Unifying Count-Based Exploration and Intrinsic Motivation

Marc G. Bellemare
bellemare@google.com

Sriram Srinivasan
srsrinivasan@google.com

Georg Ostrovski
ostrovski@google.com

Tom Schaul
schaul@google.com

David Saxton
saxton@google.com

Rémi Munos
munos@google.com

Google DeepMind
London, United Kingdom

Abstract

We consider an agent's uncertainty about its environment and the problem of generalizing this uncertainty across states. Specifically, we focus on the problem of exploration in non-tabular reinforcement learning. Drawing inspiration from the intrinsic motivation literature, we use density models to measure uncertainty, and propose a novel algorithm for deriving a pseudo-count from an arbitrary density model. This technique enables us to generalize count-based exploration algorithms to the non-tabular case. We apply our ideas to Atari 2600 games, providing sensible pseudo-counts from raw pixels. We transform these pseudo-counts into exploration bonuses and obtain significantly improved exploration in a number of hard games, including the infamously difficult MONTEZUMA'S REVENGE.

1 Introduction

Exploration algorithms for Markov Decision Processes (MDPs) are typically concerned with reducing the agent's uncertainty over the environment's reward and transition functions. In a tabular setting, this uncertainty can be quantified using confidence intervals derived from Chernoff bounds, or inferred from a posterior over the environment parameters. In fact, both confidence intervals and posterior shrink as the inverse square root of the state-action visit count $N(x, a)$, making this quantity fundamental to most theoretical results on exploration.

Count-based exploration methods directly use visit counts to guide an agent's behaviour towards reducing uncertainty. For example, Model-based Interval Estimation with Exploration Bonuses (MBIE-EB; Strehl and Littman, 2008) solves the augmented Bellman equation

$$V(x) = \max_{a \in \mathcal{A}} \left[\hat{R}(x, a) + \gamma \, \mathbb{E}_{\hat{P}} \left[V(x') \right] + \beta N(x, a)^{-1/2} \right],$$

involving the empirical reward $\hat{R}$, the empirical transition function $\hat{P}$, and an exploration bonus proportional to $N(x, a)^{-1/2}$. This bonus accounts for uncertainties in both transition and reward functions and enables a finite-time bound on the agent's suboptimality.

In spite of their pleasant theoretical guarantees, count-based methods have not played a role in the contemporary successes of reinforcement learning (e.g. Mnih et al., 2015). Instead, most practical methods still rely on simple rules such as ϵ-greedy. The issue is that visit counts are not directly useful in large domains, where states are rarely visited more than once.

Answering a different scientific question, intrinsic motivation aims to provide qualitative guidance for exploration (Schmidhuber, 1991; Oudeyer et al., 2007; Barto, 2013). This guidance can be summarized as "explore what surprises you". A typical approach guides the agent based on change

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

in prediction error, or *learning progress*. If $e_n(A)$ is the error made by the agent at time n over some event A, and $e_{n+1}(A)$ the same error after observing a new piece of information, then learning progress is

$$e_n(A) - e_{n+1}(A).$$

Intrinsic motivation methods are attractive as they remain applicable in the absence of the Markov property or the lack of a tabular representation, both of which are required by count-based algorithms. Yet the theoretical foundations of intrinsic motivation remain largely absent from the literature, which may explain its slow rate of adoption as a standard approach to exploration.

In this paper we provide formal evidence that intrinsic motivation and count-based exploration are but two sides of the same coin. Specifically, we consider a frequently used measure of learning progress, *information gain* (Cover and Thomas, 1991). Defined as the Kullback-Leibler divergence of a prior distribution from its posterior, information gain can be related to the confidence intervals used in count-based exploration. Our contribution is to propose a new quantity, the *pseudo-count*, which connects information-gain-as-learning-progress and count-based exploration.

We derive our pseudo-count from a density model over the state space. This is in departure from more traditional approaches to intrinsic motivation that consider learning progress with respect to a transition model. We expose the relationship between pseudo-counts, a variant of Schmidhuber's compression progress we call *prediction gain*, and information gain. Combined to Kolter and Ng's negative result on the frequentist suboptimality of Bayesian bonuses, our result highlights the theoretical advantages of pseudo-counts compared to many existing intrinsic motivation methods.

The pseudo-counts we introduce here are best thought of as "function approximation for exploration". We bring them to bear on Atari 2600 games from the Arcade Learning Environment (Bellemare et al., 2013), focusing on games where myopic exploration fails. We extract our pseudo-counts from a simple density model and use them within a variant of MBIE-EB. We apply them to an experience replay setting and to an actor-critic setting, and find improved performance in both cases. Our approach produces dramatic progress on the reputedly most difficult Atari 2600 game, MONTEZUMA'S REVENGE: within a fraction of the training time, our agent explores a significant portion of the first level and obtains significantly higher scores than previously published agents.

2 Notation

We consider a countable state space $\mathcal{X}$. We denote a sequence of length n from $\mathcal{X}$ by $x_{1:n} \in \mathcal{X}^n$, the set of finite sequences from $\mathcal{X}$ by $\mathcal{X}^*$, write $x_{1:n}x$ to mean the concatenation of $x_{1:n}$ and a state $x \in \mathcal{X}$, and denote the empty sequence by ϵ. A *model* over $\mathcal{X}$ is a mapping from $\mathcal{X}^*$ to probability distributions over $\mathcal{X}$. That is, for each $x_{1:n} \in \mathcal{X}^n$ the model provides a probability distribution

$$\rho_n(x) := \rho(x\,;\, x_{1:n}).$$

Note that we do not require $\rho_n(x)$ to be strictly positive for all x and $x_{1:n}$. When it is, however, we may understand $\rho_n(x)$ to be the usual conditional probability of $X_{n+1} = x$ given $X_1 \dots X_n = x_{1:n}$.

We will take particular interest in the empirical distribution μ_n derived from the sequence $x_{1:n}$. If $N_n(x) := N(x, x_{1:n})$ is the number of occurrences of a state x in the sequence $x_{1:n}$, then

$$\mu_n(x) := \mu(x\,;\, x_{1:n}) := \frac{N_n(x)}{n}.$$

We call the N_n the *empirical count function*, or simply *empirical count*. The above notation extends to state-action spaces, and we write $N_n(x, a)$ to explicitly refer to the number of occurrences of a state-action pair when the argument requires it. When $x_{1:n}$ is generated by an ergodic Markov chain, for example if we follow a fixed policy in a finite-state MDP, then the limit point of μ_n is the chain's stationary distribution.

In our setting, a *density model* is any model that assumes states are independently (but not necessarily identically) distributed; a density model is thus a particular kind of generative model. We emphasize that a density model differs from a forward model, which takes into account the temporal relationship between successive states. Note that μ_n is itself a density model.

3 From Densities to Counts

In the introduction we argued that the visit count $N_n(x)$ (and consequently, $N_n(x, a)$) is not directly useful in practical settings, since states are rarely revisited. Specifically, $N_n(x)$ is almost always zero and cannot help answer the question "How novel is this state?" Nor is the problem solved by a Bayesian approach: even variable-alphabet models (e.g. Hutter, 2013) must assign a small, diminishing probability to yet-unseen states. To estimate the uncertainty of an agent's knowledge, we must instead look for a quantity which generalizes across states. Guided by ideas from the intrinsic motivation literature, we now derive such a quantity. We call it a *pseudo-count* as it extends the familiar notion from Bayesian estimation.

3.1 Pseudo-Counts and the Recoding Probability

We are given a density model ρ over $\mathcal{X}$. This density model may be approximate, biased, or even inconsistent. We begin by introducing the *recoding probability* of a state x:

$$\rho'_n(x) := \rho(x\,;\, x_{1:n}x).$$

This is the probability assigned to x by our density model after observing a new occurrence of x. The term "recoding" is inspired from the statistical compression literature, where coding costs are inversely related to probabilities (Cover and Thomas, 1991). When ρ admits a conditional probability distribution,

$$\rho'_n(x) = \Pr_\rho(X_{n+2} = x \,|\, X_1 \dots X_n = x_{1:n}, X_{n+1} = x).$$

We now postulate two unknowns: a *pseudo-count function* $\hat{N}_n(x)$, and a *pseudo-count total* $\hat{n}$. We relate these two unknowns through two constraints:

$$\rho_n(x) = \frac{\hat{N}_n(x)}{\hat{n}} \qquad \rho'_n(x) = \frac{\hat{N}_n(x) + 1}{\hat{n} + 1}. \tag{1}$$

In words: we require that, after observing one instance of x, the density model's increase in prediction of that same x should correspond to a unit increase in pseudo-count. The pseudo-count itself is derived from solving the linear system (1):

$$\hat{N}_n(x) = \frac{\rho_n(x)(1 - \rho'_n(x))}{\rho'_n(x) - \rho_n(x)} = \hat{n}\rho_n(x). \tag{2}$$

Note that the equations (1) yield $\hat{N}_n(x) = 0$ (with $\hat{n} = \infty$) when $\rho_n(x) = \rho'_n(x) = 0$, and are inconsistent when $\rho_n(x) < \rho'_n(x) = 1$. These cases may arise from poorly behaved density models, but are easily accounted for. From here onwards we will assume a consistent system of equations.

Definition 1 (Learning-positive density model)**.** *A density model ρ is* learning-positive *if for all $x_{1:n} \in \mathcal{X}^n$ and all $x \in \mathcal{X}$, $\rho'_n(x) \geq \rho_n(x)$.*

By inspecting (2), we see that

1. $\hat{N}_n(x) \geq 0$ if and only if ρ is learning-positive;
2. $\hat{N}_n(x) = 0$ if and only if $\rho_n(x) = 0$; and
3. $\hat{N}_n(x) = \infty$ if and only if $\rho_n(x) = \rho'_n(x)$.

In many cases of interest, the pseudo-count $\hat{N}_n(x)$ matches our intuition. If $\rho_n = \mu_n$ then $\hat{N}_n = N_n$. Similarly, if ρ_n is a Dirichlet estimator then $\hat{N}_n$ recovers the usual notion of pseudo-count. More importantly, if the model generalizes across states then so do pseudo-counts.

3.2 Estimating the Frequency of a Salient Event in FREEWAY

As an illustrative example, we employ our method to estimate the number of occurrences of an infrequent event in the Atari 2600 video game FREEWAY (Figure 1, screenshot). We use the Arcade Learning Environment (Bellemare et al., 2013). We will demonstrate the following:

1. Pseudo-counts are roughly zero for novel events,

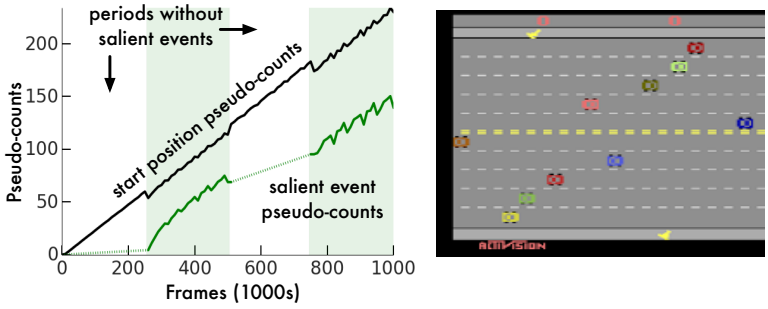

Figure 1: Pseudo-counts obtained from a CTS density model applied to FREEWAY, along with a frame representative of the salient event (crossing the road). Shaded areas depict periods during which the agent observes the salient event, dotted lines interpolate across periods during which the salient event is not observed. The reported values are 10,000-frame averages.

2. they exhibit credible magnitudes,
3. they respect the ordering of state frequency,
4. they grow linearly (on average) with real counts,
5. they are robust in the presence of nonstationary data.

These properties suggest that pseudo-counts provide an appropriate generalized notion of visit counts in non-tabular settings.

In FREEWAY, the agent must navigate a chicken across a busy road. As our example, we consider estimating the number of times the chicken has reached the very top of the screen. As is the case for many Atari 2600 games, this naturally salient event is associated with an increase in score, which ALE translates into a positive reward. We may reasonably imagine that knowing how certain we are about this part of the environment is useful. After crossing, the chicken is teleported back to the bottom of the screen.

To highlight the robustness of our pseudo-count, we consider a nonstationary policy which waits for 250,000 frames, then applies the UP action for 250,000 frames, then waits, then goes UP again. The salient event only occurs during UP periods. It also occurs with the cars in different positions, thus requiring generalization. As a point of reference, we record the pseudo-counts for both the salient event and visits to the chicken's start position.

We use a simplified, pixel-level version of the CTS model for Atari 2600 frames proposed by Bellemare et al. (2014), ignoring temporal dependencies. While the CTS model is rather impoverished in comparison to state-of-the-art density models for images (e.g. Van den Oord et al., 2016), its count-based nature results in extremely fast learning, making it an appealing candidate for exploration. Further details on the model may be found in the appendix.

Examining the pseudo-counts depicted in Figure 1 confirms that they exhibit the desirable properties listed above. In particular, the pseudo-count is almost zero on the first occurrence of the salient event; it increases slightly during the 3rd period, since the salient and reference events share some common structure; throughout, it remains smaller than the reference pseudo-count. The linearity on average and robustness to nonstationarity are immediate from the graph. Note, however, that the pseudo-counts are a fraction of the real visit counts (inasmuch as we can define "real"): by the end of the trial, the start position has been visited about 140,000 times, and the topmost part of the screen, 1285 times. Furthermore, the ratio of recorded pseudo-counts differs from the ratio of real counts. Both effects are quantifiable, as we shall show in Section 5.

4 The Connection to Intrinsic Motivation

Having argued that pseudo-counts appropriately generalize visit counts, we will now show that they are closely related to *information gain*, which is commonly used to quantify novelty or curiosity and consequently as an intrinsic reward. Information gain is defined in relation to a *mixture model* ξ over

a class of density models $\mathcal{M}$. This model predicts according to a weighted combination from $\mathcal{M}$:

$$\xi_n(x) := \xi(x\,;\,x_{1:n}) := \int_{\rho\in\mathcal{M}} w_n(\rho)\rho(x\,;\,x_{1:n})\mathrm{d}\rho,$$

with $w_n(\rho)$ the posterior weight of ρ. This posterior is defined recursively, starting from a prior distribution w_0 over $\mathcal{M}$:

$$w_{n+1}(\rho) := w_n(\rho, x_{n+1}) \qquad w_n(\rho, x) := \frac{w_n(\rho)\rho(x\,;\,x_{1:n})}{\xi_n(x)}. \tag{3}$$

Information gain is then the Kullback-Leibler divergence from prior to posterior that results from observing x:

$$\mathrm{IG}_n(x) := \mathrm{IG}(x\,;\,x_{1:n}) := \mathrm{KL}\big(w_n(\cdot, x)\,\|\,w_n\big).$$

Computing the information gain of a complex density model is often impractical, if not downright intractable. However, a quantity which we call the *prediction gain* provides us with a good approximation of the information gain. We define the prediction gain of a density model ρ (and in particular, ξ) as the difference between the recoding log-probability and log-probability of x:

$$\mathrm{PG}_n(x) := \log\rho'_n(x) - \log\rho_n(x).$$

Prediction gain is nonnegative if and only if ρ is learning-positive. It is related to the pseudo-count:

$$\hat{N}_n(x) \approx \left(e^{\mathrm{PG}_n(x)} - 1\right)^{-1},$$

with equality when $\rho'_n(x) \to 0$. As the following theorem shows, prediction gain allows us to relate pseudo-count and information gain.

Theorem 1. *Consider a sequence $x_{1:n} \in \mathcal{X}^n$. Let ξ be a mixture model over a class of learning-positive models $\mathcal{M}$. Let $\hat{N}_n$ be the pseudo-count derived from ξ (Equation 2). For this model,*

$$\mathit{IG}_n(x) \leq \mathit{PG}_n(x) \leq \hat{N}_n(x)^{-1} \qquad \text{and} \qquad \mathit{PG}_n(x) \leq \hat{N}_n(x)^{-1/2}.$$

Theorem 1 suggests that using an exploration bonus proportional to $\hat{N}_n(x)^{-1/2}$, similar to the MBIE-EB bonus, leads to a behaviour at least as exploratory as one derived from an information gain bonus. Since pseudo-counts correspond to empirical counts in the tabular setting, this approach also preserves known theoretical guarantees. In fact, we are confident pseudo-counts may be used to prove similar results in non-tabular settings.

On the other hand, it may be difficult to provide theoretical guarantees about existing bonus-based intrinsic motivation approaches. Kolter and Ng (2009) showed that no algorithm based on a bonus upper bounded by $\beta N_n(x)^{-1}$ for any $\beta > 0$ can guarantee PAC-MDP optimality. Again considering the tabular setting and combining their result to Theorem 1, we conclude that bonuses proportional to immediate information (or prediction) gain are insufficient for theoretically near-optimal exploration: to paraphrase Kolter and Ng, these methods produce explore too little in comparison to pseudo-count bonuses. By inspecting (2) we come to a similar negative conclusion for bonuses proportional to the L1 or L2 distance between ξ'_n and ξ_n.

Unlike many intrinsic motivation algorithms, pseudo-counts also do not rely on learning a forward (transition and/or reward) model. This point is especially important because a number of powerful density models for images exist (Van den Oord et al., 2016), and because optimality guarantees cannot in general exist for intrinsic motivation algorithms based on forward models.

5 Asymptotic Analysis

In this section we analyze the limiting behaviour of the ratio $\hat{N}_n/N_n$. We use this analysis to assert the consistency of pseudo-counts derived from tabular density models, i.e. models which maintain per-state visit counts. In the appendix we use the same result to bound the approximation error of pseudo-counts derived from directed graphical models, of which our CTS model is a special case.

Consider a fixed, infinite sequence $x_1, x_2, \ldots$ from $\mathcal{X}$. We define the limit of a sequence of functions $\big(f(x\,;\,x_{1:n}) : n \in \mathbb{N}\big)$ with respect to the length n of the subsequence $x_{1:n}$. We additionally assume that the empirical distribution μ_n converges pointwise to a distribution μ, and write $\mu'_n(x)$ for the recoding probability of x under μ_n. We begin with two assumptions on our density model.

Assumption 1. *The limits*

$$\text{(a) } r(x) := \lim_{n\to\infty} \frac{\rho_n(x)}{\mu_n(x)} \qquad \text{(b) } \dot{r}(x) := \lim_{n\to\infty} \frac{\rho'_n(x) - \rho_n(x)}{\mu'_n(x) - \mu_n(x)}$$

exist for all x; furthermore, $\dot{r}(x) > 0$.

Assumption (a) states that ρ should eventually assign a probability to x proportional to the limiting empirical distribution $\mu(x)$. In particular there must be a state x for which $r(x) < 1$, unless $\rho_n \to \mu$. Assumption (b), on the other hand, imposes a restriction on the learning rate of ρ relative to μ's. As both $r(x)$ and $\mu(x)$ exist, Assumption 1 also implies that $\rho_n(x)$ and $\rho'_n(x)$ have a common limit.

Theorem 2. *Under Assumption 1, the limit of the ratio of pseudo-counts $\hat{N}_n(x)$ to empirical counts $N_n(x)$ exists for all x. This limit is*

$$\lim_{n\to\infty} \frac{\hat{N}_n(x)}{N_n(x)} = \frac{r(x)}{\dot{r}(x)} \left(\frac{1 - \mu(x) r(x)}{1 - \mu(x)} \right).$$

The model's relative rate of change, whose convergence to $\dot{r}(x)$ we require, plays an essential role in the ratio of pseudo- to empirical counts. To see this, consider a sequence $(x_n : n \in \mathbb{N})$ generated i.i.d. from a distribution μ over a finite state space, and a density model defined from a sequence of nonincreasing step-sizes $(\alpha_n : n \in \mathbb{N})$:

$$\rho_n(x) = (1 - \alpha_n)\rho_{n-1}(x) + \alpha_n \mathbb{I}\{x_n = x\},$$

with initial condition $\rho_0(x) = |\mathcal{X}|^{-1}$. For $\alpha_n = n^{-1}$, this density model is the empirical distribution. For $\alpha_n = n^{-2/3}$, we may appeal to well-known results from stochastic approximation (e.g. Bertsekas and Tsitsiklis, 1996) and find that almost surely

$$\lim_{n\to\infty} \rho_n(x) = \mu(x) \qquad \text{but} \qquad \lim_{n\to\infty} \frac{\rho'_n(x) - \rho_n(x)}{\mu'_n(x) - \mu_n(x)} = \infty.$$

Since $\mu'_n(x) - \mu_n(x) = n^{-1}(1 - \mu'_n(x))$, we may think of Assumption 1(b) as also requiring ρ to converge at a rate of $\Theta(1/n)$ for a comparison with the empirical count N_n to be meaningful. Note, however, that a density model that does not satisfy Assumption 1(b) may still yield useful (but incommensurable) pseudo-counts.

Corollary 1. *Let $\phi(x) > 0$ with $\sum_{x\in\mathcal{X}} \phi(x) < \infty$ and consider the count-based estimator*

$$\rho_n(x) = \frac{N_n(x) + \phi(x)}{n + \sum_{x'\in\mathcal{X}} \phi(x')}.$$

If $\hat{N}_n$ is the pseudo-count corresponding to ρ_n then $\hat{N}_n(x)/N_n(x) \to 1$ for all x with $\mu(x) > 0$.

6 Empirical Evaluation

In this section we demonstrate the use of pseudo-counts to guide exploration. We return to the Arcade Learning Environment, now using the CTS model to generate an exploration bonus.

6.1 Exploration in Hard Atari 2600 Games

From 60 games available through the Arcade Learning Environment we selected five "hard" games, in the sense that an ϵ-greedy policy is inefficient at exploring them. We used a bonus of the form

$$R^+_n(x, a) := \beta(\hat{N}_n(x) + 0.01)^{-1/2}, \tag{4}$$

where $\beta = 0.05$ was selected from a coarse parameter sweep. We also compared our method to the optimistic initialization trick proposed by Machado et al. (2015). We trained our agents' Q-functions with Double DQN (van Hasselt et al., 2016), with one important modification: we mixed the Double Q-Learning target with the Monte Carlo return. This modification led to improved results both with and without exploration bonuses (details in the appendix).

Figure 2 depicts the result of our experiment, averaged across 5 trials. Although optimistic initialization helps in FREEWAY, it otherwise yields performance similar to DQN. By contrast, the

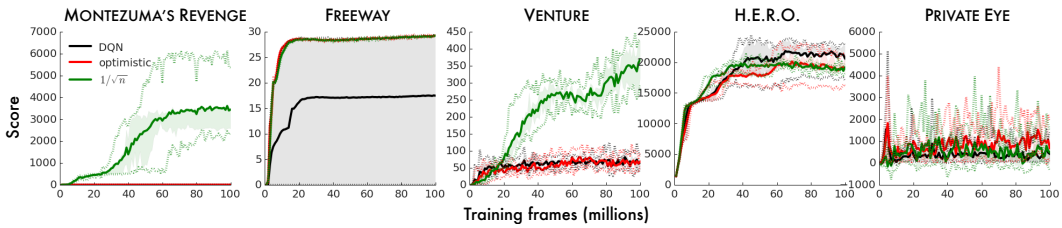

Figure 2: Average training score with and without exploration bonus or optimistic initialization in 5 Atari 2600 games. Shaded areas denote inter-quartile range, dotted lines show min/max scores.

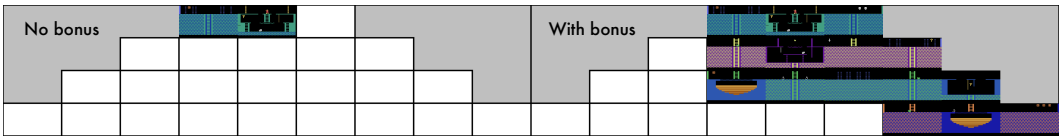

Figure 3: "Known world" of a DQN agent trained for 50 million frames with (**right**) and without (**left**) count-based exploration bonuses, in MONTEZUMA'S REVENGE.

count-based exploration bonus enables us to make quick progress on a number of games, most dramatically in MONTEZUMA'S REVENGE and VENTURE.

MONTEZUMA'S REVENGE is perhaps the hardest Atari 2600 game available through the ALE. The game is infamous for its hostile, unforgiving environment: the agent must navigate a number of different rooms, each filled with traps. Due to its sparse reward function, most published agents achieve an average score close to zero and completely fail to explore most of the 24 rooms that constitute the first level (Figure 3, top). By contrast, within 50 million frames our agent learns a policy which consistently navigates through 15 rooms (Figure 3, bottom). Our agent also achieves a score higher than anything previously reported, with one run consistently achieving 6600 points by 100 million frames (half the training samples used by Mnih et al. (2015)). We believe the success of our method in this game is a strong indicator of the usefulness of pseudo-counts for exploration.[1]

6.2 Exploration for Actor-Critic Methods

We next used our exploration bonuses in conjunction with the A3C (Asynchronous Advantage Actor-Critic) algorithm of Mnih et al. (2016). One appeal of actor-critic methods is their explicit separation of policy and Q-function parameters, which leads to a richer behaviour space. This very separation, however, often leads to deficient exploration: to produce any sensible results, the A3C policy must be regularized with an entropy cost. We trained A3C on 60 Atari 2600 games, with and without the exploration bonus (4). We refer to our augmented algorithm as A3C+. Full details and additional results may be found in the appendix.

We found that A3C fails to learn in **15** games, in the sense that the agent does not achieve a score 50% better than random. In comparison, there are only **10** games for which A3C+ fails to improve on the random agent; of these, **8** are games where DQN fails in the same sense. We normalized the two algorithms' scores so that 0 and 1 are respectively the minimum and maximum of the random agent's and A3C's end-of-training score on a particular game. Figure 4 depicts the in-training median score for A3C and A3C+, along with 1st and 3rd quartile intervals. Not only does A3C+ achieve slightly superior median performance, but it also significantly outperforms A3C on at least a quarter of the games. This is particularly important given the large proportion of Atari 2600 games for which an ϵ-greedy policy is sufficient for exploration.

7 Related Work

Information-theoretic quantities have been repeatedly used to describe intrinsically motivated behaviour. Closely related to prediction gain is Schmidhuber (1991)'s notion of compression progress,

[1] A video of our agent playing is available at `https://youtu.be/0yI2wJ6F8r0`.

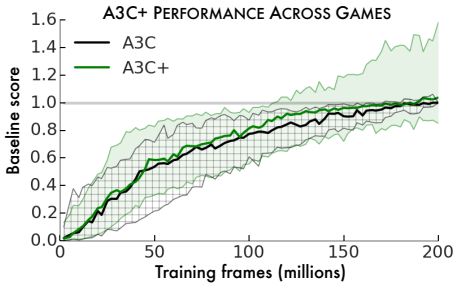

Figure 4: Median and interquartile performance across 60 Atari 2600 games for A3C and A3C+.

which equates novelty with an agent's improvement in its ability to compress its past. More recently, Lopes et al. (2012) showed the relationship between time-averaged prediction gain and visit counts in a tabular setting; their result is a special case of Theorem 2. Orseau et al. (2013) demonstrated that maximizing the sum of future information gains does lead to optimal behaviour, even though maximizing immediate information gain does not (Section 4). Finally, there may be a connection between sequential normalized maximum likelihood estimators and our pseudo-count derivation (see e.g. Ollivier, 2015).

Intrinsic motivation has also been studied in reinforcement learning proper, in particular in the context of discovering skills (Singh et al., 2004; Barto, 2013). Recently, Stadie et al. (2015) used a squared prediction error bonus for exploring in Atari 2600 games. Closest to our work is Houthooft et al. (2016)'s variational approach to intrinsic motivation, which is equivalent to a second order Taylor approximation to prediction gain. Mohamed and Rezende (2015) also considered a variational approach to the different problem of maximizing an agent's ability to influence its environment.

Aside for Orseau et al.'s above-cited work, it is only recently that theoretical guarantees for exploration have emerged for non-tabular, stateful settings. We note Pazis and Parr (2016)'s PAC-MDP result for metric spaces and Leike et al. (2016)'s asymptotic analysis of Thompson sampling in general environments.

8 Future Directions

The last few years have seen tremendous advances in learning representations for reinforcement learning. Surprisingly, these advances have yet to carry over to the problem of exploration. In this paper, we reconciled counts, the fundamental unit of uncertainty, with prediction-based heuristics and intrinsic motivation. Combining our work with more ideas from deep learning and better density models seems a plausible avenue for quick progress in practical, efficient exploration. We now conclude by outlining a few research directions we believe are promising.

Induced metric. We did not address the question of *where* the generalization comes from. Clearly, the choice of density model induces a particular metric over the state space. A better understanding of this metric should allow us to tailor the density model to the problem of exploration.

Compatible value function. There may be a mismatch in the learning rates of the density model and the value function: DQN learns much more slowly than our CTS model. As such, it should be beneficial to design value functions compatible with density models (or vice-versa).

The continuous case. Although we focused here on countable state spaces, we can as easily define a pseudo-count in terms of probability density functions. At present it is unclear whether this provides us with the right notion of counts for continuous spaces.

Acknowledgments

The authors would like to thank Laurent Orseau, Alex Graves, Joel Veness, Charles Blundell, Shakir Mohamed, Ivo Danihelka, Ian Osband, Matt Hoffman, Greg Wayne, Will Dabney, and Aäron van den Oord for their excellent feedback early and late in the writing, and Pierre-Yves Oudeyer and Yann Ollivier for pointing out additional connections to the literature.

References

Barto, A. G. (2013). Intrinsic motivation and reinforcement learning. In *Intrinsically Motivated Learning in Natural and Artificial Systems*, pages 17–47. Springer.

Bellemare, M., Veness, J., and Talvitie, E. (2014). Skip context tree switching. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1458–1466.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. (2013). The Arcade Learning Environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279.

Bellemare, M. G., Ostrovski, G., Guez, A., Thomas, P. S., and Munos, R. (2016). Increasing the action gap: New operators for reinforcement learning. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific.

Cover, T. M. and Thomas, J. A. (1991). *Elements of information theory*. John Wiley & Sons.

Houthooft, R., Chen, X., Duan, Y., Schulman, J., De Turck, F., and Abbeel, P. (2016). Variational information maximizing exploration.

Hutter, M. (2013). Sparse adaptive dirichlet-multinomial-like processes. In *Proceedings of the Conference on Online Learning Theory*.

Kolter, Z. J. and Ng, A. Y. (2009). Near-bayesian exploration in polynomial time. In *Proceedings of the 26th International Conference on Machine Learning*.

Leike, J., Lattimore, T., Orseau, L., and Hutter, M. (2016). Thompson sampling is asymptotically optimal in general environments. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Lopes, M., Lang, T., Toussaint, M., and Oudeyer, P.-Y. (2012). Exploration in model-based reinforcement learning by empirically estimating learning progress. In *Advances in Neural Information Processing Systems 25*.

Machado, M. C., Srinivasan, S., and Bowling, M. (2015). Domain-independent optimistic initialization for reinforcement learning. *AAAI Workshop on Learning for General Competency in Video Games*.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Mohamed, S. and Rezende, D. J. (2015). Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems 28*.

Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. G. (2016). Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*.

Ollivier, Y. (2015). Laplace's rule of succession in information geometry. *arXiv preprint arXiv:1503.04304*.

Orseau, L., Lattimore, T., and Hutter, M. (2013). Universal knowledge-seeking agents for stochastic environments. In *Proceedings of the Conference on Algorithmic Learning Theory*.

Oudeyer, P., Kaplan, F., and Hafner, V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2):265–286.

Pazis, J. and Parr, R. (2016). Efficient PAC-optimal exploration in concurrent, continuous state MDPs with delayed updates. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*.

Schmidhuber, J. (1991). A possibility for implementing curiosity and boredom in model-building neural controllers. In *From animals to animats: proceedings of the first international conference on simulation of adaptive behavior*.

Schmidhuber, J. (2008). Driven by compression progress. In *Knowledge-Based Intelligent Information and Engineering Systems*. Springer.

Singh, S., Barto, A. G., and Chentanez, N. (2004). Intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems 16*.

Stadie, B. C., Levine, S., and Abbeel, P. (2015). Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*.

Strehl, A. L. and Littman, M. L. (2008). An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309 – 1331.

Van den Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. (2016). Pixel recurrent neural networks. In *Proceedigns of the 33rd International Conference on Machine Learning*.

van Hasselt, H., Guez, A., and Silver, D. (2016). Deep reinforcement learning with double Q-learning. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*.

Veness, J., Bellemare, M. G., Hutter, M., Chua, A., and Desjardins, G. (2015). Compress and control. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305.

Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K., and de Freitas, N. (2016). Sample efficient actor-critic with experience replay. *arXiv:1611.01224*.

A The Connection to Intrinsic Motivation

The following provides an identity connecting information gain and prediction gain.

Lemma 1. *Consider a mixture model ξ over $\mathcal{M}$ with prediction gain PG_n and information gain IG_n, a fixed $x \in \mathcal{X}$, and let $w'_n(x) := w_n(\rho, x)$ be the posterior of ξ over $\mathcal{M}$ after observing x. Let $w''_n(x) := w'_n(\rho, x)$ be the same posterior after observing x a second time, and let $PG^\rho_n(x)$ denote the prediction gain of $\rho \in \mathcal{M}$. Then*

$$PG_n(x) = KL(w'_n \,\|\, w_n) + KL(w'_n \,\|\, w''_n) = IG_n(x) + KL(w'_n \,\|\, w''_n) + \mathbb{E}_{w'_n}\left[PG^\rho_n(x)\right].$$

In particular, if $\mathcal{M}$ is a class of non-adaptive models in the sense that $\rho_n(x) = \rho(x)$ for all $x_{1:n}$, then

$$PG_n(x) = KL(w'_n \,\|\, w_n) + KL(w'_n \,\|\, w''_n) = IG_n(x) + KL(w'_n \,\|\, w''_n).$$

A model which is non-adaptive is also learning-positive in the sense of Definition 1. Many common mixture models, for example Dirichlet-multinomial estimators, are mixtures over non-adaptive models.

Proof. We rewrite the posterior update rule (3) to show that for any $\rho \in \mathcal{M}$ and any $x \in \mathcal{X}$,

$$\xi_n(x) = \frac{\rho_n(x) w_n(\rho)}{w_n(\rho, x)}.$$

Write $\mathbb{E}_{w'_n} := \mathbb{E}_{\rho \sim w'_n(\cdot)}$. Now

$$\begin{aligned}
PG_n(x) = \log \frac{\xi'_n(x)}{\xi_n(x)} &= \mathbb{E}_{w'_n}\left[\log \frac{\xi'_n(x)}{\xi_n(x)}\right] \\
&= \mathbb{E}_{w'_n}\left[\log \frac{w'_n(\rho)}{w''_n(\rho)} \frac{w'_n(\rho)}{w_n(\rho)} \frac{\rho'_n(x)}{\rho_n(x)}\right] \\
&= \mathbb{E}_{w'_n}\left[\log \frac{w'_n(\rho)}{w_n(\rho)}\right] + \mathbb{E}_{w'_n}\left[\log \frac{w'_n(\rho)}{w''_n(\rho)}\right] + \mathbb{E}_{w'_n}\left[\log \frac{\rho'_n(x)}{\rho_n(x)}\right] \\
&= \mathrm{IG}_n(x) + \mathrm{KL}(w'_n \,\|\, w''_n) + \mathbb{E}_{w'_n}\left[\mathrm{PG}^\rho_n(x)\right].
\end{aligned}$$

□

The second statement follows immediately.

Lemma 2. *The functions $f(x) := e^x - 1 - x$ and $g(x) := e^x - 1 - x^2$ are nonnegative on $x \in [0, \infty)$.*

Proof. The statement regarding $f(x)$ follows directly from the Taylor expansion for e^x. Now, the first derivative of $g(x)$ is $e^x - 2x$. It is clearly positive for $x \geq 1$. For $x \in [0, 1]$,

$$e^x - 2x = \sum_{i=0}^{\infty} \frac{x^i}{i!} - 2x \geq 1 - x \geq 0.$$

Since $g(0) = 0$, the second result follows. □

Proof (Theorem 1). The inequality $\text{IG}_n(x) \leq \text{PG}_n(x)$ follows directly from Lemma 1, the nonnegativity of the Kullback-Leibler divergence, and the fact that all models in $\mathcal{M}$ are learning-positive. For the inequality $\text{PG}_n(x) \leq \hat{N}_n(x)^{-1}$, we write

$$\begin{aligned}
\hat{N}_n(x)^{-1} &= (1 - \xi_n'(x))^{-1} \frac{\xi_n'(x) - \xi_n(x)}{\xi_n(x)} \\
&= (1 - \xi_n'(x))^{-1} \left(\frac{\xi_n'(x)}{\xi_n(x)} - 1 \right) \\
&\overset{(a)}{=} (1 - \xi_n'(x))^{-1} \left(e^{\text{PG}_n(x)} - 1 \right) \\
&\overset{(b)}{\geq} e^{\text{PG}_n(x)} - 1 \\
&\overset{(c)}{\geq} \text{PG}_n(x),
\end{aligned}$$

where (a) follows by definition of prediction gain, (b) from $\xi_n'(x) \in [0, 1)$, and (c) from Lemma 2. Using the second part of Lemma 2 in (c) yields the inequality $\hat{N}_n(x)^{-1/2} \geq \text{PG}_n(x)$. □

B Asymptotic Analysis

We begin with a simple lemma which will prove useful throughout.

Lemma 3. *The rate of change of the empirical distribution, $\mu_n'(x) - \mu_n(x)$, is such that*

$$n\big(\mu_n'(x) - \mu_n(x)\big) = 1 - \mu_n'(x).$$

Proof. We expand the definition of μ_n and μ_n':

$$\begin{aligned}
n\big(\mu_n'(x) - \mu_n(x)\big) &= n \left[\frac{N_n(x) + 1}{n + 1} - \frac{N_n(x)}{n} \right] \\
&= \left[\frac{n}{n+1} \big(N_n(x) + 1\big) - N_n(x) \right] \\
&= \left[1 - \frac{N_n(x) + 1}{n + 1} \right] \\
&= 1 - \mu_n'(x).
\end{aligned}$$

□

Using this lemma, we derive an asymptotic relationship between N_n and $\hat{N}_n$.

Proof (Theorem 2). We expand the definition of $\hat{N}_n(x)$ and $N_n(x)$:

$$\begin{aligned}
\frac{\hat{N}_n(x)}{N_n(x)} &= \frac{\rho_n(x)(1 - \rho_n'(x))}{N_n(x)(\rho_n'(x) - \rho_n(x))} \\
&= \frac{\rho_n(x)(1 - \rho_n'(x))}{n\mu_n(x)(\rho_n'(x) - \rho_n(x))} \\
&= \frac{\rho_n(x)(\mu_n'(x) - \mu_n(x))}{\mu_n(x)(\rho_n'(x) - \rho_n(x))} \, \frac{1 - \rho_n'(x)}{n(\mu_n'(x) - \mu_n(x))} \\
&= \frac{\rho_n(x)}{\mu_n(x)} \, \frac{\mu_n'(x) - \mu_n(x)}{\rho_n'(x) - \rho_n(x)} \, \frac{1 - \rho_n'(x)}{1 - \mu_n'(x)},
\end{aligned}$$

with the last line following from Lemma 3. Under Assumption 1, all terms of the right-hand side converge as $n \to \infty$. Taking the limit on both sides,

$$\lim_{n\to\infty} \frac{\hat{N}_n(x)}{N_n(x)} \overset{(a)}{=} \frac{r(x)}{\dot{r}(x)} \lim_{n\to\infty} \frac{1-\rho'_n(x)}{1-\mu'_n(x)} \overset{(b)}{=} \frac{r(x)}{\dot{r}(x)} \frac{1-\mu(x)r(x)}{1-\mu(x)},$$

where (a) is justified by the existence of the relevant limits and $\dot{r}(x) > 0$, and (b) follows from writing $\rho'_n(x)$ as $\mu_n(x)\rho'_n(x)/\mu_n(x)$, where all limits involved exist. □

B.1 Directed Graphical Models

We say that $\mathcal{X}$ is a *factored* state space if it is the Cartesian product of k subspaces, i.e. $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_k$. This factored structure allows us to construct approximate density models over $\mathcal{X}$, for example by modelling the joint density as a product of marginals. We write the i^{th} factor of a state $x \in \mathcal{X}$ as x^i, and write the sequence of the i^{th} factor across $x_{1:n}$ as $x^i_{1:n}$.

We will show that directed graphical models (Wainwright and Jordan, 2008) satisfy Assumption 1. A directed graphical model describes a probability distribution over a factored state space. To the i^{th} factor x^i is associated a parent set $\pi(i) \subseteq \{1, \ldots, i-1\}$. Let $x^{\pi(i)}$ denote the value of the factors in the parent set. The i^{th} factor model is $\rho^i_n(x^i \,;\, x^{\pi(i)}) := \rho^i(x^i \,;\, x_{1:n}, x^{\pi(i)})$, with the understanding that ρ^i is allowed to make a different prediction for each value of $x^{\pi(i)}$. The state x is assigned the joint probability

$$\rho_{\text{GM}}(x \,;\, x_{1:n}) := \prod_{i=1}^{k} \rho^i_n(x^i \,;\, x^{\pi(i)}).$$

Common choices for ρ^i_n include the conditional empirical distribution and the Dirichlet estimator.

Proposition 1. *Suppose that each factor model ρ^i_n converges to the conditional probability distribution $\mu(x^i \mid x^{\pi(i)})$ and that for each x^i with $\mu(x^i \mid x^{\pi(i)})$,*

$$\lim_{n\to\infty} \frac{\rho^i(x^i \,;\, x_{1:n}x, x^{\pi(i)}) - \rho^i(x^i \,;\, x_{1:n}, x^{\pi(i)})}{\mu(x^i \,;\, x_{1:n}x, x^{\pi(i)}) - \mu(x^i \,;\, x_{1:n}, x^{\pi(i)})} = 1.$$

Then for all x with $\mu(x) > 0$, the density model ρ_{GM} satisfies Assumption 1 with

$$r(x) = \frac{\prod_{i=1}^k \mu(x^i \mid x^{\pi(i)})}{\mu(x)} \qquad \textit{and} \qquad \dot{r}(x) = \frac{\sum_{i=1}^k \left(1 - \mu(x^i \mid x^{\pi(i)})\right) \prod_{j\neq i} \mu(x^j \mid x^{\pi(j)})}{1-\mu(x)}.$$

The CTS density model used in our experiments is in fact a particular kind of induced graphical model. The result above thus describes how the pseudo-counts computed in Section 3.2 are asymptotically related to the empirical counts.

Proof. By hypothesis, $\rho^i_n \to \mu(x^i \mid x^{\pi(i)})$. Combining this with $\mu_n(x) \to \mu(x) > 0$,

$$\begin{aligned} r(x) &= \lim_{n\to\infty} \frac{\rho_{\text{DGM}}(x \,;\, x_{1:n})}{\mu_n(x)} \\ &= \lim_{n\to\infty} \frac{\prod_{i=1}^k \rho^i_n(x^i \,;\, x^{\pi(i)})}{\mu_n(x)} \\ &= \frac{\prod_{i=1}^k \mu(x^i \mid x^{\pi(i)})}{\mu(x)}. \end{aligned}$$

Similarly,

$$\begin{aligned} \dot{r}(x) &= \lim_{n\to\infty} \frac{\rho'_{\text{DGM}}(x \,;\, x_{1:n}) - \rho_{\text{DGM}}(x \,;\, x_{1:n})}{\mu'_n(x) - \mu_n(x)} \\ &\overset{(a)}{=} \lim_{n\to\infty} \frac{\left(\rho'_{\text{DGM}}(x \,;\, x_{1:n}) - \rho_{\text{DGM}}(x \,;\, x_{1:n})\right)n}{1-\mu'_n(x)} \\ &= \lim_{n\to\infty} \frac{\left(\rho'_{\text{DGM}}(x \,;\, x_{1:n}) - \rho_{\text{DGM}}(x \,;\, x_{1:n})\right)n}{1-\mu(x)}, \end{aligned}$$

where in (a) we used the identity $n(\mu'_n(x) - \mu_n(x)) = 1 - \mu'_n(x)$ derived in the proof of Theorem 2. Now

$$\begin{aligned}\dot{r}(x) &= (1-\mu(x))^{-1} \lim_{n\to\infty} \big(\rho'_{\text{DGM}}(x\,;\, x_{1:n}) - \rho_{\text{DGM}}(x\,;\, x_{1:n})\big)n \\ &= (1-\mu(x))^{-1} \lim_{n\to\infty} \big(\prod_{i=1}^{k} \rho^i(x^i\,;\, x_{1:n}x, x^{\pi(i)}) - \prod_{i=1}^{k} \rho^i(x^i\,;\, x_{1:n}, x^{\pi(i)})\big)n.\end{aligned}$$

Let $c_i := \rho^i(x^i\,;\, x_{1:n}, x^{\pi(i)})$ and $c'_i := \rho^i(x^i\,;\, x_{1:n}x, x^{\pi(i)})$. The difference of products above is

$$\begin{aligned}\big(\prod_{i=1}^{k} \rho^i(x^i\,;\, x_{1:n}x, x^{\pi(i)}) - \prod_{i=1}^{k} \rho^i(x^i\,;\, x_{1:n}, x^{\pi(i)})\big) &= \big(c'_1 c'_2 \dots c'_k - c_1 c_2 \dots c_k\big) \\ &= (c'_1 - c_1)(c'_2 \dots c'_k) + c_1(c'_2 \dots c'_k - c_2 \dots c_k) \\ &= \sum_{i=1}^{k} (c'_i - c_i)\Big(\prod_{j<i} c_j\Big)\Big(\prod_{j>i} c'_j\Big),\end{aligned}$$

and

$$\dot{r}(x) = (1-\mu(x))^{-1} \lim_{n\to\infty} \sum_{i=1}^{k} n(c'_i - c_i)\Big(\prod_{j<i} c_j\Big)\Big(\prod_{j>i} c'_j\Big).$$

By the hypothesis on the rate of change of ρ^i and the identity $n\left(\mu(x^i\,;\, x_{1:n}x, x^{\pi(i)}) - \mu(x^i\,;\, x_{1:n}, x^{\pi(i)})\right) = 1 - \mu(x^i \,|\, x^{\pi(i)})$, we have

$$\lim_{n\to\infty} n(c'_i - c_i) = 1 - \mu(x^i \,|\, x^{\pi(i)}).$$

Since the limits of c'_i and c_i are both $\mu(x^i \,|\, x^{\pi(i)})$, we deduce that

$$\dot{r}(x) = \frac{\sum_{i=1}^{k} \left(1 - \mu(x^i \,|\, x^{\pi(i)}\right) \prod_{j\neq i} \mu(x^j \,|\, x^{\pi_j(x)})}{1 - \mu(x)}.$$

Now, if $\mu(x) > 0$ then also $\mu(x^i\,;\, x^{\pi(i)}) > 0$ for each factor x^i. Hence $\dot{r}(x) > 0$. □

B.2 Tabular Density Models (Corollary 1)

We shall prove the following, which includes Corollary 1 as a special case.

Lemma 4. *Consider $\phi : \mathcal{X} \times \mathcal{X}^* \to \mathbb{R}^+$. Suppose that for all $(x_n : n \in \mathbb{N})$ and every $x \in \mathcal{X}$*

1. $\lim_{n\to\infty} \frac{1}{n} \sum_{x\in\mathcal{X}} \phi(x, x_{1:n}) = 0$*, and*
2. $\lim_{n\to\infty} \big(\phi(x, x_{1:n}x) - \phi(x, x_{1:n})\big) = 0$.

Let $\rho_n(x)$ be the count-based estimator

$$\rho_n(x) = \frac{N_n(x) + \phi(x, x_{1:n})}{n + \sum_{x\in\mathcal{X}} \phi(x, x_{1:n})}.$$

If $\hat{N}_n$ is the pseudo-count corresponding to ρ_n then $\hat{N}_n(x)/N_n(x) \to 1$ for all x with $\mu(x) > 0$.

Condition 2 is satisfied if $\phi_n(x, x_{1:n}) = u_n(x)\phi_n$ with ϕ_n monotonically increasing in n (but not too quickly!) and $u_n(x)$ converging to some distribution $u(x)$ for all sequences $(x_n : n \in \mathbb{N})$. This is the case for most tabular density models.

Proof. We will show that the condition on the rate of change required by Proposition 1 is satisfied under the stated conditions. Let $\phi_n(x) := \phi(x, x_{1:n})$, $\phi'_n(x) := \phi(x, x_{1:n}x)$, $\phi_n := \sum_{x\in\mathcal{X}} \phi_n(x)$ and $\phi'_n := \sum_{x\in\mathcal{X}} \phi'_n(x)$. By hypothesis,

$$\rho_n(x) = \frac{N_n(x) + \phi_n(x)}{n + \phi_n} \qquad \rho'_n(x) = \frac{N_n(x) + \phi'_n(x) + 1}{n + \phi'_n + 1}.$$

Note that we do not require $\phi_n(x) = \phi'_n(x)$. Now

$$\begin{aligned}
\rho'_n(x) - \rho_n(x) &= \frac{n+\phi_n}{n+\phi_n}\rho'_n(x) - \rho_n(x) \\
&= \frac{n+1+\phi'_n}{n+\phi_n}\rho'_n(x) - \rho_n(x) - \frac{(1+(\phi'_n-\phi_n))\rho'_n(x)}{n+\phi_n} \\
&= \frac{1}{n+\phi_n}\Big[(N_n(x)+1+\phi'_n(x) - (N_n(x)+\phi_n(x)) - (1+(\phi'_n-\phi_n))\rho'_n(x)\Big] \\
&= \frac{1}{n+\phi_n}\Big[1 - \rho'_n(x) + \big(\phi'_n(x)-\phi_n(x)\big) - \rho'_n(x)\big(\phi'_n-\phi_n\big)\Big].
\end{aligned}$$

Using Lemma 3 we deduce that

$$\frac{\rho'_n(x)-\rho_n(x)}{\mu'_n(x)-\mu_n(x)} = \frac{n}{n+\phi_n}\,\frac{1-\rho'_n(x)+\phi'_n(x)-\phi_n(x)+\rho'_n(x)(\phi'_n-\phi_n)}{1-\mu'_n(x)}.$$

Since $\phi_n = \sum_x \phi_n(x)$ and similarly for ϕ'_n, then $\phi'_n(x) - \phi_n(x) \to 0$ pointwise implies that $\phi'_n - \phi_n \to 0$ also. For any $\mu(x) > 0$,

$$\begin{aligned}
0 \le \lim_{n\to\infty} \frac{\phi_n(x)}{N_n(x)} &\overset{(a)}{\le} \lim_{n\to\infty} \frac{\sum_{x\in\mathcal{X}}\phi_n(x)}{N_n(x)} \\
&= \lim_{n\to\infty} \frac{\sum_{x\in\mathcal{X}}\phi_n(x)}{n}\frac{n}{N_n(x)} \\
&\overset{(b)}{=} 0,
\end{aligned}$$

where a) follows from $\phi_n(x) \ge 0$ and b) is justified by $n/N_n(x) \to \mu(x)^{-1} > 0$ and the hypothesis that $\sum_{x\in\mathcal{X}} \phi_n(x)/n \to 0$. Therefore $\rho_n(x) \to \mu(x)$. Hence

$$\lim_{n\to\infty} \frac{\rho'_n(x)-\rho_n(x)}{\mu'_n(x)-\mu_n(x)} = \lim_{n\to\infty} \frac{n}{n+\phi_n}\,\frac{1-\rho'_n(x)}{1-\mu'_n(x)} = 1.$$

Since $\rho_n(x) \to \mu(x)$, we further deduce from Theorem 2 that

$$\lim_{n\to\infty} \frac{\hat{N}_n(x)}{N_n(x)} = 1. \qquad \square$$

The condition $\mu(x) > 0$, which was also needed in Proposition 1, is necessary for the ratio to converge to 1: for example, if $N_n(x)$ grows as $O(\log n)$ but $\phi_n(x)$ grows as $O(\sqrt{n})$ (with $|\mathcal{X}|$ finite) then $\hat{N}_n(x)$ will grow as the larger $\sqrt{n}$.

C Experimental Methods

C.1 CTS Density Model

Our state space $\mathcal{X}$ is the set of all preprocessed Atari 2600 frames.[2] Each raw frame is composed of 210×160 7-bit NTSC pixels (Bellemare et al., 2013). We preprocess these frames by first converting them to grayscale (luminance), then downsampling to 42×42 by averaging over pixel values (Figure 5).

Aside from this preprocessing, our model is very similar to the model used by Bellemare et al. (2014) and Veness et al. (2015). The CTS density model treats $x \in \mathcal{X}$ as a factored state, where each (i, j) pixel corresponds to a factor $x^{i,j}$. The parents of this factor are its upper-left neighbours, i.e. pixels $(i-1, j)$, $(i, j-1)$, $(i-1, j-1)$ and $(i+1, j-1)$ (in this order). The probability of x is then the product of the probability assigned to its factors. Each factor is modelled using a location-dependent CTS model, which predicts the pixel's colour value conditional on some, all, or possibly none, of the pixel's parents (Figure 6).

[2]Technically, the ALE is partially observable and a frame is an observation, not a state. In many games, however, the current frame is sufficiently informative to guide exploration.

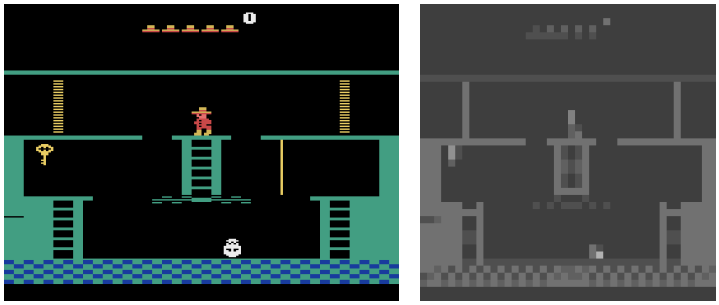

Figure 5: Sample preprocessed image provided to the CTS model (**right**), along with the original frame (**left**). Although details are lost, objects can still be made out.

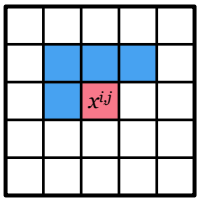

Figure 6: Depiction of the CTS "filter". Each downsampled pixel is predicted by a location-specific model which can condition on the pixel's immediate neighbours (in blue).

C.2 Reward Function and Monte Carlo Return

We added the extrinsic and intrinsic rewards together to produce a *combined reward* at each step. We clipped the resulting sum so that it lies within $[-1, 1]$ to ensure stable learning behaviour in DQN. Intrinsic rewards were computed at the end of each episode using a backward pass through the most recent episode; the resulting combined reward then replaces the extrinsic reward in the experience replay buffer. As a result, we did not allow DQN to update from incomplete episodes. We note that the resulting intrinsic rewards are slightly smaller than they would be if computed immediately after the transition. However, these particular choices were made for simplicity of implementation and should not meaningfully affect our experimental results.

Agents trained on Atari 2600 games benefit from eligibility traces and other mechanisms that propagate rewards from multiple steps ahead (Munos et al., 2016; Mnih et al., 2016; Wang et al., 2016). In our experiments we used a poor man's approximation to these ideas, namely we mix in 10% of the Monte Carlo return (computed from the experience replay buffer) together with the regular Double-DQN target. The target is thus

$$\textsc{Target}_{\textsc{mix}} = 0.9 \times \textsc{Target}_{\textsc{Double DQN}} + 0.1 \times \textsc{Target}_{\textsc{Monte Carlo}}.$$

The Monte Carlo return is the discounted sum of combined rewards along the episode, and is computed during the same backward pass used to compute the intrinsic rewards. This implementation has the advantage of computational and implementational simplicity, but we believe more elaborate schemes should improve our agents' performance.

C.3 A Taxonomy of Exploration

We provide in Table 1 a rough taxonomy of the Atari 2600 games available through the ALE in terms of the difficulty of exploration.

We first divided the games into two groups: those for which local exploration (e.g. ϵ-greedy) is sufficient to achieve a high scoring policy (*easy*), and those for which it is not (*hard*). For example, SPACE INVADERS versus PITFALL!. We further divided the *easy* group based on whether an ϵ-greedy scheme finds a *score exploit*, that is maximizes the score without achieving the game's

Easy Exploration			Hard Exploration	
Human-Optimal		Score Exploit	Dense Reward	Sparse Reward
Assault	Asterix	Beam Rider	Alien	Freeway
Asteroids	Atlantis	Kangaroo	Amidar	Gravitar
Battle Zone	Berzerk	Krull	Bank Heist	Montezuma's Revenge
Bowling	Boxing	Kung-Fu Master	Frostbite	Pitfall!
Breakout	Centipede	Road Runner	H.E.R.O.	Private Eye
Chopper Cmd	Crazy Climber	Seaquest	Ms. Pac-Man	Solaris
Defender	Demon Attack	Up n Down	Q*Bert	Venture
Double Dunk	Enduro	Tutankham	Surround	
Fishing Derby	Gopher		Wizard of Wor	
Ice Hockey	James Bond		Zaxxon	
Name this Game	Phoenix			
Pong	River Raid			
Robotank	Skiing			
Space Invaders	Stargunner			

Table 1: A rough taxonomy of Atari 2600 games according to their exploration difficulty.

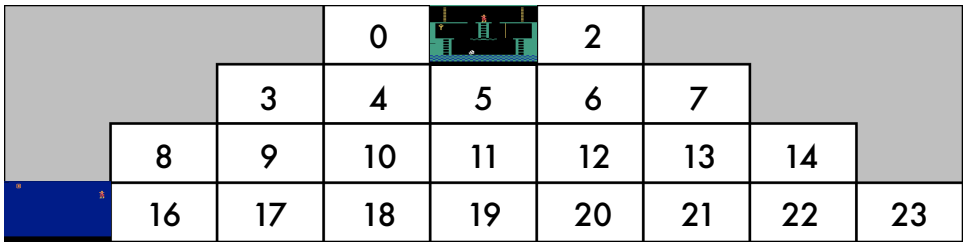

Figure 7: Layout of levels in Montezuma's Revenge, with rooms numbered from 0 to 23. The agent begins in room 1 and completes the level upon reaching room 15 (depicted).

stated objective. For example, Kung-Fu Master versus Boxing. While this distinction is not directly used here, score exploits lead to behaviours which are optimal from an ALE perspective but uninteresting to humans. We divide the games in the *hard* category into dense reward games (Ms. Pac-Man) and sparse reward games (Montezuma's Revenge).

C.4 Exploration in Montezuma's Revenge

Montezuma's Revenge is divided into three levels, each composed of 24 rooms arranged in a pyramidal shape (Figure 7). As discussed above, each room poses a number of challenges: to escape the very first room, the agent must climb ladders, dodge a creature, pick up a key, then backtrack to open one of two doors. The number of rooms reached by an agent is therefore a good measure of its ability. By accessing the game RAM, we recorded the location of the agent at each step during the course of training.[3] We computed the visit count to each room, averaged over epochs each lasting one million frames. From this information we constructed a map of the agent's "known world", that is, all rooms visited at least once. The agent's current room number ranges from 0 to 23 (Figure 7) and is stored at RAM location 0x83. Figure 8 shows the set of rooms explored by our DQN agents at different points during training.

Figure 8 paints a clear picture: after 50 million frames, the agent using exploration bonuses has seen a total of 15 rooms, while the no-bonus agent has seen two. At that point in time, our agent achieves an average score of **2461**; by 100 million frames, this figure stands at **3439**, higher than anything previously reported. We believe the success of our method in this game is a strong indicator of the usefulness of pseudo-counts for exploration.

We remark that without mixing in the Monte Carlo return, our bonus-based agent still explores significantly more than the no-bonus agent. However, the deep network seems unable to maintain a sufficiently good approximation to the value function, and performance quickly deteriorates. Comparable results using the A3C method provide another example of the practical importance of eligibility traces and return-based methods in reinforcement learning.

[3]We emphasize that the game RAM is not made available to the agent, and is solely used here in our behavioural analysis.

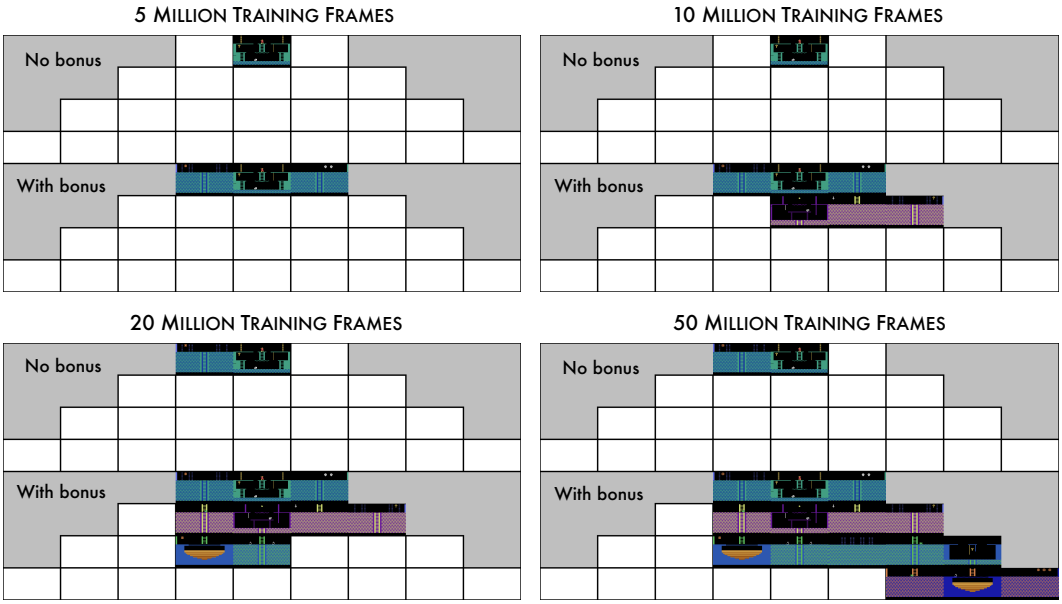

Figure 8: "Known world" of a DQN agent trained over time, with (**bottom**) and without (**top**) count-based exploration bonuses, in MONTEZUMA'S REVENGE.

C.5 Improving Exploration for Actor-Critic Methods

Our implementation of A3C was along the lines mentioned in Mnih et al. (2016) and uses 16 threads. Each thread corresponds to an actor learner and maintains a copy of the density model. All the threads are synchronized with the master thread at regular intervals of 250,000 steps. We followed the same training procedure as that reported in the A3C paper with the following additional steps: We update our density model with the states generated by following the policy. During the policy gradient step, we compute the intrinsic rewards by querying the density model and add it to the extrinsic rewards before clipping them in the range $[-1, 1]$ as was done in the A3C paper. This resulted in minimal overhead in computation costs and the memory footprint was manageable ($<$ 32 GB) for most of the Atari games. Our training times were almost the same as the ones reported in the A3C paper. We picked $\beta = 0.01$ after performing a short parameter sweep over the training games. The choice of training games is the same as mentioned in the A3C paper.

The games on which DQN achieves a score of 150% or less of the random score are: ASTEROIDS, DOUBLE DUNK, GRAVITAR, ICE HOCKEY, MONTEZUMA'S REVENGE, PITFALL!, SKIING, SURROUND, TENNIS, TIME PILOT.

The games on which A3C achieves a score of 150% or less of the random score are: BATTLE ZONE, BOWLING, ENDURO, FREEWAY, GRAVITAR, KANGAROO, PITFALL!, ROBOTANK, SKIING, SOLARIS, SURROUND, TENNIS, TIME PILOT, VENTURE.

The games on which A3C+ achieves a score of 150% or less of the random score are: DOUBLE DUNK, GRAVITAR, ICE HOCKEY, PITFALL!, SKIING, SOLARIS, SURROUND, TENNIS, TIME PILOT, VENTURE.

Our experiments involved the stochastic version of the Arcade Learning Environment (ALE) without a terminal signal for life loss, which is now the default ALE setting. Briefly, the stochasticity is achieved by accepting the agent action at each frame with probability $1 - p$ and using the agents previous action during rejection. We used the ALE's default value of $p = 0.25$ as has been previously used in Bellemare et al. (2016). For comparison, Table 2 also reports the deterministic + life loss setting also used in the literature.

Anecdotally, we found that using the life loss signal, while helpful in achieving high scores in some games, is detrimental in MONTEZUMA'S REVENGE. Recall that the life loss signal was used by

Mnih et al. (2015) to treat each of the agent' lives as a separate episode. For comparison, after 200 million frames A3C+ achieves the following average scores: 1) Stochastic + Life Loss: 142.50; 2) Deterministic + Life Loss: 273.70 3) Stochastic without Life Loss: 1127.05 4) Deterministic without Life Loss: 273.70. The maximum score achieved by 3) is 3600, in comparison to the maximum of 500 achieved by 1) and 3). This large discrepancy is not unsurprising when one considers that losing a life in MONTEZUMA'S REVENGE, and in fact in most games, is very different from restarting a new episode.

C.6 Comparing Exploration Bonuses

In this section we compare the effect of using different exploration bonuses derived from our density model. We consider the following variants:

- no exploration bonus,
- $\hat{N}_n(x)^{-1/2}$, as per MBIE-EB (Strehl and Littman, 2008);
- $\hat{N}_n(x)^{-1}$, as per BEB (Kolter and Ng, 2009); and
- $\text{PG}_n(x)$, related to compression progress (Schmidhuber, 2008).

The exact form of these bonuses is analogous to (4). We compare these variants after 10, 50, 100, and 200 million frames of training, again in the A3C setup. To compare scores across 60 games, we use inter-algorithm score distributions (Bellemare et al., 2013). Inter-algorithm scores are normalized so that 0 corresponds to the worst score on a game, and 1, to the best. If $g \in \{1, \ldots, m\}$ is a game and $z_{g,a}$ the inter-algorithm score on g for algorithm a, then the score distribution function is

$$f(x) := \frac{|\{g : z_{g,a} \geq x\}|}{m}.$$

The score distribution effectively depicts a kind of cumulative distribution, with a higher overall curve implying better scores across the gamut of Atari 2600 games. A higher curve at $x = 1$ implies top performance on more games; a higher curve at $x = 0$ indicates the algorithm does not perform poorly on many games. The scale parameter β was optimized to $\beta = 0.01$ for each variant separately.

Figure 10 shows that, while prediction gain initially achieves strong performance, by 50 million frames all three algorithms perform equally well. By 200 million frames, the $\hat{N}^{-1/2}$ exploration bonus outperforms both prediction gain and no bonus. The prediction gain achieves a decent, but not top-performing score on all games. This matches our earlier argument that using prediction gain results in too little exploration. We hypothesize that the poor performance of the $\hat{N}^{-1}$ bonus stems from too abrupt a decay from a large to small intrinsic reward, although more experiments are needed. As a whole, these results show how using PG offers an advantage over the baseline A3C algorithm, which is furthered by using our count-based exploration bonus.

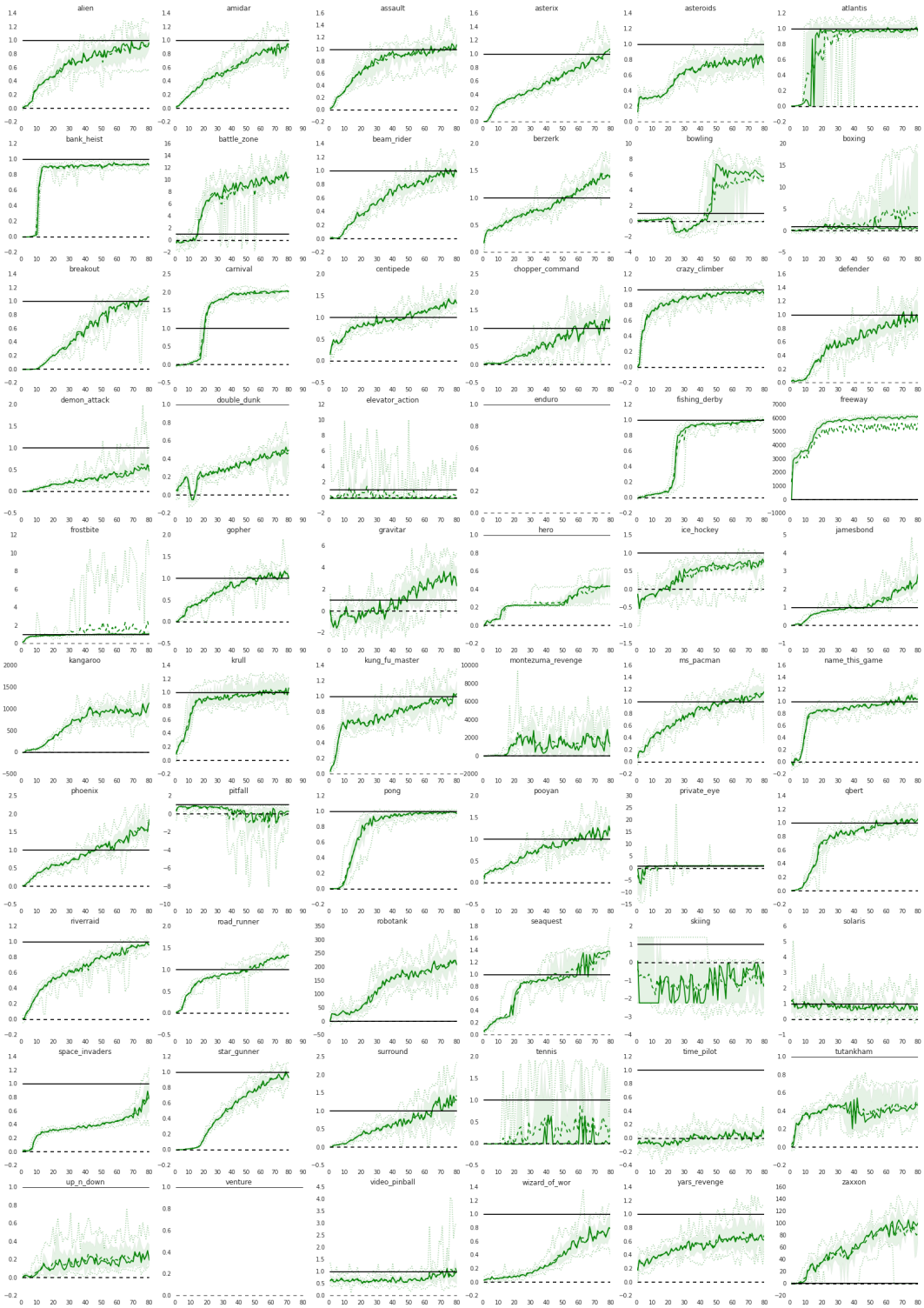

Figure 9: Average A3C+ score (solid line) over 200 million training frames, for all Atari 2600 games, normalized relative to the A3C baseline. Dotted lines denote min/max over seeds, inter-quartile range is shaded, and the median is dashed.

	Stochastic ALE			Deterministic ALE		
	A3C	A3C+	DQN	A3C	A3C+	DQN
ALIEN	**1968.40**	1848.33	1802.08	1658.25	**1945.66**	1418.47
AMIDAR	**1065.24**	964.77	781.76	**1034.15**	861.14	654.40
ASSAULT	**2660.55**	2607.28	1246.83	**2881.69**	2584.40	1707.87
ASTERIX	7212.45	**7262.77**	3256.07	**9546.96**	7922.70	4062.55
ASTEROIDS	**2680.72**	2257.92	525.09	**3946.22**	2406.57	735.05
ATLANTIS	**1752259.74**	1733528.71	77670.03	1634837.98	**1801392.35**	281448.80
BANK HEIST	**1071.89**	991.96	419.50	**1301.51**	1182.89	315.93
BATTLE ZONE	3142.95	7428.99	**16757.88**	3393.84	7969.06	**17927.46**
BEAM RIDER	**6129.51**	5992.08	4653.24	7004.58	6723.89	**7949.08**
BERZERK	1203.09	**1720.56**	416.03	1233.47	**1863.60**	471.76
BOWLING	32.91	**68.72**	29.07	35.00	**75.97**	30.34
BOXING	4.48	13.82	**66.13**	3.07	15.75	**80.17**
BREAKOUT	322.04	**323.21**	85.82	432.42	**473.93**	259.40
CENTIPEDE	4488.43	**5338.24**	4698.76	5184.76	**5442.94**	1184.46
CHOPPER COMMAND	4377.91	**5388.22**	1927.50	3324.24	**5088.17**	1569.84
CRAZY CLIMBER	**108896.28**	104083.51	86126.17	111493.76	**112885.03**	102736.12
DEFENDER	**42147.48**	36377.60	4593.79	**39388.08**	38976.66	6225.82
DEMON ATTACK	**26803.86**	19589.95	4831.12	**39293.17**	30930.33	6183.58
DOUBLE DUNK	**0.53**	-8.88	-11.57	**0.19**	-7.84	-13.99
ENDURO	0.00	**749.11**	348.30	0.00	**694.83**	441.24
FISHING DERBY	**30.42**	29.46	-27.83	**32.00**	31.11	-8.68
FREEWAY	0.00	27.33	**30.59**	0.00	**30.48**	30.12
FROSTBITE	290.02	506.61	**707.41**	283.99	325.42	**506.10**
GOPHER	5724.01	**5948.40**	3946.13	**6872.60**	6611.28	4946.39
GRAVITAR	204.65	**246.02**	43.04	201.29	**238.68**	219.39
H.E.R.O.	**32612.96**	15077.42	12140.76	**34880.51**	15210.62	11419.16
ICE HOCKEY	**-5.22**	-7.05	-9.78	**-5.13**	-6.45	-10.34
JAMES BOND	424.11	**1024.16**	511.76	422.42	**1001.19**	465.76
KANGAROO	47.19	**5475.73**	4170.09	46.63	4883.53	**5972.64**
KRULL	7263.37	**7587.58**	5775.23	7603.84	**8605.27**	6140.24
KUNG-FU MASTER	**26878.72**	26593.67	15125.08	**29369.90**	28615.43	11187.13
MONTEZUMA'S REVENGE	0.06	**142.50**	0.02	0.17	**273.70**	0.00
MS. PAC-MAN	2163.43	2380.58	**2480.39**	2327.80	**2401.04**	2391.89
NAME THIS GAME	6202.67	**6427.51**	3631.90	6087.31	**7021.30**	6565.41
PHOENIX	12169.75	**20300.72**	3015.64	13893.06	**23818.47**	7835.20
PITFALL	**-8.83**	-155.97	-84.40	**-6.98**	-259.09	-86.85
POOYAN	3706.93	**3943.37**	2817.36	4198.61	**4305.57**	2992.56
PONG	**18.21**	17.33	15.10	**20.84**	20.75	19.17
PRIVATE EYE	94.87	**100.00**	69.53	97.36	**99.32**	-12.86
Q*BERT	15007.55	**15804.72**	5259.18	19175.72	**19257.55**	7094.91
RIVER RAID	**10559.82**	10331.56	8934.68	**11902.24**	10712.54	2365.18
ROAD RUNNER	36933.62	**49029.74**	31613.83	41059.12	**50645.74**	24933.39
ROBOTANK	2.13	6.68	**50.80**	2.22	7.68	**40.53**
SEAQUEST	1680.84	**2274.06**	1180.70	1697.19	2015.55	**3035.32**
SKIING	-23669.98	**-20066.65**	-26402.39	**-20958.97**	-22177.50	-27972.63
SOLARIS	2156.96	**2175.70**	805.66	2102.13	**2270.15**	1752.72
SPACE INVADERS	**1653.59**	1466.01	1428.94	**1741.27**	1531.64	1101.43
STAR GUNNER	**55221.64**	52466.84	47557.16	**59218.08**	55233.43	40171.44
SURROUND	-7.79	**-6.99**	-8.77	**-7.10**	-7.21	-8.19
TENNIS	**-12.44**	-20.49	-12.98	-16.18	-23.06	**-8.00**
TIME PILOT	**7417.08**	3816.38	2808.92	**9000.91**	4103.00	4067.51
TUTANKHAM	**250.03**	132.67	70.84	**273.66**	112.14	75.21
UP AND DOWN	**34362.80**	8705.64	4139.20	**44883.40**	23106.24	5208.67
VENTURE	0.00	0.00	**54.86**	**0.00**	**0.00**	**0.00**
VIDEO PINBALL	53488.73	35515.91	**55326.08**	68287.63	**97372.80**	52995.08
WIZARD OF WOR	**4402.10**	3657.65	1231.23	**4347.76**	3355.09	378.70
YAR'S REVENGE	**19039.24**	12317.49	14236.94	**20006.02**	13398.73	15042.75
ZAXXON	121.35	**7956.05**	2333.52	152.11	**7451.25**	2481.40
Times Best	26	24	8	26	25	9

Table 2: Average score after 200 million training frames for A3C and A3C+ (with $\hat{N}_n^{-1/2}$ bonus), with a DQN baseline for comparison.

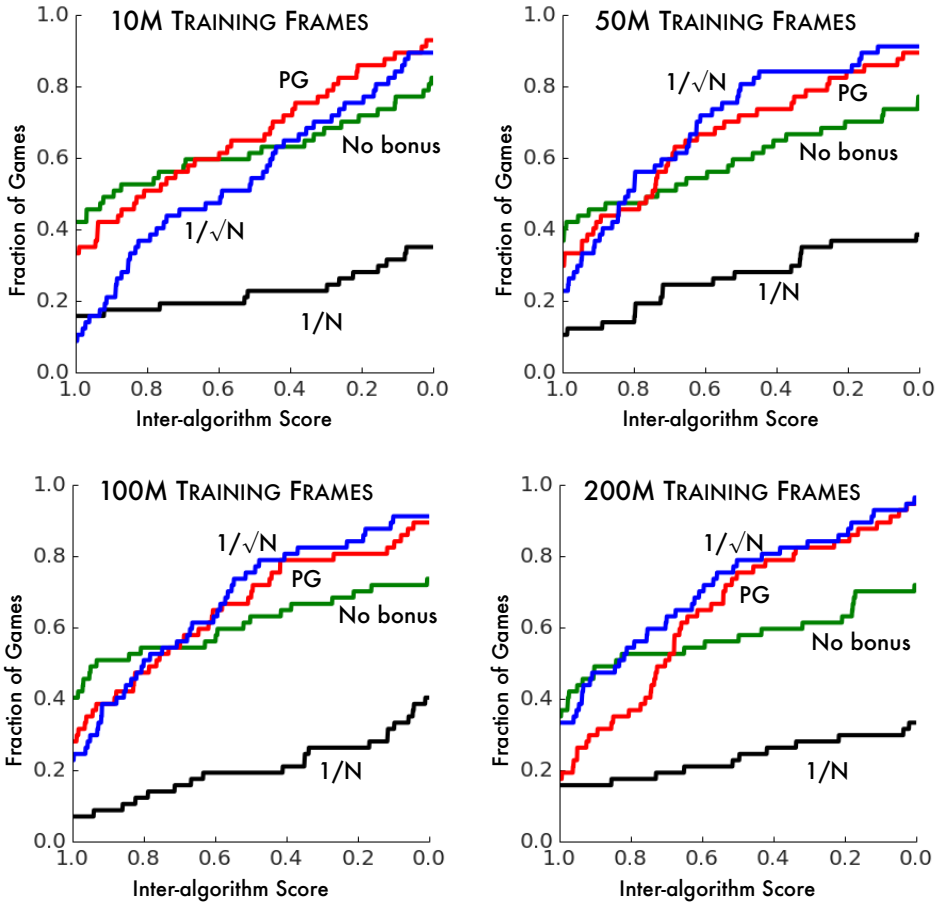

Figure 10: Inter-algorithm score distribution for exploration bonus variants. For all methods the point $f(0) = 1$ is omitted for clarity. See text for details.