# Hyper-Accelerated Learning for Brain-Computer Interfaces via Partial Target-Aware Optimal Transport[*]

Ekansh Gupta
egupta8@gatech.edu
Georgia Institute of Technology
Atlanta, GA, USA

Cheng-Yeh Chen
cchen847@gatech.edu
Georgia Institute of Technology
Atlanta, GA, USA

Raghupathy Sivakumar
siva@gatech.edu
Georgia Institute of Technology
Atlanta, GA, USA

## ABSTRACT

Brain-computer interfaces (BCIs) have surfaced as a powerful modality in human-machine interaction and wearable technology with powered futuristic applications like virtual reality, robot control, gaming, etc. Using BCIs, the brain's intent can be harnessed without explicit communication. Despite the vast promise, systems designed for BCIs generalize poorly to new or unseen individuals due to high variability in brain signals among different subjects, resulting in long retraining/calibration sessions. This lack of generalization is typically attributed to a covariate shift of signals in the probability space, which manifests itself as disparate marginal and class conditional distributions. In this paper, we overview the factors contributing to poor generalization on a more granular level by analyzing a specific brain signal called the Error Potential (ErrP), a signal well-known for its noisy characteristics and high variability, and propose a novel algorithm to mitigate the associated covariate shift using partial target-aware optimal transport. We demonstrate our method on an ErrP dataset collected in our lab. Our method outperforms state-of-the-art models for cross-user generalization which translates to a reduction in calibration time by an order of magnitude.

## CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; • **Computing methodologies → Machine learning**; **Semi-supervised learning settings**.

## KEYWORDS

Brain-Computer Interfaces, ErrP, generalization, transfer learning, optimal transport

**ACM Reference Format:**
Ekansh Gupta, Cheng-Yeh Chen, and Raghupathy Sivakumar. 2023. Hyper-Accelerated Learning for Brain-Computer Interfaces via Partial Target-Aware Optimal Transport. In *The 2nd Workshop on Smart Wearable Systems and Applications (SmartWear '23), October 2–6, 2023, Madrid, Spain.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3570361.3614208

## 1 INTRODUCTION

Brain-Computer Interfaces, or BCIs, are a direct communication pathway between the activity inside the brain and an external device or an AI system. By tapping directly into the brain, BCIs bypass the physical limitations of the body, like pressing buttons on a keyboard or giving commands using speech. Current research and availability of user-grade BCI headsets have unlocked significant possibilities for commercial BCI usage in entertainment [1], wellness [2], security [3], and other interactive wearable applications. BCIs work by sensing the electrical/magnetic activity in the brain, decoding the user's intent by signal processing and machine learning algorithms, and taking action in accordance with the interpreted activity.

Most BCI applications use non-invasive techniques, like EEG, MEG, NIRS, etc., to record brain signals. While each of these techniques has its own utility, EEG remains by far the most popular option, partially because it lies in a unique sweet spot of cost-effectiveness, portability, and user-friendliness [4]. Despite the huge promise of BCIs and EEG in general as a crucial enabling technology for smart wearables, it faces some characteristic disadvantages. Systems designed for EEG suffer from low signal-to-noise ratio and high variance of user-specific brain signals, making its detection challenging. EEG data, while being very noisy, also exhibits a lot of variation and difference among different subjects, tasks, environments, and even different sessions for the same subject. A useful analogy for brain signals is that they are like fingerprints, in the sense that they are universal to all humans but they also have abundant individual differences. This hampers the generalization capabilities of algorithms

[*]This work was funded in part by the Wayne J Holman Endowed Chair

*SmartWear '23, October 2–6, 2023, Madrid, Spain*

ACM ISBN 978-1-4503-9990-6/23/10…$15.00
https://doi.org/10.1145/3570361.3614208

that process these signals on new or unseen data. A common solution is calibrating a BCI device for every new user. As a real-world example, a mood detection wearable may work for one user's brain signals but require hours of calibration on a new user to adapt to the latter's brain signals.

Typically, covariate shift [5] is the contributing factor for poor generalization. It is described as a shift in the probability distribution of a target/test dataset relative to the source/training dataset. Numerous solutions have been proposed to mitigate the effects of covariate shift which range from spatial filtering and domain adaptation to deep learning methods. While these approaches attempt to estimate cleaner versions of these signals and marginally improve generalization, they do not fundamentally align the two distributions in the probability space. In this paper, we look at different factors that contribute to poor generalization and propose an algorithm that not only aligns a high dimensional source and target probability distribution but also matches the positive and negative class labels with the target dataset, achieving seamless generalization that approaches the performance of a regular classification model, thereby significantly accelerating model adaptation and reducing calibration time by an order of magnitude. We demonstrate our algorithm on a real-world dataset of the error potential signal (ErrP), a brain signal that is well known for its poor generalization accuracy [6]. Our research contributions are as follows:

1. We closely examine the specific factors behind poor generalization and estimate an upper bound for the generalization accuracy for different scenarios.
2. We propose a novel algorithm that uses partially estimated class centroids to adapt to a target domain and demonstrate our results on an ErrP dataset. The proposed algorithm comes within 95.6% of the accuracy of a label-assisted classifier while only using 5% of the labeled samples, thereby accelerating the process of adapting to a new user by an order of magnitude.

The rest of the paper is organized as follows. Section 2 provides a background on the ErrP signal, its poor generalization, and the related work done to address it. Section 3 provides an overview of poor generalization as a function of covariate shift and class separation. Section 4 talks about our proposed approach to solve this problem and provides promising results on a target dataset, and then finally, section 5 talks about future work and concludes the paper.

## 2 BACKGROUND, PROBLEM DEFINITION, AND DATASET

As previously mentioned, poor generalization adversely impacts the usability of a wearable, which makes its mitigation crucial. We address the generalization problem by looking at the probability space of a signal and formulating our problem as a method to reconcile disparate distributions. Mathematically, we aim to reconcile $P(L_i|X_S)$ and $P(L_i|X_T)$, where $X_S$ and $X_T$ represent source and target distributions. $P(L_i|X_S)$ is the class conditional probability for our source dataset (the dataset on which we train our classifier), and $P(L_i|X_T)$ is the class conditional probability for the target dataset (where we evaluate our model). $L_i$ stands for our $i^{th}$ class label.

In this paper, we evaluate our generalization algorithms on the ErrP signal. ErrP is a measure of the brain detecting/processing an error (for instance, seeing a robot perform a task incorrectly), which is extremely valuable for BCI applications as it provides a generalized notion of error detection in a diverse set of tasks. ErrP has been used in applications for improving the performance and reliability of BCI spellers [7], correcting and adapting systems to accelerate reinforcement learning for AI agents [8], etc. However, their generalization accuracy is inadequate owing to their high variability among individuals. Traditionally, spatial filtering techniques [9] have been used to mitigate this problem. The state-of-the-art method for cross-user ErrP detection uses XDAWN spatial filtering [10] and Riemannian Geometry [11]. This method works equally well for other ERPs (Event-related potentials) such as P300 [12]. [13] improved upon this method by introducing affine transforms which make the data reference across different users identical which improves the accuracy of cross-user generalization. More recently, deep learning models have emerged which outperform traditional spatial filtering approaches for specific tasks. There have been RNN-based zero-shot learning methods [14] for classifying object classes using EEG. DeepConvNet and ShallowConvNet proposed by [15] are deep learning models for EEG decoding and visualization. With EEGNet [16], the authors created a shallow deep learning model with 1082 or 2290 parameters (depending on the configuration) and showed promising numbers for ErrP detection. [17] used ErrP detection from multiple observers and few-shot learning to improve the generalization accuracy of ErrP signals.

While these methods provide incremental performance improvements, they do not address the fundamental reason for poor generalization in ErrP signals, which is the incongruous probability distributions in the train and test domains. To address this challenge, several works have used domain adaptation techniques like optimal transport [18] to modify the probability distribution of the signal data in the feature space. Regularized optimal transport [19] is an effective technique for aligning disparate probability distributions by keeping similar labels close to each other. [20] used regularized optimal transport with class labels to improve the transfer learning of P300 signals. Similarly, [21] used regularized optimal transport in the domain of semi-positive definite (SPD) matrices and used the Riemannian distance

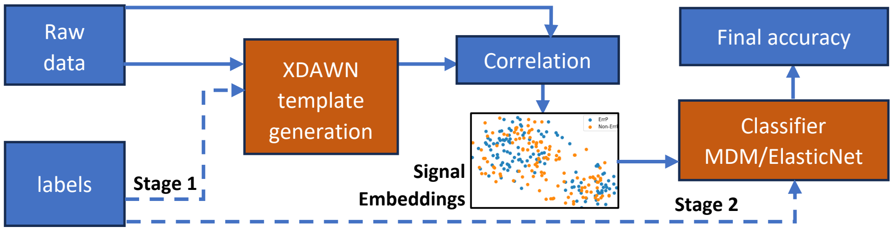


**Figure 1: Schematic of the xRG model with supervised blocks highlighted (orange).**

metric to align ERP (Event-related potential) signal distributions. However, these methods often do not provide sufficient cross-user accuracy for ErrP signals (more details in section 4) due to a lack of target class distribution information. In this work, we address this problem by estimating the target class centroids and using this information to minimize the disparity between a source and a target ErrP signal distribution, in terms of the marginal as well as the class-conditional distribution.

For this purpose, we use an ErrP-dataset collected in our lab (more details about the dataset can be found in [8]). This study was approved by the Institute Review Board (IRB) and included the EEG data of 10 human subjects recorded from 16 electrodes/channels over their scalp in a non-invasive manner. The data was sampled at 125Hz and each signal instance comprised a time window of 1.5 seconds. To remove high-frequency noise, we pass the signals through a 4-th order Butterworth filter with frequency ranging from 0.5Hz to 40Hz and select 10 channels (C3, C4, Cz, P3, P4, Pz, F3, F4, Fz and Fp2) located near the angulate cingulate cortex region of the brain as they are more relevant to the ErrP signal. The total number of samples for the 10 users is 4350. We use balanced accuracy for our evaluation since it penalizes models which are overly specific and sensitive to any class label and thus, favors a more robust model.

## 3 FACTORS LIMITING GENERALIZATION

In this section, we investigate the reasons for the poor generalization accuracy of ErrP detection algorithms. We start with the xDAWN + Riemannian Geometry (referred to as xRG from hereon) based supervised model that obtains state-of-the-art performance for ErrP generalization [11]. Figure 1 shows the general schematic of the xRG model. It contains two stages (denoted by dashed lines) that use supervised learning, namely the template generation stage (stage 1), and the classification stage (stage 2), and hence requires the ground-truth labels of a user's data. The template generation stage generates signal embeddings from raw signal data using a template estimated for each class label in a supervised manner. These embeddings are in the form of covariance matrices and are classified as belonging to ErrP or non-ErrP classes by a classifier. The aim of generalization is to systematically replace these supervised stages with label-free stages so as to generate signal embeddings without using a target user's labels as well as using a classifier that is completely blind to the target distribution. We iteratively turn these two stages from target-label-assisted to target-label-free. For any given target user, the embeddings are generated either with or without the labels of the target user, and the classifier is either trained with or without the target user's distribution. We analyze the decline in accuracy in the 4 combinations as follows:

(1) Label-assisted stages 1 and 2: for a user $U_i$, labels from $U_i$ are used for creating embeddings, which are used with labels from $U_i$ for training the classifier.
(2) Label-free stage 1 + label-assisted stage 2: for $U_i$, labels from $U_{j \neq i}$ are used for creating embeddings which are used with labels from $U_i$ for training the classifier.
(3) Label-assisted stage 1 + label-free stage 2: $U_i$'s embeddings are generated using $U_i$'s labels, which are classified by a classifier trained on $U_{j \neq i}$'s embeddings.
(4) Label-free stage 1 + label-free stage 2: $U_i$'s embeddings are generated using $U_{j \neq i}$'s labels, which are classified by a classifier trained on $U_{j \neq i}$'s embeddings.

We apply two classifiers for evaluation: the minimum-distance-to-mean (MDM) classifier [22] and the ElasticNet classifier [23]. The mean per-user detection accuracies for these 4 scenarios are shown in Table 1. We used 5-fold cross-validation for this analysis, wherein we partition the training data into 5 equal folds and each fold is iteratively used for model validation and the remainder of the data for training. When a model is used in a completely label-assisted manner, it achieves an overall 5-fold cross-validation accuracy of 76% and 78.8% for MDM and ElasticNet, respectively. Such accuracy serves as an upper bound of the possible accuracy (for the respective classifiers) as all the steps are label-assisted. Since we make the stages label-free one by one, we observe a drop in the overall detection accuracy. In the second case, the overall detection accuracy drops down to 65.4% and 71.3% for MDM and ElasticNet, respectively. Covariate shift is not responsible for the drop in accuracy since the training and

| MDM / ElasticNet | Label-assisted stage 1 | Label-free stage 1 |
|---|---|---|
| Label-assisted stage 2 | 76.0% / 78.8% | 64.8% / 71.3% |
| Label-free stage 2 | 57.1% / 60.3% | 55.8% / 59.1% |
| Silhouette score | 0.0202 | 0.0116 |

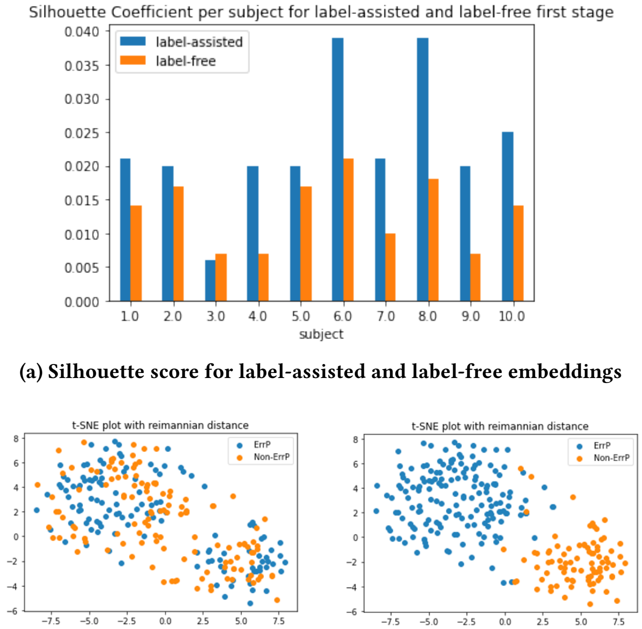
**Table 1: Balanced accuracy for label-assisted/label-free stage 1 and 2 for xRG algorithm and two classifiers.**

test data are sampled from the same distribution (the classifier is label-assisted). The decline in accuracy is attributed to diminished class discrimination in the label-free embeddings compared to the label-assisted case. This can also be empirically shown by measuring the extent of class discrimination in our embeddings using a supervised metric like the Silhouette score [24]. This score measures how closely grouped the points in the same class are compared to points from other classes. For each data point in a distribution, the Silhouette score for that data point is defined as follows:

$$s(i) = \frac{b(i) - a(i)}{max\{b(i), a(i)\}}, \tag{1}$$

where $a(i)$ is the average distance between the data point and all other points in its own class, and $b(i)$ is the average distance between the data point and all the other points belonging to another class. The total Silhouette score is then a mean of all the $s(i)$'s calculated for each point in the distribution. The amount of class discrimination present in a dataset imposes a fundamental lower bound on the classifier error trained on that dataset. The Silhouette score, which is a measure of class discrimination in a dataset is a good predictor of the maximum classifier accuracy achievable on that dataset [25]. A higher Silhouette score correlates with higher classification accuracy for that user. In our experiments, we calculate this score for every user on their embeddings, for both instances, where the embeddings are generated in a label-assisted and a label-free manner, and show their graph in Figure 2. Note that stage 2 being label-free or label-assisted has no bearing on the Silhouette score as it does not change the embeddings. Figure 2 shows the per-user Silhouette scores for two cases, when the embeddings are generated with labels vs when they are generated label-free. The label-free embeddings have a lower Silhouette score and thus, lower class discrimination, resulting in poor classification performance even with supervised classifiers. Therefore, in this case, diminished class discrimination (not covariate shift) is the contributor to low accuracy.

In the third scenario, since the embeddings are generated in a label-assisted manner, the decline in performance is a result of using a label-free classifier, i.e., a classifier that was trained on data with a different distribution as compared to the test data. In order to overcome this, we need to adapt the source dataset (consisting of embeddings that the classifier is trained on) to the target dataset such that not only the marginal distributions of both datasets achieve parity, but the class conditional distributions also equalize.



**(a) Silhouette score for label-assisted and label-free embeddings**

**(b) Label-free embeddings with lower class discrimination**

**(c) Label-assisted embeddings with higher class discrimination**

**Figure 2: Per user Silhouette scores for label-free vs label-assisted embeddings and class discrimination**

# 4 OPTIMAL TRANSPORT

## 4.1 Problem formulation

Optimal transport is the general problem of adapting one distribution to another as efficiently as possible. It requires a cost matrix that denotes the cost of moving a specific sample from the source distribution to another sample from the target distribution. Given two distributions, their associated cost matrix, and their marginal probabilities, the transport map for the source distribution is obtained by minimizing the objective function outlined below. Supposed that the densities of the source and target measures are sampled at $n_s$ and $n_t$ discrete points, we can denote the source and target probability densities by $\mathbf{a} \in \mathbb{R}^{n_s}$ and $\mathbf{b} \in \mathbb{R}^{n_t}$, respectively. Since they are sampled from a vector space, we set $[\mathbf{a}_i = 1/n_s \forall 0 \leq i \leq n_s]$ and $[\mathbf{b}_i = 1/n_t \forall 0 \leq i \leq n_t]$ Define $\langle \cdot, \cdot \rangle_\mathrm{F}$ as the Frobenius inner product. The optimal transport problem is to find a transport plan $\gamma \in \mathbb{R}^{n_s \times n_t}$ from the source domain to the target domain that is the most efficient with respect to a cost matrix $\mathbf{M} \in \mathbb{R}^{n_s \times n_t}$. In this paper, we consider the entropy regularized optimal transport, which can be computed using Sinkhorn distances [26]. The optimal transport plan $\gamma$ for this problem is obtained by minimizing:

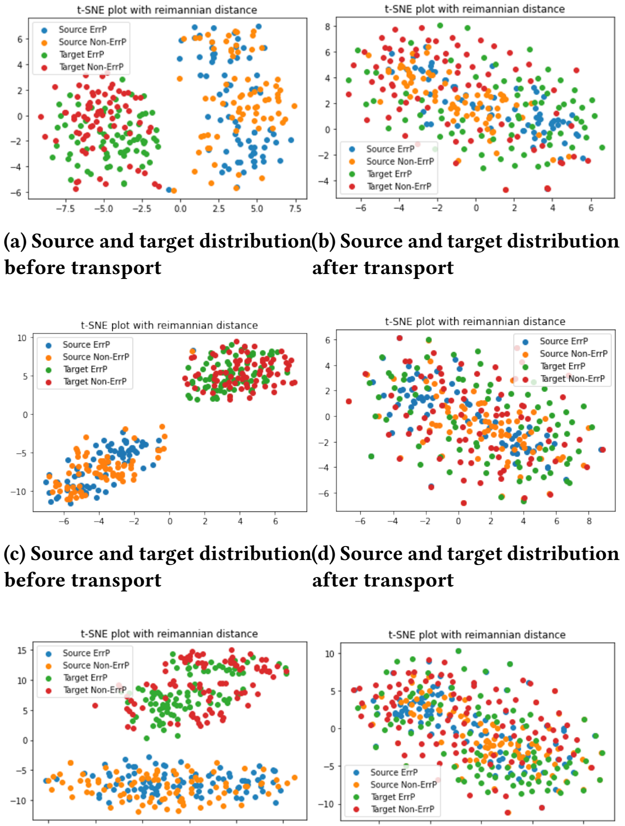
$$\gamma = \underset{\gamma}{\arg\min}\ \langle \gamma, \mathbf{M} \rangle_{\mathrm{F}} + \lambda \Omega_e(\gamma) + \eta \Omega_g(\gamma) \tag{2}$$

$$s.t.\ \gamma \mathbf{1} = \mathbf{a},\ \gamma^T \mathbf{1} = \mathbf{b},\ \gamma \geq 0 \tag{3}$$

We set $\mathbf{M}_{i,j}$, the transport cost between the $i$-th point $\mathbf{S}_i$ in the source domain and the $j$-th point $\mathbf{T}_j$ in the target domain, as the square of the Riemannian distance between these two points: $\mathbf{M}_{i,j} = \left\| \log(\mathbf{T}_j^{-1/2} \mathbf{S}_i \mathbf{T}_j^{-1/2}) \right\|_2^2$ as defined in [22], since the source and target points are both covariance matrices lying in a Riemannian manifold in our scenario. There are two regularization terms in the OT problem: the entropic regularization term, $\Omega_e(\gamma) = \sum_{i,j} \gamma_{i,j} \log(\gamma_{i,j})$, and the group lasso regularization term, $\Omega_g(\gamma) = \sum_{i,c} \parallel \gamma_{i,\mathcal{I}_c} \parallel_2$. $\Omega_e(\gamma)$ is used to ensure that the source data is transported smoothly to the target domain instead of abruptly in only a few locations. $\Omega_g(\gamma)$ is used to ensure that the source data maintains its class discrimination after transportation (labels of the same class are transported close together). We set the value of $\lambda$ to $0.02 * Median(\mathbf{M})$ and set the value of $\eta$ to be 5.

## 4.2 Types of transport maps

Equation (2) can be solved in accordance with the matrix scaling algorithm outlined in [26] to get the transport map $\gamma$. However, directly solving (2) does not yield a good transport plan as the ground-truth relationship may not be necessarily captured by the minimization of transport cost in (2). During our experiments, we observed three distinct cases of transports, namely, positive, neutral, and negative transports. Visualizations for all these are shown in Figure 3. For positive transport, the source domain maintains its class discrimination after transport, and its ErrP points are adjacent to the target ErrP points and vice versa. This is the desirable scenario as the classifier trained on the transported source data generalizes seamlessly to the target dataset and is characterized by high training accuracy on the source and high test accuracy on the target dataset. For a neutral transfer, the source ErrP and non-ErrP points do not necessarily show an affinity towards a specific class' points in the target dataset. This case is characterized by high training accuracy on the source and close to random accuracy (50%) on the target dataset. Finally, for negative transport, the source ErrP points are adjacent to the non-ErrP points in the target dataset and vice versa. Under this scenario, the respective labels of the source and target dataset are negatively matched. This case is characterized by high training accuracy on the source and very low (<50%) test accuracy on the target dataset. These three kinds of transfer occur uniformly in our dataset, which results in the overall performance of OT-based generalization being roughly equal to 50% (accuracy of a random classifier). In the next subsection, we propose an algorithm that maximizes positive transport while suppressing neutral and negative transport.



**(a) Source and target distribution before transport**
**(b) Source and target distribution after transport**

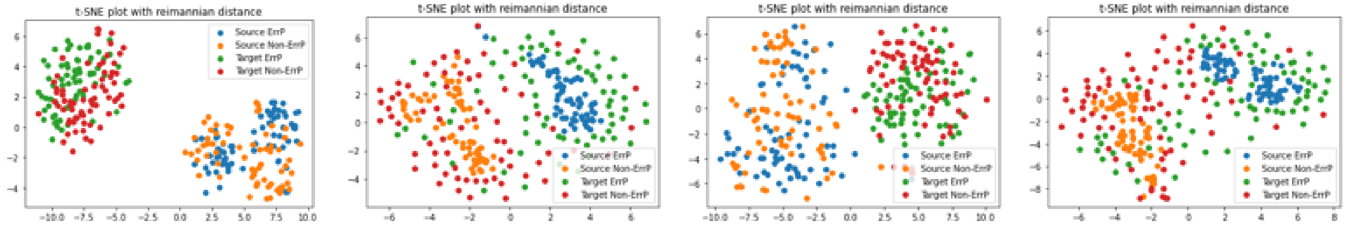
**(c) Source and target distribution before transport**
**(d) Source and target distribution after transport**

**(e) Source and target distribution before transport**
**(f) Source and target distribution after transport**

**Figure 3: Different kinds of transports, namely positive (a ⟶ b), neutral (c ⟶ d), and negative (e ⟶ f) transports**

## 4.3 Partial target-aware optimal transport

In order to mitigate negative transfer, we propose "partial target-aware optimal transport" by modifying the cost matrix $\mathbf{M}$ to establish the desired relationship between the source and target points. The full algorithm is detailed in Algorithm 1. In line 3, we first calculate the Riemannian mean of the centroids of the target data class by using only a few labeled samples from the target dataset ($m_0 = 10$ from class 0 and $m_1 = 10$ from class 1 in our experiments). After obtaining these approximate centroids, we bias the transport map to avoid transporting source labels to an area that is



**(a) Source and target distribution before transport** **(b) Source and target distribution after transport** **(c) Source and target distribution before transport** **(d) Source and target distribution after transport**

**Figure 4: Partial target-aware optimal transport visualizations and examples of transport maps**

| | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| xRG MDM | 59.2% | 53.8% | 54.8% | 60.4% | 54.2% | 56.7% | 54.3% | 55.2% | 53.5% | 55.5% | 55.8% |
| PTA-OT MDM | 61.2% | 63.0% | 61.4% | 65.8% | 59.9% | 60.6% | 59.0% | 64.4% | 60.9% | 63.8% | **62.0%** |
| xRG ElasticNet | 62.6% | 56.2% | 61.4% | 61.4% | 58.6% | 59.4% | 57.9% | 53.8% | 58.0% | 62.0% | 59.1% |
| PTA-OT ElasticNet | 66.6% | 67.6% | 63.2% | 63.9% | 68.5% | 64.1% | 60.1% | 74.0% | 62.2% | 71.54% | **66.2%** |

**Table 2: Subject-wise cross-user transfer learning accuracy for label-free xRG vs our algorithm.**

**Algorithm 1** Partial target-aware optimal transport

```
1:  Input: Source set $\{S_i | i = 1, \cdots, n_s\}$ with its density $\mathbf{a} \in \mathbb{R}^{n_s}$ and target set $\{T_i | i = 1, \cdots, n_t\}$ with its density $\mathbf{b} \in \mathbb{R}^{n_t}$. Few-shot class labeled target sets $\{{L_0}^i | i = 1, \cdots, m_0 || m_0 \ll n_t\}$ and $\{{L_1}^i | i = 1, \cdots, m_1 | m_1 \ll n_t\}$
2:  Initialization: $M_{i,j} = \| \log(T_j^{-1/2} S_i T_j^{-1/2}) \|_2^2$.
3:  $C_0 = mean(L_0), C_1 = mean(L_1)$, the initial approximation of target class centroids.
4:  for $i = 1, \cdots, n_s$ do
5:      for $j = 1, \cdots, n_t$ do
6:          $D_{j0} = dist(T_j, C_0), D_{j1} = dist(T_j, C_1)$
7:          if $class(S_i) == 0$ then
8:              $\Delta = D_{j0}/D_{j1}$
9:          else
10:             $\Delta = D_{j1}/D_{j0}$
11:         end if
12:         $M'_{i,j} = M_{i,j} * \Delta$
13:     end for
14: end for
15: $\gamma = \arg\min_\gamma \langle \gamma, M' \rangle_F + \lambda \Omega_e(\gamma) + \eta \Omega_g(\gamma)\ s.t.\ (3)$.
16: Return: $\gamma$.
```

close to the centroids of another class. From line 6 to line 12, we engineer the cost matrix **M** by increasing the distance with a dynamically calculated factor between a source point and a target point which is closer to the centroid of another class than the centroid of the source point's class. Similarly, we decrease the distance by a dynamically calculated factor between a source point and a target point that is closer to the centroid of the same class as the source point than the centroid of another class. Once we obtain $\gamma$, the transport map for the source distribution, the barycentric mapping for a source point $S_i$ is calculated as the weighted Riemannian mean of the $n_t$ target points, with the weight factor equal to $\gamma(i, :)$ which refers to the $i^{th}$ row of the transport map.

## 4.4 Performance evaluation

Table 2 details the per-user test accuracy when using partial target-aware optimal transport. Our preliminary results show that the total mean accuracy for our algorithm is equal to 62.0% for the MDM classifier (accuracy using fully label-assisted MDM classifier was 64.8%) and 66.2% for the ElasticNet classifier (accuracy using fully label-assisted ElasticNet was 71.3%). In the event a better classifier is used for ErrP detection, we expect its benefits to distribute equally to the fully label-assisted as well as our method. If we express our accuracies as percentages with respect to the accuracy achieved by label-assisted classifiers, given the label-free embeddings, we are able to reach within 95.6% and 92.8% of the accuracy for MDM and ElasticNet, respectively. Figure 4 shows a few combinations of the source and target probability distributions before and after our algorithm. As we can see in the figure, the distribution after transport not only preserves the class discrimination in the source domain but also matches ErrP points from the source user (blue) to the target user (green) and non-ErrP points from the source user (orange) to the target user (red). Our results not only outperform the cross-user generalization performance of the state-of-the-art xRG model but also approach supervised classification performance for label-free embeddings while using only a small fraction of the target labels (20 as opposed to ≈400), thereby accelerating model transfer by an order of magnitude. Please note that our algorithm is a general-purpose algorithm that works with all kinds of data distributions which suffer from covariate shift and minimizes the disparity between marginal source and target distributions while also preserving the class conditional probabilities.

## 5 CONCLUSION AND FUTURE WORK

In this work, we outlined an approach to improve transfer learning performance for a noisy and difficult-to-generalize brain signal. We demonstrated different scenarios where generalization accuracy is poor and modeled the contributors to it. We then demonstrated an approach to mitigate the effects of covariate shift using partial target-aware optimal transport and obtained state-of-the-art performance on our dataset. Our preliminary results show significant potential in using partial target-aware optimal transport to mitigate the effects of covariate shifts in cases of transfer learning. As a next step, we aim to derive a mathematical model for optimal transport which maximizes positive transport. We also aim to experiment with more datasets as well as use different label-free methods to generate embeddings that preserve class discrimination.

## REFERENCES

[1] Mohit Agarwal and Raghupathy Sivakumar. Charge for a whole day: Extending battery life for bci wearables using a lightweight wake-up command. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[2] P Brunner, L Bianchi, C Guger, F Cincotti, and G Schalk. Current trends in hardware and software for brain–computer interfaces (bcis). *Journal of neural engineering*, 8(2):025001, 2011.

[3] Ekansh Gupta, Mohit Agarwal, and Raghupathy Sivakumar. Blink to get in: Biometric authentication for mobile devices using eeg signals. *International Conference on Communications*, Jun 2020.

[4] Simanto Saha, Khondaker A. Mamun, Khawza Ahmed, Raqibul Mostafa, Ganesh R. Naik, Sam Darvishi, Ahsan H. Khandoker, and Mathias Baumert. Progress in brain computer interface: Challenges and opportunities. *Frontiers in Systems Neuroscience*, 15:4, 2 2021.

[5] Abdul Satti, Cuntai Guan, Damien Coyle, and Girijesh Prasad. A covariate shift minimization method to alleviate non-stationarity effects for an adaptive brain-computer interface. *Proceedings - International Conference on Pattern Recognition*, pages 105–108, 2010.

[6] Aniana Cruz, Gabriel Pires, and Urbano J. Nunes. Spatial filtering based on riemannian distance to improve the generalization of errp classification. *Neurocomputing*, 470:236–246, 1 2022.

[7] Aline Xavier Fidêncio, Christian Klaes, and Ioannis Iossifidis. Error-related potentials in reinforcement learning-based brain-machine interfaces. *Frontiers in Human Neuroscience*, 16:392, 6 2022.

[8] Duo Xu, Mohit Agarwal, Ekansh Gupta, Faramarz Fekri, and Raghupathy Sivakumar. Accelerating reinforcement learning using eeg-based implicit human feedback. *Neurocomputing*, Oct 2021.

[9] Mahnaz Arvaneh, Cuntai Guan, Kai Keng Ang, and Hiok Chai Quek. Spatially sparsed common spatial pattern to improve bci performance. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 2412–2415, 2011.

[10] Rivet B, Souloumiac A, Attina V, and Gibert G. xdawn algorithm to enhance evoked potentials: application to brain-computer interface. *IEEE transactions on bio-medical engineering*, 56:2035–2043, 2009.

[11] Alexandre Barachant and Marco Congedo. A plug&play p300 bci using information geometry, 2014.

[12] Marco Congedo, Alexandre Barachant, and Rajendra Bhatia. Riemannian geometry for eeg-based brain-computer interfaces; a primer and a review. *Brain-Computer Interfaces*, 4:155–174, 7 2017.

[13] Paolo Zanini, Marco Congedo, Christian Jutten, Salem Said, and Yannick Berthoumieu. Transfer learning: A riemannian geometry framework with applications to brain-computer interfaces. *IEEE Transactions on Biomedical Engineering*, 65:1107–1116, 5 2018.

[14] Sunhee Hwang, Kibeom Hong, Guiyoung Son, and Hyeran Byun. Ezslgan: Eeg-based zero-shot learning approach using a generative adversarial network. *7th International Winter Conference on Brain-Computer Interface, BCI 2019*, 2 2019.

[15] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, 38:5391–5420, 11 2017.

[16] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 15:056013, 7 2018.

[17] Ekansh Gupta and Raghupathy Sivakumar. Wisdom of the crowd: Using multi-human few-shot learning to improve cross- user generalization for error potentials in bci systems. *Proceedings of the International Joint Conference on Neural Networks*, 2022-July, 2022.

[18] Cédric Villani. Optimal transport. *Springer Berlin Heidelberg eBooks*, Jan 2009.

[19] Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. pages 1–16, 2014.

[20] Nathalie T H Gayraud, Alain Rakotomamonjy, Maureen Clerc, NTH Gayraud, A Rakotomamonjy, and M Clerc. Optimal transport applied to transfer learning for p300 detection. page 6, 9 2017.

[21] Or Yair, Felix Dietrich, Ioannis G Kevrekidis, and Ronen Talmon. Domain adaptation with optimal transport on the manifold of spd matrices.

[22] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten. Multiclass brain-computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59:920–928, 4 2012.

[23] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of The Royal Statistical Society Series B-statistical Methodology*, Apr 2005.

[24] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, Nov 1987.

[25] F. Scheidegger, R. Istrate, G. Mariani, L. Benini, C. Bekas, and A. Malossi. Efficient image dataset classification difficulty estimation for predicting deep-learning accuracy. *The Visual Computer*, Jun 2021.

[26] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.