¹ [5] demonstrates the appetite for PCG which can rapidly generate interesting content from small amounts of training data.

When dealing with search-based PCG there are several (not mutually exclusive) approaches that can be applied to evaluate candidate solutions as identified in various taxonomies (e.g. [11]):

- Use some hand-designed criteria. These tend to be specific to a game or game genre;
- Use agent-based play testing. This approach has become more attractive with the development of general game AI based on statistical forward planning algorithms such as Monte Carlo Tree Search [3] and Rolling Horizon Evolution [7];
- Use game-level samples (also referred to as Procedural Content Generation via Machine Learning (PCGML) [9].

In this paper, we are specifically targeting PCGML.

Several methods have been proposed in order to evaluate content generators (see overview in [8]). They mostly revolve around the characterisation of generated content according to features such as novelty and difficulty. Some of these methods are more explorative (e.g. *expressive range*), where others are mainly focused on optimisation, as in this paper.

There is a wide range of quality indicators, but for the PCGML usecase targeted in this paper, we specify the following objectives:











- The quality of the generated content. Some aspects of quality can be quantified while others are inherently subjective;
- The amount of sample data needed for adequate performance;
- The flexibility of the generator: e.g. can it directly produce levels of any size? Can it produce consistent results for different types of content?
- The time taken to train the system;
- The time taken to generate each new level once trained.

For the purpose of demonstration in this paper, we use the generation of Mario levels as an illustrative example. We introduce this application in section 2.1, along with corresponding related work. Afterwards, we present two state-of-the-art PCG approaches in more detail, as their results are used as a comparison for the algorithm presented in this paper. We describe Wave Function Collapse in section 2.2 and MarioGAN in section 2.3.

2.1 PCGML Mario Level Generation

Super Mario Bros. is a typical platformer game, where the main character, Mario, needs to traverse multiple different levels and defeat enemies to win the game. Mario levels are usually indicated

Table 1: Tile types in Mario levels. The symbol characters from the VGLC encoding, and the numeric identities that are mapped to the corresponding tile types as implemented in the Mario AI framework to produce the visualization shown. Table reproduced from [13]

Tile type	Symbol	Identity	Visualization
Solid/Ground	X	0	
Breakable	S	1	
Empty (passable)	-	2	
Full question block	?	3	
Empty question block	Q	4	
Enemy	E	5	
Top-left pipe	<	6	
Top-right pipe	>	7	
Left pipe	[8	
Right pipe]	9	
Coin	o	10	

to belong to one of five types (overworld, underground, athletic, castle, underwater)². The type determines the aesthetic style of the level, but also affects the challenges that are posed. For example, athletic levels usually contain platforms that are spaced far apart, as well as moving platforms. They thus require precise and timed jumps to traverse.

In Super Mario Bros., the levels are also grouped into different *worlds*. Levels are named according to which world they belong to (1-8) and based on their order within the world (1-4). For example, the first level in the first world is called 1-1 (depicted in figure 1). As can be seen in the image, Mario levels consist of several different tiles that are arranged in a 2D grid to form structures such as platforms, stairs and pipes. The full set of tiles contained in Level 1-1 is listed in table 1.

The table also contains a list of symbols that are used to represent the respective tile type in the Video Game Level Corpus (VGLC) [10]. The VGLC contains a collection of levels from tile-based games in such an encoding. The Super Mario Bros. levels from the VGLC serve as the sample set for the illustrative usecase in this paper.

Specifically, level generators have to produce fixed-size matrices using the encoding from VGLC as described in table 1. In order to keep consistent with existing work on PCGML Mario Level generation, only level 1-1 is used as a training set for the generators. Since PCGML aims to learn from the training data, the generated levels should be similar to the original. Just how similar they should be is a moot point, since the aim is to generate content which is in some way novel while looking and playing like the training sample. Here, we use KL-Divergence to measure similarity, as described in detail in section 3. We validate this decision in section 3.2.

2.2 Wave Function Collapse

Wave Function Collapse is a PCG approach that has gained recent prominence, especially from its application in popular games such as Bad North³ (Raw Fury, 2018). The approach relies on fitting patterns together such that a set of constraints are fulfilled [5].

¹<https://github.com/mxgmn/WaveFunctionCollapse>

²https://www.mariowiki.com/Super_Mario_Bros.#List_of_levels

³<https://www.badnorth.com/>



Figure 1: Super Mario Bros. Level 1-1

These constraints are in the format of which patterns fit next to each other given an offset. Constraints are either handcrafted, or learned from a set of training samples.

After identification of patterns and constraints (1), all valid patterns that are still eligible for placement can thus be identified for each open spot (2). Next, the open spot with the minimal entropy of eligible patterns is identified (3) and filled using a pattern randomly selected from the set of available patterns for this particular spot (4). The random selection is often biased by the frequency of the patterns in the training sample. Steps 2-4 are repeated until either all spots are filled, or until the constraints contradict.

The description of the algorithm above as well as its implementation for this paper are based on [5].

2.3 MarioGAN

Another recently proposed PCG approach applies the concept of Generative Adversarial Networks (GANs) to creating content for games. The approach further employs a latent vector search and is thus called Exploratory Latent Search GAN (ELSGAN) within the context of this paper. It was applied to Mario levels in [13].

The generator in ELSGAN is a multi-layer neural network that, based on an input vector $[-1, 1]^{32} \in \mathbb{R}$, outputs several real-valued vectors that can be translated into Mario levels using a one-hot encoding. The generator is trained using an adversarial neural network, often dubbed the discriminator [1]. During training, the discriminator is tasked with identifying whether levels presented to it are created by the generator, or are part of the set of original training samples. The discriminator then receives feedback based on the error rate in this classification task. In contrast, the generator is rewarded if it is able to "fool" the generator. Ideally, after many iterations of alternating training of generator and discriminator, the system converges at a point where the generator produces outputs that are indistinguishable from the training samples.

3 KL-DIVERGENCE AS FITNESS MEASURE

The Kullback-Leibler Divergence D_{KL} , also called relative entropy, is the expectation of the logarithmic differences between two probability distributions P and Q where the expectation (weighted sum) is calculated using the probability distribution P .⁴

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \quad (1)$$

Compared to *ad hoc* approaches to comparing levels such as taking the absolute difference between feature histograms, the D_{KL} has two practical advantages: its asymmetry can be used to good effect, and the probabilistic nature means proper weight is given to differences in feature counts. For example, the difference between

0 and 1 occurrences is weighted much higher than the difference between 10 and 11 occurrences.

In this paper, we define D_{KL} for Mario levels (see section 2.1). In order to apply the D_{KL} , we transform levels into probability distributions over tile pattern occurrences. Given a rectangular level of size $(T_W \times T_H)$ and a rectangular filter window of size $(F_W \times F_H)$, the total number of tile patterns N is given by:

$$N = (1 + T_W - F_W)(1 + T_H - F_H) \quad (2)$$

Thus, X in equation 1 is the set of all tile patterns observed in either the set of training samples, or in the generated levels. For each experiment, we take the set of tile patterns that occur by sliding (convolving) a fixed-size window over a level, where a level is defined as a 2-d array of tiles (see table 1). The training sample shown in figure 1 has 90 distinct 2×2 patterns and 570 distinct 4×4 patterns, shown in Figures 2 and 3, respectively.

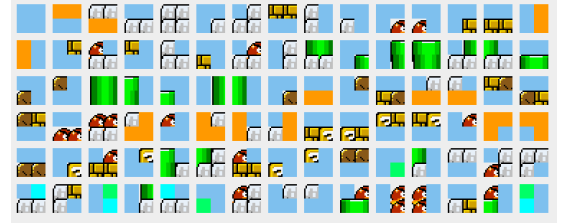


Figure 2: The 90 distinct 2×2 tile patterns in the training level, shown from most to least frequent. Top left pattern (plain blue sky) occurs 2,100 times.

We estimate the probability of a tile pattern using a frequentist approach i.e. we count the number of times pattern x occurs in a sample and divide by the total number of tile occurrences (i.e. by $|X|$). When calculating D_{KL} we ignore any patterns that occur in Q but not in P , because if $P(x) = 0$, the corresponding summand also equals 0. However, we add a small constant ϵ to each probability estimate for $Q(x)$ to avoid divide by zero errors. This also accounts for the fact that, just because a pattern was never observed, does not mean that its probability is truly zero - it just means it has not been observed yet (similar methods are used when building statistical models of natural language).

Let X_P be the set of patterns observed in probability distribution P , and $P'(x)$ and $Q'(x)$ be the epsilon-corrected probability estimates. Further, let $C(x)$ be the number of occurrences of x in a sample X and $C = \sum_{x \in X_P} C(x)$ be the sum of $C(x)$ over all $x \in X$. We then we compute $P'(x)$ as follows:

$$P'(x) = \frac{C(x) + \epsilon}{(C + \epsilon)(1 + \epsilon)} \quad (3)$$

Based on this, our safe approximation of the D_{KL} is:

$$D_{KL}(P||Q) = \sum_{x \in X_P} P'(x) \log \left(\frac{P'(x)}{Q'(x)} \right) \quad (4)$$

⁴The D_{KL} is only properly defined if $\forall x, Q(x) = 0 \implies P(x) = 0$.

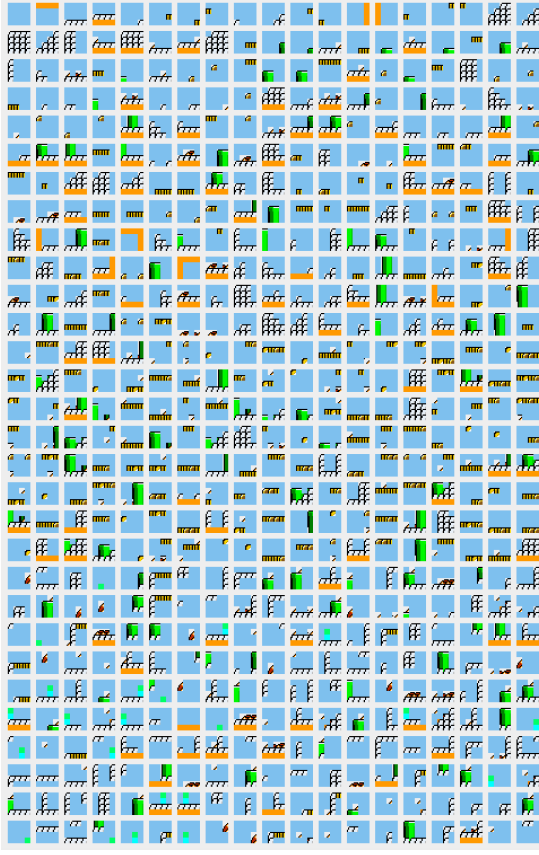


Figure 3: The 570 distinct 4x4 tile patterns in the training level, shown from most to least frequent. Top left pattern occurs 1,349 times whereas the least frequent 50% of the patterns occur only once or twice.

$D_{KL}(P||Q)$ is always greater than or equal to zero, and only zero if P is identical to Q . Since D_{KL} is an asymmetric measure, we have to decide which way around to apply it, or whether to apply it symmetrically by adding the two results together. Here we choose a weighted approach that can vary smoothly between the two extremes. We transform the resulting function into a maximisation problem by negating the D_{KL} . Hence we define the fitness of a generated level Q with respect to a set of sample levels P as:

$$F(P, Q) = -(w \cdot D_{KL}(P||Q) + (1 - w) \cdot D_{KL}(Q||P)) \quad (5)$$

3.1 Parameters

The behaviour of $F(P, Q)$ is characterised by four parameters:

- ϵ is the back-off estimation constant (see equation 3).
- Dimensions of filter: width F_w and height F_h in tiles.
- The weight w used to balance between the two asymmetric D_{KL} terms (see equation 5).

As we shall see in section 4.4, w can be used to good effect to control the degree of novelty in the generated levels. Setting $w = 0$ aims for no novelty in the tile patterns i.e. all tile patterns in the generated level should have occurred at least once in the training

sample. Note that the arrangement of tile-patterns is still likely to be novel. Satisfying this constraint may be at the expense of missing out on many patterns that did occur. On the other hand, setting $w = 1$ favours trying to include all patterns observed in the training sample in the generated level: this may be impossible if the dimensions of the generated level are smaller than the training sample, which is commonly the case.

3.2 Evaluation and Validation

The Kullback-Leibler Divergence (denoted as D_{KL} here) has been used successfully for Mario level generation in the context of MarioGAN [13]. This might not be obvious immediately, but it was shown in [4] that the training process for generative adversarial networks essentially minimises the D_{KL} between the distributions of the generator output and the original training samples. Hence, both the method used in this paper and the Mario level generation from GANs in [13] use the concept of KL-Divergence, but apply it in different ways. Here, we apply it to tile-pattern distributions, whereas GANs are computing it over individual tile-positions, but from a latent encoding that forces the tiles into structured patterns.

Below, we validate its usage as a fitness measure for generated levels. The main question is whether D_{KL} is able to correctly identify patterns in 2D tile-based levels. To this end, we conduct the following analysis using hierarchical clustering with average linkage. We test whether the D_{KL} is able to correctly identify the similarities between levels from one type and the differences to others.

We compute a similarity matrix using D_{KL} for all Super Mario Bros. levels contained in the VGLC [10]. This set of levels only contains overworld, underground and athletic levels (see section 2.1 for details). Separate experiments were conducted for D_{KL} measures based on various small filter sizes F_w and F_h between 1 and 5, and for weights $w \in \{0, 0.5, 1\}$. In all our experiments, three distinct clusters were detected, corresponding to the 3 different types of levels. A dendrogram of these results is shown in figure 4 for filter size $F_w \times F_h = 4 \times 4$. We have thus demonstrated that D_{KL} is useful to identify similarities and dissimilarities in Mario levels.

Besides this rudimentary validation of the expressiveness of the measure, a benefit of the tile-pattern D_{KL} measure is its potential for further analysis of the generator and the generated content. For example, based on the contribution of singular patterns to the D_{KL} , we are able to identify anomalies. Furthermore, the D_{KL} is able to compress information from tile-pattern occurrence histograms into a single meaningful number. By changing the filter sizes, the measure is also able to express similarity on different levels of granularities.

4 KL-DIVERGENCE GENERATOR

Using D_{KL} as the basis of a procedural level generator is conceptually simple but achieving good results requires some careful design. For the experiments in this paper, we used a Random Mutation Hill Climber (also known as a (1+1) EA). This simple algorithm has shown competitive performance across a range of problems such as evolving finite automata [6].

However, the tile-pattern D_{KL} fitness function induces a rugged fitness landscape when using a standard mutation operator because many parts of a solution have to be changed in a coordinated way

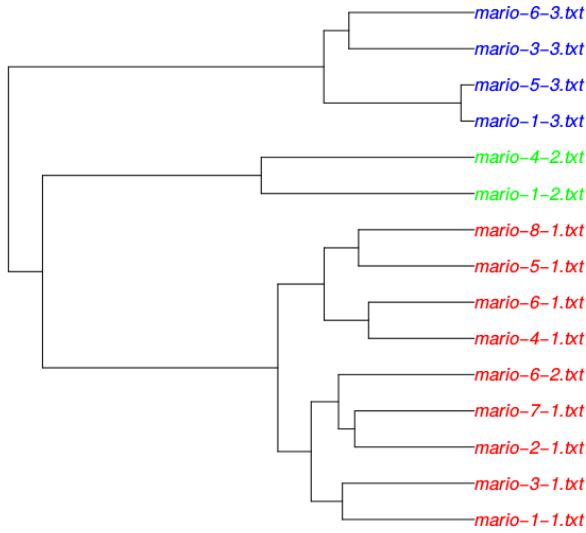


Figure 4: Dendrogram for clusters based on D_{KL} with filter $F_w \times F_h = 4 \times 4$ and weight $w = 0.5$. Colours denote level type (overworld in red, underground in green, athletic in blue).

in order to improve the solution. This problem is exacerbated for larger window sizes. In the next subsections we will first show examples evolved levels to illustrate the problem and then introduce a convolutional mutation operator that enables the $(1 + 1)$ EA to achieve acceptable performance. We judge it as acceptable in two ways: (1) the fitness values are often better than the same size snippets from the sample level, and (2) in our opinion they look like reasonable Mario levels.

4.1 Standard Mutation Operator

Figure 5 shows a level evolved using a standard “flip” mutation operator. This operator works by scanning each tile in the level, changing it to a randomly selected tile with a given (typically small) probability. We set this to on average flip 3 tiles per application of the operator. A typical result of evolving using a 2×2 filter window for 100,000 fitness evaluations is shown in Figure 5. Clearly, progress has been made compared to a uniform random level, but there are still many anomalies compared to the sample level. The regular mutation operator fails badly with larger windows such as 3×3 or bigger. Figure 6 shows an example using a 4×4 filter which fails to improve over a random point in the search space after 100,000 iterations (the figure is therefore an example of a random level).

4.2 Convolutional Mutation Operator

The problem with the standard mutation operator is that it is unlikely that any number of randomly flipped tiles will lead to an improvement in fitness. For example, improving the fitness will often involve modifying a sub-rectangle of the level so that non-matching windows will be transformed into ones that do match at least one pattern observed in the training sample. Typically, several tiles may need to be flipped in order to make a single additional match. However, many changes may also destroy existing matches with overlapping windows which further compounds the problem.



Figure 5: Sample level generated after evolving for 10,000 fitness evaluations using a 2×2 filter and a standard mutation operator.

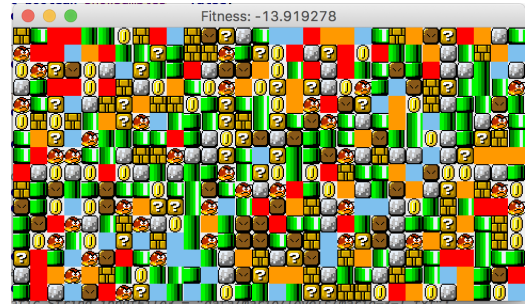


Figure 6: Sample level generated after evolving for 10,000 fitness evaluations using a 4×4 filter and a standard mutation operator. The search gets nowhere and this solution has similar fitness to a randomly generated level.

To counter this, we designed a novel mutation operator which is both simple and effective. The operator samples a filter-size rectangle from the training set and copies it to a random location in the generated level. This now has a much higher chance of leading to an improvement in fitness.

Indeed, experiments show that using this mutation operator we can evolve fitter levels of a particular size (e.g. of width 30 tiles) than are present in the training data (as measured by fitness function F). This is because the training tile distribution is drawn from a single wide level that progresses through various phases, each of which tends to favour or exclude particular pattern types. The optimiser then tries to squeeze them all into the available space. If desired, similar levels of tile densities can be achieved by simply increasing the width of levels to be generated, though larger levels lead to slower convergence.

Figures 7 and 8 show levels evolved using 2×2 and 4×4 convolutional mutation operators. The 2×2 example almost looks like a reasonable level but has some anomalies, whereas the 4×4 example is a fine looking level.

The evolutionary generator takes just under 1 second for 10,000 fitness evaluations using a 2×2 filter and just under 2 seconds when using a 4×4 filter.

Figure 9 shows typical evolutionary traces using different filter size and mutation operator combinations, each time running the $(1 + 1)$ EA for 10,000 fitness evaluations. The graph shows the fitness F of each sample for each combination. Larger filter sizes



Figure 7: Sample level generated after evolving for 10,000 fitness evaluations using a 2×2 filter using a convolutional mutation operator.



Figure 8: Sample level generated after evolving for 10,000 iterations using a 4×4 filter using a convolutional mutation operator. Note how this is more similar to the training sample now, and yet still has some novelty.

lead to more constraints due to the greater overlap of each filter position. Evolving with the smaller filter size therefore leads to more rapid improvements and better final fitness, at the expense of producing less satisfactory levels with more anomalies.

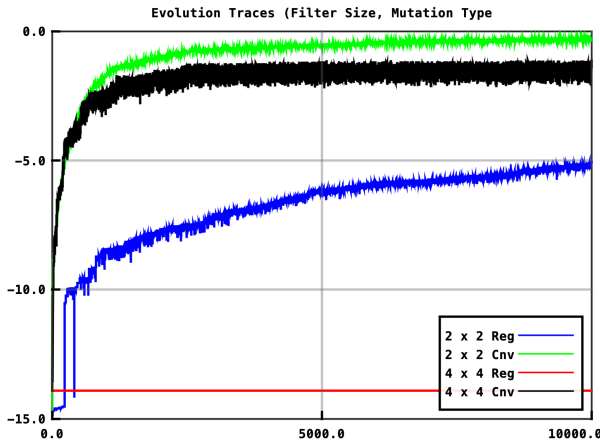


Figure 9: Evolutionary traces (plots of the fitness for each sampled candidate for evolutionary runs of 10,000 evaluations) for four combinations of filter size and mutation operator. Note that using a 4×4 filter with a regular mutation operator makes no progress at all.

4.3 Generation from Small Samples

One of the challenges mentioned by the PCGML survey [9] was generating content from small samples. One of the appeals of WFC is its ability to do this. To this end we also tested our evolutionary method on one of the tiny samples in the WFC repository, where a single 4×4 tile pattern (shown in figure 10) is used to generate interesting levels of arbitrary size. Here we used a 2×2 filter, and 10,000 fitness evaluations for the evolution. Figure 11 shows a sample tile image generated from this tiny pattern demonstrating that as with WFC, interesting images can be built from tiny samples. Unlike WFC, our method may also introduce novel sub-patterns, whether desired or not.

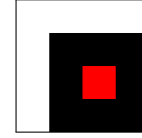


Figure 10: Small 4×4 image patch.

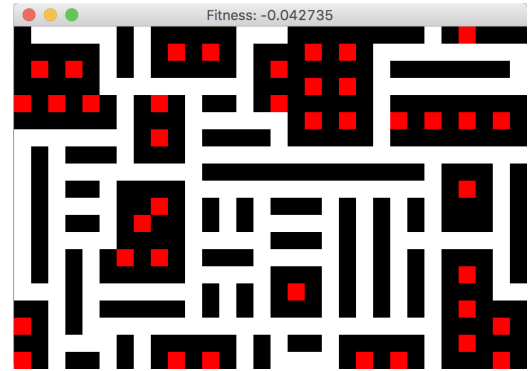


Figure 11: Sample tile image generated after evolving for 10,000 iterations using a 2×2 filter from the 4×4 image patch in figure 10 using a convolutional mutation operator. Unlike the WFC approach, this contains tile patterns that do not occur in the sample.

4.4 Exploring KL-Div Asymmetry

Here we explore the asymmetric nature of D_{KL} . If we measure the divergence of the generated level from the training sample ($w = 0.0$) we get very different results compared to making the opposite calculation. Figure 12 shows the effect of evolving a level using a 4×4 filter (our default filter size) using fitness measure F with $w = 0.0$. Note how this generates boring levels. Every tile pattern seen in the level has occurred during training, but many interesting ones have been omitted because there was insufficient incentive to include them. Hence this leads to a “lazy” effort of placing lots of easily placed patterns, leading to excessive sky when generating Mario levels.

Observe the contrast in Figure 13 that used ($w = 1.0$). Here we see that many interesting tile patterns have been included from the training sample, but put together in ways that also lead to many patterns which never occurred in the training sample.

Our default is to use the symmetric KL-Div ($w = 0.5$), which evenly weights the the two asymmetric measures. Note that any weighted sum could be used to favour inclusion of sample patterns or exclusion of novel patterns. We have not tuned this thoroughly and used ($w = 0.5$) for most of the generated levels in the paper, though values of 0.6 or 0.7 may produce even more interesting looking levels (in our opinion) at the expense of the occasional anomaly.



Figure 12: Level generated using asymmetric KL-Div ($w = 0.0$) where novel patterns (those not occurring in training sample are penalised, but there is no direct penalty for not using sample patterns.

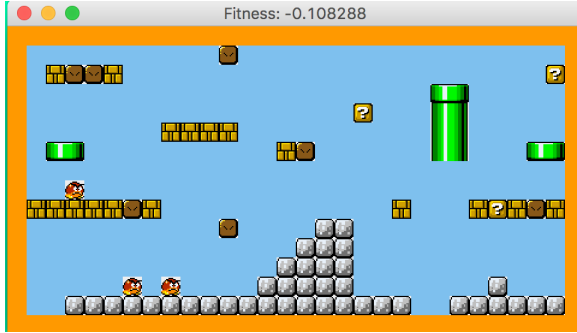


Figure 13: Level generated using asymmetric KL-Div ($w = 1.0$) where failure to use sample patterns is penalised, but any novel (unseen) patterns are not penalised.

4.5 Evaluation

Table 2 provides a summary of how the proposed method compares with other recent sample-based methods. As mentioned above, we include evaluations of Wave Function Collapse (WFC), Evolution in the Latent Space of Generative Adversarial Networks (ELSGAN)⁵ [13] and our proposed method: Evolution with Tile Pattern KL-Divergence (ETPKLDiv). We also include a GAN model without latent vector evolution (GAN). These are not meant to be an exhaustive list of general sample-based 2D generators, but both WFC and ELSGAN have gained recent prominence. All these methods potentially generalise to any 2D tile-based level (hence one-dimensional Markov chain methods are not included).

⁵<https://github.com/schrum2/GameGAN>

Table 2: A qualitative rating of the training and generation speed of each method described in this paper. For MarioGAN we decompose the rating in to the GAN part and the GAN combined with evolution (ELSGAN). *Tiny* refers to whether the method is able to generate whole levels from a single small patch (e.g. see figures 10 and 11).

Method	Training	Generation	Tiny
ETPKLDiv	Fast	Fast to Medium, Never Fails	Yes
WFC	Fast	Fast to Slow, May Fail	Yes
GAN	Slow	Always Fast, Never Fails	No
ELSGAN	Slow	Slow, Never Fails	No

In table 2 by fast we mean sub-second time, medium is the order of a few seconds, slow is of the order of minutes, hours or more. WFC has highly variable generation time compared to the other methods, and is the only one which can fail to produce a level. The other methods may produce poor levels (as in ones which differ too much from the training sample) but always produce something. Furthermore, the use of tile-pattern features together with D_{KL} leads to flexible analysis and generation, since the sliding window means the levels we generate can be of a different size compared to the training sample(s). This is in contrast to the MarioGAN method where once trained, the GAN outputs levels of a fixed size.

As we have established in section 3.2, D_{KL} is a form of valid evaluation to assess the similarity of generated levels to the original. For each of the approaches in table 2, we thus train a generator from Mario level 1-1 (see figure 1) and generate 100 levels each. For the approaches using tile patterns (WFC and ETPKLDiv), we test several versions with filter sizes (2×2 , 3×3 and 4×4). For the resulting levels, we then compute the D_{KL} values.

In order to test the effects of the parameters discussed in section 3.1, we vary the filter sizes and weights. We test all combinations of 2×2 , 3×3 and 4×4 with weights in 0, 0.5 and 1. These weights represent the symmetric (0.5) and extreme cases of asymmetry (see section 4.4).

We report the mean D_{KL} for each type of generator in figure 14. The parameters for the D_{KL} computation are displayed on the x-axis. The respective generator is indicated on the y-axis. We display the mean value rounded to 2 digits in each cell of the resulting matrix, and visualise using colour as indicated in the colour key above the heatmap.

As expected, D_{KL} values tend to be lower for small filter sizes (left three columns). However, we can also clearly see relatively good performance of ETPKLDiv with 4×4 and 3×3 filters across all D_{KL} filter sizes and weights. The ETPKLDiv with 2×2 filter performs well for D_{KL} measures with the same filter size, but worse than all other methods on larger filter sizes. We thus conclude that the optimisation of D_{KL} as suggested in section 4 was successful. However, for low filter sizes in ETPKLDiv, the successful optimisation for a smaller filter can introduce previously unseen patterns for larger filters, thus reducing the respective D_{KL} value.

Interestingly, the same is not true for the WFC approaches, with all generators performing similarly across the board. This means that in the case of Mario levels, smaller filter sizes are sufficient to express constraints for placing the tiles. Both GAN approaches tend to have larger D_{KL} values. This is to be expected, as the notion of

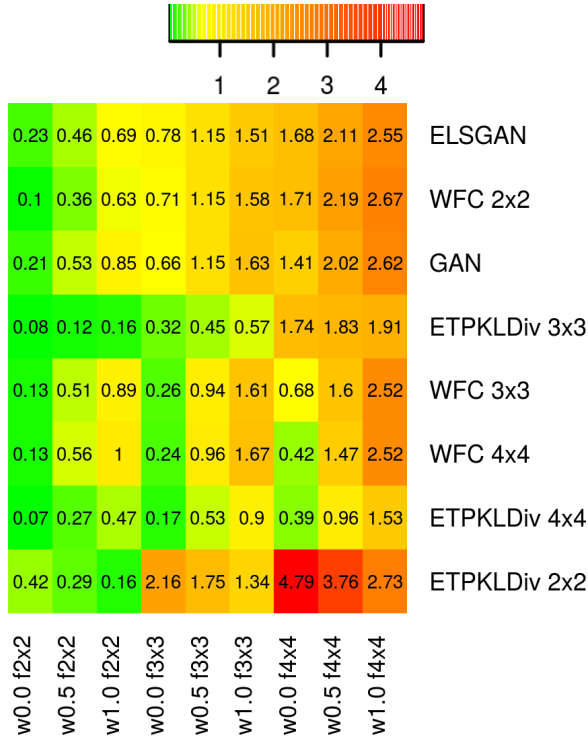


Figure 14: Heatmap depicting mean D_{KL} values for 100 levels generated by different generators

KL-Divergence is slightly different. Also, the levels in ELSGAN are not optimised for D_{KL} , but for some other, content-specific fitness (number of jumps). We thus also see no improvement in the D_{KL} values over GAN. However, in the case of ELSGAN, novel level structures are actually encouraged.

A caveat of this analysis are the large variations of the D_{KL} values for most of the levels (except for ETPKLDiv). Thus, if the generator in question is fast enough, multiple levels can (and should) be produced in order to select one with the desired properties.

Something that is harder to capture is whether the methods are predictable and reliable. Based on the rather large variances we observed, as well as from manual anecdotal observations, none of these methods are yet reliable in that they can be guaranteed to produce good (as in levels which look like the training sample and are not obviously broken) results when given a new set of sample levels from a different game to generate from, though they all work acceptably well for Mario levels. They are all at the “interesting to play with” stage.

5 CONCLUSIONS AND FUTURE WORK

In this paper we proposed a fitness function for level generation based on three steps:

- use rectangular tile patterns as features;
- interpret the feature counts as probability distributions;
- use Kullback-Leibler Divergence (D_{KL}) between generated levels and training sample levels as a measure of similarity.

The main result is that Kullback-Leibler Divergence used with tile-pattern features provides a useful and efficient way to compare game levels, providing insight into the nature of the differences. It can be applied as an objective function for evolving levels. The asymmetric nature of the measure means we can assess the novelty of the generated levels, and separately measure the extent to which features in the training set are represented in the generated levels.

The Tile Pattern KL-Div can be computed efficiently which leads to a reasonably fast evolutionary generation process. The levels generated in this paper used a budget of 10,000 fitness evaluations for the evolutionary algorithm which took less than one second to generate levels using a 2×2 filter and less than two seconds when using a 4×4 filter. Training time is negligible (around 1ms).

When used within an evolutionary algorithm we showed, not surprisingly, that a standard bit-flip mutation operator fails to find fit solutions. To counter this, we introduced a convolutional mutation operator that works by copying rectangular tile-patches from random parts of the training sample to the generated level. This was shown to provide effective search, generating fitter levels than the samples snipped out of the training data. Our convolutional mutation operator currently always copies patches of the same size as the filter. This is unlikely to be optimal, and we plan to introduce a bandit-based approach to sample different patch-sizes mutations.

While the method works well, it does not directly capture capture more holistic constraints such as the way a level may progress through different phases, and WFC also has this limitation: any progression only happens in as much as it can be expressed as an outcome of local tile-pattern constraints. Our tile-pattern implementation also has a *stride* parameter to enable gaps in the pattern which could potentially capture long-range constraints, but we have not yet had chance to experiment with this. The GAN approach is able to capture long-range constraints due to the nature of the deep neural network, so it would be interesting to quantify how effective this is which would involve (at least in the case of Mario) generating and comparing wider levels.

Despite having some attractive features, the use of tile-patterns also has a significant flaw, at least in the way we have applied it here. When sampling the tile patterns we count the number of occurrences of each exact pattern, so a pattern that differed in some small and perhaps irrelevant way is counted as being just as distinct as an entirely different pattern. Although the method still works, a way of counting patterns that focused more on their fundamental nature and ignored irrelevant differences should work even better. We are currently exploring a technique using connected component analysis that offers some promise in this direction.

There are some other obvious avenues for future work:

- Use an ensemble of filter sizes
- Apply to a wider set of games
- Validate measure with playtests (AI agents and human)

Finally, we believe the optimisation problem of finding levels with a low D_{KL} to be interesting in itself, and plan to propose it as a Game Benchmark for Evolutionary Algorithms⁶. Although the method described in this paper (using our convolutional mutation operator together with a $(1+1)$ EA) provides an acceptable solution, it is likely that better algorithms may be developed.

⁶<http://norvig.eecs.qmul.ac.uk/gbea/>

REFERENCES

- [1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML*.
- [2] Cameron B Browne. 2008. *Automatic generation and evaluation of recombination games*. PhD Thesis. Queensland University of Technology, Brisbane, Australia.
- [3] Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez-Liebana, Spyridon Samothrakis, and Simon Colton. 2012. A Survey of Monte Carlo Tree Search Methods. *IEEE TCIAIG* 4, 1 (2012), 1–43.
- [4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*. 2672–2680.
- [5] Isaac Karth and Adam M. Smith. 2017. WaveFunction Collapse is Constraint Solving in the Wild. *International Conference on the Foundations of Digital Games (FDG)*.
- [6] S. M. Lucas and T. J. Reynolds. 2005. Learning deterministic finite automata with a smart state labeling evolutionary algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 7 (July 2005), 1063–1074. <https://doi.org/10.1109/TPAMI.2005.143>
- [7] Diego Perez-Liebana, Spyridon Samothrakis, Simon Lucas, and Philipp Rohlfshagen. 2013. Rolling Horizon Evolution versus Tree Search for Navigation in Single-player Real-time Games. In *ACM Conference on Genetic and Evolutionary computation*. 351–358.
- [8] Noor Shaker, Gillian Smith, and Georgios N. Yannakakis. 2016. Evaluating content generators. In *Procedural Content Generation in Games: A Textbook and an Overview of Current Research*, Noor Shaker, Julian Togelius, and Mark J. Nelson (Eds.). Springer, 215–224.
- [9] Adam Summerville, Sam Snodgrass, Matthew Guzdial, Christoffer Holmgård, Amy K Hoover, Aaron Isaksen, Andy Nealen, and Julian Togelius. 2017. Procedural Content Generation via Machine Learning (PCGML). *arXiv preprint arXiv:1702.00539* (2017).
- [10] Adam James Summerville, Sam Snodgrass, Michael Mateas, and Santiago Ontañón Villar. 2016. The VGLC: The Video Game Level Corpus. *Proceedings of the 7th Workshop on Procedural Content Generation* (2016).
- [11] Julian Togelius and Noor Shaker. 2016. The search-based approach. In *Procedural Content Generation in Games: A Textbook and an Overview of Current Research*, Noor Shaker, Julian Togelius, and Mark J. Nelson (Eds.). Springer, 17–30.
- [12] Julian Togelius, Georgios N. Yannakakis, Kenneth O. Stanley, and Cameron Browne. 2011. Search-based procedural content generation: A taxonomy and survey. *IEEE Transactions on Computational Intelligence and AI in Games* 3, 3 (2011), 172–186.
- [13] Vanessa Volz, Jacob Schrum, Jialin Liu, Simon M. Lucas, Adam M. Smith, and Sebastian Risi. 2018. Evolving Mario Levels in the Latent Space of a Deep Convolutional Generative Adversarial Network. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2018)*. ACM, New York, NY, USA, 8. <https://doi.org/10.1145/3205455.3205517>