The Information Geometry of Local Generalization Dynamics

Emmanouil-Marios Athanasakos
National and Kapodistrian University of Athens,
Dpt. of Informatics and Telecommunications

Abstract

Information-theoretic bounds on generalization are foundational to learning theory, yet their static form offers limited insight into the dynamic, iterative nature of modern optimization. This gap is addressed herein by developing a local theory of generalization based on Euclidean Information Theory, where each update is modeled as a perturbation vector. The resulting analysis shows that the change in the generalization gap is bounded by the expected squared norm of this vector, a quantity interpreted as the local generalization cost. The proposed local bound is proven to be the first-order approximation of classic global bounds, revealing their underlying differential structure. Within this framework, it is further revealed that the optimal update direction is mathematically equivalent to the natural gradient, offering an information-geometric justification for natural gradient descent. Finally, the overall theory is validated through experiments showing that the derived bound closely tracks training dynamics.

1 INTRODUCTION

A central challenge in machine learning theory is to provide guarantees that models trained on finite datasets will generalize to unseen data. The difference between a model's performance on its training data and its expected performance on the true data distribution, is the primary measure of this capability and is termed as the generalization gap. While classical learning theory provided guarantees for simple models, the success of modern deep neural networks, which often fit their training data perfectly, has necessitated a new generation of theoretical tools to explain their remarkable generalization performance (Goldfeld and Polyanskiy, 2020; He and Goldfeld, 2025). Information-theoretic approaches have become a cornerstone of this effort, seeking to quantify generalization through the lens of information compression, as surveyed in the recent overview by Hellström et al. (2025).

The foundational works in this area, such as those by Russo and Zou (2016) and Xu and Raginsky (2017), established a deep connection between the generalization gap and the mutual information between the learned model and the training data. These global bounds provide significant insights but are often quantitatively loose for modern overparameterized models. This has spurred a fruitful line of research dedicated to deriving tighter bounds. One frontier of this research focuses on refining the information measures themselves, for example by conditioning on data subsets to remove statistical redundancies (Steinke and Zakynthinou, 2020; Hafez-Kolahi et al., 2020) or by developing sample-specific bounds to avoid the loosening effects of averaging (Bu et al., 2020). A parallel effort has achieved exactly tight bounds for canonical settings like the quadratic Gaussian problem by employing more powerful analytical tools, such as refined change-of-measure inequalities (Zhou et al., 2023; Dong et al., 2025).

The PAC-Bayesian framework (Catoni, 2007) has also proven highly effective, yielding non-vacuous bounds for deep networks by blending theory with practical estimation techniques (Dziugaite and Roy, 2017; Guedj, 2019). Extensions of this framework, such as those by Rivasplata et al. (2020), treat the generalization bound as a regularized objective to tune scalar hyperparameters to minimize the global bound. Recently, Sucker et al. (2023) and Picard-Weibel et al. (2024) developed surrogate training objectives that are explicitly optimized to satisfy PAC-Bayesian bounds. While these methods effectively regularize the final model, they

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

typically focus on the static objective rather than the dynamics of the learning process itself.

Despite these significant advances, most of this work provides a *static* characterization of the final, trained model. This creates a disconnect with the *dynamic*, iterative reality of modern learning, which is dominated by algorithms like stochastic gradient descent (SGD). A separate line of research has begun to address this by analyzing the generalization properties of entire training trajectories (Neu et al., 2021; Barnes et al., 2022), with extensions to diverse paradigms like transductive and continual learning (Tang and Liu, 2024; Wen et al., 2023). However, these full-trajectory analyses often yield complex, path-dependent bounds that do not easily reveal the immediate impact of an individual update. This gap in understanding the step-by-step nature of generalization remains a critical and challenging open problem, as highlighted by recent critiques of dynamic bounds (Haghifam et al., 2023).

This disconnect between static theory and iterative training prompts a series of fundamental questions:

1. Can we quantify the generalization impact of a *single* learning step?
2. Can a dynamic, local theory track how a model's generalization guarantee evolves step-by-step during training?
3. What constitutes a theoretically *optimal* learning step, and does this align with the behavior of practical algorithms like SGD?
4. How does such a local, dynamic theory connect to the established global, static bounds?

Answering these questions is essential for building a theory of learning that mirrors the practice of modern machine learning, potentially revealing not only why existing methods work but also how to design better ones.

In this work, a new local generalization bound is developed, derived to quantify the generalization impact of a single learning step. Unlike prior dynamic bounds which consider full training trajectories, the proposed approach is based on Euclidean Information Theory (EIT) (Borade and Zheng, 2008; Huang and Zheng, 2012; Huang et al., 2015) and focuses on the geometry of an individual update. Within this framework, a single learning step is treated as a perturbation vector in a local information space, enabling a precise analysis of small changes in model distributions. Moreover, rather than tuning scalar step-sizes, the optimal direction of the perturbation vector is derived. This provides a finer-grained, path-aware understanding of generalization that complements existing approaches.

The main contributions are as follows:

- A new, local generalization bound is derived by analyzing a single learning step through the geometric lens of EIT. The change in the generalization gap, is proven to be controlled by the expected squared Euclidean norm of a perturbation vector that characterizes the update, a quantity which is termed as the *local generalization cost*.
- This cost is shown to be the second-order approximation of the KL divergence between the pre- and post-update model distributions, which is a fundamental measure of complexity in the PAC-Bayesian framework.
- The optimal update direction—the one that maximizes empirical risk reduction for a fixed budget of the local generalization cost—is mathematically equivalent to the Natural Gradient on the probability simplex (Amari, 1998). While natural gradient methods are usually motivated by optimization efficiency, it is shown they are optimal for maintaining generalization stability.
- A formal connection is derived showing that the proposed local bound coincides with the first-order approximation of classical global bounds, thereby linking dynamic and static perspectives.
- Experimental verification of these local dynamics is provided, highlighting the framework's potential for adaptive optimization.

This work bridges the gap between static information-theoretic bounds and iterative optimization, providing a dynamic theory of generalization with both theoretical and practical implications.

2 PRELIMINARIES: GENERALIZATION AND LOCAL PERTURBATIONS

This section establishes the foundational concepts for the main results, including a definition of the generalization gap, the global information-theoretic bound that relates this gap to the information complexity of a learned representation, and the core mathematical tools of EIT used in the subsequent analysis.

2.1 Information Complexity and the Generalization Gap

Consider a standard supervised learning setting and let $Z = (X, Y)$ be a random variable representing data-

label pairs, defined on a space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The statistical properties of the data are described by a true, but unknown, probability distribution P_Z. A learning algorithm is trained on a dataset $S = \{Z_1, \ldots, Z_n\}$, consisting of n samples drawn i.i.d. from P_Z.

A model is a hypothesis h, that maps an input $x \in \mathcal{X}$ to a prediction. The quality of a prediction is measured by a loss function $l(h(x), y)$. The central challenge of learning theory is to provide guarantees that a model which performs well on the training set S will also perform well on new, unseen data from P_Z.

Definition 1. *The true risk of a hypothesis h is its expected loss on the true data distribution:*

$$R_{\text{true}}(h) \triangleq \mathbb{E}_{Z \sim P_Z}[l(h(X), Y)]. \tag{1}$$

The empirical risk of h on the training set S is its average loss on the observed data:

$$R_{\text{emp}}(h, S) \triangleq \frac{1}{n} \sum_{i=1}^{n} l(h(X_i), Y_i). \tag{2}$$

The generalization gap for hypothesis h is the difference between these two quantities:

$$G(h, S) \triangleq R_{\text{true}}(h) - R_{\text{emp}}(h, S). \tag{3}$$

In this work, models that produce an internal representation T of the input X are considered. A model is therefore defined by a stochastic map, $P_{T|X}(t|x)$. The information complexity of this representation is quantified by the mutual information $I(X;T)$, evaluated on the empirical distribution of the training data. From a key result in information-theoretic learning theory, a high-probability bound on the generalization gap is provided in terms of this complexity.

Theorem 2.1 (Xu and Raginsky (2017))**.** *With probability at least $1-\delta$ over the draw of the training set S, the generalization gap of any model $P_{T|X}$ is bounded by:*

$$|G(P_{T|X}, S)| \leq \mathcal{O}\left(\sqrt{\frac{I(X;T) + \log(1/\delta)}{n}}\right). \tag{4}$$

This result establishes $I(X;T)$ as a principled regularizer, and implies that constraining the information complexity is a direct mechanism for controlling overfitting and ensuring generalization.

2.2 Euclidean Information Theory: A Local Geometric Framework

The analysis of the *dynamics* of learning requires a tool to quantify small changes to a model. In this work, a single learning step is modeled as a local perturbation of the model's conditional distribution, and the framework of EIT is employed for this purpose.

Let $P_{T|X}^{(k)}(t|x)$ be the conditional PMF defining our model at iteration k and assume $P_{T|X}^{(k)}(t|x) > 0$ for all (t, x) in the support. An update yields a new distribution, $P_{T|X}^{(k+1)}(t|x)$. For a small update, the new distributions can be expressed as a linear perturbation:

$$P_{T|X}^{(k+1)}(t|x) = P_{T|X}^{(k)}(t|x) + \epsilon J_{T|X}(t|x), \tag{5}$$

where $\epsilon \in (0, 1)$ is a small positive scalar and $J_{T|X}(t|x)$ is a perturbation kernel. For $P_{T|X}^{(k+1)}$ to be a valid conditional PMF for any ϵ in a small neighborhood of zero, the kernel $J_{T|X}$ must satisfy two fundamental properties for each input $x \in \mathcal{X}$:

1. **Normalization Preservation:**

$$\sum_t P_{T|X}^{(k+1)}(t|x) = 1 \implies \sum_t J_{T|X}(t|x) = 0. \tag{6}$$

2. **Non-Negativity:**

$$P_{T|X}^{(k+1)}(t|x) \geq 0 \implies P_{T|X}^{(k)}(t|x) + \epsilon J_{T|X}(t|x) \geq 0. \tag{7}$$

Following the EIT methodology, it is convenient to work with a scaled and weighted version of this perturbation kernel, which has a natural geometric interpretation.

Definition 2. *For each input $x \in \mathcal{X}$, the perturbation vector $L_{T,x} \in \mathbb{R}^{|\mathcal{T}|}$ is defined as:*

$$L_{T,x}(t) \triangleq \frac{J_{T|X}(t|x)}{\sqrt{P_{T|X}^{(k)}(t|x)}}. \tag{8}$$

The normalization preservation constraint on $J_{T|X}$ translates to an orthogonality condition on $L_{T,x}$. By substituting (8)*, condition* (6) *becomes:*

$$\sum_t \sqrt{P_{T|X}^{(k)}(t|x)} L_{T,x}(t) = 0, \tag{9}$$

which implies that each perturbation vector $L_{T,x}$ must lie in the subspace orthogonal to the vector of square roots of the reference probabilities. This subspace is the tangent space of the probability simplex at the point $P_{T|X}^{(k)}(\cdot|x)$.

Additionally, the non-negativity constraint in (7) is crucial for ensuring the validity of the perturbed distribution. It is not an additional constraint on the direction of the perturbation vector $L_{T,x}$, but rather on

its magnitude ϵ. For any valid direction $L_{T,x}$ satisfying the orthogonality condition (9), one can always find a sufficiently small $\epsilon > 0$ such that the non-negativity condition holds for all $t \in \mathcal{T}$. The EIT framework is fundamentally a local analysis in the limit as $\epsilon \to 0$, where this condition is always met.

A local update is described by a set of perturbation vectors, $\{L_{T,x}\}_{x\in\mathcal{X}}$, where each $L_{T,x}$ characterizes the change in the conditional distribution $P_{T|X}(\cdot|x)$ while satisfying the necessary constraints. The fundamental insight of EIT, and the central tool for this analysis, is that the information-theoretic distance between the pre- and post-update distributions is locally approximated by a simple geometric quantity.

Lemma 2.2. *For a fixed input x, the KL divergence between the post-update and pre-update conditional distributions can be locally approximated as:*

$$\mathbb{D}_{\mathrm{KL}}\left(P_{T|X}^{(k+1)}(\cdot|x)\middle\|P_{T|X}^{(k)}(\cdot|x)\right) \approx \frac{\epsilon^2}{2}\|L_{T,x}\|_2^2, \quad (10)$$

where $\|L_{T,x}\|_2^2 \triangleq \sum_t L_{T,x}(t)^2$ is the Euclidean squared norm of the perturbation vector.

Proof. The proof is presented in Appendix A.1. □

As detailed in the proof, the local approximation error scales as $\mathcal{O}(\epsilon^3)$. For the small step sizes typically used in practical optimization regimes, this remainder is negligible compared to the $\mathcal{O}(\epsilon^2)$ generalization cost. Furthermore, because the local generalization cost scales quadratically with ϵ, maintaining a stable generalization budget as the empirical risk reduction slows down. This serves as theoretical motivation for employing learning rate decay.

The generalization bounds that are the focus of this work are controlled by the information cost averaged over the entire data distribution, i.e., the expectation of the KL-divergence from Lemma 2.2 over P_X. By the linearity of expectation, this average cost is locally approximated by the expected squared norm of the perturbation vectors, a quantity termed as the *local generalization cost*. This provides the bridge between the local, geometric analysis and the fundamental statistical quantities that govern generalization.

2.3 Assumptions and Applicability to Deep Learning

Before proceeding to the main results, it is crucial to clarify the underlying assumptions and the applicability of the developed framework to the highly non-convex settings typical of modern deep learning.

The theoretical framework herein rests on two primary assumptions. First, for the generalization bounds, the loss function is assumed to be bounded, and is normalized to the range $[0, 1]$ without loss of generality. This is a mild condition satisfied by standard classification losses such as the 0-1 loss or cross-entropy. Notably, the derivation of the local generalization cost does not require the loss to be smooth or convex.

Second, the EIT-based analysis is a local one; it characterizes the properties of an infinitesimal perturbation around a given reference distribution. This locality is a key strength when analyzing deep learning. While the global loss landscape of a deep neural network is highly non-convex, iterative optimizers like SGD operate locally by taking small steps at each iteration. The developed framework provides a characterization of the properties of these individual steps, and the results hold at any point along the optimization trajectory, regardless of the global non-convexity.

Finally, while the general theory does not assume any particular family of distributions, the analysis on the structure of the optimal update connects the optimal direction to the gradient of the loss. This specific interpretation naturally requires the loss function to be differentiable with respect to the model's outputs, an assumption that is standard for virtually all gradient-based training paradigms.

3 THE LOCAL DYNAMICS OF GENERALIZATION

The global information-theoretic bound presented in Section 2 provides a static characterization of a final, trained model. In this section, a complementary *dynamic* theory is developed, providing new local bounds that quantify the incremental change in a model's properties resulting from a single learning step. The analysis proceeds in three stages. First, a foundational lemma on the stability of the true risk is established via the EIT framework. Second, a new local generalization bound derived from PAC-Bayesian concentration inequalities. Finally, a formal connection between this local result and the bound of Theorem 2.1 is presented.

3.1 Local Risk Stability

The derivation begins by establishing a fundamental result on the stability of the true risk under small model updates. The following lemma proves that the expected squared norm of the EIT perturbation vectors directly controls the potential change in the model's performance on the true data distribution.

Lemma 3.1. *Consider a model update from iteration k to $k + 1$, described by a local perturbation of small magnitude ϵ and the set of perturbation vectors*

$\{L_{T,x}\}_{x\in\mathcal{X}}$. *The absolute change in the true risk is bounded by:*

$$|\Delta R_{\text{true}}| \leq \mathcal{O}\left(\sqrt{(\epsilon^2/2)\mathbb{E}_{P_X}\left[\|L_{T,X}\|_2^2\right]}\right). \tag{11}$$

Proof. The proof is provided in Appendix A.2 □

Lemma 3.1 provides a stability guarantee in geometric terms. It proves that a small update, as measured by the local generalization cost, can only lead to a correspondingly small change in the model's true risk. This result is the first step in the subsequent analysis; it establishes that the same geometric quantity that measures the information cost of an update also controls its impact on performance stability. This lemma therefore serves as the stepping stone for deriving the new generalization bound.

3.2 The Local Generalization Bound

Building on the stability principle of Lemma 3.1, the main theorem on the local dynamics of the generalization gap is now proven. This result provides a high-probability bound on the change in the generalization gap, controlled by the local generalization cost of the model update.

Theorem 3.2. *Consider a model update from iteration k to $k+1$, described by a local perturbation of small magnitude ϵ and the set of perturbation vectors $\{L_{T,x}\}_{x\in\mathcal{X}}$. With probability at least $1-\delta$ over the draw of a training set of size n, the absolute change in the generalization gap, $|\Delta G| \triangleq |G^{(k+1)} - G^{(k)}|$, is bounded by:*

$$|\Delta G| \leq \mathcal{O}\left(\sqrt{\frac{\epsilon^2\mathbb{E}_{P_X}\left[\|L_{T,X}\|_2^2\right]\log(2/\delta)}{n}}\right) \tag{12}$$

Proof. The proof is provided in Appendix A.3. □

Theorem 3.2 provides a dynamic, local, and geometrically grounded bound on the evolution of the generalization gap. Its significance is twofold. First, it provides a quantitative meaning to the concept of the cost of learning. It shows that the immediate impact of a learning step on the generalization guarantee is directly controlled by a simple geometric quantity: the expected squared norm of the perturbation that defines the update as a measure of complexity in the context of iterative learning. Second, it provides a new analytical tool. While global bounds characterize the properties of the final destination of the learning process, this local bound characterizes the properties of the path taken to get there. This provides new insights for analyzing the behavior of iterative optimizers. Crucially, as will be formally proven subsequently, the local generalization cost is not an arbitrary measure; it is the second-order approximation of the change in the classic mutual information, $\Delta I(X;T)$. Having established this, the next step is to formalize its connection to the established global bounds.

3.3 Connecting Local and Global bounds

The final step is to connect this new, local geometric quantity to the global bounds of learning theory. The following theorem proves that these are not separate concepts.

Theorem 3.3. *Let $f(I) \triangleq \mathcal{O}(\sqrt{I/n})$ be the function describing the global generalization bound from Theorem 2.1, where $I \triangleq I(X;T)$. The first-order Taylor expansion of this global bound around a point $I^{(k)}$ is given by:*

$$f(I^{(k+1)}) - f(I^{(k)}) \approx \left.\frac{\partial f}{\partial I}\right|_{I^{(k)}} \cdot \Delta I(X;T). \tag{13}$$

The local increase in information complexity, $\Delta I(X;T)$, is equivalent to the local generalization cost, which is also the second-order approximation of the PAC-Bayesian complexity term:

$$\begin{aligned}\Delta I(X;T) &\approx (\epsilon^2/2)\mathbb{E}_{P_X}[\|L_{T,X}\|_2^2] \\ &\approx \mathbb{E}_{P_X}[D_{KL}(P_{T|X}^{(k+1)}||P_{T|X}^{(k)})].\end{aligned} \tag{14}$$

Proof. The proof is provided in Appendix A.4. □

Theorem 3.3 provides an explicit relationship with the global bounds of works such as Xu and Raginsky (2017). It is proven that the local generalization cost is precisely the quantity that determines the local slope, or sensitivity, of the global generalization bound. This demonstrates that the presented local analysis is not an arbitrary approximation, but rather a lens that reveals the fundamental, shared first-order structure of both PAC-Bayesian and information-theoretic stability principles.

4 THE STRUCTURE OF OPTIMAL UPDATES

In Section 3, it was established that the local generalization cost of a model update is governed by the expected squared norm of the EIT perturbation vectors. This raises a crucial question for any learning algorithm: for a given budget of generalization cost, what is the optimal update direction to take in order to achieve the maximum possible reduction in empirical risk? Answering this question reveals a connection

between the presented information-geometric framework and the core mechanics of gradient-based optimization.

4.1 Characterizing the Change in Empirical Risk

The problem is formalized as a constrained optimization. The objective is to find a perturbation that maximizes the reduction in the empirical risk, which is equivalent to minimizing the change, $\Delta R_{\text{emp}} \triangleq R_{\text{emp}}^{(k+1)} - R_{\text{emp}}^{(k)}$. First, a local approximation for this change is derived. By definition, the change in empirical risk is:

$$\Delta R_{\text{emp}} = \frac{1}{n}\sum_{i=1}^{n}\Bigg(\mathbb{E}_{P_{T|X}^{(k+1)}(\cdot|x_i)}[l(T,y_i)] - \mathbb{E}_{P_{T|X}^{(k)}(\cdot|x_i)}[l(T,y_i)]\Bigg). \tag{15}$$

The term for a single training sample (x_i, y_i) is analyzed. Using (5) and the definition of perturbation vector in (8) the post-update expectation is:

$$\begin{aligned}
&\mathbb{E}_{P_{T|X}^{(k+1)}}[l(T,y_i)] = \sum_t l(t,y_i)P_{T|X}^{(k+1)}(t|x_i)\\
&\approx \sum_t l(t,y_i)\left(P^{(k)}(t|x_i) + \epsilon L_{T,x_i}(t)\sqrt{P^{(k)}(t|x_i)}\right)\\
&= \mathbb{E}_{P_{T|X}^{(k)}}[l(T,y_i)]\\
&+ \epsilon\sum_t\left(l(t,y_i)\sqrt{P^{(k)}(t|x_i)}\right)L_{T,x_i}(t).
\end{aligned} \tag{16}$$

Substituting this back into (15), the pre-update risk terms cancel, leaving the first-order approximation for the change:

$$\Delta R_{\text{emp}} \approx \frac{\epsilon}{n}\sum_{i=1}^{n}\sum_t L_{T,x_i}(t)l(t,y_i)\sqrt{P_{T|X}^{(k)}(t|x_i)}. \tag{17}$$

To reveal the geometric structure of this expression, the following are defined.

Definition 3 (Loss-Probability Vector)**.** *For each training sample $Z_i = (X_i, Y_i)$, let the vector $\mathbf{v}_i \in \mathbb{R}^{|\mathcal{T}|}$ with components $[\mathbf{v}_i]_t = l(t, Y_i)\sqrt{P_{T|X}^{(k)}(t|x_i)}$.*

The change in empirical risk is locally approximated by the average of the Euclidean inner products between the perturbation vector L_{T,X_i} and the loss-probability vector $\mathbf{v}_i$:

$$\Delta R_{\text{emp}} \approx \frac{\epsilon}{n}\sum_{i=1}^{n}\langle L_{T,X_i}, \mathbf{v}_i\rangle_2, \tag{18}$$

where $\langle \mathbf{a}, \mathbf{b}\rangle_2 \triangleq \sum_t a(t)b(t)$ is the Euclidean inner product. Maximizing the risk reduction is therefore equivalent to minimizing this average inner product. This leads to a formal optimization problem for the optimal update direction.

Problem 1. *For a fixed local generalization cost budget $C > 0$, the optimal set of perturbation vectors $\{L_{T,x_i}^*\}_{i=1}^n$ is the solution to:*

$$\textit{minimize} \quad \frac{1}{n}\sum_{i=1}^{n}\langle L_{T,X_i}, \mathbf{v}_i\rangle_2 \tag{19}$$

$$\textit{subject to} \quad \frac{1}{n}\sum_{i=1}^{n}\|L_{T,X_i}\|_2^2 = C. \tag{20}$$

The solution to this problem reveals the most efficient update direction in terms of its trade-off between performance gain and generalization cost. The following theorem provides the exact, closed-form solution.

Theorem 4.1. *Let the centered loss-probability vector for sample i be defined as $\tilde{\mathbf{v}}_i \triangleq \mathbf{v}_i - \mathbb{E}_{P_{T|X}^{(k)}(\cdot|x_i)}[\mathbf{v}_i]$. Let $S^2 = \frac{1}{n}\sum_{j=1}^{n}\|\tilde{\mathbf{v}}_j\|_2^2$ be the average squared Euclidean norm of these vectors. Assuming[1] $S^2 > 0$, the unique optimal perturbation vector L_{T,x_i}^* that solves Problem 1 is given by:*

$$L_{T,x_i}^* = -\sqrt{C}\frac{\tilde{\mathbf{v}}_i}{\sqrt{S^2}}. \tag{21}$$

Substituting this into (8), the optimal update direction J^* is proportional to the negative centered loss-probability vector $-l(t,y_i)P^{(k)}(t|x_i)$, which corresponds to the Natural Gradient Amari (1998) on the probability simplex, with its magnitude precisely determined by the budget C and the batch-wide loss variance.

Proof. The proof is provided in Appendix A.5. □

Theorem 4.1 forms a direct bridge between the information-geometric framework and the practical reality of modern machine learning. The centered loss-probability vector $\tilde{\mathbf{v}}_i$ represents the effective error signal for sample i. Thus, Theorem 4.1 proves that the most efficient way to reduce training error, for a fixed budget of generalization cost, is to perturb the model's output distribution in the direction precisely opposite to this error signal.

The significance of this result is that the gradient of the empirical risk with respect to the model's output

[1]The assumption that $S^2 > 0$ is a non-triviality condition. If $S^2 = 0$, it implies that all centered loss vectors $\tilde{\mathbf{l}}_i$ are zero. This corresponds to a state of zero empirical gradient, such as at a local minimum of the empirical risk, where no first-order improvement is possible.

probabilities, $\nabla_{P_{T|X}} R_{\text{emp}}$, is exactly this centered loss-probability vector. Therefore, the theorem provides a new insight into the nature of gradient descent. That is, gradient descent is not merely a heuristic for function minimization; it is an algorithm that, at every step, attempts to take the locally optimal path for maximizing performance improvement while satisfying a constraint on the damage to the model's generalization guarantee.

This provides a new layer of theoretical understanding to the most ubiquitous algorithm in modern AI. While the choice of learning rate (the magnitude of the step, related to budget C) remains a critical hyperparameter, this work proves that the direction chosen by gradient descent is fundamentally optimal from an information-geometric perspective. This result connects the abstract dynamics of generalization, as described in Section 3, to the mechanics of the algorithms used to train models in practice. This local, dynamic view may also offer new insights into the success of adaptive optimization methods, such as Adam (Kingma and Ba, 2015) and AdamW (Loshchilov and Hutter, 2019), which use statistics of the gradient to modulate the step size throughout training.

5 EXPERIMENTAL VALIDATION

The preceding sections have established the theoretical framework for understanding the local dynamics of generalization. In this section, a comprehensive empirical validation of the key theoretical results is provided for a standard machine learning task. Two distinct experiments were designed, each targeted at verifying a central claim of this work. First, a direct, dynamic verification of the main local generalization bound is presented by tracking its evolution during training. Second, an intuitive, visual confirmation of the optimal update direction is provided, demonstrating its connection to the practical behavior of SGD.

5.1 Experimental Setup

Task and Dataset: The MNIST dataset of handwritten digits (LeCun et al., 1998) was used. To ensure a clear and controlled experimental setting, the analysis was focused on the binary classification task of distinguishing the digits '3' and '8'. The dataset was split into a training set of 10,000 samples and a held-out test set of 2,000 samples.

Model Architecture: A simple Multi-Layer Perceptron (MLP) with one hidden layer of 128 neurons and a ReLU activation function was employed. The internal representation T is the 128-dimensional output of this hidden layer. The model was trained using the standard cross-entropy loss function.

Implementation Details: The model was implemented in PyTorch. For optimization, a standard SGD optimizer with a fixed learning rate of 0.01 and a batch size of 64 was used. To estimate the mutual information term $I(X;T)$ for the verification of the bound, a k-Nearest Neighbors (k-NN) based estimator (Kraskov et al., 2004) was employed, which is a standard non-parametric method for this task. All experiments were conducted on a machine with an Intel Core i7-3635QM processor and 16GB of RAM memory.

5.2 Verifying the Local Generalization Bound

The first experiment is designed to provide quantitative evidence for the main theoretical result. Theorem 3.2 predicts that the epoch-to-epoch change in the generalization gap, $|\Delta G|$, is controlled by the local generalization cost, which is approximated by the change in the mutual information, $\Delta I(X;T)$.

To verify this, the MLP model was trained for 20 epochs. The mutual information was estimated at the end of each epoch using the k-NN estimator with $k = 5$ on a random subset of 3000 training samples. From these measurements, the epoch-wise change in the generalization gap, $|\Delta G^{(k)}_{\text{meas}}|$, and the theoretical bound, which scales as $\sqrt{|\Delta I|/n}$, were computed.

The results are presented in Figure 1. The plot reveals two distinct phases of learning, both of which are consistent with the theory. In the initial epochs, the model undergoes rapid, large-scale learning. This is characterized by a significant drop in both the measured generalization gap change (blue solid line) and the theoretical bound (red dashed line). The strong correlation in this phase is a critical validation of the theory, showing that the large initial changes in the model's representations incur a correspondingly high generalization cost, as predicted. Following this, the optimizer enters a more stable regime where the change in the generalization gap becomes small on average. The theoretical bound correctly reflects this stabilization by also becoming smaller in magnitude. The spiky behavior observed in both curves, particularly in the bound, is an expected outcome. This variability can be attributed to a combination of factors, including the inherent stochasticity of the SGD updates, which causes occasional larger changes in the model's representations, and the known high variance of non-parametric mutual information estimators in high-dimensional spaces.

Ultimately, the experiment provides strong validation for the developed theory. By tracking the dynamics using a finite learning rate ($\eta = 0.01$), it provides empirical evidence that the theory remains robust and non-

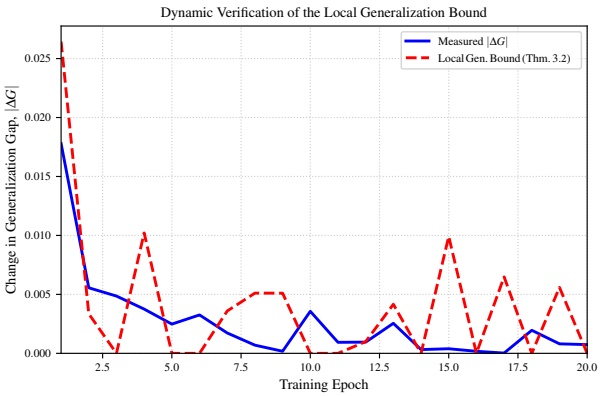

Figure 1: The solid blue line shows the empirically measured absolute change in the generalization gap, $|\Delta G|$, at each epoch. The dashed red line shows the predicted theoretical bound, which scales as $\mathcal{O}(\sqrt{|\Delta I|/n})$.

vacuous in practical training regimes where strictly infinitesimal assumptions are relaxed. The theoretical bound is shown to be both valid, as it is not violated by the empirical measurements, and non-vacuous, as it successfully captures the different orders of magnitude of the generalization dynamics across the training process. This provides evidence that the proposed local, information-geometric bound correctly characterizes the dynamics of generalization during iterative learning.

5.3 Visualizing the Principle of Aligned Updates

The second experiment validates the predicted optimal update direction. Theorem 4.1 proves that the optimal update for maximizing risk reduction is opposite to the centered loss vector, corresponding to the Natural Gradient in the representation space.

To visualize this, training was paused at an early epoch and a single mini-batch was considered. Two high-dimensional vectors were then computed: (i) the theoretical optimal direction, calculated as the negative natural gradient $-\tilde{\nabla}_T R_{\text{emp}}$, which is equivalent to the centered loss vector; and (ii) the actual SGD update direction, calculated as the average change in the representation, ΔT, resulting from a single SGD step on the network's weights. These 128-dimensional vectors were then projected into a 2D space for visualization using Principal Component Analysis (PCA), where the principal components were determined by the distribution of the representations in the mini-batch.

Figure 2 (a) shows the distribution of the mini-batch data in the projected 2D representation space, providing context for the update. Whereas sub-figure (b) provides a magnified view of the origin of this space, showing the two computed direction vectors. Visually, the two vectors appear strongly aligned. The quantitative measure, calculated in the full 128-dimensional space, confirms this, yielding a high cosine similarity of 0.867. This provides visual confirmation that the practical heuristic of SGD, which operates by updating the network's weights, is a remarkably effective implementation of the theoretically optimal, generalization-aware update strategy derived from the information-geometric principles developed herein. The slight deviation from perfect alignment is expected, reflecting the complex, non-linear mapping from the weight space to the representation space.

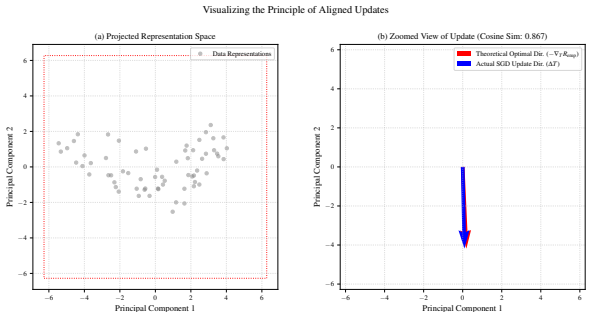

Figure 2: (a) A 2D PCA projection of the 128-dimensional representation space for a mini-batch of data. (b) A magnified view of the origin, showing that the Actual SGD Update Direction (blue) is strongly aligned with the Theoretical Optimal Direction (red).

6 CONCLUSION

This work has addressed the disconnect between the static nature of classic generalization bounds and the dynamic, iterative process of modern machine learning. A new, local generalization bound was introduced, which quantifies the information cost of a single learning step. The analysis was performed through the geometric lens of EIT, revealing that this cost is governed by the expected squared norm of a perturbation vector. This local bound was proven to be the first-order approximation of the established global bounds, thereby formally connecting the dynamic and static perspectives. The primary contribution of this work is the introduction of a new analytical tool—a dynamic, local theory of generalization—that allows for a more fine-grained understanding of the learning process. Within this framework, it is proven that the optimal update direction is mathematically equivalent to the Natural Gradient on the probability simplex, providing a new, information-geometric justification for the success of gradient-based optimization and was di-

rectly verified by experiments that successfully tracked its evolution during the training of a neural network. This local perspective opens up several exciting avenues for future research. The quantitative tools developed herein could form the basis for new, more principled adaptive optimization algorithms that explicitly control the local generalization cost at each step.

Acknowledgements

The author thanks the anonymous reviewers and the Area Chair for their feedback and constructive comments. The review process significantly improved the clarity and theoretical precision of this work.

References

Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276.

Barnes, L. P., Dytso, A., and Poor, H. V. (2022). Improved information-theoretic generalization bounds for distributed, federated, and iterative learning. *Entropy*, 24(9):1178.

Borade, S. and Zheng, L. (2008). Euclidean information theory. In *2008 International Zurich Seminar on Communications*, pages 112–115. IEEE.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Bu, Y., Zou, S., and Veeravalli, V. V. (2020). Tightening mutual information based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1):121–130.

Catoni, O. (2007). *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. HAL-Inria.

Dong, Y., Guo, H., Gong, T., Wen, W., and Li, C. (2025). Exactly tight information-theoretic generalization bounds via binary jensen-shannon divergence. In *Forty-second International Conference on Machine Learning*.

Dziugaite, G. K. and Roy, D. M. (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*.

Goldfeld, Z. and Polyanskiy, Y. (2020). Information bottleneck: Applications in machine learning. *Foundations and Trends® in Communications and Information Theory*, 17(4):394–537.

Guedj, B. (2019). A primer on pac-bayesian learning. *arXiv preprint arXiv:1901.05353*.

Hafez-Kolahi, H., Golgooni, Z., Kasaei, S., and Soleymani, M. (2020). Conditioning and processing: Techniques to improve information-theoretic generalization bounds. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 10344–10354.

Haghifam, M., Rodríguez-Gálvez, B., Thobaben, R., Skoglund, M., Roy, D. M., and Dziugaite, G. K. (2023). Limitations of information-theoretic generalization bounds for gradient descent methods in stochastic convex optimization. In *International Conference on Algorithmic Learning Theory (ALT)*, pages 663–706. PMLR.

He, H. and Goldfeld, Z. (2025). Information-Theoretic Generalization Bounds for Deep Neural Networks. *IEEE Transactions on Information Theory*. To appear. DOI: 10.1109/TIT.2025.3568697.

Hellström, F., Durisi, G., Guedj, B., and Raginsky, M. (2025). Generalization bounds: Perspectives from information theory and pac-bayes. *Foundations and Trends® in Machine Learning*, 18(1):1–223.

Huang, S., Suh, C., and Zheng, L. (2015). Euclidean information theory of networks. *IEEE Trans. on Inform. Theory*, 61(12):6795–6814.

Huang, S.-L. and Zheng, L. (2012). Linear information coupling problems. In *Proc. IEEE International Symposium on Information Theory*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015*. Conference Track.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69(6):066138.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Neu, G., Dziugaite, G. K., Haghifam, M., and Roy, D. M. (2021). Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory (COLT)*, pages 3326–3361. PMLR.

Picard-Weibel, A., Moscoviz, R., and Guedj, B. (2024). Learning via surrogate PAC-Bayes. In *Advances in Neural Information Processing Systems*, volume 37.

Rivasplata, O., Szepesvári, C., Shawe-Taylor, J., Malmgren, E., and Hazan, E. (2020). Pac-bayes generalization bounds for training algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 20303–20314.

Russo, D. and Zou, J. (2016). Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1232–1240. PMLR.

Steinke, T. and Zakynthinou, L. (2020). Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory (COLT)*, pages 3437–3452. PMLR.

Sucker, M., Fadili, J., and Ochs, P. (2023). Pac-bayesian learning of optimization algorithms. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3109–3134. PMLR.

Tang, H. and Liu, Y. (2024). Information-theoretic generalization bounds for transductive learning and its applications. *The Journal of Machine Learning Research*, 25(407):19902 – 19970.

Wen, W., Gong, T., Zhang, Y., Gao, Z., Zhang, W., and Liu, Y.-J. (2023). Information-theoretic generalization bounds of replay-based continual learning. *arXiv preprint arXiv:2305.12059*.

Xu, A. and Raginsky, M. (2017). Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Zhou, R., Tian, C., and Liu, T. (2023). Exactly tight information-theoretic generalization error bound for the quadratic gaussian problem. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 903–908.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes- Section 2 provides the mathematical setting and defines the core concepts. Section 2.3 explicitly consolidates all assumptions made in the paper. The model and algorithm used for experiments are detailed in Section 5.1.]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable- This is a theoretical paper focused on deriving new bounds and principles, a full complexity analysis is not part of this work's contribution.]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes- The source code to reproduce all experimental results can be provided.]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes-Section 2.3 provides a summary.]
   (b) Complete proofs of all theoretical results. [Yes-All theorems and lemmas presented in the main text are accompanied by full proofs in the appendices provided in the supplemental material.]
   (c) Clear explanations of any assumptions. [Yes-Each assumption is accompanied by a discussion explaining its context.]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes-The source code and instructions to reproduce the experiments will be provided as a URL.]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes-Section 5.1 explicitly details the dataset, data splits, model architecture, optimizer, learning rate, and batch size used for all experiments.]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes-The specific measures, such as the change in generalization gap $|\Delta G|$ and the mutual information $I(X;T)$, are formally defined and their estimation procedure is described. As this is a proof-of-concept validation, experiments were run once; error bars over multiple seeds are noted as a direction for future work.]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes-The computing hardware is described in Section 5.1.]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Yes-The MNIST dataset is used and is cited appropriately (LeCun et al. (1998)). The k-NN estimator method is also cited (Kraskov et al. (2004)).]

(b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Yes-The source code is provided as a new asset as a URL.]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary material

A Detailed Proofs

A.1 Proof of Lemma 2.2

Proof. Let $P^{(k)}(t) \triangleq P^{(k)}_{T|X}(t|x)$ and $P^{(k+1)}(t) \triangleq P^{(k+1)}_{T|X}(t|x)$. The KL-divergence is defined as:

$$D_{\mathrm{KL}}(P^{(k+1)}||P^{(k)}) = \sum_{t\in\mathcal{T}} P^{(k+1)}(t) \log\left(\frac{P^{(k+1)}(t)}{P^{(k)}(t)}\right). \tag{22}$$

The EIT framework parameterizes the post-update distribution as a local perturbation of the pre-update distribution by a small scalar magnitude ϵ:

$$P^{(k+1)}(t) = P^{(k)}(t) + \epsilon\sqrt{P^{(k)}(t)}L_{T,x}(t). \tag{23}$$

where the perturbation vector $L_{T,x}$ must satisfy the orthogonality condition in (9) to preserve normalization.

Let $f(\epsilon) \triangleq D_{\mathrm{KL}}(P^{(k+\epsilon)}||P^{(k)})$, the second-order Taylor series expansion around $\epsilon = 0$ is given by:

$$f(\epsilon) = f(0) + f'(0)\epsilon + \frac{1}{2}f''(0)\epsilon^2 + \mathcal{O}(\epsilon^3). \tag{24}$$

At $\epsilon = 0$, it holds that $P^{(k+0)} = P^{(k)}$, thus $D_{\mathrm{KL}}(P^{(k)}||P^{(k)}) = 0$ and

$$f(0) = 0. \tag{25}$$

The first derivative of $f(\epsilon)$ is given by:

$$\begin{aligned} f'(\epsilon) &= \frac{d}{d\epsilon}\sum_t P^{(k+\epsilon)}(t)\log\left(\frac{P^{(k+\epsilon)}(t)}{P^{(k)}(t)}\right) \\ &= \sum_t \left[\frac{dP^{(k+\epsilon)}(t)}{d\epsilon}\log\left(\frac{P^{(k+\epsilon)}(t)}{P^{(k)}(t)}\right) + P^{(k+\epsilon)}(t)\frac{P^{(k)}(t)}{P^{(k+\epsilon)}(t)}\frac{1}{P^{(k)}(t)}\frac{dP^{(k+\epsilon)}(t)}{d\epsilon}\right] \\ &= \sum_t \frac{dP^{(k+\epsilon)}(t)}{d\epsilon}\left[\log\left(\frac{P^{(k+\epsilon)}(t)}{P^{(k)}(t)}\right) + 1\right]. \end{aligned} \tag{26}$$

From (23), the derivative is $\frac{dP^{(k+\epsilon)}(t)}{d\epsilon} = \sqrt{P^{(k)}(t)}L_{T,x}(t)$. Evaluating at $\epsilon = 0$:

$$\begin{aligned} f'(0) &= \sum_t \sqrt{P^{(k)}(t)}L_{T,x}(t)\left[\log\left(\frac{P^{(k)}(t)}{P^{(k)}(t)}\right) + 1\right] \\ &= \sum_t \sqrt{P^{(k)}(t)}L_{T,x}(t)\left[\log(1) + 1\right] \\ &= \sum_t \sqrt{P^{(k)}(t)}L_{T,x}(t). \end{aligned} \tag{27}$$

Due to (9), the sum in (27) is exactly zero, that is

$$f'(0) = 0. \tag{28}$$

The second derivative of $f(\epsilon)$ is given by:

$$f''(\epsilon) = \frac{d}{d\epsilon}\left(\sum_t \frac{dP^{(k+\epsilon)}(t)}{d\epsilon}\left[\log\left(\frac{P^{(k+\epsilon)}(t)}{P^{(k)}(t)}\right)+1\right]\right). \tag{29}$$

Since $\frac{dP^{(k+\epsilon)}(t)}{d\epsilon}$ is a constant with respect to ϵ, it only remains to differentiate the term in the brackets:

$$f''(\epsilon) = \sum_t \frac{dP^{(k+\epsilon)}(t)}{d\epsilon}\left[\frac{1}{P^{(k+\epsilon)}(t)}\frac{dP^{(k+\epsilon)}(t)}{d\epsilon}\right] = \sum_t \frac{\left(\frac{dP^{(k+\epsilon)}(t)}{d\epsilon}\right)^2}{P^{(k+\epsilon)}(t)}. \tag{30}$$

Now, evaluating at $\epsilon = 0$:

$$\begin{aligned} f''(0) &= \sum_t \frac{\left(\sqrt{P^{(k)}(t)}L_{T,x}(t)\right)^2}{P^{(k)}(t)} \\ &= \sum_t \frac{P^{(k)}(t)L_{T,x}(t)^2}{P^{(k)}(t)} = \sum_t L_{T,x}(t)^2. \end{aligned} \tag{31}$$

By definition, (31) is the Euclidean squared norm of the perturbation vector:

$$f''(0) = \|L_{T,x}\|_2^2. \tag{32}$$

Substituting (25), (28) and (32) in (24) yields:

$$D_{\mathrm{KL}}(P^{(k+\epsilon)}||P^{(k)}) = \frac{\epsilon^2}{2}\|L_{T,x}\|_2^2 + \mathcal{O}(\epsilon^3). \tag{33}$$

The EIT framework provides a local approximation by considering the dominant term in this expansion for an infinitesimal perturbation, i.e., as $\epsilon \to 0$. By neglecting terms of order $\mathcal{O}(\epsilon^3)$ and higher, the final local approximation, as stated in the Lemma 2.2, is:

$$D_{\mathrm{KL}}\left(P_{T|X}^{(k+1)}(\cdot|x)\middle\|P_{T|X}^{(k)}(\cdot|x)\right) \approx \frac{\epsilon^2}{2}\|L_{T,x}\|_2^2. \tag{34}$$

This completes the proof. □

A.2 Proof of Lemma 3.1

Proof. The change in true risk, $\Delta R_{\mathrm{true}} \triangleq R_{\mathrm{true}}^{(k+1)} - R_{\mathrm{true}}^{(k)}$, is the change in the expectation of the loss function $l(T,Y)$. Using the perturbation formula $P_{T|X}^{(k+1)}(t|x) \approx P_{T|X}^{(k)}(t|x) + \epsilon L_{T,x}(t)\sqrt{P_{T|X}^{(k)}(t|x)}$, the change in risk can be expressed to first order in ϵ as:

$$\Delta R_{\mathrm{true}} \approx \epsilon \cdot \mathbb{E}_{P_{X,Y}}\left[\sum_{t\in\mathcal{T}} l(t,Y)L_{T,X}(t)\sqrt{P_{T|X}^{(k)}(t|X)}\right]. \tag{35}$$

To analyze this expression geometrically, its absolute value should be bounded.

$$\begin{aligned} |\Delta R_{\mathrm{true}}| &\approx \epsilon\left|\mathbb{E}_{P_{X,Y}}\left[\sum_{t\in\mathcal{T}} l(t,Y)L_{T,X}(t)\sqrt{P_{T|X}^{(k)}(t|X)}\right]\right| \\ &\leq \epsilon \cdot \mathbb{E}_{P_{X,Y}}\left[\left|\sum_{t\in\mathcal{T}}\left(l(t,Y)\sqrt{P_{T|X}^{(k)}(t|X)}\right)\cdot L_{T,X}(t)\right|\right]. \end{aligned} \tag{36}$$

where (36) follows from applying the Jensen inequality.

For a fixed data point (X, Y), the inner summation is bounded, by applying the Cauchy-Schwarz inequality ($|\mathbf{u} \cdot \mathbf{v}| \leq \|\mathbf{u}\|_2 \|\mathbf{v}\|_2$), to the vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{|\mathcal{T}|}$ defined with components $u_t = l(t, Y)\sqrt{P^{(k)}_{T|X}(t|X)}$ and $v_t = L_{T,X}(t)$:

$$\begin{aligned}\left|\sum_t \left(l(t, Y)\sqrt{P^{(k)}_{T|X}(t|X)}\right) L_{T,X}(t)\right| &\leq \left(\sqrt{\sum_t l(t, Y)^2 P^{(k)}_{T|X}(t|X)}\right)\left(\sqrt{\sum_t L_{T,X}(t)^2}\right) \\ &= \sqrt{\mathbb{E}_{P^{(k)}_{T|X}(\cdot|X)}[l(T, Y)^2]} \cdot \|L_{T,X}\|_2. \end{aligned} \tag{37}$$

Because the loss function $l(t, y)$ is normalized to the range $[0, 1]$, its square is bounded by 1, i.e. $l(t, y)^2 \leq 1$. Thus, the expected squared loss under the conditional distribution is bounded:

$$\mathbb{E}_{P^{(k)}_{T|X}(\cdot|X)}[l(T, Y)^2] = \sum_t l(t, Y)^2 P^{(k)}_{T|X}(t|X) \leq \sum_t P^{(k)}_{T|X}(t|X) = 1. \tag{38}$$

This implies $\|\mathbf{u}\|_2 \leq 1$ for all (X, Y). Substituting this upper bound and the Euclidean norm $\|L_{T,X}\|_2$ back into (36) yields:

$$|\Delta R_{\text{true}}| \leq \epsilon \cdot \mathbb{E}_{P_{X,Y}}\left[1 \cdot \|L_{T,X}\|_2\right] = \epsilon \cdot \mathbb{E}_{P_X}\left[\|L_{T,X}\|_2\right]. \tag{39}$$

Finally, applying Jensen's inequality to the concave square-root function, where $\mathbb{E}[W] \leq \sqrt{\mathbb{E}[W^2]}$, yields:

$$\begin{aligned}|\Delta R_{\text{true}}| &\leq \epsilon \cdot \sqrt{\mathbb{E}_{P_X}\left[\|L_{T,X}\|_2^2\right]} \\ &= \sqrt{2 \cdot \left(\frac{\epsilon^2}{2}\mathbb{E}_{P_X}\left[\|L_{T,X}\|_2^2\right]\right)}. \end{aligned} \tag{40}$$

Absorbing the constant factor of $\sqrt{2}$ into the $\mathcal{O}(\cdot)$ notation yields the desired result. This completes the proof. □

A.3 Proof of Theorem 3.2

Proof. Let the training samples $\{Z_i\}_{i=1}^n$ be drawn i.i.d. from the true distribution $P_{X,Y}$. The absolute change in the generalization gap is defined as the deviation of the empirical change in risk from the true change in risk:

$$|\Delta G| = |\Delta R_{\text{true}} - \Delta R_{\text{emp}}|. \tag{41}$$

Let the change-in-loss function for a single data sample $Z = (X, Y)$ be defined as $\phi(Z) \triangleq \mathbb{E}_{P^{(k+1)}_{T|X}(\cdot|X)}[l(T, Y)] - \mathbb{E}_{P^{(k)}_{T|X}(\cdot|X)}[l(T, Y)]$. The quantity to be bounded is the deviation of the empirical mean of n i.i.d. samples of $\phi(Z)$ from its true mean.

The perturbation formula is $P^{(k+1)}_{T|X}(t|x) = P^{(k)}_{T|X}(t|x) + \epsilon L_{T,x}(t)\sqrt{P^{(k)}_{T|X}(t|x)} + \mathcal{O}(\epsilon^2)$, where the remainder term is uniform over (t, x) assuming the perturbation is sufficiently smooth. The change-in-loss function can then be written as:

$$\begin{aligned}\phi(Z) &\approx \sum_{t\in\mathcal{T}} l(t, Y)\left[P^{(k)}_{T|X}(t|X) + \epsilon L_{T,X}(t)\sqrt{P^{(k)}_{T|X}(t|X)}\right] - \sum_{t\in\mathcal{T}} l(t, Y)P^{(k)}_{T|X}(t|X) \\ &= \mathbb{E}_{P^{(k)}_{T|X}(\cdot|X)}[l(T, Y)] + \epsilon\sum_{t\in\mathcal{T}}\left(l(t, Y)\sqrt{P^{(k)}_{T|X}(t|X)}\right) L_{T,X}(t) - \mathbb{E}_{P^{(k)}_{T|X}(\cdot|X)}[l(T, Y)] \\ &= \epsilon\sum_{t\in\mathcal{T}}\left(l(t, Y)\sqrt{P^{(k)}_{T|X}(t|X)}\right) L_{T,X}(t) \\ &= \epsilon\langle \mathbf{v}_Y, L_{T,X}\rangle_2, \end{aligned} \tag{42}$$

where $\mathbf{v}_Y$ is the loss-probability vector for the given label Y. This establishes that $\phi(Z)$ is well-approximated by the linear functional of the perturbation vector.

Let $W_i = \phi(Z_i) - \mathbb{E}_{P_Z}[\phi(Z)]$ be the centered i.i.d. random variables. Since the loss function l is normalized to $[0, 1]$, the change in its expectation, $\phi(Z)$, is bounded almost surely, i.e., $|\phi(Z)| \leq 1$. Therefore the Bernstein's inequality(Boucheron et al., 2013) for bounded random variables can be applied. With probability at least $1 - \delta$:

$$|\Delta G| = \left| \frac{1}{n} \sum_{i=1}^{n} W_i \right| \leq \sqrt{\frac{2\mathrm{Var}_{P_Z}(\phi(Z)) \log(2/\delta)}{n}} + \frac{2\log(2/\delta)}{3n}. \tag{43}$$

For large n, the leading term is the one involving the variance. The proof thus reduces to finding an EIT-based bound for $\mathrm{Var}_{P_Z}(\phi(Z))$.

The variance is bounded by the second moment: $\mathrm{Var}_{P_Z}(\phi(Z)) \leq \mathbb{E}_{P_Z}[\phi(Z)^2]$. Using the first-order approximation from (42) and applying the Cauchy-Schwarz inequality, the second moment is bounded by:

$$\begin{aligned}
\mathbb{E}_{P_Z}[\phi(Z)^2] &\approx \epsilon^2 \mathbb{E}_{P_{X,Y}} \left[\left(\sum_t \left(l(t, Y) \sqrt{P^{(k)}_{T|X}(t|X)} \right) L_{T,X}(t) \right)^2 \right] \\
&\leq \epsilon^2 \mathbb{E}_{P_{X,Y}} \left[\left(\sum_t l(t, Y)^2 P^{(k)}_{T|X}(t|X) \right) \left(\sum_t L_{T,X}(t)^2 \right) \right] + \mathcal{O}(\epsilon^3).
\end{aligned} \tag{44}$$

Since the loss function is normalized to $[0, 1]$, the expected squared loss $\sum_t l(t, Y)^2 P^{(k)}_{T|X}(t|X)$ is bounded by 1. Substituting this, and recognizing the Euclidean norm $\|L_{T,X}\|_2^2 = \sum_t L_{T,X}(t)^2$, yields:

$$\mathrm{Var}_{P_Z}(\phi(Z)) \leq \epsilon^2 \mathbb{E}_{P_X} \left[\|L_{T,X}\|_2^2 \right] + \mathcal{O}(\epsilon^3). \tag{45}$$

The variance of the change-in-loss is thus bounded by a quantity directly proportional to the expected squared norm of the EIT perturbation vectors, up to higher-order terms.

Substituting the variance bound from (45) into the dominant term of Bernstein's inequality (43):

$$|\Delta G| \leq \sqrt{\frac{2\left(\epsilon^2 \mathbb{E}_{P_X}[\|L_{T,X}\|_2^2] + \mathcal{O}(\epsilon^3)\right) \log(2/\delta)}{n}} + \mathcal{O}\left(\frac{1}{n}\right). \tag{46}$$

Absorbing the constant factor of 2 into the $\mathcal{O}(\cdot)$ notation, but retaining the $\log(2/\delta)$ factor to preserve precision for the finite-sample high-probability statement, it holds that:

$$|\Delta G| \leq \mathcal{O}\left(\sqrt{\frac{\epsilon^2 \mathbb{E}_{P_X}\left[\|L_{T,X}\|_2^2\right] \log(2/\delta)}{n}} \right). \tag{47}$$

This completes the proof.

□

A.4 Proof of Theorem 3.3

Proof. The first part of the theorem states the local equivalence:

$$\frac{\epsilon^2}{2} \mathbb{E}_{P_X} \left[\|L_{T,X}\|_2^2 \right] \approx \mathbb{E}_{P_X} \left[D_{\mathrm{KL}}(P^{(k+1)}_{T|X} || P^{(k)}_{T|X}) \right]. \tag{48}$$

The term on the right-hand side is the fundamental complexity measure in PAC-Bayesian analyses of single-step updates. This equivalence is a direct result of Lemma 2.2, which provides the local, second-order approximation for the KL-divergence for a fixed input x:

$$D_{\mathrm{KL}} \left(P^{(k+1)}_{T|X}(\cdot|x) \middle\| P^{(k)}_{T|X}(\cdot|x) \right) = \frac{\epsilon^2}{2} \|L_{T,x}\|_2^2 + \mathcal{O}(\epsilon^3). \tag{49}$$

To obtain the expected value over the entire data distribution, the expectation with respect to the input distribution P_X is taken on both sides. By the linearity of the expectation operator:

$$\begin{aligned}
\mathbb{E}_{P_X} \left[D_{\mathrm{KL}}(P^{(k+1)}_{T|X} || P^{(k)}_{T|X}) \right] &= \mathbb{E}_{P_X} \left[\frac{\epsilon^2}{2} \|L_{T,X}\|_2^2 + \mathcal{O}(\epsilon^3) \right] \\
&= \frac{\epsilon^2}{2} \mathbb{E}_{P_X} \left[\|L_{T,X}\|_2^2 \right] + \mathcal{O}(\epsilon^3).
\end{aligned} \tag{50}$$

Neglecting the higher-order terms establishes the local approximation in (48) and proves the first part of the theorem.

The second part of the theorem connects the local generalization cost to the global information-theoretic bound of Xu and Raginsky (2017). Let the bounding function be $f(I)$. From Theorem 2.1:

$$f(I) \triangleq \sqrt{\frac{2(I + \log(2|\mathcal{T}|/\delta))}{n}}, \tag{51}$$

where $I \triangleq I(X;T)$ is the global mutual information. The first-order Taylor expansion of this function describes its local change in response to a small change in mutual information, $\Delta I = I^{(k+1)} - I^{(k)}$:

$$f(I^{(k+1)}) - f(I^{(k)}) \approx f'(I^{(k)}) \cdot \Delta I. \tag{52}$$

The derivative of the bounding function $f(I)$ with respect to I is:

$$f'(I) = \frac{1}{\sqrt{2n(I + \log(2|\mathcal{T}|/\delta))}}. \tag{53}$$

It remains to show that the change in mutual information, ΔI, is itself locally approximated by the local generalization cost.

Let $g(\epsilon) \triangleq I(X;T)|_{P^{(k+\epsilon)}}$, a second-order Taylor expansion of the mutual information functional $g(\epsilon)$ with respect to ϵ gives:

$$g(\epsilon) = g(0) + \epsilon g'(0) + \frac{\epsilon^2}{2} g''(0) + \mathcal{O}(\epsilon^3). \tag{54}$$

The first derivative $g'(0)$ is zero because the normalization constraint $\sum_t J(t|x) = 0$ for all x implies that the marginal distribution P_T is invariant to the perturbation. Specifically, since the perturbed marginal is $P_T^{(k+1)}(t) = \sum_x [P_{T|X}^{(k)}(t|x) + \epsilon J(t|x)] P_X(x)$, the derivative $\frac{d}{d\epsilon} P_T^{(k+1)}(t) = \sum_x J(t|x) P_X(x) = 0$. Given that the mutual information $I(X;T)$ is a function of the marginal P_T, and the marginal is invariant to the perturbation, it follows that $g'(0) = 0$. Consequently, the second derivative at $\epsilon = 0$ is given by the variance of the perturbation vector:

$$g''(0) = \mathbb{E}_{P_X}\left[\|L_{T,X} - \mathbb{E}_{P_X}[L_{T,X}]\|_2^2\right]. \tag{55}$$

To ensure the perturbation J is a valid conditional distribution change that preserves the marginal P_X, the average perturbation across the input space must vanish. Specifically, the condition $\mathbb{E}_{P_X}[J] = 0$ implies:

$$\mathbb{E}_{P_X}\left[L_{T,X}\sqrt{P_{T|X}^{(k)}}\right] = 0. \tag{56}$$

Under the condition $\mathbb{E}_{P_X}[L_{T,X}] = 0$, the second derivative simplifies to the expected unweighted squared norm:

$$g''(0) = \mathbb{E}_{P_X}\left[\|L_{T,X}\|_2^2\right]. \tag{57}$$

By the second-order Taylor expansion, the change in mutual information is $\Delta I = g(\epsilon) - g(0) \approx \frac{\epsilon^2}{2} g''(0)$. Therefore,

$$\Delta I(X;T) \approx \frac{\epsilon^2}{2} \mathbb{E}_{P_X}\left[\|L_{T,X}\|_2^2\right]. \tag{58}$$

This establishes that the local generalization cost is the local increase in the global mutual information. Substituting this into the Taylor expansion (52) yields the second statement of the theorem. This completes the proof.

□

A.5 Proof of Theorem 4.1

Proof. The constrained optimization problem is:

$$\text{minimize} \quad \frac{1}{n}\sum_{i=1}^{n}\langle L_{T,x_i}, \mathbf{v}_i\rangle_2 \tag{59}$$

$$\text{subject to} \quad \frac{1}{n}\sum_{i=1}^{n}\|L_{T,x_i}\|_2^2 = C, \tag{60}$$

$$\langle L_{T,x_i}, \sqrt{\mathbf{P}_i}\rangle_2 = 0, \quad \forall i, \tag{61}$$

where $[\mathbf{v}_i]_t = l(t,y_i)\sqrt{P^{(k)}_{T|X}(t|x_i)}$, and denoting the probability root vector as $\sqrt{\mathbf{P}_i}$ with components $\sqrt{P^{(k)}_{T|X}(t|x_i)}$. The orthogonality constraint ensures the perturbed probabilities sum to 1. Let $\lambda \in \mathbb{R}$ be the Lagrange multiplier for the budget constraint, and let $\{\mu_i\}_{i=1}^n \in \mathbb{R}$ be the multipliers for the n orthogonality constraints. The Lagrangian, $\mathcal{L}$, is:

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}\langle L_{T,x_i}, \mathbf{v}_i\rangle_2 + \lambda\left(\frac{1}{n}\sum_{i=1}^{n}\|L_{T,x_i}\|_2^2 - C\right) + \sum_{i=1}^{n}\mu_i\langle L_{T,x_i}, \sqrt{\mathbf{P}_i}\rangle_2. \tag{62}$$

An optimal solution $\{L^*_{T,x_i}\}$ must be a stationary point of the Lagrangian. Setting the derivative of $\mathcal{L}$ with respect to the vector L_{T,x_i} to zero yields:

$$\frac{\partial \mathcal{L}}{\partial L_{T,x_i}} = \frac{1}{n}\mathbf{v}_i + \frac{2\lambda}{n}L^*_{T,x_i} + \mu_i\sqrt{\mathbf{P}_i} = \mathbf{0}. \tag{63}$$

Rearranging for L^*_{T,x_i}:

$$L^*_{T,x_i} = -\frac{1}{2\lambda}\mathbf{v}_i - \frac{n\mu_i}{2\lambda}\sqrt{\mathbf{P}_i}. \tag{64}$$

The multiplier μ_i is determined by enforcing the orthogonality constraint $\langle L^*_{T,x_i}, \sqrt{\mathbf{P}_i}\rangle_2 = 0$. Substituting (64) into the constraint:

$$-\frac{1}{2\lambda}\langle \mathbf{v}_i, \sqrt{\mathbf{P}_i}\rangle_2 - \frac{n\mu_i}{2\lambda}\langle \sqrt{\mathbf{P}_i}, \sqrt{\mathbf{P}_i}\rangle_2 = 0. \tag{65}$$

Note that $\langle \sqrt{\mathbf{P}_i}, \sqrt{\mathbf{P}_i}\rangle_2 = \sum_t P^{(k)}_{T|X}(t|x_i) = 1$. Furthermore, the inner product $\langle \mathbf{v}_i, \sqrt{\mathbf{P}_i}\rangle_2 = \sum_t l(t,y_i)P^{(k)}_{T|X}(t|x_i) = \mathbb{E}_{P^{(k)}}[\mathbf{l}_i]$, which is the expected loss for sample i. Assuming $\lambda \neq 0$ for a non-trivial budget and solving for the multiplier term:

$$n\mu_i = -\mathbb{E}_{P^{(k)}}[\mathbf{l}_i]. \tag{66}$$

Substituting this back into (64) reveals that the optimal perturbation is proportional to the projection of the negative objective vector onto the orthogonal subspace:

$$L^*_{T,x_i} = -\frac{1}{2\lambda}\left(\mathbf{v}_i - \mathbb{E}_{P^{(k)}}[\mathbf{l}_i]\sqrt{\mathbf{P}_i}\right) \triangleq -k\cdot\tilde{\mathbf{v}}_i, \tag{67}$$

where $\tilde{\mathbf{v}}_i$ is the centered loss-probability vector, and $k = \frac{1}{2\lambda} > 0$ is a global scaling constant.

The constant k is determined by the budget constraint, substituting (67) into the budget constraint gives:

$$\frac{1}{n}\sum_{i=1}^{n}\|-k\cdot\tilde{\mathbf{v}}_i\|_2^2 = k^2\left(\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\mathbf{v}}_i\|_2^2\right) = C. \tag{68}$$

Let $S^2 = \frac{1}{n}\sum_{j=1}^{n}\|\tilde{\mathbf{v}}_j\|_2^2$ be the average squared Euclidean norm of the centered vectors. Assuming $S^2 > 0$, it follows that $k^2S^2 = C$, which implies $k = \sqrt{C/S^2}$.

Substituting k back into (67) provides the final closed-form solution:

$$L^*_{T,x_i} = -\sqrt{C}\frac{\tilde{\mathbf{v}}_i}{\sqrt{S^2}}. \tag{69}$$

This completes the proof. □