

ENHANCING DEEP IMBALANCED REGRESSION VIA FROBENIUS NORM REGULARIZATION

Anonymous authors

Paper under double-blind review

ABSTRACT

Deep Imbalanced Regression (DIR) aims to train a deep neural network (DNN) model specified for the regression tasks from an imbalanced training distribution and generalize well on an unseen balanced testing distribution. While modern solutions have achieved significant progress in DIR, the performance of the samples still varies a lot across the different shots. For instance, the samples in the majority-shot always outperform the underrepresented (median and few-shot) samples, which motivates us to investigate whether we can leverage the well-trained majority-shot samples to help the other under-trained samples. Empirically, we observe that previous solutions in DIR often produce ordinal feature Frobenius norms across the majority-shot samples and considerably lower training Mean-Absolute-Error (MAE). Meanwhile, the underrepresented samples often violate the ordinality of the majority-shot Frobenius norms and exhibit a high training MAE. As a result, this demonstrates that compared to the majority-shot samples, the underrepresented samples are still under-fitted during training and the ordinality of the Frobenius norm can also be treated as an indicator to identify the training performance. Motivated by this observation, we first analyze why the ordinality of the Frobenius norm can result in good training performance across the labels. Then, we introduce a feature regularization to encourage the feature Frobenius norms to be ordinal for all labels during the training process. Moreover, we propose a novel model training strategy that incorporates the knowledge from the well-trained majority samples to help the underrepresented samples. By training a linear model from the majority-shot samples to predict the feature Frobenius norm of underrepresented samples, we fine-tune the previously trained model to enhance the outcomes of underrepresented samples. Extensive experiments over the real-world datasets also validate the effectiveness of our proposed method. Code can be found in :Here

1 APPENDIX

1.1 DISCUSSIONS

Distance metrics. In the loss function \mathcal{L}_{KL} , the distance function which aims to measure the distance between the two arbitrary labels is defined as $d(\cdot, \cdot)$. Since the elements in $d(\cdot, \cdot)$ is always two integer, it is not possible to use the cosine similarity in d . We adopt MAE in our main paper, the reason why we did not use MSE is that the MSE would enlarge the distance of d in the quadratic scale. Therefore, it would lead the values in distance set with huge difference and consequently make the probability distribution sparse, which limits the effectiveness of utilizing distribution divergence.

More implementation details. For the first phase, we use the learning rate as $1e-3$ with weight decay. For the second phase, we use the learning rate as $1e-4$ and we use $5e-5$ as the learning rate for the phase three. To make a fair comparison between our method and other works, following Yang et al. (2021), we use the MAE and Geometric Mean (GM) to evaluate the AgeDB-DIR and IMDB-WIKI-DIR. We use the MSE and Pearson Correlation to evaluate the STS-B-DIR. We report results for the four subsets: All, Many, Median, and Few. for each dataset to identify the effectiveness of our method.

Discussion on our method and the Lim et al. (2025). Compared to Lim et al. (2025), although we both used the probability distribution alignment to achieve the ordinality, but our work differs a lot. Firstly, our work differs from Lim et al. (2025) in the motivation, based on our empirical observation, we try to encourage the feature Frobenius norm to maintain the ordinal as the majority-shot samples while Lim et al. (2025) tried to regularize on the feature space. Secondly, Lim et al. (2025) introduced the additional parameters to the model (e.g. ResNet) training in implementation. Therefore, our method is distinct from Lim et al. (2025).

1.2 ABLATION STUDY ON DIFFERENT SETTING OF THE EPOCH

To investigate the impact of the epoch settings across the three phases, we provide an ablation study on the different settings of the epoch on the regression model training (e.g., the ResNet-18 on AgeDB-DIR) e_{main} , the Frobenius linear model f_F training e_F and the fine-tuning phase on the previously trained main model (e.g., the ResNet-18 on AgeDB-DIR) e_{SFT} . To make a fair comparison, we set the total number of training epochs $e_{total} = e_{main} + e_F + e_{SFT}$ as the $e = 100$, which is the same experimental setting as the previous methods. To conduct the ablation study on the three-phase epoch, we strictly constrain the total training epoch to be $e = 100$, e.g., $e_{main} = 70$, $e_F = 20$ and $e_{SFT} = 10$. To better observe the varying of MAE across different epochs, we provide the ablation study on the setting of different epochs in Fig. 1 for AgeDB-DIR and Fig. 2 for IMDB-WIKI-DIR.

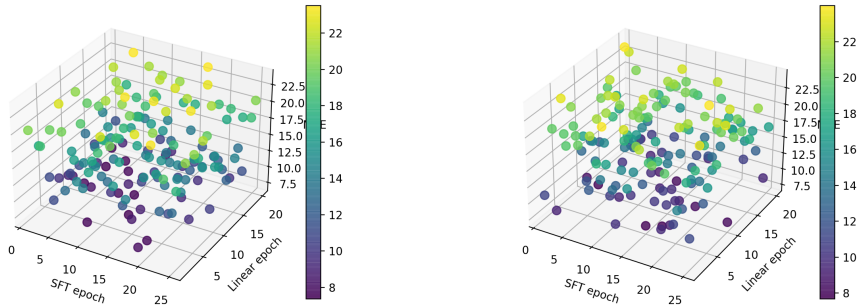


Figure 1: Ablation study on the different setting of epochs on AgeDB-DIR Figure 2: Ablation study on the different setting of epochs on IMDB-WIKI-DIR

As we can observe from Fig. 1 and Fig. 2, higher e_F and e_{SFT} do not necessarily produce a better MAE. This is because that the higher e_F and e_{SFT} are, the lower e_{main} is. As a result, the main model is not well-trained. Meanwhile, the shallow linear Frobenius network model would easily

108 overfit with the high e_F . Therefore, the feature Frobenius norm they output would not have a high
 109 quality, which would then affect the performance of the latter phase. However, the lower e_F and
 110 e_{SFT} are also not guarantee a good overall MAE performance. The reason for this is that when the
 111 main model is well-trained at the Phase 1, there would leave less time for the f_F and fine-tuning.
 112 Consequently, they would have a higher probability of getting under-trained, especially for the Phase
 113 3 which fine-tunes the main model.

114 1.3 ABLATION STUDY ON HYPER-PARAMETERS : α AND β .

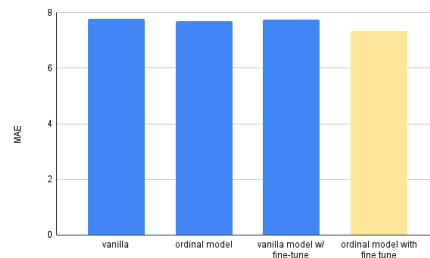
115
 116 The hyper-parameter in our method is the trade-off between the \mathcal{L}_{KL} and \mathcal{L}_{OE} : α . The goal of
 117 this hyper-parameter is to balance the impact of maintaining the ordinal feature Frobenius norm and
 118 the discriminative between different feature prototypes. Higher α would lead the model to focus
 119 more on learning discriminative feature prototypes, while lower α would lead the model to learn
 120 more ordinal information for the feature Frobenius norm. The hyper-parameter β aims to balance
 121 the ordinal penalty with the regression loss (e.g., MAE in our main paper). Higher β would lead the
 122 model to focus more on the ordinal characteristics while lower β would make the model difficult to
 123 obtain the ordinality. We provide the ablation study on the hyper-parameters in Fig. 1.3.
 124
 125



126
 127
 128
 129
 130
 131
 132
 133
 134
 135
 136
 137
 138 Figure 3: Ablation study on the different α on AgeDB-DIR
 139 Figure 4: Ablation study on the different β on AgeDB-DIR

140 1.4 ABLATION STUDY ON EACH PART OF THE TRAINING PHASE.

141
 142
 143 In this section, we compare the effectiveness between different training sections.
 144 Specifically, we compare the performance of four models on AgeDB-DIR as follows: vanilla model
 145 (MAE loss only), the ordinal model (MAE + ordinal without fine-tune), the vanilla model with fine-tune
 146 (MAE + with fine-tune on majority shots), and ordinal model with fine-tune (MAE + ordinal + with fine-tune
 147 on majority shots : Our method). We provide the ablation study on the effectiveness of each training phase in
 148 Fig. 1.4. We can observe that the although each part of the training phase can somewhat improve the perform-
 149 ance, the enhancement is very limited. The reason why each separate training phase can not significantly
 150 improve the performance is: only the ordinality in the Frobenius norm does not have a direct impact on the
 151 regression result. In the meantime, merely using the majority-shot samples would easily make the trained model to over-fit on the majority-shot and can-
 152 not improve the performance on the underrepresented samples. Therefore, all of the training phases
 153 are necessary in our proposed method.
 154
 155
 156
 157
 158
 159
 160
 161



162
 163
 164
 165
 166
 167
 168
 169
 170
 171
 172
 173
 174
 175
 176
 177
 178
 179
 180
 181
 182
 183
 184
 185
 186
 187
 188
 189
 190
 191
 192
 193
 194
 195
 196
 197
 198
 199
 200
 201
 202
 203
 204
 205
 206
 207
 208
 209
 210
 211
 212
 213
 214
 215
 216
 217
 218
 219
 220
 221
 222
 223
 224
 225
 226
 227
 228
 229
 230
 231
 232
 233
 234
 235
 236
 237
 238
 239
 240
 241
 242
 243
 244
 245
 246
 247
 248
 249
 250
 251
 252
 253
 254
 255
 256
 257
 258
 259
 260
 261
 262
 263
 264
 265
 266
 267
 268
 269
 270
 271
 272
 273
 274
 275
 276
 277
 278
 279
 280
 281
 282
 283
 284
 285
 286
 287
 288
 289
 290
 291
 292
 293
 294
 295
 296
 297
 298
 299
 300
 301
 302
 303
 304
 305
 306
 307
 308
 309
 310
 311
 312
 313
 314
 315
 316
 317
 318
 319
 320
 321
 322
 323
 324
 325
 326
 327
 328
 329
 330
 331
 332
 333
 334
 335
 336
 337
 338
 339
 340
 341
 342
 343
 344
 345
 346
 347
 348
 349
 350
 351
 352
 353
 354
 355
 356
 357
 358
 359
 360
 361
 362
 363
 364
 365
 366
 367
 368
 369
 370
 371
 372
 373
 374
 375
 376
 377
 378
 379
 380
 381
 382
 383
 384
 385
 386
 387
 388
 389
 390
 391
 392
 393
 394
 395
 396
 397
 398
 399
 400
 401
 402
 403
 404
 405
 406
 407
 408
 409
 410
 411
 412
 413
 414
 415
 416
 417
 418
 419
 420
 421
 422
 423
 424
 425
 426
 427
 428
 429
 430
 431
 432
 433
 434
 435
 436
 437
 438
 439
 440
 441
 442
 443
 444
 445
 446
 447
 448
 449
 450
 451
 452
 453
 454
 455
 456
 457
 458
 459
 460
 461
 462
 463
 464
 465
 466
 467
 468
 469
 470
 471
 472
 473
 474
 475
 476
 477
 478
 479
 480
 481
 482
 483
 484
 485
 486
 487
 488
 489
 490
 491
 492
 493
 494
 495
 496
 497
 498
 499
 500
 501
 502
 503
 504
 505
 506
 507
 508
 509
 510
 511
 512
 513
 514
 515
 516
 517
 518
 519
 520
 521
 522
 523
 524
 525
 526
 527
 528
 529
 530
 531
 532
 533
 534
 535
 536
 537
 538
 539
 540
 541
 542
 543
 544
 545
 546
 547
 548
 549
 550
 551
 552
 553
 554
 555
 556
 557
 558
 559
 560
 561
 562
 563
 564
 565
 566
 567
 568
 569
 570
 571
 572
 573
 574
 575
 576
 577
 578
 579
 580
 581
 582
 583
 584
 585
 586
 587
 588
 589
 590
 591
 592
 593
 594
 595
 596
 597
 598
 599
 600
 601
 602
 603
 604
 605
 606
 607
 608
 609
 610
 611
 612
 613
 614
 615
 616
 617
 618
 619
 620
 621
 622
 623
 624
 625
 626
 627
 628
 629
 630
 631
 632
 633
 634
 635
 636
 637
 638
 639
 640
 641
 642
 643
 644
 645
 646
 647
 648
 649
 650
 651
 652
 653
 654
 655
 656
 657
 658
 659
 660
 661
 662
 663
 664
 665
 666
 667
 668
 669
 670
 671
 672
 673
 674
 675
 676
 677
 678
 679
 680
 681
 682
 683
 684
 685
 686
 687
 688
 689
 690
 691
 692
 693
 694
 695
 696
 697
 698
 699
 700
 701
 702
 703
 704
 705
 706
 707
 708
 709
 710
 711
 712
 713
 714
 715
 716
 717
 718
 719
 720
 721
 722
 723
 724
 725
 726
 727
 728
 729
 730
 731
 732
 733
 734
 735
 736
 737
 738
 739
 740
 741
 742
 743
 744
 745
 746
 747
 748
 749
 750
 751
 752
 753
 754
 755
 756
 757
 758
 759
 760
 761
 762
 763
 764
 765
 766
 767
 768
 769
 770
 771
 772
 773
 774
 775
 776
 777
 778
 779
 780
 781
 782
 783
 784
 785
 786
 787
 788
 789
 790
 791
 792
 793
 794
 795
 796
 797
 798
 799
 800
 801
 802
 803
 804
 805
 806
 807
 808
 809
 810
 811
 812
 813
 814
 815
 816
 817
 818
 819
 820
 821
 822
 823
 824
 825
 826
 827
 828
 829
 830
 831
 832
 833
 834
 835
 836
 837
 838
 839
 840
 841
 842
 843
 844
 845
 846
 847
 848
 849
 850
 851
 852
 853
 854
 855
 856
 857
 858
 859
 860
 861
 862
 863
 864
 865
 866
 867
 868
 869
 870
 871
 872
 873
 874
 875
 876
 877
 878
 879
 880
 881
 882
 883
 884
 885
 886
 887
 888
 889
 890
 891
 892
 893
 894
 895
 896
 897
 898
 899
 900
 901
 902
 903
 904
 905
 906
 907
 908
 909
 910
 911
 912
 913
 914
 915
 916
 917
 918
 919
 920
 921
 922
 923
 924
 925
 926
 927
 928
 929
 930
 931
 932
 933
 934
 935
 936
 937
 938
 939
 940
 941
 942
 943
 944
 945
 946
 947
 948
 949
 950
 951
 952
 953
 954
 955
 956
 957
 958
 959
 960
 961
 962
 963
 964
 965
 966
 967
 968
 969
 970
 971
 972
 973
 974
 975
 976
 977
 978
 979
 980
 981
 982
 983
 984
 985
 986
 987
 988
 989
 990
 991
 992
 993
 994
 995
 996
 997
 998
 999
 1000

162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215

REFERENCES

Jongin Lim, Suchoel Lee, Daeho Um, Sung-Un Park, and Jinwoo Shin. PRIME: Deep imbalanced regression with proxies. In *Forty-second International Conference on Machine Learning, 2025*. URL <https://openreview.net/forum?id=hvaPL2DMpQ>.

Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *International conference on machine learning*, pp. 11842–11851. PMLR, 2021.