Supplementary Material for DARE-GRAM : Unsupervised Domain Adaptation Regression by Aligning Inverse Gram Matrices

Ismail Nejjar
EPFL, Switzerland
ismail.nejjar@epfl.ch

Qin Wang
ETH Zurich, Switzerland
qwang@ethz.ch

Olga Fink
EPFL, Switzerland
olga.fink@epfl.ch

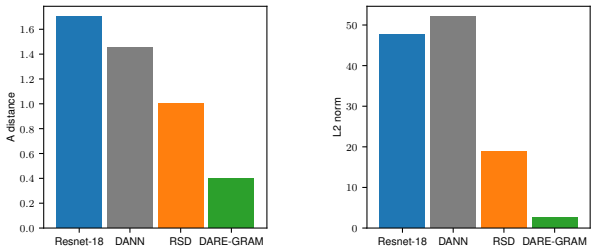
In this supplementary material for Unsupervised Domain Adaptation Regression by Aligning Inverse Gram Matrices, we provide additional discussions on the transferability, implementation, and limitation of our method.
In Section 1, we first present an additional analysis of the representation learned from our method. We then provide in Section 2, a short explanation on the algebraic details of the Moore-Penrose pseudo-inverse. An additional ablation study is provided to compare an alternative way of OLS motivated alignment, in Section 3. More dataset informration and implementation details are also provided in Section 5 and 6. In Section 7, we discuss the limitations of our method. Our code for the main paper is also attached.

1. Transferability of the representation

To better understand the learned representation of our model, we visualize the A-distance [1] of the representations trained by different methods in Figure A1a. The A-distance is a measure for distribution discrepancy and is used to evaluate the transferability of the representations [2]. As shown in Figure A1a, our method achieves the lowest A-distance, indicating that representations trained by DARE-GRAM can lead to better transferability.

In addition, we present the representation scale difference between the source and target in terms of the L2 norm eigenvalue differences. This scale difference is defined the same as Equation 11 in the main paper. As shown in Figure A1b, our method aligns the scales better than other methods.

We further conduct PCA analysis and visualize the most dominant principal components explaining the Z^Z matrix (squared eigenvalues of Z), and study their correlation with the target output y. Figure A2 shows that the two PCA principal components on the target test data are highly correlated with one dimension of the output and almost match the ground truth for DARE-GRAM on two of the three regression tasks in the dSprites dataset.



(a) A-distance comparison for different methods.

(b) Scale difference of the representation between source and target.

Figure A1. Transferability of the representation trained by different methods on the dSprite task **C** → **S**. Our method achieves the lowest A-distance. The scales are also well aligned by our method.

2. Moore-Penrose Pseudo-Inverse

In this section, we provide more details on the characteristics of the pseudo-inverse. The discussion in this section is well-studied and we use heavily the notations and conclusions from [10].

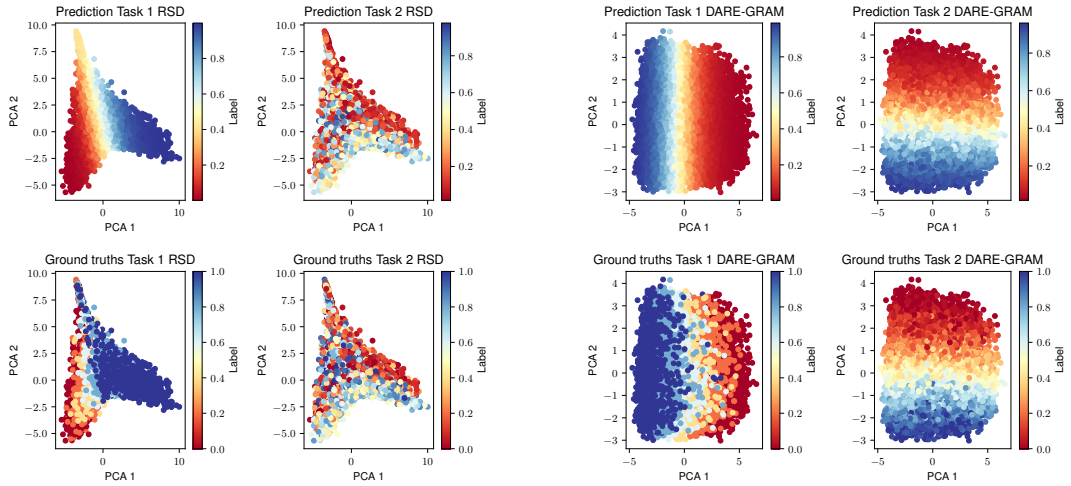
In the main paper, we proposed to use the Ordinary least squares (OLS) formulation in the context of Unsupervised Domain Adaptation. To find a regressor β for the relationship $Z\beta = Y$, where Z is the feature matrix, and Y is the ground truth. The closed-form solution for the ordinary least squares is given by

$$\beta = (Z^T Z)^{-1} Z^T Y. \tag{1}$$

In the main paper, we discussed how $Z^T Z$ could be non-invertible, and the pseudo-inverse can be of use in this case. Here we provide more details of the Moore-Penrose pseudo-inverse.

Definition : Given a matrix $A \in \mathcal{R}^{p,b}$. The Moore-Penrose pseudo inverse of A is denoted as A^+ and the following properties hold [10]:

(i) $AA^+A = A$



(a) RSD (b) DARE-GRAM (ours)

Figure A2. PCA Projection of embedding from different techniques on the dSprites data set **C** → **S**.

(ii) $A^+AA^+ = A^+$

(iii) AA^+ and A^+A are symmetrical matrices.

Proposition 1 Every real matrix has a unique Moore-Penrose pseudo-inverse matrix.

Proposition 2 Given a nonsingular matrix A, then $A^+ = A^{-1}$

Proposition 3 The pseudoinverse of the pseudoinverse is the original A, $(A^+)^+ = A$

Proposition 4 The pseudoinverse of the product of matrices if the produce of the pseudoinverse of each individual matrix $(AB)^+ = B^+A^+$

Theorem 1 The minimum of $||Az-y||^2$, can be obtained such as $z = A^+y$, and is unique.

$$A^+ = (A^TA)^+A^T \tag{2}$$

Proof of the above theorem can be found in [10].

Algorithms

While the Moore-Penrose pseudo-inverse of an A matrix can be computed using the SVD decomposition, as shown in the main paper. An exact calculation can be derived using the QR decomposition [4] of matrix A. The Gram–Schmidt method and Householder transformation are two methods to obtain the QR decomposition of A. Given an orthogonal matrix Q calculated from the Householder transformation and an upper triangular matrix R defined as:

$$R = \left(\begin{array}{c|c} R_1 & 0 \\ \hline 0 & 0 \end{array}\right) \tag{3}$$

with R_1 an upper triangular invertible matrix of shape $k \times k$ (where k was introduced in the main paper). Then the pseudo-inverse can be found :

$$A^+ = \left(\begin{array}{c|c} R_1^{-1} & 0 \\ \hline 0 & 0 \end{array}\right) Q^T \tag{4}$$

The existence and uniqueness of the Moore-Penrose pseudo-inverse provide an ideal candidate solution for the subspace alignment presented in this paper.

Based on the characteristics described in this section, in the next section, we show that there are alternative ways to align the domains other than the proposed one in the main paper. We then ablate them in the next section by empirical experiments.

3. Ablation: What if we use Z^+ instead of the Inverse Gram Matrix for the alignment

DARE-GRAM In the main paper, we proposed to align the angle of $(Z^TZ)^{-1}$ and the scale of Z. Intuitively, regression analysis aims at estimating the relationships between the output variable to the dimensional features. Given a calibrated linear regressor on the source domain, and without hurting the model generalizability the target features should have similar feature interaction and intensity to the source features. The Gram matrix can retrieve this information since it describes the activation for each variable of the embedding dimensions and summarises the pairwise interactions between the different features. While being a square matrix of the same dimension as the embedding dimension.

Method	C → N	C → S	N → C	N → S	S → C	S → N
Z allignement (RSD)	0.37	0.36	0.17	0.61	0.07	0.10
Z^+ allignment (alternative)	0.98	0.89	did not converge	0.65	0.11	0.30
DARE-GRAM (ours)	**0.30**	**0.20**	**0.11**	**0.25**	**0.05**	**0.07**

Table A1. Comparison results between aligning Z , aligning Z^+, or our proposed inverse Gram matrix $(Z^TZ)^{-1}$ on the dSprites regression tasks. All results are shown in sum of MAE with the ResNet-18.

Therefore, given the OLS closed-form solution [5]:

$$\hat{\beta} = (Z^TZ)^{-1}Z^TY, \tag{5}$$

we proposed to align the angle of $(Z^TZ)^{-1}$ and the scale of Z.

Alternative View An alternative choice would be using both terms $(Z^TZ)^{-1}Z^T$ together for the alignment. As shown in the previous section from the Equation 2, this term corresponds to the the pseudo-inverse of Z.

We ablate this choice here. Empirical results in Table A1 show that the use of the pseudo-inverse Z^+ as a basis for alignment does not lead to good performance. This formulation projects the Gram matrix onto samples of Z of size $b \times p$. Therefore, permutations of the matrix samples would lead to a different matrix, which is not the desired behavior for aligning the source and target domains.

4. Is Z aligned when $(Z^TZ)^{-1}$ is aligned

We now present the cosine similarity of the subspaces of $\{Z_s, Z_t\}$, as well as $\{(Z_s^TZ_s)^{-1}, (Z_t^TZ_t)^{-1}\}$ on dSprites dataset. We present the empirical results after deploying the two different strategies (align Z or $(Z^TZ)^{-1}$).

As shown in Table A2, when aligning Z, the inverse Gram is generally not well aligned. This is consistent with our synthetic example in the main paper. An interesting observation is that the mismatch is significant on the task of $N \rightarrow S$, which is also the task where our method has the largest improvement.

In contrast, *using our method to align the inverse Gram* $(Z^TZ)^{-1}$ *can lead to a well-aligned* Z for dSprites tasks, as also suggested by Figure A2.

Align. Strategy	Represent.	C → N	C → S	N → C	N → S	S → C	S → N	Avg
Z	Z	0.89	0.97	0.85	0.91	1.00	0.95	0.93
	$(Z^TZ)^{-1}$	0.53	0.44	0.38	0.21	1.00	0.90	0.58
$(Z^TZ)^{-1}$	Z	0.96	0.98	0.92	0.96	0.99	0.99	0.97
	$(Z^TZ)^{-1}$	0.58	0.97	0.90	0.98	0.99	1.00	0.90

Table A2. Cosine similarity on the alignment of Z or $(Z^TZ)^{-1}$.

5. Additional Dataset Information

We evaluated our proposed method in the main paper on three domain adaptations for regression benchmark datasets used in previous works [2, 7]: dSprites [11], MPI3D [6] and Biwi Kinect [3]. We now provide the variation factors for each of the datasets used in this paper, in Table A3 to Table A5.

Factor	Values	Task
Scale	6 values in [0.5, 1]	Regression
Orientation	40 values in [0, 2π]	Regression
Position X	32 values in [0, 1]	Regression
Position Y	32 values in [0, 1]	Regression
Shape Y	3 values	Classification

Table A3. Variations factor values on dSprites dataset

Factor	Values	Task
Horizontal Axis	40 values in $[0, 39]$	Regression
Vertical Axis	40 values in $[0, 39]$	Regression
Object color	6 values	Classification
Object shape	6 values	Classification
Object size	2 values	Classification
Camera height	3 values	Classification
Background Color	3 values	Classification

Table A4. Variations Factors values in MPI3D dataset

Factor	Values	Task
Yaw	Values ∈ [-92.044,231.352]	Regression
Pitch	Values ∈ [-87.7066,246.684]	Regression
Roll angle	Values ∈ [754.182,1297.45]	Regression

Table A5. Variations Factors values in Biwi Kinect dataset

6. Implementation Details

In an ideal scenario, and without any resource limitations, one would have to train the model with a batch size at least twice the integration dimension to obtain a fully ranked Gram matrix and directly use its inverse for the alignment processes. We have proposed a method that allows us to align the two distributions with reasonable batch sizes, which requires us to introduce an additional hyperparameter T, in addition to the two scaling factors for scaling and angle losses. Following [2], we use [13] for model selection to determine the hyper-parameters.

7. Limitation

One limitation of our work is the strong focus on regression tasks. While the inverse Gram Matrix brings improvement for unsupervised domain adaptation for regression and brings its motivation from the closed form of linear regression problems, extending this work to classification can be interesting for future work. In addition, many benchmark datasets used in this work are synthetic. MPI3D [6]and Biwi Kinect [3] contains limited real-world data. How the method behaves on even more complicated tasks in the real world remains a future work, but can be potentially solved by considering the setups in [7–9, 12]. In addition, real-world scenarios may face the problem of unaligned or missing output values. Partial set and open set UDA for regression settings were not addressed in this paper.

References

[1] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010. 1

[2] Xinyang Chen, Sinan Wang, Jianmin Wang, and Mingsheng Long. Representation subspace distance for domain adaptation regression. In *International Conference on Machine Learning*, pages 1749–1759. PMLR, 2021. 1, 3

[3] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3d face analysis. *Int. J. Comput. Vision*, 101(3):437–458, February 2013. 3, 4

[4] Walter Gander. Algorithms for the qr decomposition. *Res. Rep*, 80(02):1251–1268, 1980. 2

[5] Arthur Stanley Goldberger et al. Econometric theory. *Econometric theory.*, 1964. 3

[6] Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 3, 4

[7] Junguang Jiang, Yifei Ji, Ximei Wang, Yufeng Liu, Jianmin Wang, and Mingsheng Long. Regressive domain adaptation for unsupervised keypoint detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6780–6789, 2021. 3, 4

[8] Donghyun Kim, Kaihong Wang, Kate Saenko, Margrit Betke, and Stan Sclaroff. A unified framework for domain adaptive pose estimation. *Proceedings of European Conference on Computer Vision*, 2022. 4

[9] Chen Li and Gim Hee Lee. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1482–1491, 2021. 4

[10] Ross MacAusland. The moore-penrose inverse and least squares. *Math 420: Advanced Topics in Linear Algebra*, pages 1–10, 2014. 1, 2

[11] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017. 3

[12] Takehiko Ohkawa, Yu-Jhe Li, Qichen Fu, Ryosuke Furuta, Kris M Kitani, and Yoichi Sato. Domain adaptive hand keypoint and pixel localization in the wild. In *European Conference on Computer Vision*, pages 68–87. Springer, 2022. 4

[13] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007. 3