Loss Landscape Analysis for Reliable Quantized ML Models for Scientific Sensing

Tommaso Baldi [1] **Javier Campos** [2] **Olivia Weng** [3] **Caleb Geniesse** [4] **Nhan Tran** [2] **Ryan Kastner** [3]
Alessandro Biondi [1]

[1]Department of Excellence in Robotics and AI, Scuola Superiore Sant'Anna, Pisa, Italy
[2]Fermi National Accelerator Laboratory, Batavia, IL, USA
[3]University of California San Diego, San Diego, CA, USA
[4]Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Abstract

In this paper, we propose a method to perform empirical analysis of the loss landscape of machine learning (ML) models. The method is applied to two ML models for scientific sensing, which necessitates quantization to be deployed and are subject to noise and perturbations due to experimental conditions. Our method allows assessing the robustness of ML models to such effects as a function of quantization precision and under different regularization techniques—two crucial concerns that remained underexplored so far. By investigating the interplay between performance, efficiency, and robustness by means of loss landscape analysis, we both established a strong correlation between gently-shaped landscapes and robustness to input and weight perturbations and observed other intriguing and non-obvious phenomena. Our method allows a systematic exploration of such trade-offs *a priori*, i.e., without training and testing multiple models, leading to more efficient development workflows. This work also highlights the importance of incorporating robustness into the Pareto optimization of ML models, enabling more reliable and adaptive scientific sensing systems.

1. Introduction

Advances in sensing technology drive the frontiers of scientific exploration, enabling breakthroughs across disciplines. Scientific sensing challenges can far outpace industrial applications, with inference latency on the scale of nanoseconds and microseconds, and extreme data rates (Duarte et al., 2022; Wei et al., 2024). This requires new methodologies for ultra-fast and ultra-compact edge processing.

Machine learning (ML) has emerged as a transformative tool across several scientific domains. In low-latency applications, scientists can enhance the capabilities of their instruments, including adaptiveness to dynamic conditions and the extraction of deeper insights from raw data (Deiana et al., 2022). Sensing and control with ML at unprecedented spatial and temporal scales can enable real-time analytics in scientific systems to accelerate scientific discovery.

Although significant advances have been made in methods to co-design efficient ML algorithms to meet the performance and resource demands of scientific systems, such as quantization (Gholami et al., 2022; Rokh et al., 2023), pruning (Vadera & Ameen, 2022; Cheng et al., 2024), and neural architecture search (Elsken et al., 2019), the link between *reliability* and, more in general, *robustness* of models and their performance and resource optimization has been much less studied. However, this issue is crucial for the unique demands of scientific sensing, in which raw and unprocessed instrument data are typically exposed to harsh environments that make ML processing prone to noise and perturbations.

Contribution. This paper addresses the interplay between performance, efficiency, and robustness in scientific sensing. Specifically, we explore the robustness of state-of-the-art (SoTA) neural networks under quantization and data corruption, focusing on two distinct applications: **(i)** autoencoders for sensor lossy data compression in particle physics (Di Guglielmo et al., 2021) and **(ii)** computer vision regression tasks for fusion energy diagnostics (Wei et al., 2024). An overview of their workflow is shown in Figure 1, which will be later discussed in Section 4.

Specifically, the paper makes the following contributions:

- We introduce *loss landscapes analysis* (Sun et al., 2020) methods for scientific sensing capable of identifying robust configurations of ML models *a priori*, i.e., without requiring time-consuming exploration cam-

This manuscript has been authored by Fermi Forward Discovery Group, LLC under Contract No. 89243024CSC000002 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

paigns with training and testing in the loop.

- We study how different regularization methods can mitigate noise and perturbations in quantized ML models for scientific sensing.

- We unveil a strong correlation between gently-shaped landscapes, both locally and globally, and robustness to data corruption. Furthermore, we observe non-obvious phenomena that suggest the need for a careful trade-off exploration in quantizing ML models to balance precision with robustness.

This study emphasizes the importance of including robustness to Pareto optimization of ML models, in addition to performance and efficiency, when designing real-time models for scientific sensing. By providing insights on robustness *a priori*, independently of the source of noise and perturbations that may affect a model, this work paves the way for more adaptive experimental capabilities, thereby enabling more capable experiments at unprecedented timescales.

Paper structure. The paper is structured as follows. We first provide basic concepts of quantization and loss landscape analysis in Section 2. We proceed by presenting our method based on multiple loss landscape metrics in Section 3. Then, we introduce the setup used in this work in Section 4, illustrating the models, the benchmarks, and the mitigation techniques involved. Experimental results are presented in Section 5, while Section 6 concludes the paper.

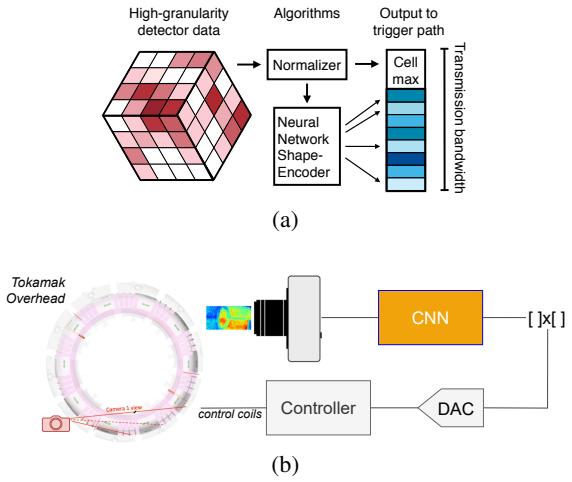


Figure 1. Workflow of the models in this study. (a) The ECON-T model workflow (Di Guglielmo et al., 2021), demonstrating the lossy data compression pipeline designed for deployment in the high-radiation environment of the Large Hadron Collider (LHC). (b) The Fusion model workflow (Wei et al., 2024), illustrating active feedback control in magnetic confinement fusion devices.

2. Related Work

Quantization. Quantizing deep neural networks is a widely used technique to reduce memory usage and enhance inference speed, making it particularly valuable for deployment on resource-constrained devices. However, these benefits often come at the expense of performance degradation and increased instability. To address this issue, a popular approach is Quantization-Aware Training (QAT), where we re-train the Neural Network (NN) model with quantized parameters so that the model can recover part of the performance by converging to a better loss point. Since it is not always possible to re-train the model due to computational costs or unavailability of the dataset, an alternative approach, called Post-Training Quantization (PTQ), allows quantizing all parameters without re-training the model, with limited overhead at the cost of lower accuracy, especially for low-precision quantization (Nagel et al., 2020; Li et al., 2021; Zheng et al., 2022; Wei et al., 2022). We focus this work on QAT mainly for two reasons: first, we are interested in studying the impact of quantization on the training of NN models; and second, the training time of the target models is low enough to favor the performance advantages provided by QAT. Regardless of the quantization method, in this work, we chose uniform integer quantization due to its superior hardware efficiency compared to Floating Point (FP) representation, as evidenced by (Jacob et al., 2018) and (van Baalen et al., 2023).

Loss landscape Analysis. Loss landscapes and the connections to training optimization techniques have been crucial research paths in ML for years. (Choromanska et al., 2015; Keskar et al., 2016; Fort et al., 2019) are works on the connection between the loss landscapes and the Stochastic Gradient Descent (SGD) optimization. (Fort et al., 2020; Yang et al., 2021) propose empirical analysis to better understand how several factors, such as quality of the data, number of parameters and hyperparameter tuning impact the generalization capability of the model. Our work took inspiration from the experiments of (Yang et al., 2021), but with a different aim. As far as we know, this is the first work that looks for correlations between loss landscape topology and model robustness in science.

3. Method

This section presents the method used in this work to visualize and analyze the loss landscape of ML models. A collection of metrics is presented for analysis purposes. Our notation aligns with the conventions established by (Yang et al., 2021).

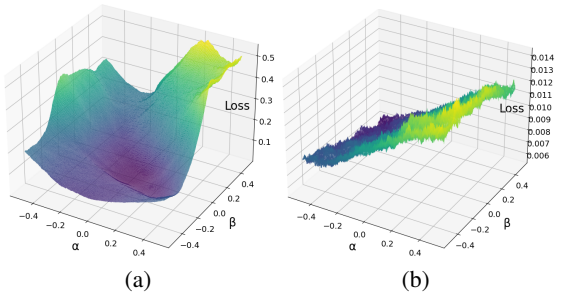
3.1. Loss Landscape Visualization

This work presents plots that approximate the surface of the loss landscape, which are useful to interpret the results we obtained. Various techniques to generate this kind of plot were proposed in previous work (Goodfellow et al., 2014; Im et al., 2016; Dinh et al., 2017; Keskar et al., 2017; Li et al., 2018). To generate our plots, we took inspiration from the approach of (Li et al., 2018), where the parameters of the model are perturbed along one random direction and its orthogonal, normalizing the weights filter-wise. Formally, given the parameters θ of a model, the resulting plots depict the following function:

$$f(\alpha, \beta) = \mathcal{L}(\theta + \alpha\sigma + \beta\eta), \tag{1}$$

where σ and η are the two directions and α and β are the steps in these directions. A series of N steps can be computed as $\alpha_i, \beta_i = \nu_{\text{min}} + i \cdot (\nu_{\text{max}} - \nu_{\text{min}})/(N-1)$, for $i = 0, 1, \ldots, N-1$, where ν_{max} and ν_{min} are the maximum and minimum perturbation module, respectively. 2D plots can be obtained by fixing either α or β.

However, when considering models with thousands of parameters, picking one random direction may lead to a gross approximation of the loss landscape which is not practically representative (e.g., Figure 2b). For this reason, we propose a novel approach that selects σ and η as the directions of the top-2 eigenvectors of the converged model. In this way, we can explore the two directions where the loss landscape faces the maximum curvature, leading to a more informative topology approximation of the surroundings of the model parameters θ (e.g., Figure 2a). A detailed comparison of the two approaches is discussed in Appendix B.1.



(a) (b)

Figure 2. Comparison of 3D loss landscape visualization methods: (a) uses the top-2 eigenvectors of model parameters, while (b) uses two random orthogonal directions.

3.2. CKA Similarity

In the context of loss landscape analysis, the Centered Kernel Alignment (CKA) similarity (Kornblith et al., 2019; Nguyen et al., 2021) is used to determine whether multiple instances of the same model, trained separately with randomly initialized parameters, tend to converge towards similar minima. This is assessed by examining the CKA similarity of their outputs, i.e., given NN f_θ and $m \in \mathbb{N}$ data samples randomly picked from the test set, we can define their concatenation as $F_\theta = [f_\theta(x_1) \cdots f_\theta(x_m)]^T \in \mathbb{R}^{m \times d_{\text{out}}}$.

The CKA similarity between two NNs with different parameter configurations θ and θ' can hence be measured by

$$\text{CKA}(\theta, \theta') = \frac{\text{Cov}(F_\theta, F'_\theta)}{\sqrt{\text{Cov}(F_\theta, F_\theta)\text{Cov}(F'_\theta, F'_\theta)}}. \tag{2}$$

In the above equation, given two matrices $X, Y \in \mathbb{R}^{m \times d}$, it holds

$$\text{Cov}(X, Y) = (m-1)^{-2} tr(XX^T H_m Y Y^T H_m), \tag{3}$$

where $H_m = I_m - \frac{1}{m}\mathbf{1}\mathbf{1}^T$ is the centering matrix.[1] Further details are discussed in the Appendix B.2.

A high CKA value indicates that the models are likely converging closely to each other, whereas a low value suggests that different parameter initializations can lead to convergence in different regions, as also empirically demonstrated by (Yang et al., 2021). This information helps describe the morphology of the loss landscape: in a globally smooth and flat loss landscape, models initialized with different parameters are expected to converge to similar, nearby minima. In contrast, low similarity may indicate a rugged loss landscape, where models risk getting trapped in suboptimal local minima.

3.3. Hessian Metrics

The Hessian is a square matrix that characterizes the curvature of the loss function at a specific point. The eigenvalues of the Hessian provide scalar values that offer insights on the curvature type at that point. Positive eigenvalues suggest local convexity of the loss, indicating a single minimum or maximum. In contrast, negative eigenvalues suggest local concavity, which implies the presence of a saddle point, often resulting from unfavorable training conditions. Zero eigenvalues signify a flat loss at that point, indicating the absence of both a minimum and a maximum. Intuitively, it is desirable to have both the top eigenvalue and the sum of the traces as close to zero as possible, because it suggests that the model has converged to a smooth, flat minimum.

Computing the Hessian matrix can be challenging from a computational perspective. In fact, it involves evaluating second-order partial derivatives of the loss function with respect to each pair of the model parameters. However, our analysis focuses on models deployed on edge devices, which are characterized by a limited number of parameters

[1] We use I_m to denote the identity matrix of size m and $\mathbf{1}$ represents the *indicator function*.

compared to foundational models. Furthermore, we leverage PyHessian (Yao et al., 2020) to calculate Hessian metrics, an open-source framework that approximates Hessian values by applying power methods.

3.4. Mode Connectivity

Mode connectivity is a global metric that provides insights about how well two minima are connected (Garipov et al., 2018; Draxler et al., 2018). It is capable of revealing the presence of barriers of loss between two points in which the two models converged. A simple way to compute mode connectivity is to set up a linear interpolation between two given models and sample a certain number of parameter configurations along it. Then, we can compute the loss of the models resulting from the sampled parameter configurations and see if there are barriers. However, this is a gross approximation because we do not know a priori the shape of the minima: for instance, it may happen that the two models are well connected via a curved line, whereas the linear interpolation cannot catch this information. A more convoluted method is hence required. We adopted the one proposed by (Garipov et al., 2018), which is based on a parameterized Bezier curve with $k+1$ bends. Consider $k+1$ models with parameters $\phi = \{\theta_0, \dots, \theta_k\}$, where $\theta_0 = \theta'$ and $\theta_k = \theta''$ are the parameters of the two models to be compared, while θ_j for $1 \leq j < k$ are the parameters of other models to be trained. The Bezier curve is defined as:

$$\gamma_\phi(t) = \sum_{j=0}^{k} \binom{k}{j} (1-t)^{k-j} \cdot t^j \cdot \theta_j, \tag{4}$$

where $t \in [0,1]$ is a scalar to move along the curve. The training of the models related to θ_j for $1 \leq j < k$ is performed as follows: first, the parameters are initialized using a linear interpolation between θ_0 and θ_k; then each of such models is trained to reach convergence.

At this point, we can now sample $m > 2$ different parameter configurations by picking m values of t, denoted by $\mathcal{T} = \{t_i\}_{i=0}^{m-1}$, including the two boundaries $t_1 = 0$ and $t_{m-1} = 1$. Intermediate values are computed by $t_i = i/(m-1)$, for $i = 1, \dots, m-2$.

We define the distance between the average loss of the two models being compared and the loss of a sampled one by

$$d(t, \theta', \theta'') = \frac{1}{2}(\mathcal{L}(\theta') + \mathcal{L}(\theta'')) - \mathcal{L}(\gamma_\phi(t)). \tag{5}$$

Finally, we can now define mode connectivity as the maximum deviation from the average loss of the two boundaries in the selected sampling points t_i, i.e., $mc(\theta', \theta'') = d(t^*)$ where $t^* = \arg\max_{t \in \mathcal{T}} \{|d(t, \theta', \theta'')|\}$.

This metric can be interpreted as follows:

$$\begin{cases} mc(\theta', \theta'') > 0, & \text{There are better minima.}^2 \\ mc(\theta', \theta'') < 0, & \text{There are barriers.} \\ mc(\theta', \theta'') \approx 0, & \text{Loss landscape is well connected.} \end{cases}$$

In this work, we computed mode connectivity with 3 bends ($k = 2$), training the models used to shape the Bezier curve for 30 epochs. A graphical example of mode connectivity is available in Appendix A, while an ablation study about how we set the right number of bends and training epochs is presented in Appendix B.3.

4. Experimental Setup

This section provides a detailed overview of the models employed in this study as representative benchmarks, and the noise mitigation techniques we tested.

4.1. Benchmark Models

Both models employed in our analysis are used for scientific sensing and are designed to be deployed on resource-constrained devices such as Field Programmable Gate Arrays (FPGAs) and Application Specific Integrated Circuits (ASICs), where the number of model parameters and the architectural design play a pivotal role in achieving efficiency.

4.1.1. ECON-T MODEL

The ECON-T model introduced by (Di Guglielmo et al., 2021) is an autoencoder for lossy data compression created for the Large Hadron Collider (LHC) and its high luminosity upgrade (HL-LHC) at CERN. Figure 1a shows an example compression flow employed by ECON-T. We focus our analysis on the encoder composed of 2288 parameters, and deployed on the detector in a high-radiation environment. The size and complexity of the model are constrained by area, on-chip memory, and power (≤ 100 mW). The performance of the autoencoder is measured via Earth Mover's Distance (EMD) (Rubner et al., 2000), which is not differentiable, so it is not used during training. A physics-inspired loss similar to Mean Square Error (MSE) called "telescoping MSE" is used during training, and EMD is used for measuring performance. Differentiable EMD loss (Shenoy et al., 2023) has been studied but is not benchmarked here.

4.1.2. FUSION MODEL

Active feedback control in thermonuclear fusion devices based on magnetic confinement is required to mitigate plasma instabilities and enable robust operation, preventing damage to the reactor. (Wei et al., 2024) combined effi-

[2]The training failed to locate a reasonable optimum, i.e., $\mathcal{L}(\theta')$ and $\mathcal{L}(\theta'')$ are large.

cient processing of FPGAs with high-speed imaging camera diagnostic and convolutional neural networks (CNNs) for magnetohydrodynamic (MHD) mode control on a tokamak device. The workflow of this system is illustrated in Figure 1b. The model inputs a camera image and predicts the $n = 1$ MHD mode amplitude and phase, where n is the *number density* used to describe the degree of concentration of countable objects in physical space. As CNN are not the SoTA for phase predictions due to periodicity of the task, between $-\pi$ and π, this work focuses on amplitude only.

4.2. Scientific Corruptions

Neural network quantization may not be the only noise facing the model at the stage of deployment. Indeed, models used in scientific experiments may operate in harsh environments, such as space or particle accelerators, where they typically experience significant performance degradation due to noise in the input data and weight perturbation, such as Single Event Upsets (SEU). To simulate the perturbations in the input data we adopt the injection of two different types of noise:

- **Gaussian noise**: It appears as random variations in pixel intensity that follow a Gaussian distribution. It typically arises from electronic noise in imaging sensors or during transmission.
- **Salt and Pepper noise**: It commonly arises from defects in imaging sensors, transmission errors, or faulty pixels in digital cameras. Unlike Gaussian and random noise, salt-and-pepper noise introduces localized disruptions in image content, which can severely degrade image quality. The target pixels are randomly selected without following any particular distribution and their value is either maximized or minimized.

We evaluated the performance of the models under varying levels of noisy perturbations to better understand their sensitivity to distorted input data. It is important to note that, in a practical setting, the nature of the perturbation is typically unknown a priori. As we will demonstrate later, in section 5, in such cases, mitigation techniques that are agnostic to the type of noise are generally more effective.

In addition to input corruptions, we also study weight corruptions from SEUs. On the software side, we adopted the FKeras (Weng et al., 2024) methodology, which ranks bits approximately from most to least sensitive to flipping. This allows us to simulate worst-case scenarios by flipping the top-k most sensitive bits and then evaluating the model's performance under these conditions. The sensitive bit ranking is done by first sorting the weights by a sensitive score computed as $H' = \sum_{i=1}^{k} \lambda_i (v_i \cdot \theta) v_i \quad \in \mathbb{R}^n$, where k indicates the number of top eigenvectors of the model, λ_i is the i-th eigenvalue, v_i is the i-th eigenvector of the model, and θ is the vector of model parameters (n is the number of parameters). Once the parameters are sorted by sensitivity, we then sort the bits of each parameter from the most significant bit (MSB) to the least significant one (LSB).

In Appendix C.1, we demonstrate the effectiveness of FKeras with respect to random bit flipping.

Note: Adversarial attacks were not considered, as models used in scientific experiments are deployed in controlled environments where they are not exposed to this threat.

4.3. Noise mitigation methods

To demonstrate the correlation between loss landscape analysis and robustness, we evaluated different versions of the target models, trained with different regularization methods. These methods aim to increase the robustness of the models against input and parameter perturbations by modifying the model's loss function and thus also its loss landscape. However, we will see in section 5, that they are not always effective, and how the morphology of the loss landscape can help us to understand which one is more effective.

In this work, we compare two noise mitigation techniques: Jacobian regularization (Sokolić et al., 2017; Hoffman et al., 2019) and orthogonal regularization (Cisse et al., 2017; Wang et al., 2020; Eryilmaz & Dundar, 2022).

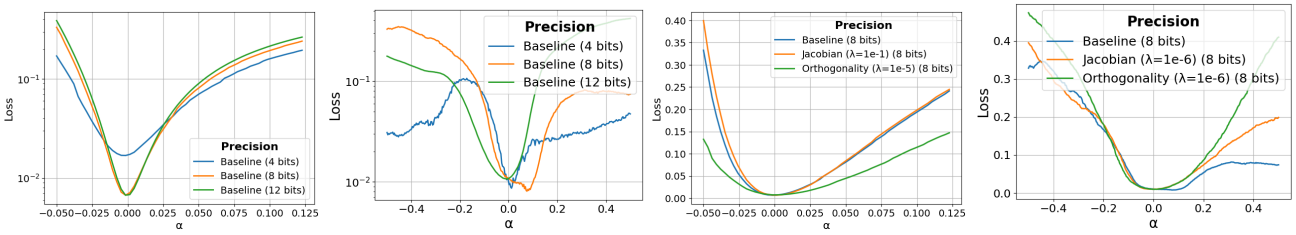
Jacobian regularization aims to limit the impact of input perturbations by adding a penalty to the loss function, which is proportional to the Frobenius norm (denoted by $\|\cdot\|_F$) of the model's Jacobian matrix $J(x)$, i.e.,

$$\delta \|J(x)\|_F^2, \tag{6}$$

where δ is a scalar used to weigh the impact of the regularization on the training loss.

This method controls the magnitude of the components of the Jacobian matrix, which are partially correlated with the contribution of noise to the model output. The aim of this method is to increase the margin between the input space and the decision boundaries of the target class. In our experiments, we use an efficient approximation of the Frobenius norm of the Jacobian matrix proposed by (Hoffman et al., 2019), which allows us to implement this method with negligible overhead.

The relationship between orthogonality and quantization has previously been studied by (Eryilmaz & Dundar, 2022). Although they demonstrated the beneficial effects of enforcing orthogonality among neural network weights during QAT, our goal is to provide a more detailed analysis of the effect of orthogonality on the loss landscape and reveal a possible correlation of it with a model's robustness. Various approaches to promote weight orthogonality have been



(a) (b) (c) (d)

Figure 3. Comparison of loss curves computed by perturbing the models along the top eigenvector of the Hessian matrix (i.e., varying parameter α in Eq. (1), reported on the x-axis, while keeping $\beta = 0$). (a) and (b) compare the loss line of models trained with different precision respectively for ECON-T and Fusion models, while (c) and (d) compare models fine-tuned with different regularization techniques respectively for ECON-T and Fusion models.

proposed in the literature (Miyato et al., 2018; Bansal et al.; Brock, 2018; Wang et al., 2020). In this work, we use a soft orthogonal regularization based on the Frobenius norm, formulated as:

$$\delta \|W^T W - I\|_F. \tag{7}$$

More complex regularization formulations could be applied, but given the nature of the model and the objectives of this study, this technique provides sufficiently effective results.

Moreover, in this work we did not take into account defensive approaches such as adversarial training (Kurakin et al., 2016) or noise injection during training for mainly two reasons: first, we have no guarantees that these methods will enhance reliability against noise of different nature; and second, the overhead introduced during training it is not negligible, especially when combined with QAT.

In Appendix C.2, we study how different values of δ impact the performances of the model.

5. Experimental Results

In this section, we first present results to evaluate the metrics introduced above and then assess the potential correlation between these metrics and the robustness of the tested models (ECON-T and Fusion). We studied robustness in the presence of network quantization. The models were quantized using integer uniform quantization implemented in Brevitas (Pappalardo, 2023) library, and all experiments were conducted on an NVIDIA A100 GPU using QAT. We evaluated *three versions* of each model: **(i)** a baseline version fine-tuned without regularization, **(ii)** one incorporating *Jacobian regularization* (with $\delta = 0.1$ for the ECON-T Model and $\delta = 10^{-6}$ for the Fusion model), and **(iii)** another one employing *orthogonal regularization* (with $\delta = 10^{-5}$ for the ECON-T Model and $\delta = 10^{-6}$ for the Fusion model). This approach assesses how a quantization scheme and noise mitigation influence the performance and reliability of the models. Each model version was then trained three times under different precisions, i.e., for bit widths ranging from 3 to 12. Unless otherwise stated, the results presented in this section represent the average of these three model versions.

Codes are available at `https://github.com/balditommaso/PyLandscape`.

5.1. Loss Landscape analysis

Visualizing the Loss Landscape. Visualization techniques provide an approximate representation of the shape of the loss landscape. In this work, we utilized 2D plots. The intrinsic regularization effect of quantization is illustrated in Figure 3a. The 4-bit version of the baseline model exhibits a higher minimum compared to the 8-bit and 12-bit configurations, due to the performance degradation typically associated with low-bit quantization. However, examining the entire loss curve reveals that the convex portion of the loss landscape is wider and flatter compared to models trained with higher precision. In contrast, Figure 3b demonstrates that low-precision quantization achieves a minimum comparable to the one of higher-precision configurations, but at the cost of sharper and more jagged minima. This results in a harsher loss landscape, which can complicate model training. As also discussed later, this has a significant impact on model robustness. For the ECON-T model, orthogonal regularization proves to be more effective than Jacobian regularization, resulting in a smoother and wider convex loss landscape (Fig. 3c). Conversely, for the Fusion model, the impact of regularization on the loss landscape is limited (Fig. 3d).

CKA similarity. We are interested in understanding if different instances of the same model converge in a close area of the loss landscape by looking at their CKA similarity. Figures 4a and 4b report the average CKA similarity between all pairs of parameter configurations (resulting from the three trainings) of the three model versions as a function of the quantization precision in bits. The figures show that the baseline versions have limited CKA similarity, which

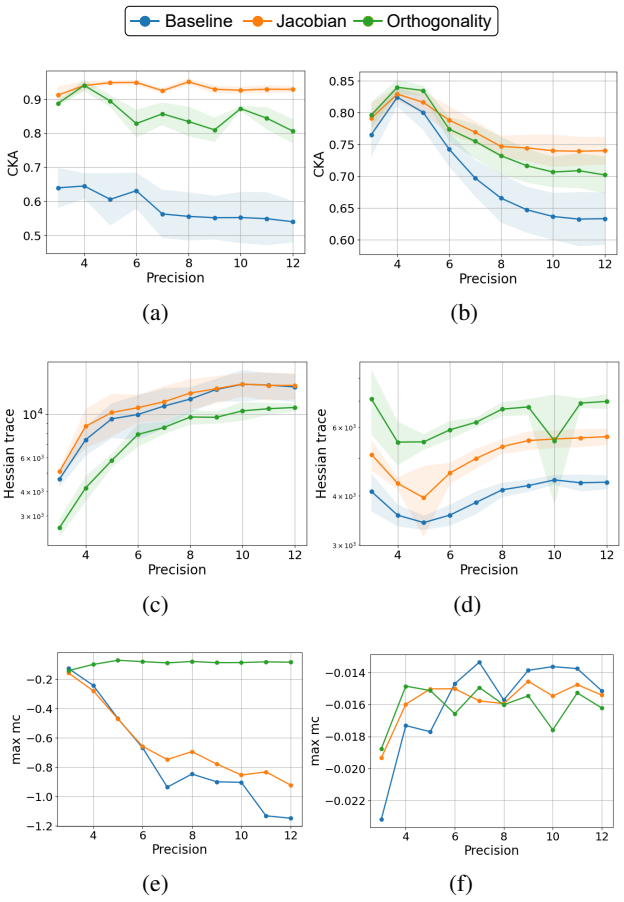


Figure 4. Analysis of loss landscape metrics for ECON-T (left column) models and Fusion models (right column) fine-tuned with different regularization strategies across varying precision levels. Subplots show: (a) and (b) **CKA similarity**, which evaluates representational alignment among models; (c) and (d) **Hessian trace**, capturing the overall curvature of the loss landscape where the model is converged; and (e) and (f) **mode connectivity**, indicating the presence of barriers among different minima. Regularization methods include Baseline (no regularization), Jacobian regularization, and orthogonal regularization.

tends to decrease as the precision increases, probably due to the implicit regularization effect resulting from low-bit quantization. This is not the case for the other two model variants, where, especially for Jacobian regularization, they converge to CKA-similar models. In these cases, the results suggest the presence of a smoother loss landscape, where the models do not get trapped in suboptimal minima during training. The drop in CKA similarity as the precision increases is more significant for the Fusion model (Fig. 4b) than the ECON-T model (Fig. 4a): this was expected because the Fusion model has more parameters, leading to a more complex loss landscape).

Hessian trace. The curvature of the loss landscape where the model lands at the end of QAT can be analyzed with the Hessian trace. The average Hessian trace values are reported in Figures 4c and 4d. From the figures we can see that all models tend to follow the same pattern in which the slope of the loss landscape increases with precision ≥ 5 bits. Furthermore, while the ECON-T model (inset (c)) benefits from orthogonal regularization, which allows converging to flatter minima, regularization is instead letting the Fusion model (inset (d)) converge to steeper minima. These behaviors are further investigated later in this section.

Mode connectivity. The presence of barriers in the loss landscape is undesirable as they hinder the optimization process, making it difficult for algorithms to efficiently converge to a global or near-global minimum. Weight perturbations caused by bit errors can shift the model's position in the loss landscape, with the magnitude of the shift determined by the difference between the original and perturbed parameters and the influenced direction. Intuitively, if the region surrounding the model is free of barriers, the impact of perturbations on performance is likely to be lower. We studied the maximum mode connectivity (Max mc) obtained as follows. Given the three models for each version, we sampled $m = 60$ points on the corresponding Bezier curve (Eq. (4)) and denote by T the set of model parameters corresponding to those 60 points. Max mc is hence given by the maximum mode connectivity of all pairs in T, i.e., $\max_{(\Theta',\Theta'')\in T\times T}\{mc(\Theta',\Theta'')\}$. Figures 4e and 4f illustrate how the presence of barriers is influenced by precision. Note that not all regularization methods effectively mitigate these barriers. The two models exhibit different behaviors: for ECON-T (Fig. 4e), Jacobian regularization provides only slight improvements in connectivity for precisions ranging from 6 to 12 bits, whereas models fine-tuned with orthogonal regularization are well connected, demonstrating the absence of significant barriers. In contrast, for Fusion (Fig. 4f), the presence of barriers between minima decreases as precision increases, with regularization mainly offering benefits in low-precision configurations.

5.2. Performance under perturbations

Input Perturbations. The robustness of the ECON-T model against input perturbations is critical to ensuring the reliability of processed data. Regularization techniques are promising candidates to improve model robustness; however, they may degrade performance on clean data, as shown in Figures 5a and 5e. While Jacobian regularization introduces negligible performance degradation, the degradation caused by orthogonal regularization may be unacceptable depending on the use case. The regularization weight δ in the optimization loss (see Eqs. (6) and (7)) can be adjusted to manage the trade-off between clean-data performance and model robustness. Fine-tuning this trade-off often requires several iterations to test various noise types and magnitudes to guarantee reliability. The loss landscape metrics previ-

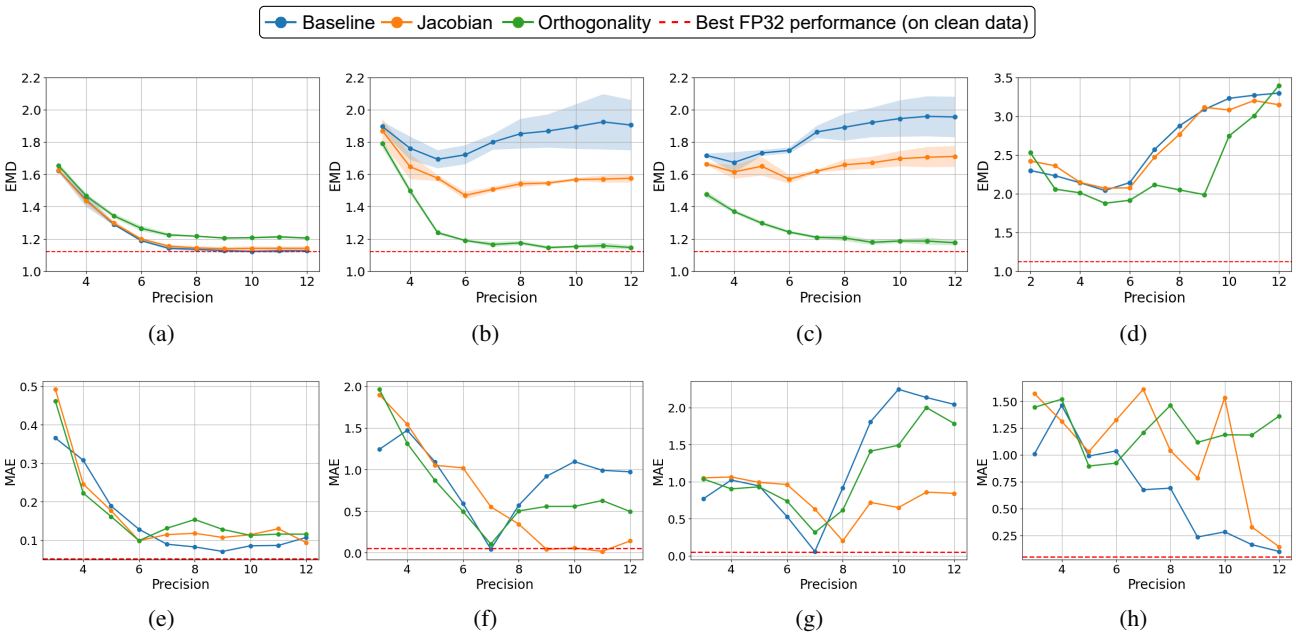


Figure 5. Evaluation of ECON-T models (top row) and Fusion models (bottom row) robustness under different input or weight perturbations. Each subplot represents performance benchmarks on specific scenarios: (a and e) clean data, (b and f) perturbed data with Gaussian noise, (c and g) perturbed data with salt-and-pepper noise, and (d and h) flipping the five most sensitive bits. The models are trained with three regularization methods: Baseline (no regularization), Jacobian regularization, and orthogonal regularization.

ously introduced provide valuable insights that align with the performance results in Figure 5. Noise significantly affects the performance of baseline models, as shown in Figures 5b and 5f for Gaussian noise, and Figures 5c and 5g for salt-and-pepper noise. However, mitigation techniques are not always effective in increasing robustness, and the reasons can be identified by looking at the loss landscape. Indeed, by matching the performance of the ECON-T model with the analysis of the loss landscape of discussed above, we can note that improvements in robustness provided by orthogonal regularization (Figs. 5b-5c) correspond to a flatter and smoother landscape (Fig. 4e). In contrast, Jacobian regularization is less effective in mitigating noise, which correlates with poorer minimum connectivity and steeper slopes at convergence points in the landscape (Fig. 4e). For the Fusion model, the effectiveness of mitigation techniques is evident for high precisions only. Although baseline models converge to flatter minima in these cases, they tend to produce highly divergent representations (Fig. 4b), which indicate obstacles in the optimization process that obstruct the reach of lower minima. Interestingly, both models exhibit higher robustness to input perturbations under low-bit quantization, a behavior that was also observed in previous studies (Lin et al., 2019).

Weight Perturbations. Given the size of the analyzed models, the impact of weight perturbations (bit flips) on their performance is highly destructive. Nevertheless, the results of Figures 5d and 5h confirm the observations made above also for this issue. Specifically, ECON-T models fine-tuned with low-bit configurations demonstrate greater robustness to weight perturbations, even with fewer bits per parameter. This behavior strongly correlates with the Hessian trace (Fig. 4c) and mode connectivity (Fig. 4e) analyses. Models with fewer bits per parameter tend to have a lower Hessian trace, i.e., indicating convergence to flatter minima, and reduced barriers between minima. For most configurations, except for extreme low-bit settings (e.g., 3 or 4 bits) where quantization and performance degradation are more pronounced, these phenomena are strongly correlated. Additionally, models trained with orthogonal regularization exhibit greater robustness to weight perturbations across most quantization configurations. This result aligns with their favorable Hessian trace and mode connectivity characteristics. In contrast, the Fusion model exhibits a different behavior. Performance degradation decreases as precision increases, consistently with the trends observed in Figures 4d and 4f. Low-precision Fusion models show more barriers between minima, and the baseline version of the model has the lowest Hessian trace compared to the regularized versions, explaining the results in Figure 5h.

6. Conclusion

We proposed a method to conduct a comprehensive empirical analysis of the loss landscape of machine learning models and applied the method to two representative, yet diverse

models for scientific applications. These models require quantization to be deployed and are subject to noise and bit flips in the model parameters. Two regularization techniques were considered to mitigate noise and bit flips, complementing the intrinsic regularization provided by quantization.

Most interestingly, contrary to what one may expect, we found that increasing quantization precision does not always provide benefits in terms of robustness to noise and bit flips. Furthermore, we found that different models may benefit from different regularization techniques.

Our method allows efficient exploration of the trade-offs between robustness and performance without calling for tedious and time-consuming training campaigns for design space exploration. It does so without assuming prior knowledge of the perturbations. Automatic Pareto optimization of model configurations can hence be enabled by building on our method and represents our prominent future work.

Acknowledgements

We thank Ryan Forelli for his help in this work. We also acknowledge the Fast Machine Learning collective as an open community of multi-domain experts and collaborators. This community was important for the development of this project.

TB and AB were supported by SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU, and OPERAND PRIN-PNRR project.

JC and NT were supported by the U.S. Department of Energy (DOE), Office of Science, Office of Advanced Scientific Computing Research under the "Real-time Data Reduction Codesign at the Extreme Edge for Science" Project (DE-FOA-0002501).

OW was supported by the NSF Graduate Research Fellowship Program under Grant No. DGE-2038238. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

CG was supported by the U.S. Department of Energy (DOE), Office of Science, Advanced Scientific Computing Research (ASCR) program under Contract Number DE-AC02-05CH11231 to Lawrence Berkeley National Laboratory ("Visualizing High-dimensional Functions in Scientific Machine Learning").

Impact Statement

This work contributes to the advancement of ML research by being the first reliability analysis of ML models for scientific sensing applications based on the loss landscape, and also copes with their training strategy to make them more reliable against unknown corruptions. Using loss landscape analysis, we provide actionable a-priori insights that minimize the need for brute-force testing, which typically requires extensive iterative experimentation. This approach can significantly reduce training resources and provide robustness against out-of-training-distribution corruptions. Furthermore, our analysis is grounded in models that are actively used in scientific experiment – improving control of complex systems and accelerating scientific research and, thus, potential discoveries. By enhancing the efficiency and reliability of ML models in these critical domains, our methodology aligns with the broader goals of sustainable and impactful AI research and development.

References

Bansal, N., Chen, X., and Wang, Z. Can we gain more from orthogonality regularizations in training deep cnns? arxiv 2018. *arXiv preprint arXiv:1810.09102*.

Brock, A. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Cheng, H., Zhang, M., and Shi, J. Q. A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pp. 192–204. PMLR, 2015.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. In *International conference on machine learning*, pp. 854–863. PMLR, 2017.

Deiana, A. M. et al. Applications and Techniques for Fast Machine Learning in Science. *Front. Big Data*, 5:787421, 2022. doi: 10.3389/fdata.2022.787421.

Di Guglielmo, G., Fahim, F., Herwig, C., Valentin, M. B., Duarte, J., Gingu, C., Harris, P., Hirschauer, J., Kwok, M., Loncar, V., et al. A reconfigurable neural network asic for detector front-end data compression at the hl-lhc. *IEEE Transactions on Nuclear Science*, 68(8):2179–2186, 2021.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

Draxler, F., Veschgini, K., Salmhofer, M., and Hamprecht, F. Essentially no barriers in neural network energy land-

scape. In *International conference on machine learning*, pp. 1309–1318. PMLR, 2018.

Duarte, J., Tran, N., Hawks, B., Herwig, C., Muhizi, J., Prakash, S., and Reddi, V. J. FastML Science Benchmarks: Accelerating Real-Time Scientific Edge Machine Learning. In *5th Conference on Machine Learning and Systems*, 7 2022.

Elsken, T., Metzen, J. H., and Hutter, F. Neural architecture search: A survey. *Journal of Machine Learning Research*, 20(55):1–21, 2019.

Eryilmaz, S. B. and Dundar, A. Understanding how orthogonality of parameters improves quantization of neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10737–10746, 2022.

Fort, S., Hu, H., and Lakshminarayanan, B. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Fort, S., Dziugaite, G. K., Paul, M., Kharaghani, S., Roy, D. M., and Ganguli, S. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. *Advances in Neural Information Processing Systems*, 33: 5850–5861, 2020.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., and Keutzer, K. A survey of quantization methods for efficient neural network inference. In *Low-Power Computer Vision*, pp. 291–326. Chapman and Hall/CRC, 2022.

Goodfellow, I. J., Vinyals, O., and Saxe, A. M. Qualitatively characterizing neural network optimization problems. *arXiv preprint arXiv:1412.6544*, 2014.

Hoffman, J., Roberts, D. A., and Yaida, S. Robust learning with jacobian regularization. *arXiv preprint arXiv:1908.02729*, 5(6):7, 2019.

Im, D. J., Tao, M., and Branson, K. An empirical analysis of deep network loss surfaces. *ArXiv*, abs/1612.04010, 2016. URL https://api.semanticscholar.org/CorpusID:13651606.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., and Kalenichenko, D. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2704–2713, 2018.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima, 2017. URL https://arxiv.org/abs/1609.04836.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMLR, 2019.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Li, Y., Gong, R., Tan, X., Yang, Y., Hu, P., Zhang, Q., Yu, F., Wang, W., and Gu, S. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*, 2021.

Lin, J., Gan, C., and Han, S. Defensive quantization: When efficiency meets robustness. *arXiv preprint arXiv:1904.08444*, 2019.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Nagel, M., Amjad, R. A., Van Baalen, M., Louizos, C., and Blankevoort, T. Up or down? adaptive rounding for post-training quantization. In *International Conference on Machine Learning*, pp. 7197–7206. PMLR, 2020.

Nguyen, T., Raghu, M., and Kornblith, S. Do wide and deep networks learn the same things. *Uncovering How Neural Network Representations Vary with Width and Depth Cs. Lg: arXiv: 2010.15327*, 2021.

Pappalardo, A. Xilinx/brevitas, 2023. URL https://doi.org/10.5281/zenodo.3333552.

Rokh, B., Azarpeyvand, A., and Khanteymoori, A. A comprehensive survey on model quantization for deep neural networks in image classification. *ACM Trans. Intell. Syst. Technol.*, 14(6), November 2023. ISSN 2157-6904. doi: 10.1145/3623402. URL https://doi.org/10.1145/3623402.

Rubner, Y., Tomasi, C., and Guibas, L. J. The Earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.*, 40:99, 2000. doi: 10.1023/A:1026543900054.

Shenoy, R., Duarte, J., Herwig, C., Hirschauer, J., Noonan, D., Pierini, M., Tran, N., and Suarez, C. M. Differentiable earth mover's distance for data compression at the high-luminosity lhc. *Machine Learning: Science and Technology*, 4(4):045058, 2023.

Sokolić, J., Giryes, R., Sapiro, G., and Rodrigues, M. R. D. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, 2017. doi: 10.1109/TSP.2017.2708039.

Sun, R., Li, D., Liang, S., Ding, T., and Srikant, R. The global landscape of neural networks: An overview. *IEEE Signal Processing Magazine*, 37(5):95–108, 2020.

Vadera, S. and Ameen, S. Methods for pruning deep neural networks. *IEEE Access*, 10:63280–63300, 2022.

van Baalen, M., Kuzmin, A., Nair, S. S., Ren, Y., Mahurin, E., Patel, C., Subramanian, S., Lee, S., Nagel, M., Soriaga, J., et al. Fp8 versus int8 for efficient deep learning inference. *arXiv preprint arXiv:2303.17951*, 2023.

Wang, J., Chen, Y., Chakraborty, R., and Yu, S. X. Orthogonal convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11505–11515, 2020.

Wei, X., Gong, R., Li, Y., Liu, X., and Yu, F. Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization. *arXiv preprint arXiv:2203.05740*, 2022.

Wei, Y., Forelli, R. F., Hansen, C., Levesque, J. P., Tran, N., Agar, J. C., Di Guglielmo, G., Mauel, M. E., and Navratil, G. A. Low latency optical-based mode tracking with machine learning deployed on FPGAs on a tokamak. *Rev. Sci. Instrum.*, 95(7):073509, 2024. doi: 10.1063/5.0190354.

Weng, O., Meza, A., Bock, Q., Hawks, B., Campos, J., Tran, N., Duarte, J. M., and Kastner, R. Fkeras: A sensitivity analysis tool for edge neural networks. *Journal on Autonomous Transportation Systems*, 2024.

Yang, Y., Hodgkinson, L., Theisen, R., Zou, J., Gonzalez, J. E., Ramchandran, K., and Mahoney, M. W. Taxonomizing local versus global structure in neural network loss landscapes. *Advances in Neural Information Processing Systems*, 34:18722–18733, 2021.

Yao, Z., Gholami, A., Keutzer, K., and Mahoney, M. W. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE international conference on big data (Big data)*, pp. 581–590. IEEE, 2020.

Zheng, D., Liu, Y., Li, L., et al. Leveraging inter-layer dependency for post-training quantization. *Advances in Neural Information Processing Systems*, 35:6666–6679, 2022.

A. Example of mode connectivity computation

In this section, we provide a more detailed example of how mode connectivity is computed. Please refer to Section 3.4 of the paper.

Once the Bezier curve between two models, with parameters θ' and θ'', is defined, we can then sample an arbitrary number of points along this curve by picking a value $t \in [0, 1]$. The extreme cases, $t = 0$ and $t = 1$, correspond to θ' and θ'', respectively. Each intermediate point t_i for $i = 1, ..., m - 2$ serves as input to the Bezier curve, yielding a possible configuration of model parameters. These parameters can be used to evaluate the loss along the Bezier curve, as illustrated by the blue line in Figure 6).

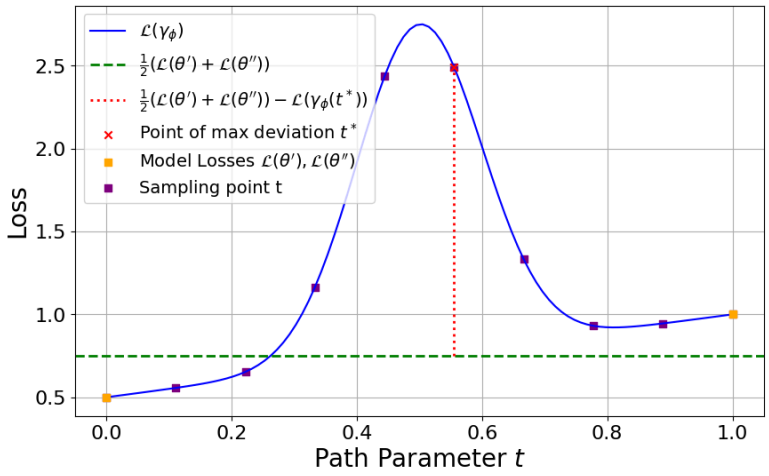


Figure 6. Example of mode connectivity computation. The blue line represent the loss computed along the Bezier curve γ_ϕ. We sample m points along this curve, and then we look for the point which is maximizing the deviation from the average loss (t^* in Section 3.4) between the two extreme model parameters θ' and θ''.

B. Ablation study on metrics

In this section, we first compare the loss landscapes visualization method proposed by (Li et al., 2018) with our approach, and then investigate different configurations when measuring the CKA similarity and the mode connectivity.

B.1. Ablation study on loss landscape visualization

We start by showing why visualizing the loss landscape using the top eigenvector as the perturbation direction (Figure 7c), our novel proposal that differs from the approach in (Li et al., 2018), provides more informative insights compared to using random directions (Figure 7f). Although both plots depict the loss landscape of the same Fusion models, they exhibit notable differences.

First, by examining the y-axis, we observe that the scales of the two plots are completely different. When using random directions—computed within the same range and with the same resolution (i.e., number of steps)—the loss variation is negligible. As a result, this approach fails to provide meaningful insights into suboptimal minima and their sharpness.

The only notable feature in Figure 7f is the jagged shape introduced by low-bit quantization. However, this effect can be disregarded, as we have no guarantee that the model will follow this direction during training. In contrast, when adopting the top eigenvector as the perturbation direction, we ensure that the visualization captures the model's behavior along the direction of maximum curvature. This is particularly relevant because most training optimizers, such as SGD, tend to explore this direction, making it a more reliable choice for loss landscape analysis.

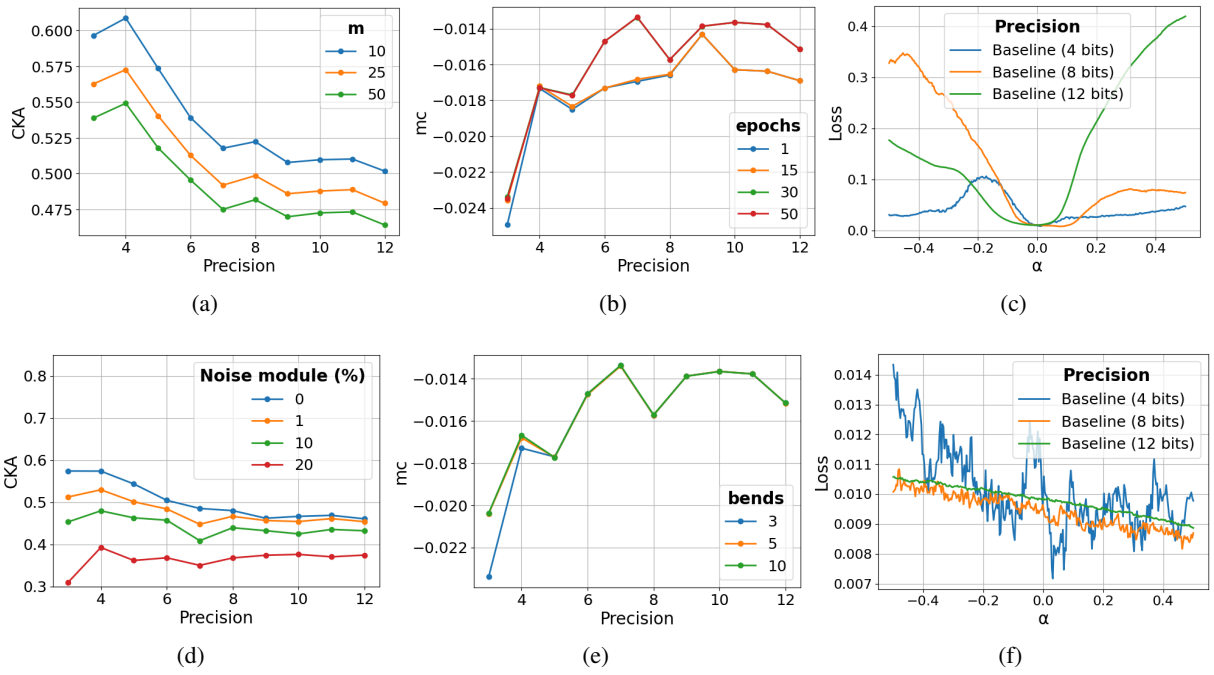


Figure 7. Ablation studies of the loss landscape metrics. Subplots show: (a) and (d) **CKA similarity** of the ECON-T model where we respectively explore the impact of changing the number of concatenated outputs m, and the noise intensity; (b) and (e) **mode connectivity** of the Fusion model, comparing results obtained tuning the number of training epochs (b) and the number of bends (e); (c) and (f) **loss landscape visualization** methods, where the directions are computed respectively with top eigenvector of model parameters and a random direction.

B.2. Ablation study on measuring CKA similarity

In this subsection, we provide a detailed analysis of how the number of concatenated outputs and the intensity of input perturbations impact the CKA results. As shown in Figure 7a, increasing the number of concatenated outputs leads to lower CKA similarity. This outcome is expected, as increasing the dimensionality of the matrices being compared raises the likelihood of differences between them. However, we observe that the overall patterns remain consistent, demonstrating the robustness of the metric across different configurations. In this work we used $m = 10$ to save computation time. Additionally, it is often preferable to conduct this analysis on perturbed data distributions, which can be achieved by adding uniformly distributed noise, such as Gaussian noise, to the inputs. This approach is particularly useful when models are trained to achieve near-zero training loss, as it enhances the informativeness of the metric. However, Figure 7d indicates that this is not necessary for the ECON-T model. Even without noise injection, we can extract meaningful insights regarding the CKA similarity of the models.

B.3. Ablation study on measuring mode connectivity

In this subsection, we analyze the optimal configuration for generating the mode connectivity plots shown in Figures 4e and 4f. The first key parameter to consider is the number of training epochs for the models used to construct the Bezier curve. This hyperparameter is particularly important because the characteristic curved shape of the Bezier curve emerges only when the intermediate models begin converging to their respective minima. Otherwise, the sampled models will lie along the linear interpolation between the two models, leading to a coarse approximation.

Figure 7b highlights this effect. Specifically, training the intermediate models for 1 and 15 epochs produces similar results, whereas training for 30 and 50 epochs leads to significantly different outcomes. Since the curves for 30 and 50 epochs are nearly identical, we opted for 30 epochs to reduce computational cost in subsequent experiments.

Another crucial hyperparameter in mode connectivity analysis is the number of bends $k+1$, which determines the complexity of the Bezier curve. Figure 7e shows that increasing the number of bends has little impact, likely due to the relatively low

number of parameters in the model under analysis. This results in a non-trivial loss landscape morphology that does not require highly parameterized Bezier curves. Therefore, we chose to use three bends to balance accuracy and computational efficiency.

C. Ablation study on benchmarks and mitigation techniques

In this section, we study different configurations of regularization methods and benchmarks.

C.1. Ablation study on benchmarks

In this subsection, we provide a comprehensive evaluation of the performance of the baseline version of the ECON-T model under varying noise magnitudes (Figures 8a and 8b) and bit error rates (Figures 8c and 8d), analogous to the analysis conducted in Figure 5.

Regarding input perturbations, both noise types exhibit similar behavior: for noise intensities below 20%, performance degradation is observed, but the model still follows the same overall pattern. However, at higher noise intensities, the destructive effects become irrecoverable and unpredictable.

For parameter perturbations, we validate the effectiveness of the FKeras approach as a benchmark for the worst-case scenario by comparing Figures 8c and 8d. Notably, not all bits contribute equally to model performance, as evidenced by the fact that flipping the most sensitive bit, as identified by FKeras, leads to significantly more destructive effects than randomly flipping 100 bits.

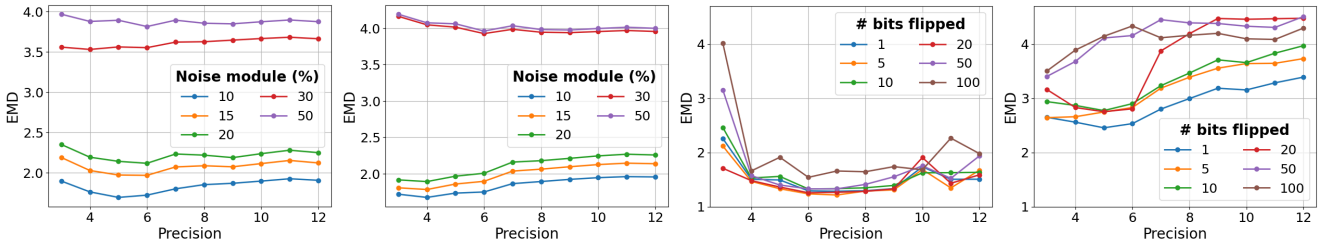


Figure 8. Evaluation of the ECON-T model under different stress conditions: (a) and (b) shows respectively the performances of the model where the input is corrupted with Gaussian and salt-and-pepper noise, focusing the attention on the models reliability respect to different noise intensity; (c) and (d), instead, compare the degradation of performances of the models where the parameters are perturbed by different numbers of bit errors,the bits to be flipped are picked randomly in (c) and adopting FKeras methodology in (d).

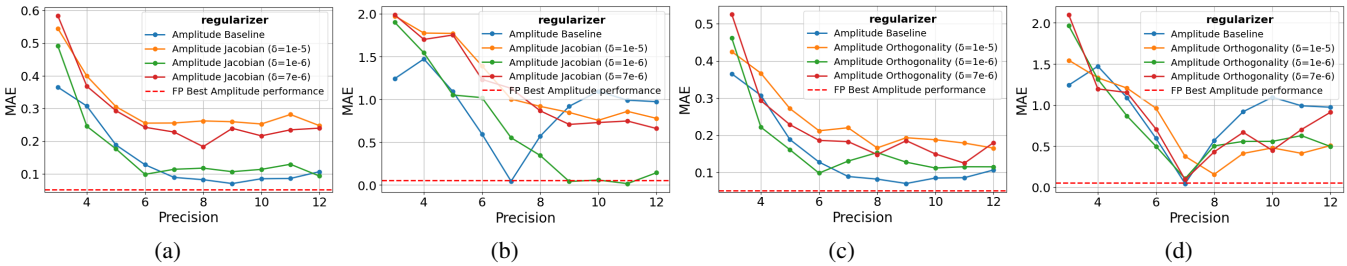


Figure 9. Evaluation of the Fusion model fine-tuned with different values of the coefficient δ for the regularization part of the loss: (a) and (b) shows respectively the performances of the model on clean and perturbed input (10% Gaussian noise), changing the δ of the Jacobian regularization; (c) and (d), instead, shows respectively the performances of the model on clean and perturbed input (10% Gaussian noise), changing the δ of the orthogonal regularization.

C.2. Ablation study on mitigation techniques tuning

In this subsection, we provide a comprehensive evaluation of the impact of the regularization techniques proposed in this work, comparing their performance on both clean data (Figures 9a and 9c) and perturbed data (Figures 9b and 9d). The model under analysis is the Fusion model, trained with different regularization coefficients δ for each regularizer.

The coefficient δ controls the weight of the regularization term in the loss function, determining the trade-off between performance on clean and perturbed data. Intuitively, a lower δ results in minimal performance degradation on clean data while offering limited robustness improvement. Conversely, a higher δ may significantly enhance robustness but at the cost of greater performance degradation on clean data.

In this study, we tested different values of δ and selected the one that provided the best trade-off ($\delta = 10^{-6}$ in this case). While more sophisticated tuning methods could be explored, they are beyond the scope of this work.