A Theoretical Analysis of Information Bottlenecks for Zero-Shot Transfer in Reinforcement Learning

Kenzo Clauw, Daniel Polani, Nicola Catenacci Volpi
Adaptive Systems Research Group
University of Hertfordshire
United Kingdom
{k.clauw,d.polani,n.catenacci-volpi}@herts.ac.uk

Abstract

This work studies zero-shot policy transfer in reinforcement learning (RL), where a policy trained on multiple tasks is deployed on new tasks without fine-tuning. We study the Information Bottleneck (IB) framework as a probabilistic state abstraction, compressing observations into latent variables that retain task-relevant information. Our goal is to theoretically understand when and why such compressed representations enable policy transfer. We make three contributions: (1) an encoder-free metric that measures a priori task transferability using the Shannon divergence between action distributions, jointly with bisimulation distances comparing dynamics and rewards, (2) an encoder-dependent diagnostic metric, Latent-Action Divergence (LAD), which compares action distributions in the latent space to assess how well the learned abstraction preserves transferable behavior, and (3) a generalization bound showing that excess information, information retained in the latent representation that is not predictive of the policy, bounds the generalization error. Finally, we discuss empirically testable ideas motivated by our theoretical framework.

1 Introduction

Reinforcement learning (RL) has achieved remarkable successes in controlled benchmark environments, yet its deployment in real-world settings where trial-and-error interactions are costly or unsafe remains challenging Dulac-Arnold et al. [2021]. A key requirement for practical RL is *zero-shot policy transfer*: the ability to apply learned policies to novel tasks without further training Kirk et al. [2023]. However, most RL agents often overfit to spurious correlations or task-specific cues that fail to generalize even under slight distributional shifts Packer et al. [2018], Cobbe et al. [2019], Farebrother et al. [2020].

A promising approach to mitigating overfitting is to learn compressed state abstractions that retain only task-relevant information. Information-theoretic objectives, such as mutual information (MI) maximization, are commonly used for state representation learning in RL Mazoure et al. [2020], Laskin et al. [2022], as they encourage representations to capture relevant dependencies between states and tasks. However, maximizing MI alone often leads to representations that encode many irrelevant details, and it does not ensure that similar states are grouped together in a meaningful way Rakelly et al. [2021].

The information bottleneck (IB) provides a principled framework for learning probabilistic state abstractions that both compress and aggregate states into soft clusters based on their task relevance Tishby et al. [2000], Tishby and Zaslavsky [2015]. By filtering out irrelevant information while grouping states with shared functional similarity, IB enables agents to extract reusable knowledge that can facilitate transfer across tasks and policies. IB-based regularization has been widely applied

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

in RL to improve generalization and knowledge transfer Igl et al. [2019], Goyal et al. [2019], You and Liu [2024], Yingjun and Xinwen [2019], Lu et al. [2020a], Fan and Li [2022], Islam et al. [2022]. Despite these empirical successes, a rigorous theoretical understanding of *why* and *when* IB-regularized representations enables transfer remains largely missing. In practice, IBs are often applied and evaluated ad hoc through extensive hyperparameter tuning, and without guarantees on performance, where deploying such policies in novel tasks can be risky.

In this paper, we provide, to our knowledge, the first theoretical study of why information bottlenecks (IBs) improve transfer in reinforcement learning (RL). Our analysis builds on the empirical setup in [Igl et al., 2019], where the IB is applied as a policy regularizer to enhance generalization. To make theoretical analysis tractable, we focus on a simpler problem: zero-shot policy transfer, in which a policy trained on one or more source MDPs is evaluated directly on a target MDP without further training. Our central hypothesis is that transfer arises because the IB soft clusters action distributions in a compressed latent representation. This clustering allows different abstract policies to emerge during evaluation, which are sub-optimal yet sufficient to solve the task. It is precisely the trade-off between compression and sub-optimality that we aim to characterize. To formalize this phenomenon, we leverage rate–distortion theory Cover and Thomas [1991], Tishby et al. [2000], Zaslavsky and Tishby [2019] to identify critical points along the compression–performance trade-off where clusters emerge, providing a bound on the achievable sub-optimality that still enables successful transfer.

Contributions We provide three use cases grounded in rate–distortion theory, offering information-theoretic tools to diagnose, evaluate, and understand zero-shot policy transfer when using IB policy-level regularization:

Use Case 1: Diagnosing Task Transferability Before Training When designing benchmarks or evaluating transfer methods, fully training agents on every candidate task is often costly. We propose the Jensen–Shannon divergence of actions to quantify differences in policy distributions, which we connect to a bisimulation metric capturing behavioral similarity. These measures define a *transfer radius*, specifying the range over which a policy trained on one task is expected to transfer to others. This enables diagnosing which tasks are inherently transferable, independent of optimization or exploration effects, and supports designing benchmarks of tasks with controllable levels of transfer difficulty.

Use Case 2: Evaluating Representation Generalization After Training After training, it is crucial to know whether an agent has captured task-general structure or merely memorized the training tasks. To assess this theoretically, we introduce the Latent-Action Divergence (LAD) to measure the consistency of the latent policy across tasks, and pair it with a latent-space bisimulation metric computed in the same bottleneck latent space. Together, these metrics define a theoretical "generalization envelope" that can be visualized across a range of compression levels, indicating for each level which tasks the latent representation is expected to transfer to.

Use Case 3: Linking Compression to Generalization Error Practitioners need a formal understanding of how the compression of a latent bottleneck affects transferability with guarantees on its performance. We introduce a new quantity, the excessive information, which measures the extra information retained in the bottleneck that does not contribute to the policy's predictions and can lead to overfitting. Our theoretical framework relates excessive information to the LAD and the latent-space bisimulation metric from use case 2, showing that low values of these quantities are expected to be necessary for good transfer performance. From this, we derive a PAC-Bayes generalization bound that provides formal guarantees on performance transferability between tasks.

Outlook The use cases in this paper provide theoretical tools for applying information bottleneck principles in RL. While our focus is on theoretical analysis, we plan to empirically validate these measures in ongoing work. Additionally, our theoretical framework suggests several promising directions for future empirical studies.

2 Related Work

Information Bottleneck Prior work has demonstrated that the information bottleneck (IB) can enhance transferability in various RL settings, including policy regularization for generalization [Igl et al., 2019], goal-conditioned RL [Goyal et al., 2019], robust control [Eysenbach et al., 2021], and dynamics model learning [Lu et al., 2020b,a]. However, these contributions are primarily empirical, whereas our work provides a theoretical perspective on transfer.

Abel et al. [2018] uses the information bottleneck to provide performance guarantees when compressing an abstract policy to match a single demonstrator policy. In contrast, our work studies a zero-shot transfer setting, where the IB is used to compress states arising from multiple policies into abstract stochastic representations, enabling knowledge transfer to a target policy. While Abel et al. [2018] briefly mentions grouping multiple abstract policies and its potential benefit for transfer, we provide a theoretical framework to justify and formalize this process.

Bisimulation metrics Ferns and Precup [2014], Zhang et al. [2021], Castro [2019] provide a principled approach to state abstraction by quantifying behavioral equivalence between states. These metrics have been used to improve generalization in RL by identifying states that induce similar future dynamics and rewards. To our knowledge, no prior work has considered the use of stochastic encoders to learn bisimulation-based abstractions or connected them explicitly to the IB.

3 Preliminaries

3.1 Information Bottleneck

The Information Bottleneck (IB) method [Tishby et al., 2000] is a principled framework for extracting relevant information from a signal random variable X about a target variable Y. It seeks a stochastic encoding $p(Z|X)$ that produces a compressed representation Z while retaining relevant information about Y. The IB optimization can be written as a trade-off:

$$\min_{p(Z|X)} \mathcal{L}_{\text{IB}} = I(X;Z) - \beta I(Z;Y),$$

where the parameter $\beta \geq 0$ balances compression (via $I(X;Z)$) and predictive power (via $I(Z;Y)$).

3.2 State-Policy Information Bottleneck in RL

Let us denote with $S \in \mathcal{S}$ the state random variable and with $A^\pi \in \mathcal{A}$ the action random variable distributed according to the policy π . We consider the *state-policy information bottleneck* $Z \Leftarrow S \to A^\pi$, with source variable S, relevancy variable A^π and bottleneck variable Z. In this context, the IB method provides us with an optimal encoder $p^*(z \mid s)$ that compresses the state space S as much as possible by minimizing the mutual information $I(S;Z)$, while keeping a given amount of information about the policy $I(Z;A^\pi)$ specified by the value of β.

4 Quantifying generalization performance

What is the role of β in the state-policy IB? For the limit $\beta \to \infty$ all s that induce the same $p(a|s)$ share the same code value z and the resulting representation is a hard partition by policy invariances. For large β, two states with slightly different $p(a|s)$ may still be assigned the same z with high probability. By decreasing β further, the partition lines blur and a coarser grouping increases the action prediction error. The "distance" that decides whether two states fall into the same soft cluster is the IB information-distortion $D_{\text{KL}}\big(p(a|s_1)||p(a|s_2)\big)$, where as β grows, the effective tolerance β^{-1} shrinks.[1]

To formally study how the IB latent space policy leads to better transfer than a full-state policy, we need to study what MDPs fall in the generalisation offered by the IB method. Considering how the IB assign states to latent variable via the similarity of their action distributions, we want to define metrics that quantify the extent the same IB latent code z (almost) identifies states up to their action distribution $p(a|s)$ and investigate how this is related to successful transfer. In the following section we will define encoder-free metrics that justify why the IB policy should generalise, and encoder-dependent metrics that indicate whether the IB actually generalised.

[1] When β crosses a critical value the optimum bifurcates: a cluster splits because keeping those two subgroups together would now cost more in prediction accuracy than the β-weighted penalty for storing an extra bit. This deterministic-annealing picture is worked out in Zaslavsky and Tishby [2019]

4.1 Use case 1: Compatibility via Encoder-free Metrics for analysis before training

Since we know what determines the soft or hard partitioning of the states s in latent variables z (i.e., the similarity of their $p(a|s)$), we can define MDP metrics that predict when the representation computed by the IB on a MDP will transfer over another MDP. Given two MDPs $\mathcal{M}_1$ and $\mathcal{M}_2$, we want a numerical distance $D(\mathcal{M}_1, \mathcal{M}_2)$ that quantifies how much similar the two MDPs are when the IB transfer between the two is successful. Two complementary families of metrics have emerged in the recent literature and they slot neatly into the IB viewpoint. At the policy level, the expected *Jensen–Shannon Divergence (JSD) of Actions* measures how different two MDPs' policy distributions are. At dynamics level, the *Wasserstein bisimulation distance* Ferns and Precup [2014], Zhang et al. [2021], Castro [2019], measures how different rewards and transitions of the two MDPs are.[2]

The expected JSD of actions is defined as

$$D_{\text{JS}} \doteq \mathbb{E}_{s\sim\bar{p}}\big[\text{JSD}\big(p_1(a|s) \,\|\, p_2(a|s)\big)\big] \tag{1}$$

where, given two probability distributions P and Q defined over the same domain, the JSD is defined as $\text{JSD}(P\|Q) \doteq \frac{1}{2}D_{\text{KL}}(P\|M) + \frac{1}{2}D_{\text{KL}}(Q\|M)$, with $M \doteq \frac{1}{2}(P+Q)$ being the average (mixture) distribution, and $D_{\text{KL}}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence. In Equation 4.1, $\bar{p}(s) \doteq \frac{1}{2}(p_1(s) + p_2(s))$, with $p_k(s)$ denoting the state occupancy distribution generated by the action distribution $p_2(a|s)$ for $k = 1, 2$. The metric D_{JS} is zero if and only if the two tasks have identical $p(a \mid s)$ for every sampled state s, i.e. their hard IB partitions coincide. Note that this quantity can be estimated before training the IB.

The trade-off parameter β is the slope of the IB rate-distortion curve (i.e., they are related by the dual relation $\beta = R'(D)$), where the rate is $R \doteq I(S;Z)$ and the distortion is $D \doteq E_{z|s,s}\Big[D_{\text{KL}}\big(p(a \mid s) \,\|\, q(a \mid z)\big)\Big]$. Here, β indirectly controls an information-distortion radius via the allowed distortion budget $D_{\text{max}}(\beta)$: the IB encoder merges all states whose action distributions obey $D \leq D_{\text{max}}(\beta)$, where $D_{\text{max}}(\beta)$ in practice is chosen by setting β. Starting from these observations, in Equation (5) we bound the D_{JS} of two tasks in order to guarantee that the very same soft clustering will meet the distortion budget in both MDPs, so that transfer is guaranteed without re-tuning β. Conversely, when D_{JS} is above the provided bound, one must raise β (keep more bits) or learn a new task-dependent encoder.

Proposition 1(D_{JS}-radius transfer guarantee)

Given β and the corresponding distortion budget $D_{\text{max}}(\beta)$, let an encoder $p(z \mid s)$ and decoder $q(a \mid z)$ be trained on task $\mathcal{M}_1$ with average IB distortion

$$D_1 := \mathbb{E}_{s\sim p_1,\, z|s} D_{KL}\big[p_1(a \mid s) \| q(a \mid z)\big] \leq D_{\max}(\beta) \tag{2}$$

Suppose the target occupancy satisfies the density-ratio bound

$$r_{\max} := \sup_s \frac{p_2(s)}{p_1(s)} < \infty \tag{3}$$

Then, for the same frozen encoder,

$$D_2 \;\leq\; r_{\max} D_1 + 4D_{\text{JS}}, \tag{4}$$

where $D_2 := \mathbb{E}_{s\sim p_2,\, z|s} D_{KL}\big[p_2(a \mid s) \| q(a \mid z)\big]$. Consequently,

$$D_2 \;\leq\; D_{\max}(\beta) \quad \text{whenever} \quad D_{\text{JS}} \;\leq\; \frac{D_{\max}(\beta) - r_{\max} D_1}{4} \tag{5}$$

The proof of Proposition 1 is provided in Appendix A.

[2] Both metrics are easy to estimate from roll-outs in the two environments.

Hence, when D_{JS} satisfies inequality (5), the soft clustering found for MDP $\mathcal{M}_1$ automatically satisfy the distortion constraint in the second $\mathcal{M}_2$, and no β retuning is needed. If (5) fails, then the bound allows $D_2 > D_{\max}(\beta)$ and to enforce the budget one can raise β (store more bits) or relearn the encoder. The occupancy ratio $r_{\max}$ quantifies the extent $\mathcal{M}_2$ is a subset or truncation of $\mathcal{M}_1$ (i.e., when tasks differ only by removing states or shortening episodes). If task 2 is never denser than task 1 then $r_{\max} \leq 1$ and one obtains a tighter margin that enlarges the radius correspondingly. In this case one can show that the condition in Equation (5) simplifies to $D_{\text{JS}} \leq \frac{1}{4}\big(D_{\max} - D_1\big)$ and not extra constraint is needed. Here, when $r_{\max} \leq 1$, the positivity of the radius $D_{\max}(\beta) - r_{\max} D_1$ in Equation (5) is guaranteed by the the IB, whose optimal solutions satisfy $D_1 \leq D_{\max}(\beta)$. On the contrary, if $r_{\max} D_1 \geq D_{\max}$ then the bound is vacuous and a larger $D_{\max}$ (or stronger compression) is needed.

The bottleneck is only safe if two states that share a latent code also behave alike after the action. For this reason, we integrate the D_{JS} metric with the following behavioural bisimulation metric Ferns et al. [2004]. Both must be small for zero-shot generalisation. The one-step Wasserstein bisimulation metrics is defined as

$$D_{\text{Dyn}} = \max_a \Big| \mathbb{E}_{s\sim\bar{p}}\big[R_1(s,a) - R_2(s,a)\big] \Big| + \gamma\, \mathbb{E}_{s\sim\bar{p}}\big[W_1\big(P_1(\cdot \mid s,a), P_2(\cdot \mid s,a)\big)\big] \tag{6}$$

where W_1 is the 1-Wasserstein (earth-mover) distance. If two environments have ε-bisimilar dynamics in this sense, then any aggregation of states (in our case the IB encoder of $\mathcal{M}_1$) preserves value functions up to $\frac{\varepsilon}{1-\gamma}$ when ported to $\mathcal{M}_2$ Zhang et al. [2020], Ferns et al. [2004]. Hence a policy that is near-optimal on $\mathcal{M}_1$ stays near-optimal on $\mathcal{M}_2$ provided D_{JS} is also small. In other words, if $D_{\text{JS}} \leq \frac{D_{\max}(\beta) - r_{\max} D_1}{4}$ and $D_{\text{Dyn}} \leq \varepsilon$, then the IB encoder trained on $\mathcal{M}_1$ with that β yields a policy that is $\leq \varepsilon/(1-\gamma)$ sub-optimal on task $\mathcal{M}_2$. D_{JS} is a policy-centric metric, D_{Dyn} is an environment-centric metric and both must be small for painless zero-shot transfer. The pair $(D_{\text{JS}}, D_{\text{Dyn}})$ can be used as a pre-screen for task selection, answering the question: "which tasks would benefit from IB?".

4.2 Use case 2: Post-hoc diagnostics: encoder-dependent metrics

In this section, we employ aggregated metrics to enable post-hoc diagnostics of the quality of the latent representation Z learned from the source task. These can be used to provide an interpretation of why a transfer succeeds or fails. We want to quantify how much the same latent Z induces (almost) the same action distribution in the two MDPs, so that the latent policy $p(a|z)$ transfers. At the representation level, the *latent-action divergence* measures how different the action distributions conditioned on the latent are. At the dynamics level, the *Wasserstein bisimulation distance*, measures how different rewards and transitions after the latent abstraction are.

Given the same IB encoder $p(z \mid s)$ and policy $\pi(a \mid z)$ trained on $\mathcal{M}_1$, but two different environment-induced state distributions $p_k(s),\ k = 1, 2$, the Latent-Action Divergence (LAD) is the expectation (weighted by any non-zero mixture $\bar{p}(z)$) of the JSD between the two latent action distributions:

$$D_{\text{LAD}}(\mathcal{M}_1, \mathcal{M}_2) := \sum_z \bar{p}(z)\ \text{JSD}\big(p_1(a \mid z) \,\|\, p_2(a \mid z)\big) \tag{3}$$

where $p_k(a \mid z) = \frac{\sum_s p_k(s)\, p_\theta(z|s)\, \pi_\theta(a|s)}{p_k(z)}$ and $p_k(z) = \sum_s p_k(s)\, p_\theta(z \mid s)$. If $D_{\text{LAD}} = 0$, then $p_1(a \mid z) = p_2(a \mid z)$ for every z, hence the same decoder is optimal in both tasks. One can expect that a small LAD implies small loss in return when one transfers the policy.

Similarly, one can define the latent Wasserstein bisimulation metric as

$$d_B = \max_{z,a} \Big\{ |R_1(z,a) - R_2(z,a)| + \gamma\, W_1\big(P_1(\cdot \mid z,a), P_2(\cdot \mid z,a)\big) \Big\}, \tag{4}$$

where each MDP has been aggregated in the latent space using $R_k(z,a) = \sum_s w_k(s,z)\, R_k(s,a)$, $P_k(z' \mid z,a) = \sum_s w_k(s,z) \sum_{s'} P_k(s' \mid s,a)\, p_\theta(z' \mid s')$ and weights $w_k(s,z) = \dfrac{p_k(s)\, p_\theta(z \mid s)}{p_k(z)}$.

The pair (D_{LAD}, d_B) is a task-agnostic, quantitative gauge of whether two environments fall inside the "generalisation envelope" carved out by an IB representation. Armed with these two metrics one can map the β–transfer frontier that contains the test MDPs where the IB-compressive policy is guaranteed (up to the theoretical bounds) to work as well as it did on the source MDP. A low β implies large latent clusters and small average D_{LAD} and d_B, which in turns imply wide transfer radius but lower in-task accuracy. On the contrary, a large β implies more fine-grained clusters but possibly large cross-MDP distances — brittle transfer.

Let us consider a source MDP $\mathcal{M}_1$, a set of candidate target MDPs $\mathcal{M}$, and the distances D_{LAD}, d_B of each target MDP with the source. For each β, given the (D_{LAD}, d_B) frontiers (τ, τ') of MDPs that score a good transferred return $J_{\text{tr}}(\mathcal{M})$ (e.g., at least 90% of the source return $J_{\text{src}}(\mathcal{M})$), one can define the set of MDPs that reside within the frontier as

$$\mathcal{D}_{\beta,\,\tau,\tau'} \;=\; \big\{\,\mathcal{M} \;:\; D_{\text{LAD}}(\mathcal{M}_1, \mathcal{M}) \le \tau,\; d_B(\mathcal{M}_1, \mathcal{M}) \le \tau'\big\} \tag{7}$$

with $\tau = \max_{\mathcal{M}:\, J_{\text{tr}} \ge 0.9 J_{\text{src}}} D_{\text{LAD}}(\mathcal{M}_1, \mathcal{M})$, and $\tau' = \max_{\mathcal{M}:\, J_{\text{tr}} \ge 0.9 J_{\text{src}}} d_B(\mathcal{M}_1, \mathcal{M})$.

Almost all MDPs inside the area $[0, \tau] \times [0, \tau']$ of the $D_{\text{LAD}}\breve{}d_B$ plane yield $J_{\text{tr}} \ge 0.9 J_{\text{src}}$; points outside fail to do so.

4.3 Excess Information

Let us consider the IB compression term $I(S; Z)$ and its predictive power $I(Z; A^\pi)$. We define their difference as $g \doteq I(S; Z) - I(Z; A^\pi)$, which can be considered the the *excess information* that is stored about S but not needed for predicting A^π. If this gap is large and positive, the IB is keeping information about the state that is not useful for choosing actions (i.e., over-fitting to environment idiosyncrasies). In other words, g tells how much mutual information one is "paying for nothing." In general, to increase β pushes the solution towards sufficiency/minimality, hence reduces g[3]. A large β builds a latent Z that (almost) identifies states only up the similarity of their action distribution $p(a|s)$. For instance, all map layouts, colours, or other nuisance factors that do not change the action probabilities are lumped together — exactly the kind of invariance that helps when you move the policy to a new environment with the same dynamics but different visuals. As soon as β increases further, although $I(Z; A^\pi)$ almost plateaued, Z starts to remember extra facts about the raw state — e.g., wall textures, precise coordinates, random seeds — because those details still slightly improve the action prediction. They are useless (and even harmful) once the environment changes, hence generalisation degrades. This "redundant" capacity is exactly g, which is zero in the limiting invariant partition for $\beta \to \infty$ (Z is a minimal sufficient statistic for A and $I(Z; A) = I(S; A^\pi) = I(S; Z)$).[4]

4.4 Use Case 3: Excessive Information and Generalization Bound

In this section we connect the excess information g to out-of-distribution regret.

Achille and Soatto [2018b] derive an IB-based generalisation bound for supervised learning with n training samples,

$$\big|\,\text{Test loss} - \text{Train loss}\,\big| \;\le\; \sqrt{\frac{2\,I(S; Z)}{n}} \;+\; \text{higher-order terms.} \tag{8}$$

The same statement can be applied to a latent representation of a RL policy once one fixes a data-generating distribution $p(s, a)$ (e.g. the advantage-weighted occupancy used in Igl et al. [2019]). Equation 8 tells us that the absolute compression $I(S; Z)$ controls the worst-case generalisation gap. In other words, every extra bit of $I(S; Z)$ that is not needed for predicting A directly hurts the out-of-distribution performance. A large g therefore inflates the right-hand side of 8 without improving training performance—a clean, formal reason to expect poor transfer when the g is big. Conversely, driving g towards zero makes the bound tighter while preserving in-task reward. So, while Achille

[3] until the critical value where a cluster splits; after that both $I(S; Z)$ and $I(Z; A^\pi)$ drop together

[4] In practice we optimise a parametric variational lower-bound (VIB), with finite latent dimension, finite data and stochastic optimisation. These restrictions make it impossible to reach the exact sufficient/minimal point, so experiments usually show a gap that shrinks but that does not vanish.

and Soatto [2018a] bound links $I(S;Z)$ to the generalisation gap, in the case of generalization in RL, g sharpens the picture by isolating the useless part of that information. Matching it to the smallest $I(S;Z)$ that still preserves high $I(Z;A)$ is therefore a principled route to robustness. The following Proposition 2 formalises these concepts.

Proposition 2 after Achille and Soatto [2018a]. Let $S \to Z \to A$ be the encoder–decoder chain trained on n i.i.d. state samples from an MDP $\mathcal{M}_k$. Denote with g_k the excess information resulting from MDP $\mathcal{M}_k$. Indicate with $\text{Gap}(\mathcal{M}_k) \doteq \big|\ \mathcal{L}_{\text{test}} - \mathcal{L}_{\text{emp}}\ \big|$, where $\mathcal{L}_{\text{emp}}$ is the average negative advantage loss of the samples and $\mathcal{L}_{\text{test}}$ is the population negative advantage loss under the true state distribution $p_k(s)$ of MDP $\mathcal{M}_k$. Then, with probability > 0.95,

$$\text{Gap}(\mathcal{M}_k) \ \leq\ \sqrt{\tfrac{2}{n}}\ \big[I(Z;A) + g_k\big]^{1/2} \tag{9}$$

The proof of Proposition 2 is provided in Appendix B.

If the two metrics D_{LAD} or d_B are small, then $g_1 \approx g_2$ and the bound for the target task is close to that of the source task, predicting good transfer. Conversely large metrics show up exactly as an increase in g_2, loosening the generalisation guarantee.

5 Conclusion and Discussion

In this work, we introduced information-theoretic measures for analyzing information bottlenecks in RL. Before training, we proposed comparing action distributions via Jensen–Shannon divergence and bisimulation metrics to evaluate task transferability. After training, we introduced latent-action divergences to assess whether transfer was successful, and identified a new phenomenon, excessive information, which captures when compression exceeds what is necessary for generalization. Finally, we derived a generalization bound linking excessive information to policy generalization, providing a formal framework for understanding when and how bottlenecks support transfer. As this is ongoing work, we outline the following research directions:

5.1 Empirical Validation

So far, our framework has only been justified in theory. In ongoing work, we aim to validate its practical utility through controlled experiments. For empirical validation, we use the IBAC framework [Igl et al., 2019], which employs a Variational Information Bottleneck (VIB) to regularize the policy during training. We focus on the gridworld experiment defined in [Igl et al., 2019], where agents are trained on procedurally generated environments containing 1–3 connected rooms. We adapt this setup to a zero-shot policy transfer setting, training agents on procedurally generated single-room layouts and evaluating them on a diverse set of test tasks.

Test tasks are selected to vary in complexity, allowing us to systematically probe transferability. We start with in-distribution (IID) transfer, where room layouts are fixed but initial or goal positions vary. Next, we evaluate transfer across layouts with previously seen positions to isolate transfer to structural changes. Finally, we consider out-of-distribution (OOD) transfer, including inverted goals within the single-room layout and transfer from single to two-room layouts. To validate our framework, we progressively assess each use case in order of increasing complexity:

Use Case 1: Diagnosing Task Transferability Before Training As a feasibility test, we start by validating the pre-training measures and transfer radius. In this setting, we assume that the ground-truth expert policies are known, which allows us to evaluate the bottleneck encoder trained on N tasks by comparing its induced policy distributions against these experts. Concretely, we aim to estimate the transfer radius and the quantities in Equations 4.1 (D_{JS}) and 4.1 (D_{Dyn}) from rollouts generated by the expert policies in both tasks. To validate these bounds, we then train IBAC agents with different values of β and assess whether the observed transfer aligns with the predicted radius. While assuming access to the ground-truth policy is a strong simplification, this approach is valuable in its own right and could form the basis of a standalone study focused on systematically designing and evaluating benchmark tasks for zero-shot transfer.

Use Case 2: Evaluating Representation Generalization After Training Once feasibility is established, we turn to post-training validation. Here we measure whether the learned latent space captures task-general structure rather than memorizing training tasks. We validate the Latent-Action Divergence and a latent-space bisimulation metric within the bottleneck space. In addition, we compare these measures against direct visualizations of the latent space to test whether geometric structure aligns with our theoretical quantities. Together, these metrics define a *generalization envelope*, which we will empirically probe across compression levels.

Use Case 3: Linking Compression to Generalization Error Finally, we test our most challenging case: linking information compression to generalization. Here, we validate the notion of *excessive information*, quantifying information in the bottleneck that does not contribute to action prediction. We will compare excessive information across different β values of the bottleneck, evaluating its correspondence with LAD and the latent-space bisimulation metric. This use case directly probes our PAC-Bayes bound, testing whether low values of these quantities indeed predict stronger transfer performance.

Bridging theory and practice We anticipate several challenges in bridging theory and practice. To keep comparisons consistent across use cases, we replace the Wasserstein distance with the Jensen–Shannon divergence, putting all measures on a comparable scale. Since IBAC may not exactly match theoretical predictions due to approximations in the VIB, we first compute a near-optimal bottleneck using Blahut-Arimoto [Tishby et al., 2000] to estimate the measures. If this approach is insufficient, we fall back to a tabular setting with a discrete latent space similar to [Abel et al., 2018].

5.2 Adapting Theory for Empirical Validation

It is unclear whether our assumptions suffice for feasible empirical validation. On the theoretical side, we consider the following extensions to bridge this gap. First, extending the analysis to contextual MDPs and/or block MDPs is crucial. These settings capture structured variations in observations and latent states Diuk et al. [2008], Agarwal et al. [2020], which can simplify the analysis and reduce complexity when studying transfer across tasks. Second, the PAC-Bayes bound of use case 3 currently assumes a supervised learning setting. It is based on the loss, while in RL tasks are typically evaluated based on their performance. Similar to Abel et al. [2018], we aim to find the relationship between the distortion and the value error.

6 Future Work

The following section proposes empirical strategies originating from our theory, which may require practical adjustments to implement.

Excessive Information for Model Selection and Regularization Beyond quantifying generalization bounds, excessive information could have practical uses in RL. One direction is to use it as a regularizer in existing bottleneck methods, where it could act as a hyperparameter to be automatically tuned. A more challenging but potentially impactful direction is to leverage excessive information for model selection, particularly in offline RL, where overfitting is a significant concern. We hypothesize that this quantity may indicate when a model captures spurious correlations or memorizes the dataset, providing a criterion to select models that generalize better.

Probabilistic State Abstractions A key future direction is the empirical validation of probabilistic bottlenecks that capture both policy comparison and bisimulation metrics within a single latent representation. In our work, these objectives are derived from the same latent but are currently assumed to be optimized separately. To our knowledge, no existing methods use probabilistic embeddings to jointly represent policy and bisimulation information. Optimizing for both objectives simultaneously represents a promising research avenue. We expect this approach to be particularly effective in multi-task or meta-learning settings, where a shared probabilistic latent can help generalize across tasks and handle partial observability, a known challenge for generalization in RL Ghosh et al. [2021].

Task Transfer Benchmarks A natural future direction is to use the pre-training measures from Use Case 1 as a tool to compare MDPs before training and estimate their intrinsic transfer complexity. By evaluating tasks gradually in this way, one could systematically construct benchmarks that span a range of transfer difficulties. Such an approach has been instrumental in computer vision for guiding knowledge transfer across tasks [Zamir et al., 2018], but, to our knowledge, it has not yet been applied to RL. This is particularly important for RL generalization and multi-task learning, where it is often unclear a priori which tasks will transfer and benchmark suites are typically constructed in an ad hoc manner.

Value-Based RL Our analysis focused on policy-level bottlenecks, leaving open the question of how they might regularize value-based learning. Value-based RL faces persistent issues such as primacy bias[Nikishin et al., 2022], plasticity loss [Abbas et al., 2023], policy churn [Schaul et al.], and overestimation bias[Hasselt, 2010, Fujimoto et al., 2018]. These problems could arise from unstable or poorly generalized value representations. We hypothesize that bottlenecks could mitigate these pathologies by filtering out irrelevant features, stabilizing value estimates, and promoting information-efficient representations. Extending our framework to Q-functions, by relating excessive information to Bellman error [Fujimoto et al., 2022], may yield new regularization strategies and insights that improve stability and transfer in value-based RL.

References

Zaheer Abbas, Rosie Zhao, Joseph Modayil, Adam White, and Marlos C. Machado. Loss of Plasticity in Continual Deep Reinforcement Learning. In *Proceedings of The 2nd Conference on Lifelong Learning Agents*, pages 620–636. PMLR, November 2023. URL https://proceedings.mlr.press/v232/abbas23a.html. ISSN: 2640-3498.

David Abel, Dilip Arumugam, Lucas Lehnert, and Michael Littman. State Abstractions for Lifelong Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 10–19. PMLR, July 2018. URL https://proceedings.mlr.press/v80/abel18a.html. ISSN: 2640-3498.

Alessandro Achille and Stefano Soatto. Emergence of Invariance and Disentanglement in Deep Representations. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9, San Diego, CA, February 2018a. IEEE. ISBN 978-1-7281-0124-8. doi: 10.1109/ITA.2018.8503149. URL https://ieeexplore.ieee.org/document/8503149/.

Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40 (12):2897–2905, 2018b. Publisher: IEEE.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: Structural Complexity and Representation Learning of Low Rank MDPs. In *Advances in Neural Information Processing Systems*, volume 33, pages 20095–20107. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e894d787e2fd6c133af47140aa156f00-Abstract.html.

Pablo Samuel Castro. Scalable methods for computing state similarity in deterministic Markov Decision Processes, November 2019. URL http://arxiv.org/abs/1911.09291. arXiv:1911.09291 [cs, stat].

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying Generalization in Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1282–1289. PMLR, May 2019. URL https://proceedings.mlr.press/v97/cobbe19a.html. ISSN: 2640-3498.

Thomas M. Cover and Joy A. Thomas. Rate distortion theory. *Elements of Information Theory*, pages 336–373, 1991. URL https://static.ias.edu/pitp/archive/2012files/Cover_and_Thomas_Chptr13.pdf.

Carlos Diuk, Andre Cohen, and Michael L. Littman. An object-oriented representation for efficient reinforcement learning. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 240–247, Helsinki, Finland, 2008. ACM Press. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390187. URL http://portal.acm.org/citation.cfm?doid=1390156.1390187.

Gabriel Dulac-Arnold, Nir Levine, Daniel J. Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, September 2021. ISSN 1573-0565. doi: 10.1007/s10994-021-05961-4. URL https://doi.org/10.1007/s10994-021-05961-4.

Benjamin Eysenbach, R. Salakhutdinov, and S. Levine. Robust Predictable Control. September 2021. URL https://www.semanticscholar.org/paper/734d45f1f4aa62c9e629df90879fef259f6abbe9.

Jiameng Fan and Wenchao Li. DRIBO: Robust Deep Reinforcement Learning via Multi-View Information Bottleneck. In *Proceedings of the 39th International Conference on Machine Learning*, pages 6074–6102. PMLR, June 2022. URL https://proceedings.mlr.press/v162/fan22b.html. ISSN: 2640-3498.

Jesse Farebrother, Marlos C. Machado, and Michael Bowling. Generalization and Regularization in DQN, January 2020. URL http://arxiv.org/abs/1810.00123. arXiv:1810.00123 [cs].

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for Finite Markov Decision Processes. In *UAI*, volume 4, pages 162–169, 2004. URL https://cdn.aaai.org/AAAI/2004/AAAI04-124.pdf.

Norman Ferns and Doina Precup. Bisimulation Metrics are Optimal Value Functions. In *UAI*, pages 210–219, 2014. URL https://normferns.com/assets/documents/uai2014paper.pdf.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing Function Approximation Error in Actor-Critic Methods. In *Proceedings of the 35th International Conference on Machine Learning*, July 2018. URL https://proceedings.mlr.press/v80/fujimoto18a.html.

Scott Fujimoto, David Meger, Doina Precup, Ofir Nachum, and Shixiang Shane Gu. Why Should I Trust You, Bellman? The Bellman Error is a Poor Replacement for Value Error, June 2022. URL http://arxiv.org/abs/2201.12417. arXiv:2201.12417 [cs].

Dibya Ghosh, Jad Rahme, Aviral Kumar, Amy Zhang, Ryan P Adams, and Sergey Levine. Why Generalization in RL is Difficult: Epistemic POMDPs and Implicit Partial Observability. In *Advances in Neural Information Processing Systems*, volume 34, pages 25502–25515. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/d5ff135377d39f1de7372c95c74dd962-Abstract.html.

Anirudh Goyal, Riashat Islam, Daniel Strouse, Zafarali Ahmed, Matthew Botvinick, Hugo Larochelle, Yoshua Bengio, and Sergey Levine. InfoBot: Transfer and Exploration via the Information Bottleneck, April 2019. URL http://arxiv.org/abs/1901.10902. arXiv:1901.10902 [cs, stat].

Hado Hasselt. Double Q-learning. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/hash/091d584fced301b442654dd8c23b3fc9-Abstract.html.

Maximilian Igl, Kamil Ciosek, Yingzhen Li, Sebastian Tschiatschek, Cheng Zhang, Sam Devlin, and Katja Hofmann. Generalization in Reinforcement Learning with Selective Noise Injection and Information Bottleneck, October 2019. URL http://arxiv.org/abs/1910.12911. arXiv:1910.12911 [cs].

Riashat Islam, Hongyu Zang, Manan Tomar, Aniket Didolkar, Md Mofijul Islam, Samin Yeasar Arnob, Tariq Iqbal, Xin Li, Anirudh Goyal, and Nicolas Heess. Representation Learning in Deep RL via Discrete Information Bottleneck. *arXiv preprint arXiv:2212.13835*, 2022.

Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A Survey of Zero-shot Generalisation in Deep Reinforcement Learning. *Journal of Artificial Intelligence Research*, 76: 201–264, January 2023. ISSN 1076-9757. doi: 10.1613/jair.1.14174. URL https://www.jair.org/index.php/jair/article/view/14174.

Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. CIC: Contrastive Intrinsic Control for Unsupervised Skill Discovery, March 2022. URL http://arxiv.org/abs/2202.00161. arXiv:2202.00161.

Xingyu Lu, Kimin Lee, Pieter Abbeel, and Stas Tiomkin. Dynamics Generalization via Information Bottleneck in Deep Reinforcement Learning, August 2020a. URL http://arxiv.org/abs/2008.00614. arXiv:2008.00614 [cs].

Xingyu Lu, Stas Tiomkin, and Pieter Abbeel. *Generalization via Information Bottleneck in Deep Reinforcement Learning*. PhD Thesis, Master's thesis. University of California at Berkeley, 2020b.

Bogdan Mazoure, Remi Tachet des Combes, Thang Long Doan, Philip Bachman, and R Devon Hjelm. Deep Reinforcement and InfoMax Learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 3686–3698. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/26588e932c7ccfa1df309280702fe1b5-Abstract.html.

Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. The Primacy Bias in Deep Reinforcement Learning. In *Proceedings of the 39th International Conference on Machine Learning*, pages 16828–16847. PMLR, June 2022. URL https://proceedings.mlr.press/v162/nikishin22a.html. ISSN: 2640-3498.

Charles Packer, Katelyn Gao, Jernej Kos, Philipp Krähenbühl, V. Koltun, and D. Song. Assessing Generalization in Deep Reinforcement Learning. *ArXiv*, September 2018. URL https://www.semanticscholar.org/paper/caea502325b6a82b1b437c62585992609b5aa542.

Kate Rakelly, Abhishek Gupta, Carlos Florensa, and Sergey Levine. Which Mutual-Information Representation Learning Objectives are Sufficient for Control?, June 2021. URL http://arxiv.org/abs/2106.07278. arXiv:2106.07278 [cs].

Tom Schaul, Andre Barreto, John Quan, and Georg Ostrovski. The Phenomenon of Policy Churn. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/114292cf3f930ba157ed33f66997fee2-Abstract-Conference.html.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE, 2015. URL https://ieeexplore.ieee.org/abstract/document/7133169/.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, April 2000. URL http://arxiv.org/abs/physics/0004057. arXiv:physics/0004057.

Pei Yingjun and Hou Xinwen. Learning Representations in Reinforcement Learning:An Information Bottleneck Approach, November 2019. URL http://arxiv.org/abs/1911.05695. arXiv:1911.05695 [cs].

Bang You and Huaping Liu. Multimodal information bottleneck for deep reinforcement learning with multiple sensors. *Neural Networks*, 176:106347, August 2024. ISSN 0893-6080. doi: 10.1016/j.neunet.2024.106347. URL https://www.sciencedirect.com/science/article/pii/S0893608024002715.

Amir Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling Task Transfer Learning, April 2018. URL http://arxiv.org/abs/1804.08328. arXiv:1804.08328 [cs].

Noga Zaslavsky and Naftali Tishby. Deterministic annealing and the evolution of Information Bottleneck representations, 2019. URL https://www.nogsky.com/publication/2019-evo-ib/2019-evo-IB.pdf.

Amy Zhang, Clare Lyle, Shagun Sodhani, Angelos Filos, Marta Kwiatkowska, Joelle Pineau, Yarin Gal, and Doina Precup. Invariant Causal Prediction for Block MDPs. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11214–11224. PMLR, November 2020. URL https://proceedings.mlr.press/v119/zhang20t.html. ISSN: 2640-3498.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning Invariant Representations for Reinforcement Learning without Reconstruction, April 2021. URL http://arxiv.org/abs/2006.10742. arXiv:2006.10742 [cs].

A Proof of Jensen-Shannon Divergence of Actions Radius Transfer Guarantees

For any s, z the following triangle-type inequality holds:

$$D_{\mathrm{KL}}\big(p_2 \| q\big) \;\le\; D_{\mathrm{KL}}\big(p_1 \| q\big) + 2\,\mathrm{JSD}\big(p_1,\, p_2\big). \tag{10}$$

Take the conditional expectation over $p(z \mid s)$

$$\mathbb{E}_{z|s}\big[D_{\mathrm{KL}}\big(p_2 \| q\big)\big] \;\le\; \mathbb{E}_{z|s}\big[D_{\mathrm{KL}}\big(p_1 \| q\big)\big] + 2\,\mathrm{JSD}(p_1, p_2). \tag{11}$$

Then, take the expectation over the state distribution $s \sim p_2$, given by following the action distribution of MDP $\mathcal{M}_2$

$$D_2 \;\le\; \mathbb{E}_{s\sim p_2}\mathbb{E}_{z|s}\big[D_{\mathrm{KL}}\big(p_1 \| q\big)\big] + 2\mathbb{E}_{s\sim p_2}\big[\mathrm{JSD}\big(p_1, p_2\big)\big] \tag{12}$$

The first term of the rhs of (12) can be bound using the density ration as follows

$$\begin{aligned}
\mathbb{E}_{s\sim p_2}\mathbb{E}_{z|s}\big[D_{\mathrm{KL}}\big(p_1 \| q\big)\big] &= \sum_s p_2(s)\,\mathbb{E}_{z|s}\big[D_{\mathrm{KL}}\big(p_1 \| q\big)\big] \\
&= \sum_s r(s)\,p_1(s)\,\mathbb{E}_{z|s}\big[D_{\mathrm{KL}}\big(p_1 \| q\big)\big] \\
&\le r_{\max} \sum_s p_1(s)\,\mathbb{E}_{z|s}\big[D_{\mathrm{KL}}\big(p_1 \| q\big)\big] \\
&= r_{\max}\, D_1.
\end{aligned} \tag{13}$$

The second term of the rhs of (12) can be bound via the mixture $\bar{p} = \frac{1}{2}(p_1 + p_2) \ge \frac{1}{2} p_2$

$$\mathbb{E}_{s\sim p_2}\big[\mathrm{JSD}\big(p_1, p_2\big)\big] \;\le\; 2\,\mathbb{E}_{s\sim\bar{p}}\big[\mathrm{JSD}\big(p_1, p_2\big)\big] = 2\,D_{\mathrm{JS}}. \tag{14}$$

By combining the bounds (13) and (14) in (12) one obtains

$$D_2 \;\le\; r_{\max}\, D_1 + 4\,D_{\mathrm{JS}}. \tag{15}$$

Since we want the transfer bound $D_2 \le D_{\max}(\beta)$, a sufficient condition for the encoder to respect the same distortion budget $D_{\max}(\beta)$ in MDP $\mathcal{M}_2$ is

$$D_{\mathrm{JS}} \;\le\; \frac{D_{\max}(\beta) - r_{\max}\, D_1}{4} \qquad (\text{provided } r_{\max}\, D_1 < D_{\max}(\beta)). \tag{16}$$

B Proof of proposition 2

For any stochastic encoder $p_\theta(z \mid s)$ and deterministic decoder f_θ trained by empirical risk minimisation on i.i.d. samples $\mathcal{D}_n = \{s_i\}_{i=1}^{n}$,

$$\big|\,\mathcal{L}_{\text{test}} - \mathcal{L}_{\text{emp}}\big| \;\le\; \sqrt{\frac{2}{n}\, I(S;Z)} \tag{17}$$

The proof of 17 follows the same PAC-Bayes argument reported in Achille and Soatto [2018a]: with probability ≥ 0.95 over draws of $\mathcal{D}_n$,

$$\underbrace{\mathbb{E}_{s\sim p_k}\Big[\mathcal{L}\big(f_\theta(Z), s\big)\Big]}_{\text{test risk}} - \underbrace{\tfrac{1}{n}\sum_{i=1}^{n}\mathcal{L}\big(f_\theta(Z_i), s_i\big)}_{\text{empirical}} \;\le\; \sqrt{\tfrac{2}{n}\, I(S;Z)}\;, \tag{18}$$

where Z_i is the code of s_i. Although in the RL setting we set the negative advantage loss $\mathcal{L} = -A^\pi(s, a)$, the proof is identical to Achille and Soatto [2018a], because only the boundedness of the loss counts.

Because

$$I(S;Z) = \underbrace{I(Z;A)}_{\text{useful bits}} + \underbrace{I(S;Z) - I(Z;A)}_{=: g_k \quad \text{“excess” bits}}. \tag{19}$$

Insert 19 into 17 and take square roots:

$$\text{Gap}(\mathcal{M}_k) := \left|\mathcal{L}_{\text{test}} - \mathcal{L}_{\text{emp}}\right| \quad \sqrt{\frac{2}{n}\,[\,I(Z;A) + g_k\,]}, \tag{20}$$

which is exactly the inequality in 9 (the loose simply hides the 0.95 confidence constant).

C Proof of inequality (10)

Let $p \equiv p_2$, $r \equiv p_1$, $q \equiv q(a \mid z)$. Write

$$D_{\text{KL}}(p\|q) = \sum_a p(a) \log \frac{p(a)}{q(a)} = \sum_a p(a) \log \frac{p(a)}{m(a)} + \sum_a p(a) \log \frac{m(a)}{q(a)}, \tag{21}$$

with $m = \frac{1}{2}(p + r)$. The first term is $2\,D_{JS}(p, r)$. The second term can be upper-bounded by replacing p with r (which only increases the logarithm, because $x \mapsto x \log x$ is convex) and using the definition of KL:

$$\sum_a p(a) \log \frac{m(a)}{q(a)} \;\leq\; \sum_a r(a) \log \frac{r(a)}{q(a)} = D_{\text{KL}}(r\|q). \tag{22}$$

Putting the pieces together yields (10).