Balanced Sharpness-Aware Minimization for Imbalanced Regression

Yahao Liu[1] Qin Wang[2] Lixin Duan[1] Wen Li[1]*
[1] University of Electronic Science and Technology of China [2] ETH Zürich
{lyhaolive, wangqin.ee, lxduan, liwenbnu}@gmail.com

Abstract

Regression is fundamental in computer vision and is widely used in various tasks including age estimation, depth estimation, target localization, etc. *However, real-world data often exhibits imbalanced distribution, making regression models perform poorly especially for target values with rare observations (known as the imbalanced regression problem). In this paper, we reframe imbalanced regression as an imbalanced generalization problem. To tackle that, we look into the loss sharpness property for measuring the generalization ability of regression models in the observation space. Namely, given a certain perturbation on the model parameters, we check how model performance changes according to the loss values of different target observations. We propose a simple yet effective approach called Balanced Sharpness-Aware Minimization (BSAM) to enforce the uniform generalization ability of regression models for the entire observation space. In particular, we start from the traditional sharpness-aware minimization and then introduce a novel targeted reweighting strategy to homogenize the generalization ability across the observation space, which guarantees a theoretical generalization bound. Extensive experiments on multiple vision regression tasks, including age and depth estimation, demonstrate that our BSAM method consistently outperforms existing approaches. The code is available here.*

1. Introduction

Regression tasks are fundamental in computer vision [16, 18, 30, 34], encompassing various applications from age estimation to depth estimation. Unlike classification which predicts discrete categories, regression tasks require models to learn more precise continuous mappings, inherently more challenging to optimize and generalize. This challenge is further compounded in real-world scenarios where data imbalance is prevalent. For instance, in age estimation tasks [21, 26], data from elderly subjects is typically more scarce compared to middle-aged and young subjects. Similar patterns of imbalance persist across various vision tasks, from depth estimation [29] where certain depth ranges dominate the data distribution, to image quality assessment [19] where extreme quality scores are rarely observed.

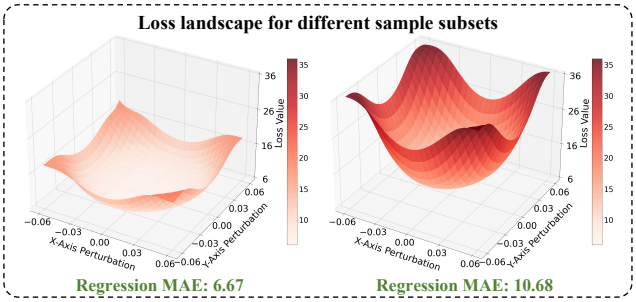

Figure 1. Visualization of loss landscapes for different sample subsets in AgeDB-DIR dataset. **Left**: Loss landscape computed over the entire dataset, exhibiting relatively smooth geometry with a lower MAE of 6.67, indicating better generalization. **Right**: Loss landscape for samples from low-density regions shows significantly increased sharpness and higher sensitivity to parameter perturbations, resulting in degraded generalization performance (MAE: 10.68). This stark contrast in both landscape geometry and quantitative metrics demonstrates the inherent challenge of maintaining consistent generalization ability across different density regions in imbalanced regression.

Prior research has established that regression tasks necessitate careful consideration of label continuity and inter-label relationships [14, 33, 35]. Various strategies have been proposed to address this challenge: [33] propose using Gaussian kernel smoothing during sample reweighting to maintain continuity between label weights. Others emphasize the importance of feature continuity during model optimization, employing contrastive learning [14, 35] and rank-based [35] approaches to impose additional constraints on model features.

While existing methods have shown promise, a fundamental challenge persists: the degraded model generalization under imbalanced data distributions, where test performance consistently deteriorates for low-density values that are underrepresented during training. To analyze this phe-

*The corresponding author

nomenon, we visualize the loss landscape of imbalanced regression models following the visualization methodology proposed in [17], as illustrated in Figure 1. Namely, by applying a controlled perturbation on the parameters of a regression model trained from imbalanced data, we can visualize the resulting loss variation of the test observations. When focusing specifically on samples from low-density regions of the data distribution, we observe significantly degraded performance in MAE, indicating poor model generalization in these regions, along with markedly increased sharpness in loss landscapes compared to the overall dataset. This observation aligns with established studies [7, 15] that demonstrate strong connections between loss landscape geometry and model generalization, where flatter minima often correlate with superior generalization performance.

Motivated by these findings, we propose Balanced Sharpness-Aware Minimization (BSAM) for imbalanced regression, a novel optimization framework that addresses the challenge of imbalanced regression through the lens of loss landscape geometry. While traditional Sharpness-Aware Minimization (SAM) [7] has demonstrated significant improvements in model generalization by seeking flat minima of the loss landscape, our analysis reveals that its direct application to imbalanced regression tasks is susceptible to distributional biases. Specifically, SAM's uniform treatment of all samples in the perturbation step leads to optimization trajectories biased toward high-density regions of the target distribution.

To address this, BSAM extends the theoretical foundations of SAM in two crucial aspects. First, rather than simply pursuing flat minima, BSAM aims to achieve consistent flatness of loss landscape across the entire observation space. Second, we introduce a targeted reweighting mechanism that dynamically adjusts each sample's influence during the perturbation calculation process. This approach effectively balances the contribution of samples across the entire target distribution spectrum, preventing the dominance of frequently observed target values while simultaneously ensuring robust generalization across all regions. Our framework maintains algorithmic simplicity while avoiding the complexity often associated with specialized loss functions or distribution smoothing techniques. Through extensive experimentation across multiple vision regression tasks, including age estimation and depth prediction, we demonstrate that BSAM consistently achieves state-of-the-art performance. Our empirical results validate the effectiveness of combining sharpness-aware optimization with targeted reweighting strategies. The main contributions of our work can be summarized as follows:

- We reframe the imbalanced regression problem from a novel perspective of generalization ability, revealing the connection between loss landscape geometry and model performance of the target distribution, especially in low-density regions.
- We propose Balanced Sharpness-Aware Minimization (BSAM), a simple yet effective framework that addresses the limitations of conventional SAM in imbalanced regression by integrating targeted reweighting mechanisms, derived from our generalization analysis.
- We validate the effectiveness of BSAM through comprehensive experiments including age estimation and depth prediction, achieving superior performance across multiple vision regression tasks.

2. Related Work

2.1. Regression for Vision Tasks

Regression tasks are fundamental in computer vision, underpinning a wide range of applications, including age estimation [21, 27, 34], depth estimation [6, 29], pose estimation [18, 20], and image quality assessment [16, 19]. Traditional regression methods typically employ l_1, l_2, and Huber loss [12] to learn a continuous mapping between input images and target values. Recent research has focused on the regression-specific characteristics such as the ordinal relationships inherent to the dataset have been utilized to design more effective loss functions and decompose the regression task into multiple binary classification tasks [8, 22, 26, 28]. Additionally, in the field of pose estimation, [2, 30] reformulate the regression problem as a segmentation task by generating heatmaps to represent continuous variables spatially. However, these approaches typically assume balanced data distributions, which rarely hold in real-world scenarios.

2.2. Imbalanced Regression

Imbalanced regression has received comparatively less attention than its classification counterpart. However, many real-world vision tasks such as depth estimation [29], age estimation [21, 27], and image quality assessment [19] often exhibit long-tail distributions, resulting in a severe imbalance where certain target values are significantly underrepresented. Early methods largely relied on SMOTE-based algorithms [5], which use linear interpolation for data augmentation. Recently, significant progress has been made in tackling this challenge. [33] proposed a comprehensive benchmark for imbalanced regression and introduced label and feature smoothing techniques based on local similarities. The connection between regression and classification losses has been investigated in several studies [23, 32, 36]; for instance, [36] demonstrated that enforcing constraints to encourage high-entropy feature spaces can enhance regression performance. Statistical approaches have also emerged in this area [25, 31]. For example, [25] modeled regression predictions as a Gaussian distribution to design a balanced

mean squared error (MSE) loss, addressing the imbalance in a probabilistic framework. Additionally, feature-label consistency constraints for contrastive learning [14, 35] and rank-based constraints [9] have shown promising results in this domain. Among these, the Rank-N-contrast method [35] stands out, achieving state-of-the-art performance through a two-stage training process: initially training a feature extractor using advanced contrastive learning techniques, followed by a separate regressor training phase. Despite these advances, we reframe imbalanced regression as a generalization problem, addressing the inconsistent model behavior between training and test distributions across different value ranges.

2.3. Loss Landscape

The geometry of the loss landscape plays a crucial role in understanding the generalization capabilities of deep neural networks. Extensive research has demonstrated that flatter minima typically correlate with superior generalization performance [4, 13, 15]. Leveraging this insight, Sharpness-Aware Minimization (SAM) [7] has been introduced to explicitly encourage flatness by minimizing the worst-case perturbation in the loss landscape. SAM has demonstrated significant improvements in robustness and generalization across various tasks [1, 3, 24]. Recent adaptations of SAM have extended its application to imbalanced learning scenarios. Notably, ImbSAM [37] and CC-SAM [38] have tailored the approach for class-imbalanced datasets. However, these methods primarily address classification challenges, leaving the challenging domain of imbalanced regression relatively unexplored (detailed analysis in Section 3.3).

3. Methodology

In imbalanced regression tasks, we are provided with a training set $\mathcal{S}_{tr} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ drawn from a distribution $\mathcal{D}_{tr}$, where $\mathbf{x}_i \in \mathbb{R}^{H\times W\times C}$ represents an RGB image of height H, width W, and channels C, and $y_i \in \mathbb{R}$ denotes the corresponding continuous label. Unlike classification tasks, the label space $\mathcal{Y}$ in regression is continuous, bounded by a lower bound L and an upper bound U, such that: $\mathcal{Y} = \{y \mid L \leq y \leq U\}$, where y is the continuous target value associated with each input data point. In general, a regression network can be described as a function $f_\theta : \mathbb{R}^{H\times W\times C} \to \mathbb{R}$, which maps the input $\mathbf{x}$ to a continuous output y. Here, θ represents the model parameters, which are learned by minimizing the loss function $\mathcal{L}_{tr}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell(f_\theta(\mathbf{x}_i), y_i)$ over the training set. The common choices for ℓ are l_1 or l_2. To quantify the imbalance in regression tasks, we use a histogram over the label space $\mathcal{Y}$ to represent the distribution of target values. The histogram divides the continuous label range from L to U into a series of intervals $b_1, ..., b_K$, each accumulating the frequency of samples falling within that interval.

This systematic partitioning enables us to quantitatively assess the distribution of samples across the continuous label space. Imbalance occurs when substantial variation in sample density exists across different intervals, manifesting as regions that are either sparsely populated (rare values) or densely populated (common values) with samples. The primary learning objective of imbalanced regression is to develop a model that can accurately predict values across the entire range of the target variable, ensuring low loss on $\mathcal{L}_{test}(\theta)$ balanced data distribution $\mathcal{D}_{test}$.

3.1. Generalization Analysis for Imbalanced Regression

A regression model trained with imbalanced data can suffer from deteriorated generalization ability. As shown in Figure 1, the test performance consistently deteriorates for low-density values that are underrepresented during training. To further quantify this generalization ability difference, we analyze the loss landscape for the overall dataset and the low-density region. Given a controlled perturbation on the model parameter, the loss change is much more rapid for the low-density samples (right) than the dataset average (left). All of these demonstrate the inherent challenge of maintaining consistent generalization ability across different density regions in imbalanced regression.

To address these generalization challenges, we adopt the framework of Sharpness-Aware Minimization [7] which improves model generalization by seeking parameter values whose entire neighborhoods have uniformly low training loss. The theoretical foundation for this approach is established through a theorem that bounds generalization ability based on neighborhood-wise training loss:

Lemma 3.1. *For any $\rho > 0$, with high probability over training set $\mathcal{S}$ generated from distribution $\mathcal{D}$,*

$$\mathcal{L}_{\mathcal{D}}(\theta) \leq \max_{\|\epsilon\|_2\leq\rho} \mathcal{L}_{\mathcal{S}}(\theta+\epsilon) + h(\|\theta\|_2^2/\rho^2), \qquad (1)$$

where $h : \mathbb{R}_+ \to \mathbb{R}_+$ is a strictly increasing function.

Examining the right side of the inequality reveals two key components. The first term describes the worst-case loss within a ρ-radius neighborhood of the current parameters θ, effectively characterizing the local sharpness of the loss landscape. The second term $h(\|\theta\|_2^2/\rho^2)$ represents a regularization component typically controlled through weight decay in deep learning practice. This theorem establishes a fundamental relationship: when the training set $\mathcal{S}$ is sampled from distribution $\mathcal{D}$, the generalization error on distribution $\mathcal{D}$ can be bounded by the maximum perturbation error within a neighborhood on the training set $\mathcal{S}$.

However, this formulation reveals a critical limitation in imbalanced regression settings. According to Lemma 3.1, when measuring model sharpness using the training set $\mathcal{S}_{tr}$,

SAM can only guarantee generalization to the training distribution $\mathcal{D}_{tr}$. This limitation becomes particularly acute in imbalanced scenarios, where our target is the balanced test distribution $\mathcal{D}_{te}$. Under such distribution shifts ($\mathcal{D}_{tr} \neq \mathcal{D}_{te}$), conventional SAM's generalization guarantees deteriorate significantly, especially for underrepresented samples. The inadequate exploration of the loss landscape in these underrepresented regions leads to unreliable sharpness estimates and compromised generalization performance. These theoretical insights highlight the necessity for a more sophisticated approach that explicitly addresses the distribution shift between training and testing environments.

3.2. Balanced Sharpness-Aware Minimization

As analyzed in the previous section, the conventional SAM approach, while effective for standard learning scenarios, fails to provide adequate generalization guarantees for the balanced test set when faced with the imbalance of the training set. To address this, we propose Balanced Sharpness-Aware Minimization (BSAM), a novel approach that explicitly accounts for the target balanced distribution in its optimization framework.

To establish our method, we provide an analytical solution for computing the maximum perturbation ϵ^* in Lemma 3.1. Specifically, given a parameter vector θ and a small perturbation radius ρ, the worst-case perturbation within the ρ-ball can be approximated through first-order Taylor expansion of $\mathcal{L}_{\mathcal{S}}(\theta + \epsilon)$:

$$\epsilon^* = \underset{\|\epsilon\|_p \leq \rho}{\operatorname{argmax}} \mathcal{L}_{\mathcal{S}}(\theta + \epsilon) \tag{2}$$

$$\approx \underset{\|\epsilon\|_p \leq \rho}{\operatorname{argmax}} \left[\mathcal{L}_{\mathcal{S}}(\theta) + \epsilon^T \nabla_\theta \mathcal{L}_{\mathcal{S}}(\theta)\right] \tag{3}$$

$$= \rho \cdot \operatorname{sign}\left(\nabla_\theta \mathcal{L}_{\mathcal{S}}\left(\theta\right)\right) \cdot \frac{\left|\nabla_\theta \mathcal{L}_{\mathcal{S}}\left(\theta\right)\right|^{q-1}}{\left\|\nabla_\theta \mathcal{L}_{\mathcal{S}}\left(\theta\right)\right\|_q^{q/p}}, \tag{4}$$

where $\rho \geq 0$ is a hyperparameter and $1/p + 1/q = 1$. The last equation gives the closed-form solution for the maximum perturbation ϵ^* that maximizes the local loss increase by the Dual Norm Problem solution. Once this worst-case perturbation ϵ^* is determined, we can update the model parameters θ_t of step t using the following optimization step:

$$\theta_{t+1} = \theta_t - \alpha_t \cdot \nabla \mathcal{L}_{\mathcal{S}}(\theta_t + \epsilon^*), \tag{5}$$

where α_t is the learning rate at step t. Here, we have omitted the explicit weight decay term for simplicity.

Notably, the primary impact of distribution discrepancy manifests in the computation of the maximum perturbation ϵ^*. Therefore, to effectively address the imbalance issue and ensure generalization to the balanced test distribution $\mathcal{D}_{te}$, we propose a simple yet effective reweighting strategy that incorporates balance-aware weights directly into the perturbation computation process.

Specifically, to bridge the gap between training and testing label distributions, we first examine the relationship between Empirical Risk Minimization (ERM) $\mathcal{L}_{erm}$[1] and Balanced Error (BE) $\mathcal{L}_{be}$. Let $P_{tr}(k)$ denote the empirical distribution of interval k in the training set and $P_{te}(k)$ represents the desired uniform distribution for our test set. Through importance reweighting, we can bridge these two objectives:

$$\mathcal{L}_{be}(\theta) = \sum_{k=1}^{K} P_{te}(k) \cdot \mathbb{E}_{(\mathbf{x},y)\sim S^{b_k}} \ell\left(f_\theta\left(\mathbf{x}\right), y\right) \tag{6}$$

$$= \sum_{k=1}^{K} P_{te}(k) \cdot \frac{P_{tr}(k)}{P_{tr}(k)} \cdot \mathbb{E}_{(\mathbf{x},y)\sim S^{b_k}} \ell\left(f_\theta\left(\mathbf{x}\right), y\right) \tag{7}$$

$$= \mathbb{E}_{(\mathbf{x},y)\sim \mathcal{S}_{tr}} w(k(y)) \cdot \ell\left(f_\theta\left(\mathbf{x}\right), y\right), \tag{8}$$

where $w(k(y)) = \frac{P_{te}(k(y))}{P_{tr}(k(y))}$ represents the importance weight for samples from bin k. Building upon this insight, we incorporate this reweighting mechanism into the computation of maximum perturbation ϵ:

$$\boldsymbol{\epsilon} = \underset{\|\boldsymbol{\epsilon}\|_p \leq \rho}{\operatorname{argmax}} \mathbb{E}_{(\mathbf{x},y)\sim \mathcal{S}_{tr}} \left[w(k(y)) \cdot \ell\left(f_{\theta+\epsilon}(\mathbf{x}), y\right)\right] \tag{9}$$

$$\approx \rho \cdot \operatorname{sign}(\nabla \mathcal{L}^w_{\mathcal{S}_{tr}}(\theta)) \cdot \frac{\left|\nabla \mathcal{L}^w_{\mathcal{S}_{tr}}(\theta)\right|^{q-1}}{\left\|\nabla \mathcal{L}^w_{\mathcal{S}_{tr}}(\theta))\right\|_q^{q/p}}, \tag{10}$$

$$\mathcal{L}^w_{\mathcal{S}_{tr}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} w\left(k(y_i)\right) \cdot \ell\left(f_\theta\left(\mathbf{x}_i\right), y_i\right). \tag{11}$$

It is important to note that BSAM differs fundamentally from traditional loss reweighting approaches. While loss reweighting modifies the optimization objective used in parameter updates, BSAM introduces the targeted reweighting specifically in the perturbation computation step, making it independent of the specific form of the loss function $\mathcal{L}$ used in parameter updates, which we will verify in the experimental section.

3.3. Overall algorithm and discussion

Our analysis reveals that there is an inherent disparity in loss landscape geometry across different density regions, leading to inconsistent generalization ability across the target distribution. This observation motivates us to develop an optimization framework that maintains uniform generalization ability across the entire observation space by seeking loss sharpness with consistent flatness regardless of the training set distribution density. Based on the analysis above, we present the complete algorithm of Balanced

[1] $\mathcal{L}_{erm} = \mathbb{E}_{(\mathbf{x},y)\sim \mathcal{S}_{tr}} \ell\left(f_\theta\left(\mathbf{x}\right), y\right)$

Algorithm 1 Balanced Sharpness-Aware Minimization (BSAM)

Require: Training set $\mathcal{S}_{tr}$, initial parameters θ_0, learning rate $\{\alpha_t\}_{t=1}^{T}$, neighborhood size ρ, number of bins K

Divide label space into K equal-width interval $b_1, ..., b_K$
Calculate bin frequencies $\{N_k\}_{k=1}^{K}$ from training set
Compute importance weights $w(k)$ for each bin k
for $t = 1$ to T **do**
  Sample a mini-batch $\mathcal{B}_t$ from $\mathcal{S}_{tr}$
  Compute weighted loss function by Eq. 11
  Calculate the perturbation ϵ^* by Eq. 10
  Update parameters:
  $\theta_{t+1} = \theta_t - \alpha_t \cdot \nabla \mathcal{L}(\theta_t + \epsilon^*)$
end for
return Final model parameters θ_T

Sharpness-Aware Minimization (BSAM) for imbalanced regression tasks. The overall procedure is summarized in Algorithm 1.

Discussion. Our BSAM differs from existing SAM variants [37, 38] for imbalanced learning in several key aspects. First, while both [37] and [38] focus on classification tasks, BSAM is specifically designed for regression problems. Second, [37] selectively computes perturbations using only minority classes based on a predefined threshold. However, as indicated by Lemma 3.1, this approach potentially compromises the generalization capability across the entire data distribution, which we empirically verify in our experiments. On the other hand, [38] calculates perturbations separately for each class, which cannot guarantee that the perturbation maximizes the loss concerning the entire target distribution $\mathcal{D}_{te}$. Moreover, its requirement to iterate through all classes during each update becomes computationally intractable for regression tasks where the label space is continuous. In contrast, BSAM employs a simple yet effective targeted weighted perturbation strategy that doesn't introduce additional computational complexity to the original SAM while effectively capturing the local geometry of the loss landscape for $\mathcal{D}_{test}$.

4. Experiments

4.1. Datasets

We conduct experiments on three public benchmarks for deep imbalanced regression, including two age estimation datasets and one depth estimation dataset. To ensure fair comparisons, we follow the dataset splits as described in [33]. For age estimation, we use bins with an interval of 1 year, while for depth estimation, bins are defined with an interval of 0.1 meters, consistent with previous related works [9, 25, 33].

- **AgeDB-DIR** is an imbalanced facial age estimation benchmark, derived from the AgeDB dataset [21]. It comprises $16,488$ manually curated noise-free labeled images, with the training set containing $12,208$ images, and both the validation and test sets containing $2,140$ images each.
- **IMDB-WIKI-DIR** is an imbalanced facial age estimation benchmark, derived from the IMDB-WIKI dataset [27]. It comprises $213,554$ images semi-automatically collected and annotated from the IMDB and Wikipedia websites. The training set consists of $191,509$ images, while both the validation and test sets contain $11,022$ images each.
- **NYUD2-DIR** is an imbalanced depth estimation benchmark, derived from the NYU Depth Dataset V2 [29]. It includes $50,688$ images for training and 654 images for testing. Notably, following the default setting of [33], the test dataset of NYUD2-DIR only considers a randomly selected $9,357$ pixels per bin from the 654 test images to ensure the test set is balanced, corresponding to the minimum number of pixels in any bin of the test set.

4.2. Implementation Details

For all experiments, unless otherwise specified, we use square-root-inverse weighting for calculating $w(k)$ and set $p = 2$ for calculating ρ. All experiments are conducted on Tesla V100 GPUs with a PyTorch implementation.

AgeDB-DIR and IMDB-WIKI-DIR Benchmark. Following the experiment setup used in RnC [35], ResNet-18 [10] is utilized as the backbone, with the same data augmentations applied across all compared methods, including random crop, resize (with random horizontal flip), and color distortions. We use the l_1 loss (Vanilla) combined with the square-root-inverse weighting variant (SQINV) as the primary optimization objective, maintaining a fixed batch size of 256. For evaluation, Mean Squared Error (MAE) and Geometric Mean (GM) are selected as metrics, quantifying the model's accuracy and fairness in predictions, respectively. Following the approach in [25], we also present the results on the balanced Mean Absolute Error (bMAE) metric. Following [33], we divide the target space into three disjoint subsets: many-shot region (intervals with over 100 training samples), medium-shot region (intervals with $20 \sim 100$ training samples), and few-shot region (intervals with under 20 training samples).

NYUD2-DIR Benchmark. Following [9, 33], we adopt ResNet-50 [10] as the backbone, integrating it within an encoder-decoder architecture [11]. The l_2 loss (Vanilla) with the square-root-inverse weighting variant (SQINV) is applied as the optimization objective, with a batch size of 32. Evaluation is conducted using root mean squared er-

Table 1. **Main results on *AgeDB-DIR* benchmark.** Results marked with ⋆ are directly quoted from their original paper while results marked with † are obtained through our reproduction and RNC [35] following the RNC training protocol.

	MAE ↓				GM ↓			
	All	Man.	Med.	Few	All	Man.	Med.	Few
LDS [33]⋆	7.42	6.83	8.21	10.79	4.85	4.39	5.80	7.03
FDS [33]⋆	7.55	6.99	8.40	10.48	4.82	4.49	5.47	6.58
RankSim [9]⋆	6.91	6.34	7.79	9.89	4.28	3.92	4.88	6.89
ConR [14]+LDS [33]⋆	7.16	6.61	7.97	9.62	4.51	4.21	4.92	5.87
ConR [14]+FDS [33]⋆	7.08	6.46	7.89	9.80	4.31	4.01	5.25	6.92
RankSim [9]†	6.51	-	-	-	-	-	-	-
RNC [35]†	6.14	-	-	-	-	-	-	-
LDS [33]†	6.350	5.925	7.078	8.355	3.963	3.721	4.441	**5.249**
Ordinal Entropy [36]†	6.360	5.778	7.059	9.921	3.987	**3.656**	4.382	6.958
Vanilla†	6.690	5.959	7.740	10.688	4.254	3.734	5.281	8.021
SQINV†	6.391	5.955	7.155	8.390	4.039	3.774	4.577	5.425
BSAM	**6.067**	**5.801**	**6.304**	**7.928**	**3.895**	3.748	**3.925**	5.473

Table 2. **Main results on *IMDB-WIKI-DIR* benchmark.** Results marked with ⋆ are directly quoted from their original paper while results marked with † are obtained through our reproduction following the RNC training protocol.

	MAE ↓				GM ↓			
	All	Man.	Med.	Few	All	Man.	Med.	Few
LDS [33]⋆	7.83	7.31	12.43	22.51	4.42	4.19	7.00	13.94
FDS [33]⋆	7.83	7.23	12.60	22.37	4.42	4.20	6.93	13.48
RankSim [9]⋆	7.42	6.84	12.12	22.13	4.10	3.87	6.74	12.78
ConR [14]+LDS [33]⋆	7.43	6.84	12.38	21.98	4.06	3.94	6.83	12.89
ConR [14]+FDS [33]⋆	7.29	6.90	12.01	21.72	4.02	3.83	6.71	12.59
Ordinal Entropy [36]†	7.322	6.629	13.154	23.235	4.000	3.708	7.977	14.961
RNC [35]†	7.466	6.757	13.511	23.168	4.043	3.729	8.516	15.540
LDS [33]†	7.214	6.686	11.491	20.659	4.030	3.835	6.127	**12.425**
Vanilla†	7.358	6.644	13.391	23.544	4.110	3.784	8.789	16.101
SQINV†	7.040	6.508	11.263	**21.301**	3.921	3.710	6.233	14.457
BSAM	**6.811**	**6.294**	**10.823**	21.339	**3.765**	**3.580**	**5.663**	13.609

ror (RMSE) and threshold accuracy δ_1.

We show the label distributions for three datasets and the detailed formulations of our evaluation metrics and reweighting strategies for w in the supplementary material.

4.3. Analysis

4.3.1. Comparisons with state-of-the-art methods

We compare our proposed methods with previous state-of-the-art imbalance regression approaches on three benchmarks in Table 1, Table 2, and Table 3 respectively. Specifically, we selected two baseline methods, *Vanilla* and *SQINV*, and multiple state-of-the-art regression learning schemes: 1) distribution rebalance methods include the distribution smoothing [33] methods (LDS and FDS) and the Balanced MSE [25], 2) feature space constraints methods include the contrastive learning [14, 35], rank-based constraints [9] and the entropy constraints [36] methods.

Given that the recent RnC method [35] established a comprehensive experimental protocol for data augmentations and model training strategies in age prediction tasks, which leads to stronger baselines, we conducted our age prediction experiments based on RnC training protocol as mentioned in Section 4.2. For a fair comparison, we report the baseline results quoted from [35], which explains the missing metrics for some methods. We also re-produce other relevant methods following the identical RnC training strategies. We also present the original results of these methods in Table 1 and 2 for comprehensive evaluation.

AgeDB-DIR Benchmark. We evaluate our method on the AgeDB-DIR benchmark and compare it with state-of-the-art approaches. As shown in Table 1, our method achieves superior performance across different metrics. Specifically, our BSAM with SQINV achieves the lowest MAE of 6.067 on All settings and GM of 3.895, surpassing previous methods by a clear margin. The improvement is particularly significant in Few-shot scenarios, where our method reduces the MAE from 10.688 to 7.928, demonstrating its effectiveness in handling data sparsity.

Table 3. **Main results on *NYUD2-DIR* benchmark.** Results marked with ⋆ are directly quoted from their original paper.

	RMSE↓				δ_1↑			
	All	Man.	Med.	Few	All	Man.	Med.	Few
FDS [33]⋆	1.442	0.615	0.940	2.059	0.681	0.760	0.695	0.596
LDS [33]⋆	1.387	0.671	0.913	1.954	0.672	0.701	0.706	0.630
Balanced MSE (BNI) [25]⋆	1.283	0.787	0.870	1.736	0.694	0.622	0.806	0.723
Balanced MSE (BNI) [25] + LDS [33]⋆	1.319	0.810	0.920	1.820	0.681	0.601	0.695	0.648
ConR [14] + FDS [33]⋆	1.299	0.613	0.836	1.825	0.696	**0.800**	0.819	0.701
ConR [14] + LDS [33]⋆	1.323	0.786	**0.823**	1.852	0.700	0.632	**0.827**	0.702
Vanilla⋆	1.477	**0.591**	0.952	2.123	0.677	0.777	0.693	0.570
SQINV	1.341	0.604	0.832	1.912	0.717	0.769	0.752	0.653
BSAM	**1.272**	0.728	1.046	**1.705**	**0.727**	0.742	0.695	**0.724**

IMDB-WIKI-DIR Benchmark. IMDB-WIKI-DIR is a particularly challenging benchmark due to its dual challenges: label noise and data imbalance. Despite these intrinsic difficulties, our method demonstrates remarkable improvements over existing approaches as shown in Table 2. Specifically, BSAM achieves the best overall performance with 6.811 MAE on all samples, substantially outperforming conventional methods like RNC. The effectiveness of our approach is consistent across different data density regions: it achieves 6.294 MAE for many-shot samples, and 10.823 for medium-shot samples while maintaining robust performance for few-shot samples. The GM metric further demonstrates the superiority of our method, achieving a GM score of 3.765, which is notably better than other methods. The results validate the effectiveness of our approach in handling complex age estimation tasks.

NYUD2-DIR Benchmark. We further evaluate our method on the NYUD2-DIR benchmark, with results shown in Table 3. Our approach achieves state-of-the-art performance with the lowest RMSE of 1.272 on All settings, outperforming strong baselines including Balanced MSE. Notably, our method demonstrates substantial improvements in the challenging Few-shot regime, indicating its effectiveness in handling the imbalance problem. For the δ_1 metric, our method achieves competitive results of 0.727 on All settings, demonstrating its capability to maintain high accuracy across different evaluation criteria.

4.3.2. The bMAE metric

As shown in Table 4, we present the results of the bMAE metric evaluation on both the AgeDB-DIR and IMDB-WIKI-DIR benchmarks. It is evident that BSAM consistently outperforms SQINV across different data density regions on both benchmarks in terms of bMAE measurement. Notably, in the few-shot region, where bMAE more effectively evaluates model performance, as stated in [25], BSAM demonstrates significant improvements over SQINV.

Table 4. The bMAE metric (lower is better) on AgeDB-DIR and IMDB-WIKI-DIR benchmark.

Datasets	Methods	All	Many	Med.	Few
AgeDB	SQINV	7.075	5.955	7.203	9.278
	BSAM	**6.734**	**5.801**	**6.340**	**8.890**
IMDB	SQINV	11.838	6.673	13.718	29.826
	BSAM	**11.282**	**6.356**	**11.333**	**26.466**

Table 5. The MAE of BSAM with vanilla regression loss on AgeDB-DIR benchmark.

Methods	All	Man.	Med.	Few
Vanilla	6.690	5.959	7.740	10.688
Vanilla + Ours	**6.427**	**5.856**	**7.116**	**9.915**

4.3.3. Combinations with different regression loss

One of the distinctive features of BSAM is its flexible design regarding the choice of regression loss functions for parameter optimization. To verify this flexibility, we combined BSAM with vanilla regression loss (l_1), denoted as "Vanilla + Our". Table 5 presents the comparative results on the AgeDB-DIR benchmark. As shown, integrating BSAM with the vanilla loss yields consistent improvements across all data density regions. Specifically, our approach reduces the overall MAE from 6.690 to 6.427. These results indicate that BSAM's reweighting mechanism during perturbation calculation can enhance model performance independent of the specific form of loss function used, making it a versatile approach that can be combined with different regression losses to improve performance across different data density regions.

4.3.4. Comparisons with different SAM

To compare the effectiveness of our improvements to SAM, we evaluate various SAM-based methods as shown in Table 6. The standard SAM model improves the baseline MAE from 6.391 to 6.230, while the ImbSAM variant achieves an MAE of 6.184. Our proposed method further

Table 6. The MAE of BSAM with different SAM on AgeDB-DIR benchmark.

Methods	All	Man.	Med.	Few
SQINV	6.391	5.955	7.155	8.390
+ SAM [7]	6.230	**5.639**	7.313	8.822
+ imbSAM [37]	6.184	5.844	6.799	**7.695**
+ BSAM	**6.067**	5.801	**6.304**	7.928

Table 7. The MAE of BSAM with different reweighting on AgeDB-DIR benchmark.

Methods	All	Man.	Med.	Few
SQINV	6.391	5.955	7.155	8.390
INV-BSAM	6.146	5.806	6.668	**7.921**
SQINV-BSAM	**6.067**	**5.801**	**6.304**	7.928

Table 8. Comparisons of λ_{max} and $Tr(H)$ averaged over three experiments on the few-shot scenarios of AgeDB-DIR benchmark.

Metric	SQINV	BSAM
$\lambda_{max} \downarrow$	141.21	**86.51**
$Tr(H) \downarrow$	653.53	**90.88**

reduces the MAE to 6.067, demonstrating superior performance. While we observe performance variations across different data distribution regions, it is important to note that BSAM is designed to enhance generalization performance across the entire distribution. In contrast, ImbSAM primarily focuses on optimization for few-shot regions, which can potentially compromise overall performance across the complete dataset.

4.3.5. Different Reweighting for the perturbation

As shown in Table 7, we validate the impact of two approaches for calculating $w(k)$ on BSAM performance: inverse-frequency weighting (*INV*-BSAM) and square-root-inverse weighting (*SQINV*-BSAM). *INV*-BSAM reduces the MAE to 6.146, while *SQINV*-BSAM further improves performance, achieving an MAE of 6.067. Both methods show notable improvement compared to the baseline. This demonstrates that our approach is not dependent on a specific reweighting method but rather focuses on balancing the influence of data from all density regions during perturbation calculation.

4.3.6. The effectiveness for low-density regions

Table 8 provides a quantitative analysis of the loss landscape geometry through two key metrics: the maximum eigenvalue λ_{max} and the trace of the Hessian matrix $Tr(H)$. Lower values for both metrics indicate smoother, flatter loss landscapes associated with better generalization. On the AgeDB-DIR benchmark in the few-shot scenarios, BSAM significantly outperforms SQINV. These dramatic improvements suggest that BSAM is particularly effective at creating flatter loss landscapes in low-density regions, which explains its superior generalization performance for underrepresented samples.

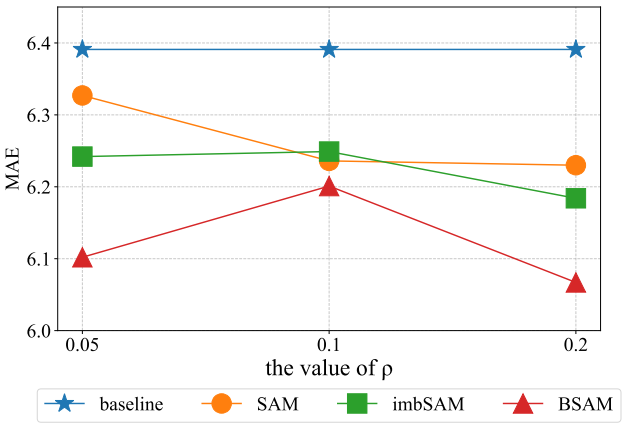

Figure 2. Ablation study for the value of $\rho \in \{0.05, 0.1, 0.2\}$ on AgeDB-DIR benchmark.

4.4. Ablation study

To evaluate the influence of the hyperparameter ρ, we conducted an ablation study on the AgeDB-DIR benchmark dataset. As shown in Figure 2, we compared the mean absolute error (MAE) across different values of $\rho \in \{0.05, 0.1, 0.2\}$ for various methods: baseline, SAM [7], imbSAM [37], and our proposed BSAM approach. The results validate that our BSAM method effectively leverages sharpness to address imbalanced regression tasks, outperforming existing methods under different values of ρ.

5. Conclusion

In this paper, we have presented Balanced Sharpness-Aware Minimization (BSAM), a novel approach that effectively addresses the challenge of imbalanced regression through principled integration of loss landscape sharpness and targeted reweighting mechanisms. Our analysis reveals the critical limitations of conventional SAM in handling underrepresented samples, leading to the development of a targeted reweighting mechanism that effectively balances model generalization across the entire data distribution. This simple yet effective approach maintains the computational efficiency of standard SAM while achieving superior performance across three challenging benchmarks including AgeDB-DIR, IMDB-WIKI-DIR, and NYUD2-DIR.

6. Acknowledgment

This work is supported by the National Natural Science Foundation of China (No.62476051, No.62176047), the Sichuan Natural Science Foundation (No.2024NSFTD0041), and the Sichuan Science and Technology Program (No.2021YFS0374).

References

[1] Dara Bahri, Hossein Mobahi, and Yi Tay. Sharpness-aware minimization improves language model generalization. *arXiv preprint arXiv:2110.08529*, 2021. 3

[2] Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 717–732. Springer, 2016. 2

[3] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34: 22405–22418, 2021. 3

[4] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019. 3

[5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 2

[6] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 2

[7] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. 2, 3, 8

[8] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 2

[9] Yu Gong, Greg Mori, and Frederick Tung. RankSim: Ranking similarity regularization for deep imbalanced regression. In *International Conference on Machine Learning*, 2022. 3, 5, 6

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[11] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *IEEE Winter Conference on Applications of Computer Vision*, 2019. 5

[12] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992. 2

[13] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019. 3

[14] Mahsa Keramati, Lili Meng, and R. David Evans. Conr: Contrastive regularizer for deep imbalanced regression. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 3, 6, 7

[15] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017. 2, 3

[16] Jun-Tae Lee and Chang-Su Kim. Image aesthetic assessment based on pairwise comparison a unified approach to score regression, binary classification, and personalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1191–1200, 2019. 1, 2

[17] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Neural Information Processing Systems*, 2018. 2

[18] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11025–11034, 2021. 1, 2

[19] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katie Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2

[20] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2640–2649, 2017. 2

[21] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017. 1, 2, 5

[22] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928, 2016. 2

[23] Silvia L Pintea, Yancong Lin, Jouke Dijkstra, and Jan C van Gemert. A step towards understanding why classification helps regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19972–19981, 2023. 2

[24] Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, Arihant Jain, and Venkatesh Babu Radhakrishnan. A closer look at smoothness in domain adversarial training. In *International conference on machine learning*, pages 18378–18399. PMLR, 2022. 3

[25] Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced mse for imbalanced visual regression. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7926–7935, 2022. 2, 5, 6, 7

[26] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE International Conference on computer vision workshops*, pages 10–15, 2015. 1, 2

[27] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2):144–157, 2018. 2, 5

[28] Deval Shah, Zi Yu Xue, and Tor Aamodt. Label encoding for regression networks. In *International Conference on Learning Representations*, 2022. 2

[29] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 1, 2, 5

[30] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems*, 27, 2014. 1, 2

[31] Ziyan Wang and Hao Wang. Variational imbalanced regression: Fair uncertainty quantification via probabilistic smoothing. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[32] Haipeng Xiong and Angela Yao. Deep imbalanced regression via hierarchical classification adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23721–23730, 2024. 2

[33] Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *International conference on machine learning*, pages 11842–11851. PMLR, 2021. 1, 2, 5, 6, 7

[34] Dong Yi, Zhen Lei, and Stan Z Li. Age estimation by multi-scale convolutional network. In *Asian conference on computer vision*, pages 144–158. Springer, 2014. 1, 2

[35] Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: Learning continuous representations for regression. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 3, 5, 6

[36] Shihao Zhang, Linlin Yang, Michael Bi Mi, Xiaoxu Zheng, and Angela Yao. Improving deep regression with ordinal entropy. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 6

[37] Yixuan Zhou, Yi Qu, Xing Xu, and Hengtao Shen. Imbsam: A closer look at sharpness-aware minimization in class-imbalanced recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11345–11355, 2023. 3, 5, 8

[38] Zhipeng Zhou, Lanqing Li, Peilin Zhao, Pheng-Ann Heng, and Wei Gong. Class-conditional sharpness-aware minimization for deep long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3499–3509, 2023. 3, 5