Improving Generalization in Imitation Learning

Gemmin

November 4, 2025

Abstract

Generalization remains a critical challenge in imitation learning (IL), where models must replicate diverse expert behaviors without overfitting to training data. While recent theoretical work has identified mutual information terms—between learned representations and inputs, and between model parameters and training data—as key factors bounding the generalization gap, most existing methods address these aspects separately and remain largely theoretical. In this work, we propose a novel regularizer that simultaneously accounts for both mutual information components, motivated by their joint role in controlling generalization. From the IL perspective, where generalization gaps critically limit performance, our approach adapts the Hessian Trace Information Bottleneck (HT-IB) framework to encourage compression of input representations while promoting flat minima in the loss landscape. Notably, the method flexibly balances these components, adapting to dataset characteristics such as output variability that influence the relative importance of compression and flatness. To empirically validate these theoretical insights, we conduct experiments on benchmark IL tasks, including autonomous hotlapping (placeholder environment). Our results demonstrate the complementary benefits of combining representation compression and flatness regularization, providing practical guidance for training robust IL policies. This study bridges the gap between recent theoretical advances and real-world performance in imitation learning.

1 Introduction

A fundamental challenge in imitation learning (IL) is generalization: models often overfit to the training demonstrations and fail to perform robustly in new or shifted environments. This generalization gap limits the practical deployment of IL systems, especially in safety-critical applications like autonomous driving.

Recent theoretical advances have identified key information-theoretic quantities that bound the generalization gap. In particular, the mutual information (MI) between the learned representation and the input, $I(Z;X)$, and between the model parameters and the training dataset, $I(\theta;D)$, have emerged as critical factors influencing generalization. Intuitively, compressing the representation to retain only task-relevant information while encouraging flat minima in the parameter space can reduce overfitting.

Despite these insights, existing works often treat these MI terms separately and remain largely theoretical, with limited empirical validation in realistic IL settings. To address this gap, we propose a novel regularizer based on the Hessian Trace Information Bottleneck (HT-IB) framework that simultaneously minimizes both $I(Z;X)$ and promotes flatness in the loss landscape. This dual approach is motivated by recent bounds on the generalization gap that incorporate both representation compression and parameter-level complexity.

We evaluate our method on benchmark IL tasks, including autonomous hotlapping (placeholder, can do Dr. Chen pathfinding bots), to empirically validate the theoretical predictions and explore the interplay between dataset conditional entropy, MI terms, and loss landscape geometry. Our results demonstrate that combining compression and flatness regularization yields improved generalization and stability, providing practical guidance for training robust imitation learning policies.

In summary, this work bridges the gap between theory and practice by integrating complementary information-theoretic regularizers within IL and empirically studying their joint impact on generalization.

2 Background and Definitions

2.1 Generalization Gap

In imitation learning, the generalization gap $\Delta(s)$ quantifies the difference between the expected loss on unseen data and the empirical loss on the training dataset s:

$$\Delta(s) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[\ell(f(X), Y)\right] - \frac{1}{n}\sum_{i=1}^{n}\ell(f(x_i), y_i),$$

where ℓ is the loss function, f the learned model, and $\mathcal{D}$ the data distribution. Minimizing $\Delta(s)$ is critical for robust performance.

Recent theoretical advances highlight key factors influencing generalization in imitation learning:

- **Representation Compression:** Compressing the intermediate representation of inputs X while preserving sufficient information to predict outputs Y reduces the upper bound on the generalization gap.
- **Encoder Robustness:** Encoders that are less dependent on the specific training dataset, yet still achieve low training loss, tend to generalize better.
- **Role of Conditional Entropy:** The conditional entropy $H(Y|X)$, which measures the intrinsic variability or multimodality of outputs given the same input, plays a central role. Larger $H(Y|X) \implies$
  - Provably leads to flatter likelihood landscapes, tightening generalization bounds by reducing encoder-dataset dependence.
  - Accelerates stochastic gradient descent's escape from sharp minima, improving optimization outcomes.

These findings suggest that promoting diversity not only in input states but also in actions conditioned on the same state fosters models that generalize more robustly. Building on these foundations, the following theorems provide formal bounds on the generalization gap in terms of information-theoretic quantities and properties of the loss landscape.

Theorem 1 If the encoder ϕ_l^s is independent of the training set s, then with probability at least $1-\delta$,

$$\Delta(s) \leq G_3^l\sqrt{\frac{1}{n}\left(I\left(X; Z_l^s \mid Y\right)\ln 2 + \mathcal{G}_2^l\right)} + \frac{G_1^l(0)}{\sqrt{n}}.$$

Theorem 2 For any $\delta > 0$, with probability at least $1-\delta$,

$$\Delta(s) \leq \min_l Q_l,$$

where Q_l depends on both $I\left(X; Z_l^s \mid Y\right)$ and $I\left(\phi_l^S; S\right)$.

Theorem 3 For parameters θ at local minimum $\hat{\theta}$ with Hessian $\mathcal{H}$, the mutual information between θ and dataset D satisfies

$$I(\theta; D) \leq \frac{1}{2}K\left[\log\|\hat{\theta}\|_2^2 + \log\operatorname{tr}(\mathcal{H}) - K\log\left(\frac{K^2\beta}{2}\right)\right].$$

2.2 Motivation

These bounds highlight that the generalization gap can be controlled by minimizing two intertwined terms: the mutual information between representations and inputs $I(Z; X)$, which encourages compression, and the mutual information between parameters and dataset $I(\theta; D)$, which relates to flatness of the loss landscape via Hessian trace. The practical goal is thus to find models that balance these information quantities to reduce overfitting while maintaining low training error.

3 Hessian Trace Information Bottleneck (HT-IB) Regularizer

Building on the theoretical motivation to jointly minimize the mutual information terms $I(Z;X)$ and $I(\theta;D)$, we propose the Hessian Trace Information Bottleneck (HT-IB) compressor. Our method encourages learning compressed, task-relevant representations while promoting flatness in the loss landscape to improve generalization in imitation learning.

3.1 Framework (Conceptual)

The HT-IB compressor integrates two complementary objectives:

1. **Representation Compression:**
   We apply an Information Bottleneck (IB) principle to the learned representation Z, encouraging the encoder $\phi(\cdot)$ to extract minimal sufficient statistics from input X relevant for predicting Y. This reduces $I(Z;X)$, limiting overfitting to irrelevant input details.

2. **Flat Minima via Hessian Trace Regularization:**
   Inspired by Theorem 3, we incorporate a regularizer based on the trace of the Hessian $\mathcal{H}$ of the loss with respect to model parameters θ. Minimizing $\mathrm{tr}(\mathcal{H})$ encourages flatter minima, thereby reducing $I(\theta;D)$ and improving robustness to training data variations.

3.2 Practical Implementation (Conceptual)

For the compression stage, a simple linear bottleneck can be employed as a proof of concept:

$$Z = WX,$$

where W is a learnable linear projection matrix.

The overall loss combines the standard imitation learning objective $\mathcal{L}_{IL}$ (e.g., behavior cloning loss) with the HT-IB regularizer:

$$\mathcal{L} = \mathcal{L}_{IL} + \lambda_1 I(Z;X) + \lambda_2 \,\mathrm{tr}(\mathcal{H}),$$

where λ_1, λ_2 control the trade-off between fitting training data, compressing representations, and encouraging flatness.

Estimating $I(Z;X)$ can be operationalized using variational approximations or upper bounds from the IB literature, while $\mathrm{tr}(\mathcal{H})$ can be efficiently approximated via Hutchinson's method or other stochastic estimators.

4 Methodology

4.1 Baseline Setup

We implement standard imitation learning algorithms such as Behavioral Cloning (BC) and DAgger on benchmark tasks relevant to our study (e.g., simulated autonomous racing or Dr. Chen's pathfinding agent). All models employ a pretrained encoder to isolate the effects of our proposed regularizers.

4.2 Estimating Dataset Condition Entropy

We estimate $\hat{H}(Y|X)$ for different datasets and data augmentation schemes to quantify output variability. We hypothesize that datasets with higher conditional entropy will naturally exhibit flatter loss landscapes, reducing the relative benefit of Hessian trace regularization while emphasizing the importance of representation compression. These hypotheses will be empirically tested by analyzing how variations in $H(Y|X)$ correlate with the effectiveness of our HT-IB regularizer components.

4.3 Training with HT-IB Regularizer Variants and Baselines

We train models under several regimes to assess the contributions of compression and flatness regularization:

- Baseline: No additional regularization beyond the imitation learning objective.
- Compression-only: Minimizing $I(X;Z)$ to encourage compact representations (e.g., using Variational Information Bottleneck (VIB) or PAC-Bayes IB methods where feasible).
- Flatness-only: Penalizing the Hessian trace $\mathrm{tr}(\mathcal{H})$ to promote flat minima (e.g., Sharpness-Aware Minimization (SAM) or CR-SAM as reference methods).
- Combined HT-IB: Jointly minimizing both $I(X;Z)$ and $\mathrm{tr}(\mathcal{H})$.

(NOTE: TONS OF PLACEHOLDERS)

4.4 Evaluation of Generalization

We evaluate the generalization gap on held-out and distribution-shifted test sets by measuring:

- Validation accuracy or task-specific performance metrics.
- Hessian trace estimates as a proxy for flatness.
- Estimates of mutual information terms to verify compression and parameter-data dependence.

4.5 Analysis of Interaction Effects and Training Dynamics

We analyze how dataset conditional entropy correlates with the effectiveness of compression and flatness regularization, identifying regimes where each regularizer individually or jointly offers the most benefit. Additionally, we measure convergence speed and training stability, as flatter loss landscapes are typically associated with more stable and faster convergence. These metrics provide practical insights into the training dynamics and robustness of each regularization approach.

5 References

Concise list of the important papers.

1. https://arxiv.org/abs/2504.18538 (base)
2. https://arxiv.org/abs/2310.01770v2 (dual perspective)
3. https://arxiv.org/abs/2208.05924 (hessian trace regularizer)
4. https://arxiv.org/abs/2505.09239 (ib regularizer)
5. https://arxiv.org/abs/1801.04062 (mi nn estimation)
6. https://arxiv.org/abs/2110.05437 (autonomous racing 1)
7. https://arxiv.org/abs/2509.16894 (autonomous racing 2)
8. https://ieeexplore.ieee.org/document/10869380 (pathfinding)