Improve Representation for Imbalanced Regression through Geometric Constraints

Supplementary Material

6. Proof of Theorem 1.

Proof. Define $y \in [0,1]$. A reparametrization of the path $l(y)$ is defined by a bijective strictly increasing function $r(y):[0,1]\to[0,1]$, denoted as $\tilde{l}(y):=(l\circ r)(y)$. Due to the fact that $\mathrm{Im}(l)=\mathrm{Im}(\tilde{l})$,

$$T(l,\epsilon)=T(\tilde{l},\epsilon')\Rightarrow\mathcal{L}_{\text{env}}(l,\epsilon)=\mathcal{L}_{\text{env}}(\tilde{l},\epsilon') \tag{14}$$

Denote r' as the derivative of r. Further we have

$$\begin{aligned}\mathcal{L}_{\text{homo}}(\tilde{l})&=\int_0^1|\nabla_y l(r(y))|^2dy\\&=\int_0^1|\nabla_r l(r)|^2|_{r=r(y)}\cdot|r'(y)|^2dy\\&=\int_0^1|\nabla_r l(r)|^2|_{r=r(y)}\cdot r'(y)^2dy\\&=\int_0^1|\nabla_r l(r)|^2\cdot r'(y)dr\\&=\int_0^1|\nabla_r l(r)|^2\cdot s(r)dr,\end{aligned} \tag{15}$$

where $s=r'\circ r^{-1}$. This separates the dependence of $\mathcal{L}_{\text{homo}}$ on the reparametrization to a single weight function $s:[0,1]\to\mathbb{R}_+$.

Then we have

$$\mathcal{L}_{\text{homo}}(\tilde{l})-\mathcal{L}_{\text{homo}}(l)=\int_0^1|\nabla_y l(y))|^2(s(y)-1)dy. \tag{16}$$

Now if the original curve is moving at constant speed, *i.e.*, $|\nabla_y l(y)|=c$, where c is a positive constant. In other words, the data is uniformly distributed. Then

$$\begin{aligned}\mathcal{L}_{\text{homo}}(\tilde{l})-\mathcal{L}_{\text{homo}}(l)&=c^2\int_0^1(s(y)-1)dy\\&=c^2\left(\int_0^1 s(y)dy-1\right),\end{aligned}$$

which means in this case the loss will increase if $\int_0^1 s(y)dy>1$ and decrease otherwise. Since r is a bijection, we have

$$\begin{aligned}\int_0^1 s(r)dr&=\int_0^1 s(r(y))r'(y)dy\\&=\int_0^1 r'(y)^2dy\end{aligned}$$

Since $(r'(t)-r'(y))^2\geq 0,\ t,y\in[0,1]$, we have

$$\begin{aligned}0&\leq\int_0^1\int_0^1(r'(y)-r'(t))^2dtdy\\&=2\int_0^1\int_0^1 r'(y)^2dydt-2\left(\int_0^1 r'(y)dy\right)^2\\&=2\int_0^1 r'(y)^2dy-2\\&\Rightarrow\int_0^1 r'(y)^2dy\geq 1,\end{aligned}$$

where the inequality holds when $r'(y)$ is a constant, since r is bijective, r should be the function: $r(y)=y$. This means $l(y)=\tilde{l}(y),\forall y$. Therefore, we have $\int_0^1 r'(y)^2dy>1$, for $\tilde{l}\neq l$, which means, the loss attains its minimum if and only if the data is uniformly distributed. □

7. Datasets

7.1. UCI-DIR

We curated UCI-DIR to evaluate the performance of imbalanced regression methods on tabular datasets. Here, we consider four regression tasks from UCI machine learning repository [2] (Airfoil, Concrete, Real Estate and Abaleone). Their input dimensions range from 5 to 8. Following the original DIR setting [25], we curated a balanced test set with balanced distribution across the label range and leave the training set naturally imbalanced (Figure 8). We partitioned the label range into three regions based on the occurrence. The threshold for [few-shot/med-shot, med-shot/many-shot] are [10, 40], [5, 15], [3, 10] and [100, 400] for Airfoil, Concrete, Real Estate and Abalone respectively.

7.2. OL-DIR

We follow Lu et al. [14] for the basic setting of operator learning. However, we change the original uniform sampling of locations in the domain of the output function to three regions: few, medium, and many regions.

For the linear operator defined in Equation (12), the input function u is generated from a Gaussian Random Field (GRF):

$$u\sim\mathcal{G}(0,k(x_1,x_2)) \tag{17}$$

$$k(x_1,x_2)=\exp\left(-\frac{\|x_1-x_2\|^2}{2l^2}\right) \tag{18}$$

where the length-scale parameter l is set to be 0.2. For x, we fix 100 locations to represent the input function u. The locations

Table 8. Overview of the six curated datasets used in our experiments

Dataset	Target type	Target range	Bin size	# Training set	# Val. set	# Test set
IMDB-WIKI	Age	0∼186*	1	191,509	11,022	11,022
AgeDB-DIR	Age	0∼101	1	12,208	2,140	2,140
STS-B-DIR	Text similarity score	0∼5	0.1	5,249	1,000	1,000

**Note: wrong labels in the original dataset.*

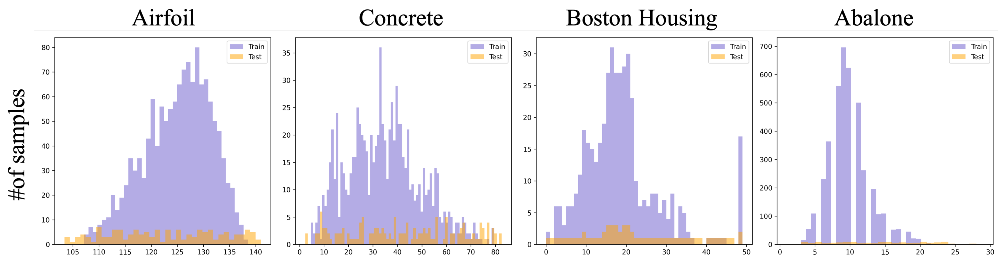

Figure 8. Overview of training and test set label distribution for UCI-DIR datasets.

in the output function ys are manually sampled from the domain of $G(u)$, such that few-shot region: $y \in [0.0, 0.2] \cup [0.8, 1.0]$; medium-shot region: $y \in [0.2, 0.4] \cup [0.6, 0.8]$; many-shot region: $y \in [0.4, 0.6]$.

We manually create an imbalanced training set with many/medium/few-shot regions of 10k samples and a balanced testing test of 100k samples.

For the nonlinear operator defined in Equation (13), the input function is defined as:

$$b(x;\omega) \sim \mathcal{GP}(b_0(x), \text{cov}(x_1, x_2)) \tag{19}$$

$$b_0(x) = 0 \tag{20}$$

$$\text{cov}(x_1, x_2) = \sigma^2 \exp\left(-\frac{\|x_1 - x_2\|^2}{2l^2}\right) \tag{21}$$

where ω is sampled from a random space with Dirichlet boundary conditions $u(0) = u(1) = 0$, $f(x) = 10$. $\mathcal{GP}$ is a Gaussian random process. The target locations are sampled in the same way as the linear task.

The number and split of the nonlinear operator dataset are the same as those of the linear one.

7.3. AgeDB-DIR, IMDB-DIR and STS-B-DIR

For the real-world datasets (AgeDB-DIR, IMDB-WIKI-DIR and STS-B-DIR), We follow the original train/val./test split from [25].

7.4. Ethic Statements

All datasets used in our experiments are publicly available and do not contain private information. All datasets (AgeDB, IMDB-WIKI, STS-B, and UCI) are accrued without any engagement or interference involving human participants and are devoid of any confidential information.

8. Experiment Detail

8.1. Implementation Detail (Table 9).

8.2. Choices of N.

We investigate how varying N (the number of uniformly distributed points on a hypersphere used to calculate enveloping loss) impacts the performance of our approach on the AgeDB-DIR and IMDB-WIKI-DIR datasets (Table 10). To achieve optimal performance, it is crucial to choose a sufficiently large N. A smaller N might fail to cover the entire hypersphere adequately, resulting in an imprecise calculation of enveloping loss.

8.3. Ablation on proposed components.

The Table 11 presents the results of an ablation study examining the impact of different loss functions on the model performance. As we mentioned before, the use of only homogeneity loss ($\mathcal{L}_{\text{homo}}$) could lead to trivial solutions due to feature collapse. Additionally, using only the enveloping loss ($\mathcal{L}_{\text{env}}$) causes the features to spread out along the trajectory, resulting in suboptimal performance. Through the contrastive loss ($\mathcal{L}_{\text{con}}$), individual representations could converge towards their corresponding locations on the surrogate. It is evident from the Table 11 that the model incorporating all loss functions outperforms the other configuration.

8.4. Computational cost

In this subsection, we compare the time consumption of the Surrogate-driven Representation Learning (SRL) framework with other baseline methods for age estimation and text similarity regression tasks. The reported time consumption, expressed in seconds, represents the average training time per mini-batch update. All experiments were conducted using a GTX 3090 GPU.

Table 9. Hyper-parameters used in SRL

Dataset	IMDB-WIKI	AgeDB-DIR	STS-B-DIR	UCI-DIR	OL-DIR
Temperature (τ)	0.1	0.1	0.1	0.1	0.1
Momentum (α)	0.9	0.9	0.9	0.9	9.9
N	2000	2000	1000	1000	1000
λ_e	1e-1	1e-1	1e-2	1e-2	1e-1
λ_h	1e-1	1e-1	1e-4	1e-2	1e-1
Backbone Network ($f(\cdot)$)	ResNet-50	ResNet-50	BiLSTM	3layer MLP	3layer MLP
Feature Dim	128	128	128	128	128
Learning Rate	2.5e-4	2.5e-4	2.5e-4	1e-3	1e-3
Batch Size	256	64	16	256	1000

Table 10. Vary the number of N

N	100	200	500	1000	**2000**	4000	10000
AgeDB	7.78	7.55	7.37	7.31	7.22	7.22	7.22
IMDB-WIKI	7.85	7.78	7.72	7.69	7.69	7.69	7.72

Table 11. Ablation Studies, best results are bold

$\mathcal{L}_{\text{env}}$	$\mathcal{L}_{\text{homo}}$	$\mathcal{L}_{\text{con}}$	MAE ↓				GM ↓			
			All	Many	Med	Few	All	Many	Med	Few
			7.67	6.66	9.30	12.61	4.85	4.17	6.51	8.98
	✓		7.87	7.01	8.99	12.90	5.12	4.56	6.11	9.39
✓			7.52	**6.63**	8.69	12.63	4.85	4.27	5.90	9.48
✓	✓		7.50	6.73	8.53	11.92	4.81	4.37	5.49	8.29
		✓	7.55	6.73	8.47	12.71	4.79	4.24	5.68	9.42
✓	✓	✓	**7.22**	7.38	**6.64**	**8.28**	**4.50**	**4.12**	**5.37**	**6.29**

Table 12 shows that SRL achieves a considerably lower training time compared to the LDS + FDS, while remaining competitive with RankSim, Balanced MSE, and Ordinal Entropy. This demonstrates SRL's ability to handle complex tasks efficiently without introducing substantial computational overhead.

Table 12. Average training time per mini-batch update (in seconds) for age estimation (AgeDB-DIR) and text similarity regression (STS-B-DIR) tasks, using a GTX 3090 GPU.

Method	AgeDB-DIR (s)	STS-B-DIR (s)
VANILLA	12.24	25.13
LDS + FDS	38.42	44.45
RankSim	16.86	30.04
Balanced MSE	16.21	28.12
Ordinal Entropy	17.29	29.37
SRL (Ours)	17.10	27.35

8.5. Impact of Bin Numbers

In our geometric framework, we employ piecewise linear interpolation to approximate the continuous path l. The granularity of this approximation is determined by the number of bins used for discretization, where finer binning naturally leads to smoother interpolation. To empirically analyze the impact of bin numbers (B) on model performance, we conducted extensive experiments across both synthetic and real-world datasets. For the synthetic OL-DIR dataset and the real-world AgeDB-DIR dataset, we varied the number of bins across the label space. Note that for AgeDB-DIR, the finest possible bin size is constrained to 1 due to the discrete nature of age labels, while OL-DIR allows for arbitrary bin sizes. The results are presented in Table 13.

Table 13. Impact of bin numbers on model performance

B	10	20	50	100	1000	2000	4000
OL-DIR (MAE $\times 10^{-3}$)	9.92	9.29	9.20	9.18	9.18	9.17	9.18
AgeDB-DIR (MAE)	7.44	7.38	7.31	7.22	-	-	-

8.6. Experiments on UCI-DIR (Table 14, 15, 16, 17)

Table 14. Complete results on UCI-DIR for Airfoil (MAE with standard deviation), the best results are **bold**.

Metrics	MAE			
Shot	All	Many	Med	Few
VANILLA	5.657(0.324)	5.112(0.207)	5.031(0.445)	6.754(0.423)
LDS + FDS	5.761(0.331)	**4.445**(0.208)	4.792(0.412)	7.792(0.499)
RankSim	5.228(0.335)	5.049(0.92)	4.908(0.786)	5.718(0.712)
BalancedMSE	5.694(0.342)	4.512(0.179)	5.035(0.554)	7.277(0.899)
Ordinal Entropy	6.270(0.415)	4.847(0.223)	5.369(0.635)	8.315(0.795)
SRL (ours)	**5.100**(0.286)	4.832(0.098)	**4.745**(0.336)	**5.693**(0.542)

Table 15. Complete results on UCI-DIR for Abalone (MAE with standard deviation), the best results are **bold**.

Metrics	MAE			
Shot	All	Many	Med	Few
VANILLA	4.567(0.211)	**0.878**(0.152)	2.646(0.349)	7.967(0.344)
LDS + FDS	5.087(0.456)	0.904(0.245)	3.261(0.435)	9.261(0.807)
RankSim	4.332(0.403)	0.975(0.067)	2.591(0.516)	7.421(0.966)
BalancedMSE	5.366(0.542)	2.135(0.335)	2.659(0.456)	9.368(0.896)
Ordinal Entropy	6.774(0.657)	2.314(0.256)	4.013(0.654)	11.610(1.275)
SRL (ours)	**4.158**(0.196)	0.892(0.042)	**2.423**(0.199)	**7.191**(0.301)

Table 16. Complete results on UCI-DIR for Real Estate (MAE with standard deviation), the best results are **bold**.

Datasets	MAE			
Shot	All	Many	Med	Few
VANILLA	0.326(0.003)	0.273(0.005)	0.376(0.003)	0.365(0.012)
LDS + FDS	0.346(0.004)	0.325(0.002)	0.400(0.002)	0.335(0.023)
RankSim	0.373(0.008)	0.343(0.004)	0.381(0.008)	0.397(0.032)
BalancedMSE	0.337(0.007)	0.313(0.004)	0.398(0.009)	0.326(0.028)
Ordinal Entropy	0.339(0.007)	0.286(0.004)	0.421(0.005)	0.351(0.031)
SRL (ours)	**0.278**(0.002)	**0.262**(0.006)	**0.296**(0.005)	**0.287**(0.023)

Table 17. Complete results on UCI-DIR for Concrete (MAE with standard deviation), the best results are **bold**.

Datasets	MAE			
Shot	All	Many	Med	Few
VANILLA	7.287(0.364)	5.774(0.289)	6.918(0.346)	9.739(0.487)
LDS + FDS	6.879(0.344)	6.210(0.310)	6.730(0.337)	7.594(0.380)
RankSim	6.714(0.336)	5.996(0.300)	5.574(0.279)	9.456(0.473)
BalancedMSE	7.033(0.352)	**4.670**(0.234)	6.368(0.318)	9.722(0.486)
Ordinal Entropy	7.115(0.356)	5.502(0.275)	6.358(0.318)	9.313(0.466)
SRL (ours)	**5.939**(0.297)	5.318(0.266)	**5.800**(0.290)	**6.603**(0.330)

8.7. Experiments on AgeDB-DIR

Training Details: In Table 18, our primary results on AgeDB-DIR encompasses the replication of all baseline models on an identical server configuration (RTX 3090), adhering to the original codebases and training recipes. We observe a performance drop in RankSim [6] and ConR [10] in comparison to the results reported in their respective studies. To ensure a fair comparison, we present the **mean and standard deviation (in parentheses)** of the performances for SRL (ours), RankSim, and ConR, based on three independent runs. We found SRL superiors performance in most categories and all Med-shot and Few-shot metrics.

We would like to note that we found self-conflict performance in the original ConR [10] paper, where they report overall MAE of 7.20 in main result (Table 1) and 7.48 in the ablation studies (Table 6). **The results in Table 6 are closed to our reported result**.

8.8. Experiment on IMDB-WIKI-DIR

Training Details: In Table 19, our primary results on IMDB-WIKI-DIR encompass the replication of all baseline models on an identical server configuration (RTX 3090), adhering to the original codebases and training receipes. We observe a performance drop of ConR [10] in comparison to the results reported in their respective studies. To ensure a fair comparison, we present the **mean and standard deviation (in parentheses)** of the performances for SRL (ours) and ConR, based on three independent runs. We found SRL superiors performance in most categories and all Med-shot and Few-shot metrics.

We would like to note that we found self-conflict performance in the original ConR [10] paper, where they report overall MAE of 7.33 in the main result (Table 2) and 7.84 in the ablation studies (Table 8), **The results in Table 8 are close to our reported result**.

8.9. Complete result on STS-B-DIR (Table 20)

8.10. Complete result on Operator Learning (Table 21)

9. Pseudo Code (Algorithm 1) for Surrogate-driven Representation Learning (SRL)

10. Broader impacts

We introduce novel geometric constraints to the representation learning of imbalanced regression, which we believe will significantly benefit regression tasks across various real-world applications. Currently, we are not aware of any potential negative societal impacts.

11. Limitation and Future Direction

In considering the limitations and future directions of our research, it's important to acknowledge that our current methodology has not delved into optimizing the feature distribution in scenarios involving regression with higher-dimensional labels. This presents a notable area for future exploration. Additionally, investigating methods to effectively handle complex label structures in imbalanced regression scenarios could significantly enhance the applicability and robustness of our proposed techniques.

Table 18. Complete Results on AgeDB-DIR

Metrics	Shot	VANILLA	LDS + FDS	RankSim	BalancedMSE	Ordinal Entropy	ConR	**SRL (ours)**
MAE↓	All	7.67	7.55	7.41(0.03)	7.98	7.60	7.41(0.02)	**7.22**(0.02)
	Many	6.66	7.03	**6.49**(0.01)	7.58	6.69	6.51(0.02)	6.64(0.01)
	Med	9.30	8.46	8.73(0.05)	8.65	8.87	8.81(0.03)	**8.28**(0.04)
	Few	12.61	10.52	12.47(0.09)	9.93	12.68	12.04(0.04)	**9.81**(0.05)
GM↓	All	4.85	4.86	4.71(0.03)	5.01	4.91	4.70(0.02)	**4.50**(0.02)
	Many	4.17	4.57	4.15(0.02)	4.83	4.28	4.13(0.02)	**4.12**(0.02)
	Med	6.51	5.38	5.74(0.04)	5.46	6.20	5.91(0.06)	**5.37**(0.02)
	Few	8.98	6.75	8.92(0.08)	6.30	9.29	8.59(0.0)	**6.29**(0.04)
MSE↓	All	100.01	97.05	94.37(0.10)	107.35	97.28	92.57(0.06)	**91.71**(0.02)
	Many	76.67	82.68	**72.00**(0.09)	95.49	74.79	72.06(0.04)	77.23(0.05)
	Med	130.21	114.00	121.38(2.15)	125.55	122.07	121.24(1.88)	**115.65**(1.42)
	Few	237.00	185.98	230.97(3.22)	169.00	241.13	207.00(3.09)	**162.22**(2.08)

Table 19. Complete Results on IMDB-WIKI-DIR

Metrics	Shot	VANILLA	LDS + FDS	RankSim	BalancedMSE	Ordinal Entropy	ConR	SRL (ours)
MAE↓	All	8.03	7.73	7.72	8.43	8.01	7.84(0.04)	**7.69**(0.02)
	Many	7.16	7.22	**6.92**	7.84	7.17	7.15(0.03)	7.08(0.01)
	Med	15.48	12.98	14.52	13.35	15.15	14.36(0.04)	**12.65**(0.04)
	Few	26.11	23.71	25.89	23.27	26.48	25.15(0.06)	**22.78**(0.06)
GM↓	All	4.54	4.40	4.29	4.93	4.47	4.43(0.04)	**4.28**(0.02)
	Many	4.14	4.17	3.92	4.68	4.07	4.05(0.03)	**4.03**(0.02)
	Med	10.84	7.87	9.72	7.90	10.56	9.91(0.05)	**7.28**(0.03)
	Few	18.64	15.77	18.02	15.51	21.11	18.55(0.06)	**15.25**(0.05)
MSE↓	All	136.04	130.56	130.95	146.19	137.50	132.41(1.22)	**129.97**(0.93)
	Many	105.72	106.93	**102.06**	121.64	107.62	105.29(0.88)	105.83(0.77)
	Med	373.07	315.92	351.22	343.12	369.88	338.30(1.99)	**311.17**(1.25)
	Few	978.00	861.15	977.82	787.71	976.56	934.12(3.03)	**859.81**(2.28)

Table 20. Complete Results on STS-B-DIR

Metrics	Shot	VANILLA	LDS + FDS	RankSim	BalancedMSE	Ordinal Entropy	SRL (ours)
MSE ↓	All	0.993	0.900	0.889	0.909	0.943	**0.877**
	Many	0.963	0.911	0.907	0.894	0.902	**0.886**
	Med	1.000	0.881	0.874	1.004	1.161	**0.873**
	Few	1.075	0.905	0.757	0.809	0.812	**0.745**
Pearson correlation ↑	All	0.742	0.757	0.763	0.757	0.750	**0.765**
	Many	0.685	0.698	0.708	0.703	0.702	**0.708**
	Med	0.693	0.723	0.692	0.685	0.679	**0.749**
	Few	0.793	0.806	0.842	0.831	0.767	**0.844**
MAE ↓	All	0.804	0.768	0.765	0.776	0.782	**0.750**
	Many	0.788	0.772	0.772	0.763	0.756	**0.748**
	Med	0.865	0.785	0.779	0.839	0.900	**0.773**
	Few	0.837	0.712	0.699	0.749	0.762	**0.694**
Spearman correlation ↑	All	0.740	0.760	0.767	0.762	0.755	**0.769**
	Many	0.650	0.670	0.685	0.677	0.669	**0.689**
	Med	0.495	0.488	0.495	0.487	0.448	**0.503**
	Few	0.843	0.819	0.862	0.867	0.845	**0.879**

Table 21. Complete results on OL-DIR with standard deviation added, best results are **bold**.

Operation	MAE(10^{-3}) ↓				MSE (10^{-4}) ↓			
Shot	All	Many	Med	Few	All	Many	Med	Few
Linear								
VANILLA	15.64(2.72)	11.86(2.20)	15.45(3.55)	27.00(5.62)	5.40(1.10)	2.81(0.75)	4.40(1.23)	14.20(2.25)
Ordinal Entropy	10.07(1.22)	9.26(0.98)	9.85(1.45)	13.01(1.92)	2.00(0.32)	1.53(0.19)	1.89(0.73)	3.42(0.82)
SRL (ours)	**9.18**(0.92)	**8.32**(0.66)	**9.47**(1.13)	**9.33**(1.89)	**1.98**(0.37)	**0.98**(0.21)	**1.72**(0.62)	**2.67**(0.99)
Nonlinear								
VANILLA	11.64(1.87)	9.89(1.25)	11.02(2.23)	19.77(2.89)	9.20(1.23)	**4.33**(0.88)	7.53(1.55)	24.70(1.99)
Ordinal Entropy	12.91(1.25)	9.93(0.93)	13.07(1.57)	21.02(1.89)	13.80(2.98)	8.82(2.25)	11.84(3.59)	30.12(5.40)
SRL (ours)	**11.25**(1.13)	**9.48**(0.75)	**9.22**(1.45)	**17.00**(1.54)	**8.60**(1.04)	7.42(0.70)	**6.41**(1.15)	**14.12**(1.39)

Algorithm 1 Pseudo Code for Surrogate-driven Representation Learning (SRL)

Require: Training set $D = \{(x_i, y_i)\}_{i=1}^N$, encoder f, regression function g, total training epochs E, momentum α, a set of uniformly distributed points U, surrogate S, batch size M.

for $e = 0$ to E **do**
  repeat
    Sample a mini-batch $\{(x_m, y_m)\}_{m=1}^M$ from D
    $\{z_m\}_{m=1}^M = f(\{x_m\}_{m=1}^M)$
    if $e = 0$ **then**
      Update the model with loss $\mathcal{L} = \mathcal{L}_{reg}(\{y_m\}_{m=1}^M, g(\{z_m\}_{m=1}^M))$
    else
      get C from $\{z_m\}_{m=1}^M$ using Equation (8)
      get $S^{e'}$ from C and S^e using Equation (9)
      Update the model with loss $\mathcal{L} = \mathcal{L}_{reg}(\{y_m\}_{m=1}^M, g(\{z_m\}_{m=1}^m)) + \mathcal{L}_G(S^{e'}, U) + \mathcal{L}_{con}(S^{e'}, \{z_m\}_{m=1}^M)$
    end if
  until iterate over all training samples at current epoch e
  // Update the surrogate
  get $\hat{S}^e$by calculate the class center for the current epoch
  if $e = 0$ **then**
    $S^1 = \hat{S}^e$
  else
    $S^{e+1} = \alpha S^e + (1-\alpha)\hat{S}^e$ # Momentum update the surrogate, Equation (9)
  end if
end for