ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230318

Distributionally Robust Cross Subject EEG Decoding

Tiehang Duan[a, c], Zhenyi Wang[b], Gianfranco Doretto[c], Fang Li[d], Cui Tao[d] and Donald Adjeroh[c]

[a]Meta AI
[b]University of Maryland, College Park
[c]West Virginia University
[d]University of Texas Health Science Center at Houston

Abstract. Recently, deep learning has shown to be effective for Electroencephalography (EEG) decoding tasks. Yet, its performance can be negatively influenced by two key factors: 1) the high variance and different types of corruption that are inherent in the signal, 2) the EEG datasets are usually relatively small given the acquisition cost, annotation cost and amount of effort needed. Data augmentation approaches for alleviation of this problem have been empirically studied, with augmentation operations on spatial domain, time domain or frequency domain handcrafted based on expertise of domain knowledge. In this work, we propose a principled approach to perform dynamic evolution on the data for improvement of decoding robustness. The approach is based on distributionally robust optimization and achieves robustness by optimizing on a family of evolved data distributions instead of the single training data distribution. We derived a general data evolution framework based on Wasserstein gradient flow (WGF) and provides two different forms of evolution within the framework. Intuitively, the evolution process helps the EEG decoder to learn more robust and diverse features. It is worth mentioning that the proposed approach can be readily integrated with other data augmentation approaches for further improvements. We performed extensive experiments on the proposed approach and tested its performance on different types of corrupted EEG signals. The model significantly outperforms competitive baselines on challenging decoding scenarios.

1 Introduction

Deep learning has found wide adoption in EEG-related clinical assistance applications in recent years, with examples such as autonomous wheelchair control [13], digital tablet interface control [2] and clinical seizure detection etc. [22]. With the signal recorded in a non-invasive way outside of human scalp, significant variance exists in the recorded signal. Researchers also observed the patterns of signal show significant deviation for different subjects [8, 9]. Cross subject EEG decoding is thus a challenging problem in that the subjects used to train the decoder is different from the subjects used for testing. The aim is for the model to perform well on arbitrary unknown subjects. The model needs to generalize well onto all subjects during training with robustness towards the variance and patterns that are subject-specific. In addition, the size of EEG dataset is relatively small given the cost and effort involved in data annotation. These pose significant challenges to the robustness of the EEG decoding model.

Data augmentation approaches perform synthetic transformations on training data. This helps the model prediction to be invariant of different forms of perturbations and improves generalization ability. It can also be seen as a regularization approach by adding specified bias and preventing model overfitting on irrelevant features. Previous works have shown augmentation operations based on domain knowledge are effective to improve EEG decoding robustness [26]. The augmentation operations are performed in the frequency domain [31], time domain [33], or spatial domain [28]. Application of such transformations needs a *priori* and the optimal choice are often dependent on model architecture, dataset processing and training setting etc., requiring manual effort in the process. Recently, explorations are also made on gradient-based automatic augmentation approaches [18] and automatic class-dependent augmentations [25]. The models introduce relaxations on the augmentation problem and enables gradient-based automatic augmentation, allowing them to exploit invariances in a broader space. For previous works, the improvements on robustness is observed and evaluated based on empirical study.

In this work, we propose a principled approach to improve the robustness of EEG decoding, by considering this as a distributionally-robust optimization (DRO) problem. It enables the design of a family of data evolution and augmentation approaches for robustness improvement in EEG decoding. The approach optimizes on the worst-case of perturbed data, which makes it robust regardless of the exact form of corruptions and variances in test data. The overall workflow is shown in fig. 1. Its functionality can be decoupled into a bi-level optimization problem that optimizes on all neighboring distributions of the training data. We formulate the distribution evolution process as a gradient flow system. More specifically, the inner sup optimization performs data evolution and augmentation with Wasserstein gradient flow on neighboring data distributions, and target function is minimized in outer optimization with gradient update of model parameters. We develop two different data evolution approaches that are approximate solutions to this DRO problem for robustness improvement, with different tradeoffs between computational efficiency, implementation simplicity, and evolution effectiveness. Intuitively, the dynamic evolution on training data generates more diverse and representative features for the EEG decoder to be robust and improves generalization. It can also be seen as filling the gap between the limited amount of labeled EEG data and the underlying data distribution, and works well to counter the variance and corruption in the EEG recordings.

We performed an extensive experimental study on model performance in addition to the theoretical analysis. We explored on the model performance with different types of corruptions that are com-

Please contact the corresponding author for any appendices or supplementary material mentioned in the paper.

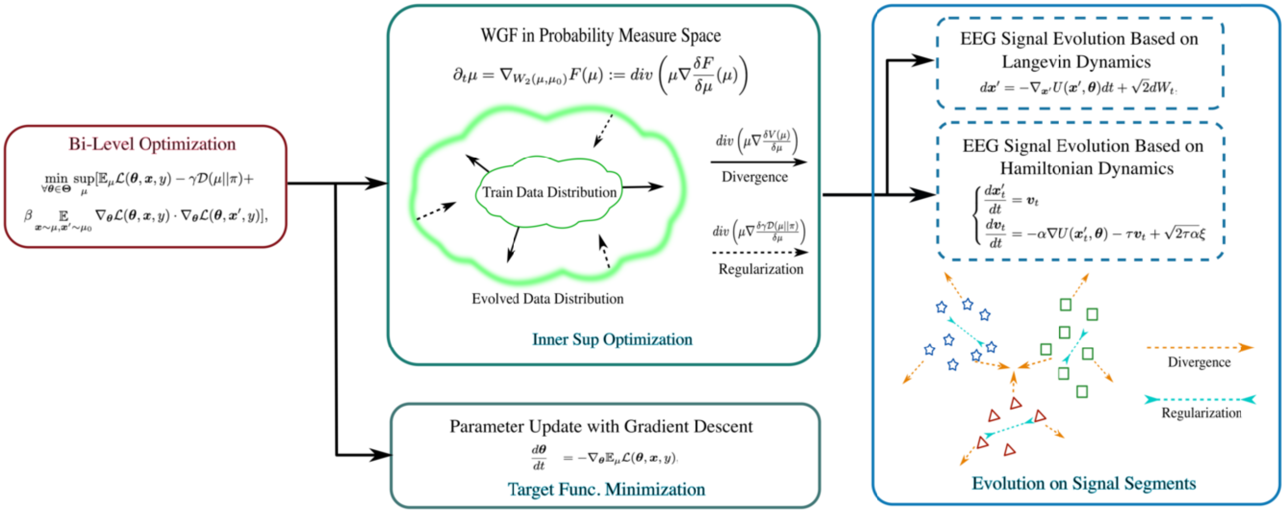

Bi-Level Optimization
$\min_{\forall\theta\in\Theta}\sup_{\mu}[\mathbb{E}_{\mu}\mathcal{L}(\theta,x,y)-\gamma\mathcal{D}(\mu||\pi)+$
$\beta\mathop{\mathbb{E}}_{x\sim\mu,x'\sim\mu_0}\nabla_{\theta}\mathcal{L}(\theta,x,y)\cdot\nabla_{\theta}\mathcal{L}(\theta,x',y)],$
WGF in Probability Measure Space
$\partial_t\mu=\nabla_{W_2(\mu,\mu_0)}F(\mu):=div\left(\mu\nabla\frac{\delta F}{\delta\mu}(\mu)\right)$
Train Data Distribution
Evolved Data Distribution
$div\left(\mu\nabla\frac{\delta V(\mu)}{\delta\mu}\right)$
Divergence
$div\left(\mu\nabla\frac{\delta\gamma\mathcal{D}(\mu||\pi)}{\delta\mu}\right)$
Regularization
Inner Sup Optimization
Parameter Update with Gradient Descent
$\frac{d\theta}{dt}=-\nabla_{\theta}\mathbb{E}_{\mu}\mathcal{L}(\theta,x,y).$
Target Func. Minimization
EEG Signal Evolution Based on Langevin Dynamics
$dx'=-\nabla_{x'}U(x',\theta)dt+\sqrt{2}dW_t.$
EEG Signal Evolution Based on Hamiltonian Dynamics
$\begin{cases}\frac{dx'_t}{dt}=v_t\\ \frac{dv_t}{dt}=-\alpha\nabla U(x'_t,\theta)-\tau v_t+\sqrt{2\tau\alpha}\xi\end{cases}$
Divergence
Regularization
Evolution on Signal Segments

Figure 1: Illustration on the overall workflow of distributionally robust decoding (DRD) of EEG signal. It optimizes on all neighboring data distributions of the EEG signal. We formulate the data distribution evolution process as a gradient flow system. Two data evolution approaches based on Langevin dynamics and Hamiltonian dynamics serve as approximate solutions to the problem.

monly incurred in recorded EEG signals due to electrodes misconnections and subject movements etc. Additionally, we studies the model's robustness towards two different types of adversarial attacks for a thorough analysis on its robustness against adversarial examples. We compared the proposed approach to competitive data augmentation approaches on different ablation settings such as varying proportions of training data, different types of corrupted data and adversarial examples etc. We also explored the influence of two different types of distribution constraints including KL-divergence and Wasserstein ball constraints in our ablation study. The distribution constraints regulate the evolved signal to not deviate too much from the original signal. Wasserstein ball constraint is able to model much richer family of distributions than KL-divergence, but it doesn't have closed form solution for the problem. We thus convert it to a surrogate loss function with proper approximation.

We summarize the contributions of this work as follows:

1) We propose a principled framework to improve robustness in EEG decoding based on distributionally robust optimization, enabling the design of a family of data evolution and augmentation approaches for robustness improvement in EEG decoding. The proposed framework enables theoretical analysis on the effectiveness of our approach to robustness improvement.

2) We formulate the data evolution process as a gradient flow system, and offer two different evolution approaches on EEG signals for solving the DRO problem with different trade-offs between efficiency and effectiveness.

3) We performed both detailed theoretical analysis and extensive empirical study on the proposed approach. The results demonstrate its effectiveness on robustness improvement in EEG decoding. The proposed approach does not require changes in the EEG decoder and can be readily integrated into current widely used BCI systems.

2 Related Work

2.1 *Robustness in EEG Decoding*

With the significant variance and signal corruption in EEG recordings, previous work have explored on the direction of robustness improvement in EEG decoding utilizing different types of data augmentation methods including generative models [11] and domain knowledge inspired approaches [4, 27]. The most straightforward augmentation is to add different types of noise to the signal [33]. Another thread of work performs time-related transformations including time shifting and time masking [5]. Similarly, spatial transformations have also been explored in recent years. Krell and Kim [14] performed rotation and shift on sensor positions to simulate the misalignment of sensor cap and scalp. Deiss et al Saeed et al [6] exploited the brain bilateral symmetry and switched left and right-sided signals. [28] proposed to randomly drop or shuffle channels for robustness improvements. Researchers also explored augmentation in the frequency domain based on expert knowledge. Schwabedal et al [31] proposed FT-surrogate transform to replace the phases of Fourier coefficients with random numbers in the range of [0, 2π]. Narrow bandstop filtering at random spectra positions are proposed in [5, 23] to prevent the model from emphasizing too much on specific frequency bands. Different from previous works that design augmentation operations based on domain knowledge with empirical analysis on effectiveness, the proposed approach offers a principled formulation on robustness improvement and enables theoretical analysis in addition to empirical evaluation on its performance.

2.2 *Distributionally Robust Optimization*

Distributionally robust optimization (DRO) aims to effectively optimize on target function across an ambiguity set of data distributions and allows the model to generalize well on decision making under uncertainty [15]. The ambiguity set is usually defined as the neighborhood of a specific distribution, with the distance between two distributions measured by probability metrics such as Wasserstein metric, in which case it is referred as WDRO [24]. DRO has been advocated to achieve robustness in noisy subpopulations [34, 35] and against adversarial examples [20], also promote stability for auto text completion of different demographic groups [12]. Previous work have also utilized DRO for problems involving group/subpopulation shift [29], class imbalance issues [38] and domain shift in meta learning. [36]. To our best knowledge, our work is the first principle approach to utilize DRO for robustness improvement in EEG decoding. We formulate the problem under the continuous dynamics perspective, and provides two different data evolution approaches to solve the problem by casting it as a gradient flow system. The proposed approach offers significant flexibility with a family of data evolution dynamics, and more evolution options are available for future explorations.

3 Method

In this section, we first present the problem setup of robustness improvement in EEG decoding, then we propose the **D**istributionally **R**obust **D**ecoding (DRD) framework for this purpose, followed by two different data evolution approaches to solve the problem based on Wasserstein gradient flows. The DRD framework is a systematic and principled approach to deal with the ambiguity in distribution of EEG signal across different subjects. It explicitly models the distribution of testing subjects to be unknown and lies in the ambiguity set of data distributions. This is particularly useful for EEG decoding with high variance in the signal and significant distribution bias across subjects, with significant uncertainty to perform decoding on the unseen subjects during testing. The proposed approach helps the model to generalize on testing subjects by simultaneously optimizing on the ambiguity set of distributions in the neighborhood of training data, and helps the model to learn features robust to EEG signal perturbations.

3.1 Problem Setup

Denote the data distribution of the recorded EEG signal as μ_0, which involves corruptions and noise in the recording process. The robustness of EEG decoding can be expressed as to achieve optimized performance with any data distribution μ that endures perturbations or corruptions and lying in the neighborhood of μ_0. The optimization process with robustness considerations is thus performed within a region of probability measure space instead of a single distribution. Formally, the optimization process with robustness considerations can be represented as

$$\min_{\forall \boldsymbol{\theta}\in\boldsymbol{\Theta}} \sup_{\mu\in\mathcal{P}} \mathbb{E}_{\mu}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{x},y) \tag{1}$$

$$\text{s.t. } \mathcal{P}=\{\mu:\mathcal{D}(\mu||\mu_0)\le\epsilon\}, \tag{2}$$

where $\{\boldsymbol{x},y\}$ is the recorded EEG data and $\boldsymbol{x}'$ is the evolved data. $\boldsymbol{\theta}$ is the model parameter of the EEG decoder, $\mathcal{D}(\cdot)$ is the distance metric between two probability distributions and ϵ is a threshold to characterize on the neighborhood of data distributions. With this problem setup on robustness, the model optimizes on the worst-case performance in the ambiguity set of neighboring distributions. This helps the model to generalize to data previously unseen and learn features that are robust to corruption and noise.

3.2 Distributionally Robust EEG Decoding

The DRD framework effectively tackles the problem setting in Section 3.1 with bi-level optimization formulation. The inner sup optimization evolves the data towards distribution that model performs worst in the ϵ-neighborhood, and outer minimization updates model parameters to improve decoding accuracy. The proposed DRD framework for robustness improvement can be expressed as:

$$\min_{\forall \boldsymbol{\theta}\in\boldsymbol{\Theta}} \sup_{\mu\in\mathcal{P}} \mathbb{E}_{\mu}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{x},y) \tag{3}$$

$$\text{s.t. } \mathcal{P}=\{\mu:\mathcal{D}(\mu||\pi)\le\mathcal{D}(\mu_0||\pi)\le\epsilon\}, \tag{4}$$

$$\mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mu_0,\boldsymbol{x}'\sim\mu} \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{x},y)\cdot\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{x}',y)\ge\lambda, \tag{5}$$

where π is the data distribution that model performs worst within the $\epsilon-$neighborhood of μ_0, i.e. the distribution with $\sup_{\mu\in\mathcal{P}}\mathbb{E}_{\mu}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{x},y)$. In this formulation, we added the constraint on the gradient dot product between original data $\boldsymbol{x}$ and evolved data $\boldsymbol{x}'$ in eq. 5. This ensures that the evolved data does not deviate too much from the original data and don't interfere with parameter update of $\boldsymbol{\theta}$. Intuitively, a negative value on the dot product indicates that the gradient direction of evolved data is contradicting with the original data. λ is a constant threshold on this constraint.

It is worth noting that exact solution for the above distributionally robust optimization problem is computationally intractable. We formulate the problem as a gradient flow system to enable gradient-based solutions with Wasserstein gradient flow. It alternatively performs evolution on the data distribution and model parameter update. The data evolution corresponds to the inner sup optimization in eq. 3, and model parameter updates is to perform the outer minimization of the target function.

3.3 Formulation of Gradient Flow System

In this section we formulate the problem into a gradient flow system and solve the inner sup optimization in eq. 3 with Wasserstein gradient flow (WGF). With $\mathcal{P}_2(\mathbb{R}^d)$ denoting the probability space on $\mathbb{R}^d$ with finite second-order moments, each $\mu\in\mathcal{P}_2(\mathbb{R}^d)$ is a probability measure defined as $\mu:\mathbb{R}^d\to\mathbb{R}$. The evolution of μ is a Wasserstein gradient flow if there exists a functional F with the following

$$\partial_t\mu=\nabla_{W_2(\mu,\mu_0)}F(\mu):=div\left(\mu\nabla\frac{\delta F}{\delta\mu}(\mu)\right) \tag{6}$$

$div(\cdot)$ is the divergence operator, ∇ is the gradient of a scalar, and $\frac{\delta F}{\delta\mu}(\mu)$ is the first derivative of F at μ.

$$\frac{\delta F}{\delta\mu}(\mu)=\lim_{\epsilon\to0}\frac{F(\mu+\epsilon\psi)-F(\mu)}{\epsilon}, \tag{7}$$

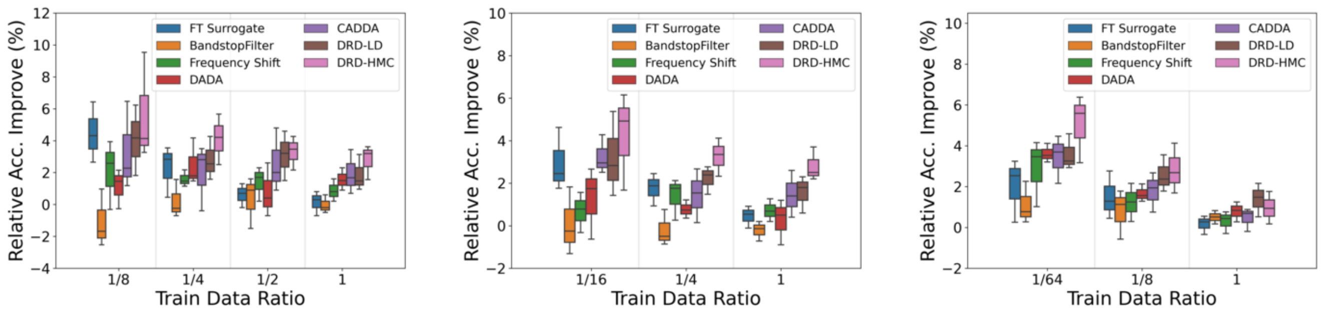
where ψ is an arbitrary function. $W_2(\mu,\mu_0)$ is the Wasserstein distance between probability measure of original data μ_0 and probability measure of evolved data μ, which is defined as

$$W_2(\mu,\mu_0)=\left(\min_{\rho(\boldsymbol{x},\boldsymbol{x}')\in\prod(\mu,\mu_0)}\int||\boldsymbol{x}-\boldsymbol{x}'||^2d\rho(\boldsymbol{x},\boldsymbol{x}')\right)^{1/2} \tag{8}$$

with $\rho(\boldsymbol{x},\boldsymbol{x}')$ being the joint probability measure of original and evolved data, $\prod(\mu,\mu_0)=\{\omega|\omega(A\times\mathbb{R}^d)=\mu(A),\omega(\mathbb{R}^d\times B)=\mu_0(B)\}$. WGF allows the data distribution μ to evolve along the steepest curve of functional $F(\mu)$ during the inner sup optimization and gradually move towards the target evolved probability measure π, starting from the initial probability measure μ_0.

For effective evolution of the signal data, we convert the optimization target based on Lagrange duality of eq. 3-eq. 5 as

$$\begin{aligned}\min_{\forall \boldsymbol{\theta}\in\boldsymbol{\Theta}} \sup_{\mu}[&\mathbb{E}_{\mu}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{x},y)-\gamma\mathcal{D}(\mu||\pi)+\\ &\beta\mathop{\mathbb{E}}_{(\boldsymbol{x},\boldsymbol{x}')\sim\prod(\mu_0,\mu)}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{x},y)\cdot\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{x}',y)],\end{aligned} \tag{9}$$



(a) BCI-IV 2a

(b) High Gamma

(c) SEED

Figure 2: Comparison on accuracy improvement with different fractions of training data, relative to the base EEG decoding model. (a) BCI-IV 2a dataset, (b) High gamma dataset, (c) SEED dataset. We observed the proposed approaches have more significant improvement on model performance in low data resource scenarios.

The target function $F(\mu)$ is defined accordingly for effective signal data evolution

$$F(\mu) = \underbrace{-\mathbb{E}_{\mu}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{x},y) - \beta\mathbb{E}_{\mu}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{x},y)\cdot\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{x}',y)}_{V(\mu)} + \gamma\mathcal{D}(\mu||\pi) \tag{10}$$

The DRO problem in eq. 9 can be solved with the following gradient flow system

$$\begin{cases} \partial_t \mu &= div\left(\mu\nabla\frac{\delta(V(\mu)+\gamma\mathcal{D}(\mu||\pi))}{\delta\mu}\right); & (11) \\ \dfrac{d\boldsymbol{\theta}}{dt} &= -\nabla_{\boldsymbol{\theta}}\mathbb{E}_{\mu}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{x},y), & (12) \end{cases}$$

Eq. 11 solves the inner sup with evolution on μ and eq. 12 solves the outer minimization with update on θ. We propose two different types of data evolution methods to effectively solve eq. 11-eq. 12. The first approach utilizes Langevin dynamics with a diffusion process to perform data evolution, then we generalize the above WGF to have better flexibility and instantiate the generalized WGF with Hamiltonian dynamics for data evolution.

Algorithm 1 DRD-LD/HMC Model

```
REQUIRE: EEG decoder parameters $\boldsymbol{\theta}$, learning rate $\eta$, evolution rate $\alpha$, evolution time $T$.
for $i = 1$ to $N$ do
    input EEG data $(\boldsymbol{x}_i, y_i)$ arrives.
    $\boldsymbol{x}' = \boldsymbol{x}$
    for $t = 1$ to $T$ do
        $(\boldsymbol{x}', y) = Transform((\boldsymbol{x}', y))$ by Langevin dynamics (Eq. (15)) or Hamiltonian dynamics (Eq. (18)).
    end for
    $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \eta\nabla_{\boldsymbol{\theta}}[\mathcal{L}(\boldsymbol{\theta}_i, \boldsymbol{x}, y) + \mathcal{L}(\boldsymbol{\theta}_i, \boldsymbol{x}', y)]$
end for
```

Evolution based on Langevin Dynamics The gradient flow in eq. 11 on probability measure corresponds to the Langevin dynamics [37] on data samples that are depicted with the following stochastic differential equation:

$$d\boldsymbol{x}' = -\alpha\nabla_{\boldsymbol{x}'}U(\boldsymbol{x}',\boldsymbol{\theta})dt + \sqrt{2\alpha}dW_t, \tag{13}$$

$$\text{where } U(\boldsymbol{x}',\boldsymbol{\theta}) = \frac{\delta(V(\mu)+\gamma\mathcal{D}(\mu||\pi))}{\delta\mu} \tag{14}$$

$$= -\mathcal{L}(\boldsymbol{\theta},\boldsymbol{x},y) - \beta\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{x},y)\cdot\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{x}',y) + \gamma(\log\frac{\mu}{\pi}+1) \tag{15}$$

with $\boldsymbol{x}' = (\boldsymbol{x}'_t)_{t\geq 0}$ the evolved data, $d\boldsymbol{x}'$ the evolution during dt and α is evolution rate. W_t is the standard Brownian motion in $\mathbb{R}^n$. Derivation details are provided in Appendix C. Intuitively, the left-hand side of eq. 15 evolves the data towards harder cases in the neighborhood of original data and makes it more challenging for model to learn. Discretize the data evolution in eq. 15, then we got the following update rule:

$$\boldsymbol{x}_{t+1} - \boldsymbol{x}_t = -\alpha(\nabla_{\boldsymbol{x}_t}U(\boldsymbol{x}_t,\boldsymbol{\theta})) + \sqrt{2\alpha}\xi. \tag{16}$$

We abbreviate this distributionally robust data evolution approach as **DRD-LD**. The first term in the right hand side of eq. 16 drives the signal segments towards the target probability distribution π, and the second term generates necessary randomness for increased diversity in the data.

General Form of Evolution with Hamiltonian Dynamics Given the fact that a continuous Markov process that produces samples following a probability measure can be written into the general form [19], the previous WGF on data evolution can similarly be represented as

$$\partial_t\mu = div(\mu(\boldsymbol{H}+\boldsymbol{J})\nabla\frac{\delta F}{\delta\mu}(\mu)) \tag{17}$$

with $\boldsymbol{H}$ being the diffusion matrix and $\boldsymbol{J}$ the skew-symmetric curl matrix. This general form of representation allows flexibility to encode prior or geometric information into the evolution process. A specific instantiation of $\boldsymbol{H}$ and $\boldsymbol{J}$ is to set

$$\boldsymbol{H} = \begin{pmatrix} 0 & 0 \\ 0 & \boldsymbol{R} \end{pmatrix}, \boldsymbol{J} = \begin{pmatrix} 0 & -\boldsymbol{I} \\ \boldsymbol{I} & 0 \end{pmatrix},$$

this WGF formulation follows Hamiltonian dynamics with $\boldsymbol{I}$ the identity matrix and $\boldsymbol{R}$ the friction matrix, and corresponds to the following data evolution

Table 1: Model performance on adversarial examples. The model is evaluated under two different types of adversarial attacks, including projected gradient descent (PGD) ℓ_∞ attack and Carlini & Wagner ℓ_2 attack. For PGD attack, We experimented with two different levels of perturbation magnitude 0.02 and 0.1 on the normalized data.

Dataset	BCI-IV 2a			High Gamma			SEED		
Method	PGD (0.02)	PGD (0.1)	C&W	PGD (0.02)	PGD (0.1)	C&W	PGD (0.02)	PGD (0.1)	C&W
DADA	$23.84_{\pm 2.16}$	$1.81_{\pm 0.31}$	$15.72_{\pm 0.54}$	$33.94_{\pm 1.35}$	$2.07_{\pm 0.61}$	$38.26_{\pm 3.67}$	$16.42_{\pm 1.78}$	$0.96_{\pm 0.23}$	$5.57_{\pm 0.49}$
CADDA	$31.50_{\pm 1.29}$	$1.03_{\pm 0.06}$	$16.44_{\pm 0.83}$	$37.78_{\pm 3.14}$	$2.62_{\pm 0.27}$	$33.53_{\pm 2.48}$	$14.75_{\pm 3.13}$	$2.18_{\pm 0.54}$	$9.31_{\pm 0.66}$
DRD-LD	$42.62_{\pm 1.73}$	$3.26_{\pm 0.65}$	$19.38_{\pm 1.28}$	$60.40_{\pm 2.39}$	$5.13_{\pm 0.84}$	$49.69_{\pm 1.52}$	$19.56_{\pm 2.10}$	$8.61_{\pm 0.72}$	$10.58_{\pm 1.49}$
DRD-HMC	$43.97_{\pm 2.85}$	$2.49_{\pm 0.08}$	$24.61_{\pm 1.46}$	$67.32_{\pm 1.68}$	$5.47_{\pm 0.35}$	$52.45_{\pm 2.13}$	$22.38_{\pm 1.27}$	$6.25_{\pm 1.68}$	$13.64_{\pm 2.23}$

Table 2: The influence of evolution steps on testing accuracy of cross subject EEG decoding. The model performance converges after more than 5 evolution steps.

Evolution Time	1		3		5		7	
Method	DRD-LD	DRD-HMC	DRD-LD	DRD-HMC	DRD-LD	DRD-HMC	DRD-LD	DRD-HMC
BCI-IV 2a	$53.12_{\pm 0.74}$	$53.39_{\pm 1.82}$	$54.01_{\pm 2.08}$	$54.56_{\pm 1.15}$	$54.20_{\pm 1.53}$	$54.85_{\pm 2.13}$	$54.48_{\pm 1.76}$	$55.24_{\pm 1.21}$
High Gamma	$81.08_{\pm 2.10}$	$81.23_{\pm 1.29}$	$81.65_{\pm 1.47}$	$82.04_{\pm 2.36}$	$81.37_{\pm 1.82}$	$82.15_{\pm 2.54}$	$81.42_{\pm 0.89}$	$82.56_{\pm 1.44}$
SEED	$69.41_{\pm 1.83}$	$70.26_{\pm 2.17}$	$70.32_{\pm 0.74}$	$71.93_{\pm 1.85}$	$69.80_{\pm 2.79}$	$72.56_{\pm 4.13}$	$70.69_{\pm 1.27}$	$72.75_{\pm 2.58}$

$$\begin{cases} \dfrac{d\boldsymbol{x}'_t}{dt} = \boldsymbol{v}_t \\ \dfrac{d\boldsymbol{v}_t}{dt} = -\alpha\nabla U(\boldsymbol{x}'_t, \boldsymbol{\theta}) - \tau\boldsymbol{v}_t + \sqrt{2\tau\alpha}\xi \end{cases} \tag{18}$$

where $\boldsymbol{v}_t$ is the momentum and τ is its update rate. The evolution rule in eq. 18 can be discretized as:

$$\begin{cases} \boldsymbol{x}_{t+1} - \boldsymbol{x}_t = \boldsymbol{v}_t, \\ \boldsymbol{v}_{t+1} - \boldsymbol{v}_t = -\alpha(\nabla_{\boldsymbol{x}_t} U(\boldsymbol{x}_t, \boldsymbol{\theta})) - \tau\boldsymbol{v}_t + \sqrt{2\tau\alpha}\xi, \end{cases} \tag{19}$$

We name this approach as **DRD-HMC**. The desirable property of this approach is that it offers flexibility to freely specify the matrixs $\boldsymbol{H}$ and $\boldsymbol{J}$ based on specific practical requirements, and encode prior information or geometric constraints into the tailored $\boldsymbol{H}$ and $\boldsymbol{J}$. Note further extensions on the evolution approaches are available under this general framework, i.e. utilize the reproducing kernel Hilbert space (RKHS) kernels built on $\boldsymbol{x}$ and $\boldsymbol{x}'$ to serve as $\boldsymbol{H}$ in eq. 17, for which we leave as future work. We provide the overall evolution algorithm in Algorithm 1.

Table 3: Comparison of different methods in terms of corruption error on all three datasets. The corruption error is the averaged error rate of model predictions across the different types of corruption operations.

Methods	BCI-IV 2a	High Gamma	SEED
FT Surrogate	$61.27_{\pm 2.31}$	$48.95_{\pm 1.86}$	$54.19_{\pm 1.02}$
BandStopFilter	$65.64_{\pm 3.08}$	$50.20_{\pm 2.34}$	$56.03_{\pm 1.86}$
Frequency Shift	$62.49_{\pm 1.58}$	$47.27_{\pm 1.30}$	$52.44_{\pm 2.10}$
DADA	$58.44_{\pm 2.47}$	$42.38_{\pm 3.06}$	$53.85_{\pm 1.47}$
CADDA	$56.91_{\pm 1.69}$	$41.46_{\pm 1.83}$	$51.92_{\pm 0.93}$
DRD-LD	$55.58_{\pm 2.17}$	$38.31_{\pm 1.02}$	$50.58_{\pm 1.80}$
DRD-HMC	**54.13±1.54**	**37.87±2.36**	**48.95±2.62**

4 Experiments

We performed extensive evaluation of cross subject EEG decoding performance in this section, with ablation study on model performance with respect to different types of signal corruptions and adversarial attacks, training with different data volumes etc. We also performed detailed analysis on model sensitivity to hyperparameters. In this section we first make an introduction on data processing and model settings, followed by detailed performance analysis.

Datasets We perform detailed evaluation on model performance with three public EEG datasets, BCI-IV 2a [32] [1], high gamma dataset [30] [2] and SEED dataset [7] [3].

BCI-IV 2a dataset involves 9 subjects performing 4 different classes of motor imagery tasks including left hand, right hand, feet and tongue. Each subject takes part in 2 sessions of 288 trials. The signals are recorded with 22 electrodes and downsampled to 250Hz.

High gamma dataset consists of 14 subjects with each performing 880 trials. The dataset is originally recorded with 128 electrodes and we used 44 channels covering the motor cortex. The dataset is also downsampled to 250Hz.

SEED dataset is formed with 15 subjects performing emotion recognition tasks. The subjects watch film clips with positive, neural and negative emotions states. The signals are recorded with 62-channel ESI NeuroScan System, originally sampled at 1000Hz and then downsampled to 200Hz.

Baselines We include a wide range of baselines on data augmentation for comparison in our experiment, which can be categorized as following:

1) Augmentation approaches based on domain expertise. We incorporated augmentation approaches currently widely used for EEG signals including **FT surrogate** [31], **BandstopFilter** [23] and **frequency shift** [10]. FT surrogate replaces the coefficients of Fourier transformation on the signal with random numbers in $[0, 2\pi]$. BandstopFilter performs narrow bandstop filtering at numerous random spectral positions and avoids the model from overfitting onto a single frequency band. Frequency shift performs an uniform offset of Δf on signal frequencies, which is sampled uniformly from range linearly set by the magnitude.

2) Gradient-based automatic data augmentation approaches including **DADA** [17] and **CADDA** [25] are incorporated in our comparison. DADA performs automatic search on augmentation policies and relaxes the discrete augmentation policy selection into a differ-

[1] http://bnci-horizon-2020.eu/database/data-sets
[2] https://github.com/robintibor/high-gamma-dataset
[3] http://bcmi.sjtu.edu.cn/~seed/downloads.html

entiable problem. CADDA is another gradient-based automatic augmentation approach leveraging class information.

4.1 *Settings*

Data Processing

For BCI-IV 2a Dataset, the trials are processed into segments of size 400×22, with a span of 400 along the time axis and 22 channels. The stride between adjacent segments is 50. We extracted the period between $t = 3s$ and $t = 6s$ in each trial for decoding purposes. This generates 8 signal segments per trial.

For high gamma dataset, we processed the trials into segments of size 400×44, with time length of 400 and 44 sensor channels used. The stride size is 100 between adjacent segments.

For SEED dataset, the trials are divided into segments of size 800×62, with a stride size of 100 between adjacent segments. This produces 472 segments per trial. Given the dataset is too large for model to digest, we downsampled it to 10% of its original size and repeat each run for 10 times to get an accurate estimation on its performance.

Model Settings The base EEG decoding model is a compact 3-layer convolutional neural network similar to EEGNet [16]. The first layer is formed with filters of size $(1, C)$, C is the number of channels for spatial convolution. Filters of second layer are of size $(32, 2)$ emphasizing on temporal convolutions. The third layer performs pointwise convolution operations for improved computational efficiency. Zero padding is performed between adjacent layers to maintain data dimensionality. For the cross subject EEG decoding scenario, we leave one subject out for testing and use the other subjects for training each time, and the performance is averaged across all subjects. The number of evolution steps is set to 5 by default, the gradient dot product factor β is set to 0.003 for BCI-IV 2a dataset, 0.001 for high gamma dataset and 0.005 for SEED dataset. The evolution rate α is set to 0.05 for BCI-IV 2a dataset, while high gamma and SEED dataset use an evolution rate of 0.01. Results are averaged across 10 runs in the experiment.

4.2 *Performance Analysis*

Results on the performance of different approaches for the three datasets are illustrated in Fig. 2. We experimented with different fractions of training data to understand model performance in low resource scenarios, in the range of [1/8, 1] for BCI-IV 2a dataset, [1/16, 1] for high gamma dataset and [1/64, 1] for SEED dataset. The varying ranges takes the different data volumes of the three datasets into consideration, with volume of SEED dataset much larger than BCI-IV 2a dataset. The proposed approaches steadily outperform other baselines in these different settings, e.g. DRD-HMC achieved a margin of more than 4% on accuracy for all three datasets with low regime of training data used. Fig. 3 visualizes the effect of evolution with different number of evolution steps on the signal segments. The evolution at sample level generates more diverse features which contributes to robustness improvement of the model. Augmentation approaches built on empirical experience such as FT surrogate and Frequency Shift leads to more than 2% accuracy improvement compared to base model on BCI-IV 2a and SEED datasets respectively, and more than 1% improvement on high gamma dataset. For gradient based approaches including DADA and CADDA, we use the learned policy to retrain the model from scratch, and we observed CADDA steadily achieved more than 2% accuracy gain for all three datasets.

Performance on Corrupted Data

We perform evaluation of model performance in terms of corruption error on different types of corrupted data, and computes the averaged error over the different types of corruptions (the list of corruptions and their parameter settings are provided in Appendix E). The result is shown in Table 3. DRD-HMC has a margin of 2.78%, 3.59% and 2.97% in terms of corruption error reduction for BCI-IV 2a, high gamma and SEED dataset respectively.

Performance on Adversarial Examples

Adversarial examples are data samples $\boldsymbol{x}'$ that are close enough to original data $\boldsymbol{x}$ as determined by some distance function $D(\boldsymbol{x}.\boldsymbol{x}') \leq \epsilon$ but divert the classifier to produce different predictions, i.e. $f_{\boldsymbol{\theta}}(\boldsymbol{x}) \neq f_{\boldsymbol{\theta}}(\boldsymbol{x}')$. We evaluated the model performance under two different types of adversarial attacks, namely, Projected Gradient Descent (PGD) ℓ_∞ attack [21] and Carlini & Wagner ℓ_2 attack [3]. The result is shown in Table 1. PGD ℓ_∞ attack forms the adversarial examples with gradient projection under ℓ_∞ norm constraint. We experimented with two different levels of perturbation magnitude, 0.02 and 0.1, on the normalized three datasets. We adopt the ℓ_2 settings in [3] for Carlini & Wagner attack. For PGD ℓ_∞ attack with perturbation magnitude of 0.02, the performance of comparison models are near random guess, and the accuracy further reduces to near zero with perturbation magnitude of 0.1. The proposed DRD-LD and DRD-HMC approaches significantly outperform baselines by at least 4.81% for PGD ℓ_∞ (0.02), and 1.46% for PGD ℓ_∞ (0.1). For Carlini & Wagner attack, the proposed approaches have a margin of 2.94% on BCI-IV 2a dataset, 11.43% on high gamma dataset and 1.27% on SEED dataset. Both results demonstrate the robustness improvement of proposed approach on adversarial examples.

4.3 *Ablation Study*

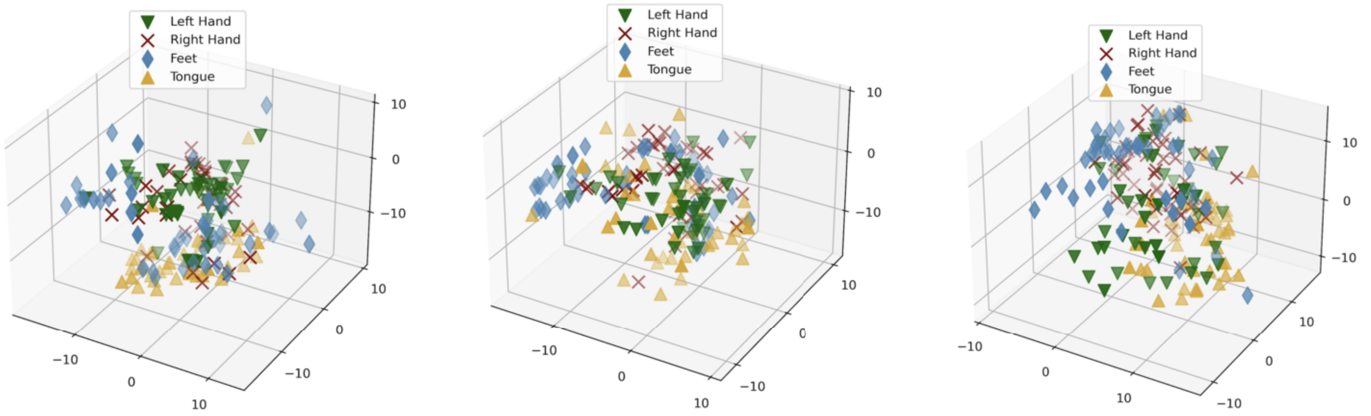
Table 4: Ablation study on influence of regularization weight γ, gradient dot product factor β and evolution rate α on testing accuracy.

γ	0.1	0.3	0.5	0.8
BCI-IV 2a	54.47±2.58	**54.85±2.13**	54.69±1.80	54.22±3.16
High Gamma	81.24±1.72	81.71±1.29	**82.15±2.54**	81.36±1.87
SEED	71.18±3.07	71.94±1.52	**72.56±4.13**	72.18±2.30
β	0.0	0.001	0.003	0.005
BCI-IV 2a	54.10±1.48	54.23±1.71	**54.85±2.13**	54.66±1.39
High Gamma	81.67±1.95	**82.15±2.54**	82.02±1.29	81.74±1.75
SEED	72.14±1.37	71.98±2.06	72.32±1.94	**72.56±4.13**
α	0.01	0.03	0.05	0.1
BCI-IV 2a	54.59±1.87	54.74±1.48	**54.85±2.13**	54.41±2.62
High Gamma	**82.15±2.54**	81.72±1.10	81.87±1.96	81.30±2.28
SEED	**72.56±4.13**	72.23±2.85	71.72±1.49	71.95±3.04

Hyperparameter Sensitivity We perform ablation study on the model hyperparameters including regularization weight γ, gradient dot product factor β and evolution rate α. The result is provided in Table 4. We observed the optimal choice of γ is 0.3 for BCI-IV 2a dataset, and 0.5 for high gamma dataset and SEED dataset. For gradient dot product factor β, we performed sensitivity analysis within the range of $[0.0, 0.005]$, with $\beta = 0$ corresponding to the case with no gradient regularization added. We also performed the ablation study on the evolution rate α which controls the data evolution speed. We observed BCI-IV 2a needs a higher evolution rate to achieve optimal performance than the other datasets. We explored the influence of different evolution time on model performance, the result is provided in Table 2. The model performance converges with evolution

Table 5: Performance comparison with different distance constraints, including KL-divergence and the distance depicted by Wasserstein ball. Wasserstein ball constraint offers more flexibility but its exact solution is computationally intractable and we adopted the approximation approach in [1]

Distance Constraint	BCI-IV 2a		High Gamma		SEED	
	DRD-LD	DRD-HMC	DRD-LD	DRD-HMC	DRD-LD	DRD-HMC
KL-divergence	54.20±1.53	54.85±2.13	81.37±1.82	82.15±2.54	69.80±2.79	72.56±4.13
WB-distance	53.86±2.29	55.02±1.07	81.15±1.38	81.73±0.91	69.54±3.35	71.12±1.57



(a) Original Data

(b) Evolution Steps $t = 3$

(c) Evolution Steps $t = 5$

Figure 3: TSNE visualization of the different motor imagery classes at feature level for BCI-IV 2a dataset. (a) original data, (b) evolved data with 3 evolution steps, (c) evolved data with 5 evolution steps. The dynamic evolution on training data generates more diverse and robust features for classes such as left hand and tongue.

steps larger than 5. We set the evolution steps to be 5 as default in our experiment, which is the tradeoff between performance and computational efficiency.

Distance Constraints We explored the effect of different types of distance constraints on model performance. In addition to KL-divergence, we also explored to instantiate the distance $\mathcal{D}(\mu||\pi)$ with Wassarstein distance $W(\mu, \pi)$ to constrain the evolved data distribution and not deviate too much from the original distribution. The comparison is summarized in Table 5, with WB-distance denote the Wasserstein ball constraint. KL-divergence and Wassarstein distance are endowed with different properties. The gradient flows of KL-divergence is straight forward to solve with calculus of variation. On the other hand, Wasserstein ball constraint incorporates more flexibility in distance definition but its gradient flow solution is computationally intractable and approximation is needed in the process. We adopt the approximation optimization approach for Wasserstein ball constraint introduced in [1] and turning it into a surrogate loss function (more details in Appendix D). This makes the computation tractable but its performance is slightly inferior than KL-divergence.

Computational Cost We performed evaluation on the computational efficiency of proposed approach compared to baselines. The result is provided in Table 6. For gradient-based auto augmentation approaches such as DADA and CADDA, we run the search algorithm to the point where it converges to a stable performance. We observed DRD-LD is more computationally efficient than DRD-HMC, showing the tradeoff between flexibility of evolution depiction and computationally complexity. In general, the proposed approach takes less time to reach optimal than automatic gradient based augmentation approaches.

Table 6: Computational cost comparison of proposed approach with other baselines. For gradient-based auto augmentation approaches such as DADA and CADDA, we run the search algorithm to the point that it converges to a stable performance.

Method	BCI-IV 2a (min)	High Gamma (hr)	SEED (hr)
base model	4.2	1.27	3.22
DADA	11.1	2.73	5.80
CADDA	16.4	3.86	9.88
DRD-LD	9.5	2.91	6.61
DRD-HMC	13.8	3.44	8.06

5 Conclusion

In this work, we proposed a principled data evolution approach for robustness improvement in decoding of EEG signals. The proposed approach utilizes distributionally robust optimization to achieve optimized performance on any data distribution lying in the neighborhood of training data distribution instead of training data itself. We formulate the proposed DRD framework into a gradient flow system to enable tractable data evolution solutions with Wasserstein gradient flow, and provide two data evolution mechanisms based on Langevin dynamics and Hamiltonian dynamics, respectively. We performed detailed evaluation on the proposed approach with different types of corrupted data and adversarial examples. The proposed approach outperforms competitive baselines by a large margin in these challenging scenarios. Numerous future extensions are available based on current work, including tailored matrix design in generalized WGF formulation to encode prior knowledge, and the utilization of kernelized WGF in the evolution process.

6 Acknowledgements

This material is based upon work supported by the National Science Foundation under Grants No. 1920920, 2125872 and 2223793.

References

[1] Jose Blanchet and Karthyek R. A. Murthy, 'Quantifying distributional model risk via optimal transport', *https://arxiv.org/abs/1604.01446*, (2017).

[2] Andrew Campbell et al., 'Neurophone: Brain-mobile phone interface using a wireless eeg headset', in *Proceedings of the Second ACM SIGCOMM Workshop on Networking, Systems, and Applications on Mobile Handhelds*, MobiHeld '10, p. 3–8, New York, NY, USA, (2010). Association for Computing Machinery.

[3] Nicholas Carlini and David Wagner, 'Towards evaluating the robustness of neural networks', *2017 IEEE Symposium on Security and Privacy (SP)*, (2017).

[4] Kaixuan Chen, Lina Yao, Dalin Zhang, Xiaojun Chang, Guodong Long, and Sen Wang, 'Distributionally robust semi-supervised learning for people-centric sensing', *CoRR*, **abs/1811.05299**, (2018).

[5] Joseph Y Cheng, Hanlin Goh, Kaan Dogrusoz, Oncel Tuzel, and Erdrin Azemi, 'Subject-aware contrastive learning for biosignals', *arXiv preprint arXiv:2007.04871*, (2020).

[6] Olivier Deiss, Siddharth Biswal, Jing Jin, Haoqi Sun, M Brandon Westover, and Jimeng Sun, 'Hamlet: interpretable human and machine co-learning technique', *arXiv preprint arXiv:1803.09702*, (2018).

[7] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu, 'Differential entropy feature for EEG-based emotion classification', in *6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pp. 81–84. IEEE, (2013).

[8] Tiehang Duan, Mohammad Abuzar Shaikh, Mihir Chauhan, Jun Chu, Rohini K. Srihari, Archita Pathak, and Sargur N. Srihari, 'Meta learn on constrained transfer learning for low resource cross subject eeg classification', *IEEE Access*, **8**, 224791–224802, (2020).

[9] Tiehang Duan, Zhenyi Wang, Sheng Liu, Yiyi Yin, and Sargur N. Srihari, 'Uncer: A framework for uncertainty estimation and reduction in neural decoding of eeg signals', *Neurocomputing*, **538**, 126210, (2023).

[10] Daniel Freer and Guang-Zhong Yang, 'Data augmentation for self-paced motor imagery classification with c-lstm', *Journal of Neural Engineering*, **17**(1), 016041, (jan 2020).

[11] Kay Gregor Hartmann, Robin Tibor Schirrmeister, and Tonio Ball, 'Eeg-gan: Generative adversarial networks for electroencephalograhic (eeg) brain signals', *arXiv preprint arXiv:1806.01875*, (2018).

[12] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang, 'Fairness without demographics in repeated loss minimization', in *International Conference on Machine Learning*, pp. 1929–1938. PMLR, (2018).

[13] I. Iturrate, J. Antelis, and J. Minguez, 'Synchronous eeg brain-actuated wheelchair with automated navigation', in *2009 IEEE International Conference on Robotics and Automation*, pp. 2318–2325, (2009).

[14] Mario Michael Krell and Su Kyoung Kim, 'Rotational data augmentation for electroencephalographic data', in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 471–474, (2017).

[15] Yongchan Kwon, Wonyoung Kim, Joong-Ho Won, and Myunghee Cho Paik, 'Principled learning method for wasserstein distributionally robust optimization with local perturbations', in *International Conference on Machine Learning*, pp. 5567–5576. PMLR, (2020).

[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**(5), 056013, (jul 2018).

[17] Yonggang Li, Guosheng Hu, Yongtao Wang, Timothy Hospedales, Neil M Robertson, and Yongxin Yang, 'Dada: Differentiable automatic data augmentation', *arXiv preprint arXiv:2003.03780*, (2020).

[18] Yonggang Li, Guosheng Hu, Yongtao Wang, Timothy Hospedales, Neil M. Robertson, and Yongxin Yang, 'Differentiable automatic data augmentation', in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII*, p. 580–595, Berlin, Heidelberg, (2020). Springer-Verlag.

[19] Yi-An Ma, Tianqi Chen, and Emily B. Fox, 'A complete recipe for stochastic gradient mcmc', in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, p. 2917–2925, Cambridge, MA, USA, (2015). MIT Press.

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, 'Towards deep learning models resistant to adversarial attacks', *arXiv preprint arXiv:1706.06083*, (2017).

[21] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, 'Towards deep learning models resistant to adversarial attacks', *Proceedings of the International Conference on Learning Representations*, (2018).

[22] Saba Moghimi et al., 'A review of eeg-based brain-computer interfaces as access pathways for individuals with severe disabilities', *Assistive Technology*, **25**(2), 99–110, (2013).

[23] Mostafa Neo Mohsenvand, Mohammad Rasool Izadi, and Pattie Maes, 'Contrastive representation learning for electroencephalogram classification', in *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 136 of *Proceedings of Machine Learning Research*, pp. 238–253. PMLR, (11 Dec 2020).

[24] Hamed Rahimian and Sanjay Mehrotra, 'Distributionally robust optimization: A review', *arXiv preprint arXiv:1908.05659*, (2019).

[25] Cédric Rommel, Thomas Moreau, Joseph Paillard, and Alexandre Gramfort, 'CADDA: Class-wise automatic differentiable data augmentation for EEG signals', in *International Conference on Learning Representations*, (2022).

[26] Cédric Rommel, Joseph Paillard, Thomas Moreau, and Alexandre Gramfort, 'Data augmentation for learning predictive models on eeg: a systematic comparison', *Journal of Neural Engineering*, **19**(6), 066020, (nov 2022).

[27] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert, 'Deep learning-based electroencephalography analysis: a systematic review', *Journal of Neural Engineering*, **16**(5), 051001, (aug 2019).

[28] Aaqib Saeed et al., 'Learning from heterogeneous eeg signals with differentiable channel reordering', in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1255–1259. IEEE, (2021).

[29] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang, 'Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization', *arXiv preprint arXiv:1911.08731*, (2019).

[30] Robin Tibor Schirrmeister et al., 'Deep learning with convolutional neural networks for eeg decoding and visualization', *Human Brain Mapping*, **38**(11), 5391–5420, (2017).

[31] Justus TC Schwabedal, John C Snyder, Ayse Cakmak, Shamim Nemati, and Gari D Clifford, 'Addressing class imbalance in classification problems of noisy signals by using fourier transform surrogates', *arXiv preprint arXiv:1806.08675*, (2018).

[32] Michael et al. Tangermann, 'Review of the bci competition iv', *Frontiers in Neuroscience*, **6**, 55, (2012).

[33] Fang et al. Wang, 'Data augmentation for eeg-based emotion recognition with deep convolutional neural networks', in *MultiMedia Modeling*, eds., Klaus Schoeffmann, Thanarat H. Chalidabhongse, Chong Wah Ngo, Supavadee Aramvith, Noel E. O'Connor, Yo-Sung Ho, Moncef Gabbouj, and Ahmed Elgammal, pp. 82–93, Cham, (2018). Springer International Publishing.

[34] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan, 'Robust optimization for fairness with noisy protected groups', *Advances in neural information processing systems*, **33**, 5190–5203, (2020).

[35] Zhenyi Wang, Tiehang Duan, Le Fang, Qiuling Suo, and Mingchen Gao, 'Meta learning on a sequence of imbalanced domains with difficulty awareness', in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8927–8937, (2021).

[36] Zhenyi Wang, Li Shen, Le Fang, Qiuling Suo, Tiehang Duan, and Mingchen Gao, 'Improving task-free continual learning by distributionally robust memory evolution', in *International Conference on Machine Learning*, pp. 22985–22998. PMLR, (2022).

[37] Max Welling and Yee Whye Teh, 'Bayesian learning via stochastic gradient langevin dynamics', in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, p. 681–688, Madison, WI, USA, (2011). Omnipress.

[38] Ziyu Xu, Chen Dan, Justin Khim, and Pradeep Ravikumar, 'Class-weighted classification: Trade-offs and robust approaches', in *International Conference on Machine Learning*, pp. 10544–10554. PMLR, (2020).