Vol.2022-AVM-116 No.15
2022/2/25

ラプラス正則化付き最適輸送問題を用いたPU学習手法の検討

影山 遼[1]　福永 拓海[2]　笠井 裕之[1,2]

概要：Positive Unlabeled (PU) 学習は機械学習の分野の一つであり，二値分類を拡張した学習手法である．二値分類との違いは，学習データとして正のラベルデータのみを与える点にある．従来，PU 学習の一手法として最適輸送問題を用いた手法が検討されてきた．PU 学習を行う際のデータの分布に関する仮定の 1 つとして，滑らかさというものが存在する．これは，近いところに位置するデータ同士は同じ正クラスに属する確率も近いとする仮定である．これを踏まえると，最適輸送問題を考える際に，輸送元となるラベルなしデータで近かった点は輸送後も近くなるようにマッピングがなされるべきであると考えられる．本稿では，ラプラス正則化を用いて輸送前後の距離関係に基づいたマッピングを行う学習法について提案する．

PU Learning using Optimal Transport with Laplacian Regularization

1. はじめに

Positive Unlabeled (PU) 学習は，二値分類の発展形として 2000 年代初頭に登場した機械学習の分野の一つである．二値分類との決定的な違いは，学習時に正ラベルデータのみが与えられ，負のクラスについて一切の情報を持たないという点である．ユーザーのコンテンツに対する関心や過去の病歴など，PU データは実世界にさまざまな形で存在している [1]．例えば，ユーザーがコンテンツに興味を持っているか否かについて分類を行いたい場合，ユーザーがアクセスしたことがあるコンテンツについては興味がある，つまり正のデータとして扱うことができる一方で，アクセスしたことがないコンテンツに関しては，ユーザーの興味があるかないかについては判断が難しく，ラベルなしのデータとして扱われる．なぜならば，クリックしていない原因には，興味がない可能性だけでなくただユーザーの目に留まっていないだけの可能性も考えられるためである．

PU 学習の手法の 1 つとして，最適輸送問題を用いる手法が存在する [2]．この手法では，PU 学習をラベルなしデータから正ラベルデータへ質量を輸送する最適輸送問題と見做し，正ラベルデータにマッピングしたものを正クラスに分類し，それ以外を負クラスに分類することで正負の分類を行うことが提案されている．この手法の中で登場する最適輸送問題の定式化ではグループラッソ正則化が用いられており，この正則化は輸送行列の不等式制約を解消するために追加されたダミーの点への質量の輸送と，正ラベルデータへの輸送が分かれることに寄与している．最適輸送問題にはそれ以外にもさまざまな正則化を適用することができ，そのような正則化の 1 つとしてラプラス正則化が存在する [3]．ラプラス正則化によって，同じクラスに属するデータ同士を輸送後も近くなるように輸送を行なったり，輸送前後でデータの位置関係を維持するように輸送を行うことが可能となる．

本稿では，距離が近いデータ同士は同じクラスに属する確率も近いとする仮定に着目し，PU 学習における最適輸送問題ついて輸送前の距離関係に基づいたマッピングを行うためにラプラス正則化を用いるアプローチを提案する．

2. 関連研究

2.1 Positive Unlabeled (PU) 学習 [1]

PU 学習は，正と負を分類する分類器を学習するということを目的とした二値分類の発展形である．従来の二値分

[1] 早稲田大学 基幹理工学部 情報通信学科
Department of Communications and Computer Engineering, School of Fundamental Science and Engineering, Waseda University
[2] 早稲田大学大学院 基幹理工学研究科 情報理工・通信専攻
Department of Computer Science and Communications Engineering, Graduate School of Fundamental Science and Engineering, Waseda University

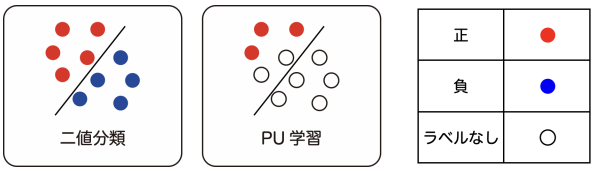

図 1　二値分類と PU 学習 [4]

類と比較して決定的に異なる点は，学習時に与えられるラベル付きデータが正クラスからのもののみであるという点である．図 1 に，二値分類と PU 学習の違いについて示す．

学習段階では負のクラスについて一切の情報がないため，学習時に仮定を立てる．立てる仮定はいくつか存在する．

まず，正クラスデータからラベルを付与するデータがどのように選択されているのかについては，SCAR (Selected Completely At Random) や SAR (Selected At Random) といった仮定が存在する．

SCAR は学習時にラベル付けされる確率が自身の特徴に依らず，正のデータの分布から完全にランダムに選択される，とする仮定で，多くの PU 学習の手法の根幹となる仮定である．

SAR はラベル付けされる正のデータが選択される確率はその特徴 $\boldsymbol{x}$ に依存する [5]，とする仮定で，ラベル付けのメカニズムにおける最も一般的な仮定である．実世界で例を挙げると，例えばネット上の広告をクリックするかどうかはその配置箇所という特徴に影響される．また他の例として，病気にかかった人間が医者にかかるかどうかも社会的・経済的地位や症状の重さといったような特徴に依存する．

データの分布に関してもいくつか仮定が存在するが，そのうちの滑らかさは互いに近いデータほど同じクラスに属する可能性が高いとする仮定である．つまり，2 つのデータの特徴 $\boldsymbol{x}_1$ と $\boldsymbol{x}_2$ が似ていて，クラス情報が y で与えられるとき，$\Pr(y=1|\boldsymbol{x}_1)$ と $\Pr(y=1|\boldsymbol{x}_2)$ の値も近いことを意味する．この仮定により，全てのラベル付き正データから遠いデータは負の確率が高いデータと見なすことが可能となる．

2.2　最適輸送問題 [2]

$\mathcal{X}=\{\boldsymbol{x}_i\}_{i=1}^n$ と $\mathcal{Y}=\{\boldsymbol{y}_j\}_{j=1}^m$ をそれぞれソース，ターゲットの点群として，以下のような 2 つの経験分布 $(\boldsymbol{p},\boldsymbol{q})$ を考える．

$$\boldsymbol{p}=\sum_{i=1}^{n} p_i\delta_{\boldsymbol{x}_i},\quad \boldsymbol{q}=\sum_{j=1}^{m} q_j\delta_{\boldsymbol{y}_j}.$$

このとき，考えられる写像 $\mathbf{T}:\mathcal{X}\to\mathcal{Y}$ の集合 $\Pi(\boldsymbol{p},\boldsymbol{q})$ は以下のように表せる．

$$\Pi(\boldsymbol{p},\boldsymbol{q})=\{\mathbf{T}\in\mathbb{R}_+^{|\boldsymbol{p}|\times|\boldsymbol{q}|}\mid \mathbf{T}\mathbf{1}_{|\boldsymbol{q}|}=\boldsymbol{p},\mathbf{T}^\top\mathbf{1}_{|\boldsymbol{p}|}=\boldsymbol{q}\}.$$

ここで，$\mathbf{T}$ は輸送行列と呼ばれ，行列の各成分 $\mathbf{T}_{ij}$ は正の実数値 $\mathbf{T}_{ij}\in\mathbb{R}_+$ であり，$\boldsymbol{x}_i$ から $\boldsymbol{y}_j$ に輸送する質量を表している．最適輸送では，$\boldsymbol{x}_i$ と $\boldsymbol{y}_j$ の間の輸送コスト $\mathbf{D}_{ij}$ が与えられたとき，$\boldsymbol{p}$ から $\boldsymbol{q}$ に最小コストで輸送する問題を扱う．より具体的には，$\mathbf{C}=\mathbf{D}^p$ が距離行列として与えられたとき，p 次 Wasserstein 距離は以下のように定義される．

$$W_p^p(\boldsymbol{p},\boldsymbol{q})=\underset{\mathbf{T}\in\Pi(\boldsymbol{p},\boldsymbol{q})}{\text{minimize}}\langle\mathbf{C},\mathbf{T}\rangle=\underset{\mathbf{T}\in\Pi(\boldsymbol{p},\boldsymbol{q})}{\text{minimize}}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{C}_{ij}\mathbf{T}_{ij}. \tag{1}$$

2.3　最適輸送問題を用いた PU 学習 [2]

通常の最適輸送問題では，輸送元と輸送先で質量の総和が等しく ($\|\boldsymbol{p}\|_1=\|\boldsymbol{q}\|_1=1$)，かつ全ての質量を輸送しなければならないため，制約が厳しく一部の問題設定で不適切となりうる．そこで，輸送質量 s を $s=1$ ではなく $0\le s\le \text{minimize}(\|\boldsymbol{p}\|_1,\|\boldsymbol{q}\|_1)$ として，輸送行列 $\mathbf{T}$ が以下のような条件を満たす集合 $\Pi^u(\boldsymbol{p},\boldsymbol{q})$ に属する部分最適輸送問題を考える．

$$\Pi^u(\boldsymbol{p},\boldsymbol{q})=\{\mathbf{T}\in\mathbb{R}_+^{|\boldsymbol{p}|\times|\boldsymbol{q}|}\mid \mathbf{T}\mathbf{1}_{|\boldsymbol{q}|}\le\boldsymbol{p},\mathbf{T}^T\mathbf{1}_{|\boldsymbol{p}|}\le\boldsymbol{q},\ \mathbf{1}_{|\boldsymbol{p}|}^T\mathbf{T}\mathbf{1}_{|\boldsymbol{q}|}=s\}. \tag{2}$$

このとき，Partial Wasserstein 距離 PW_p^p を定義すると，式 (1) の Wasserstein 距離同様以下のようになる．

$$PW_p^p(\boldsymbol{p},\boldsymbol{q})=\underset{\mathbf{T}\in\Pi^u(\boldsymbol{p},\boldsymbol{q})}{\text{minimize}}\langle\mathbf{C},\mathbf{T}\rangle. \tag{3}$$

この問題は [7] や [8] で研究されており，数値解は [9] や [10] で与えられている．[2] では，ダミーの点を追加して式 (2) に含まれている不等式制約を解消するアプローチが用いられている．具体的には，コスト行列を以下のように拡張し，$\bar{\mathbf{C}}$ とする．

$$\bar{\mathbf{C}}=\begin{bmatrix}\mathbf{C} & \xi\mathbf{1}_{|\boldsymbol{q}|}\\ \xi\mathbf{1}_{|\boldsymbol{p}|}^T & 2\xi+A\end{bmatrix}$$

ただし A はスカラーであり $A>\text{maximize}(\mathbf{C}_{ij})$ である．また，ξ は定数である．$p_{n+1}=\|\boldsymbol{q}\|_1-s$ と $q_{m+1}=\|\boldsymbol{p}\|_1-s$ を追加して新たに $\bar{\boldsymbol{p}}=[\boldsymbol{p},\|\boldsymbol{q}\|_1-s]$ と $\bar{\boldsymbol{q}}=[\boldsymbol{q},\|\boldsymbol{p}\|_1-s]$ を定義し，以下の Wasserstein 距離を最小化する輸送行列 $\hat{\mathbf{T}}$ を求める．

$$W_p^p(\bar{\boldsymbol{p}},\bar{\boldsymbol{q}})=\underset{\bar{\mathbf{T}}\in\Pi^u(\bar{\boldsymbol{p}},\bar{\boldsymbol{q}})}{\text{minimize}}\langle\bar{\mathbf{C}},\bar{\mathbf{T}}\rangle.$$

求まった $\bar{\mathbf{T}}$ から $n+1$ 行目，$m+1$ 列目を削除することにより，式 (3) の PW_p^p を最小化する $\mathbf{T}$ が求められる．

次に，最適輸送問題を PU 学習に導入することを考える．$\mathcal{X}=\{\boldsymbol{x}_i\}_{i=1}^n$ をデータ全体から取ってきたラベルなしデータの集合，$\mathcal{Y}=\{\boldsymbol{y}_j\}_{j=1}^m$ を正のクラスから取ってきて正のラベルを付加してあるデータの集合として，元のデータ全

体に占める正クラスデータの割合を π とすると，PU 学習は総輸送質量 $s = \pi$ を満たすラベルなしデータ $\mathcal{X}$ から正ラベルデータ $\mathcal{Y}$ への最適輸送問題と考えることができる．

このとき，さらに $p_i = \frac{1}{n}, q_j = \frac{s}{m}$，つまりそれぞれのデータの持つ質量は等しいとすると，輸送行列 $\mathbf{T}$ は以下のような条件を満たす集合 Π^{PU} に属する．

$$\Pi^{PU}(\boldsymbol{p}, \boldsymbol{q}) = \{\mathbf{T} \in \mathbb{R}_+^{|\boldsymbol{p}|\times|\boldsymbol{q}|} \mid \mathbf{T}\mathbf{1}_{|\boldsymbol{q}|} = \{\boldsymbol{p}, 0\}, \mathbf{T}^T\mathbf{1}_{|\boldsymbol{p}|} \leq \boldsymbol{q}, \mathbf{1}_{|\boldsymbol{p}|}^T\mathbf{T}\mathbf{1}_{|\boldsymbol{q}|} = s\}.$$

ここで，$\mathbf{T}\mathbf{1}_{|\boldsymbol{q}|} = \{\boldsymbol{p}, 0\}$ は $\forall i, \sum_j \mathbf{T}_{ij} = p_i \; or \; 0$ であることを意味する．これは，ラベルが付加されていない負のデータから正ラベルデータに質量が輸送されないようにするための制約である．また，ノイズによって正のラベル付きデータの特徴が壊れていたり，ラベル付けが誤っている可能性があるため，ノイズレベル α を $0 \leq \alpha \leq 1 - s$ の範囲で定義する．このとき，$p_i = \frac{1-\alpha}{n}$，$q_j = \frac{s+\alpha}{m}$ となる．コスト行列を拡張すると，求めるべき輸送行列 $\bar{\mathbf{T}}^*$ は

$$\bar{\mathbf{T}}^* = \underset{\bar{\mathbf{T}}\in\Pi(\boldsymbol{p},\boldsymbol{q})}{\arg\min} \sum_{i=1}^{n+1}\sum_{j=1}^{m+1} \bar{\mathbf{C}}_{ij}\bar{\mathbf{T}}_{ij} + \eta_{GL}\Omega_{GL}(\bar{\mathbf{T}}). \quad (4)$$

と表せる．ただし，$\eta_{GL} \geq 0$ は正則化パラメータである．また，$\Omega_{GL}(\mathbf{T}) = \sum_i \sum_g ||\mathbf{T}_{i\mathcal{I}_g}||_2$ であり，$\mathcal{I}_g$ は正のデータ ($g = [1, m]$) あるいは $g = m + 1$ に対応する行列 $\mathbf{T}$ のインデックスを含む．つまり，$g = [1, m]$ のとき $\mathbf{T}_{i\mathcal{I}_g} \in \mathbb{R}^m$ であり，$g = m + 1$ のとき $\mathbf{T}_{i\mathcal{I}_g} \in \mathbb{R}$ である．この項はグループラッソ正則化と呼ばれ，この項によってラベルが付加されていない負のデータから正ラベルデータにマッピングしたり，ラベルが付加されていない正のデータからダミーの点にマッピングしたりすることを防止できる．

3. 提案手法

3.1 従来手法の問題点

[2] の研究では，PU 学習をラベルなしデータから正ラベルデータへの最適輸送問題と見做すアプローチを提案している．その最適輸送の定式化では，グループラッソ正則化を用いることで意図しない質量の輸送がなされることを回避していた．しかし，ラベル付けにおける仮定の 1 つである滑らかさを考慮するならば，ただ自らから近い点に輸送するだけでなく，ソース同士やターゲット同士の距離関係も考慮されるべきである．そこで，同じクラスに属する点を近づけるように輸送するラプラス正則化を適用することも正則化の選択として適切なのではないか，また，[2] で使用されているグループラッソ正則化と併用することで更なる分類精度の向上が見込めるのではないかと考え，本手法の提案に至った．

3.2 提案手法の内容

ラプラス正則化項 $\Omega_{Lap}(\mathbf{T})$ は以下の形で表される [3].

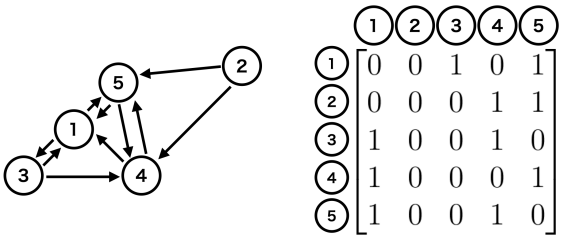

図 2 k-NN を用いた隣接行列の導出例 ($k = 2$)

$$\Omega_{Lap}(\mathbf{T}) = \mathrm{Tr}(\mathbf{Y}^T\mathbf{T}^T\mathbf{L}_s\mathbf{T}\mathbf{Y}).$$

ここで，$\mathbf{L}_s = \mathrm{diag}(\mathbf{S}_s\mathbf{1}) - \mathbf{S}_s$ はグラフ $\mathbf{S}_s$ のラプラシアン行列であり，$\mathbf{S}_s$ はソースの隣接行列である．[2] において，ソースはラベルなしデータでありクラス情報を持たないため，隣接行列を計算できない．そこで，k-NN を用いて自らに最も近い k 個のデータへのエッジを持つと考え，隣接行列を計算する．

図 2 に，$k = 2$ の場合に k-NN を用いて隣接行列を導出する例を示す．i 番目のノードから j 番目のノードにエッジが伸びている際，隣接行列 $\mathbf{S}$ の (i, j) 成分 $\mathbf{S}_{ij} = 1$ となり，逆に伸びていない場合は $\mathbf{S}_{ij} = 0$ となる．ノードは自身には繋がらないため，隣接行列の対角成分は必ず 0 である．

本研究では，式 (4) についてラプラス正則化への置換やラプラス正則化項の付加を行うことを考える．具体的には，以下の 2 通りの式を用いる．

$$\bar{\mathbf{T}}^* = \underset{\bar{\mathbf{T}}\in\Pi(\boldsymbol{p},\boldsymbol{q})}{\arg\min} \sum_{i=1}^{n+1}\sum_{j=1}^{m+1} \bar{\mathbf{C}}_{ij}\bar{\mathbf{T}}_{ij} + \eta_{Lap}\Omega_{Lap}(\bar{\mathbf{T}}). \quad (5)$$

$$\bar{\mathbf{T}}^* = \underset{\bar{\mathbf{T}}\in\Pi(\boldsymbol{p},\boldsymbol{q})}{\arg\min} \sum_{i=1}^{n+1}\sum_{j=1}^{m+1} \bar{\mathbf{C}}_{ij}\bar{\mathbf{T}}_{ij} + \eta_{GL}\Omega_{GL}(\bar{\mathbf{T}}) + \eta_{Lap}\Omega_{Lap}(\bar{\mathbf{T}}). \quad (6)$$

ここで，$\eta_{Lap} \geq 0$ はラプラス正則化項全体にかかるパラメータである．

4. 実験

4.1 実験内容

[2] と同じく，SCAR 仮定と SAR 仮定での実験を 3 手法で行う．3 手法とは，従来のグループラッソ正則化のみを用いた手法 [2]・ラプラス正則化のみを用いた手法・グループラッソ正則化とラプラス正則化の両方を用いた手法を指し，以下ではそれぞれ GL・Lap・GL+Lap と記述する．

実験では，ランダムに正データとラベルなしデータを選択しラベルなしデータの分類を行うということを 10 回繰り返し，その分類精度の平均値を算出する．

用いるデータセットは，SCAR 仮定については UCI リポジトリ[*1]から mushrooms，shuttle，pageblocks，usps，connect-4，spambase の 6 つを使用し，SAR 仮定の実験で

[*1] https://archive.ics.uci.edu/ml/datasets.php

表 1　SCAR 仮定での分類精度

データセット	π	GL [2]	Lap	GL+Lap
mushrooms	0.518	0.952	0.985	**0.991**
shuttle	0.786	0.951	**0.973**	0.972
pageblocks	0.898	0.919	0.926	**0.930**
usps	0.167	0.983	**0.990**	**0.990**
connect-4	0.658	0.610	**0.626**	0.597
spambase	0.394	0.788	**0.793**	0.785
平均値		0.867	**0.882**	0.877

表 2　SAR 仮定での分類精度

データセット	π	GL [2]	Lap	GL+Lap
mnist	0.1	0.998	**1.000**	**1.000**
colored mnist	0.1	**0.813**	0.806	0.804

は colored MNIST データセット [11] を使用した．MNIST データセットは 0 から 9 の手書き数字のデータセットで，colored MNIST データセットはそれぞれの数字が赤か緑のいずれかに，赤に色付けされたデータの数と緑に色付けされたデータの数の比が赤: 緑 $= 9 : 1$ となるように着色されている．どちらのデータセットも 0，2，4，6，8 と書かれているデータが正クラス，1，3，5，7，9 と書かれているデータが負クラスとなるように処理されている．colored MNIST データセットではさらに，最初にラベル付けされる正データは緑に色付けされたデータからのみ選択される．このアプローチによって正ラベルデータを選択する確率が色という属性に依存することとなり，SAR 仮定の実験を可能にしている．パラメータについては，式 (5) を用いた手法については $k = 43$，$\eta_{Lap} = 1$ とした．また，式 (6) を用いた手法については $k = 5$，$\eta_{Lap} = 1$，$\eta_{GL} = 10^{-3}$ とした．

4.2　実験結果

実験の結果を表 1 (SCAR) と表 2 (SAR) に示す．それぞれの表の π はデータセット全体に占める正クラスデータの割合を示している．また各データセットについて，最も精度の良かった値を太字で示している．SCAR については Lap によって全体的な分類精度の向上が見られた．一方 SAR では，提案手法は 2 手法とも従来手法 [2] と比較して精度が低下するという結果となった．この原因として，SAR 仮定のように学習データに偏りがある場合では，ラプラス正則化によって輸送後の距離が近づくように促すのはかえって逆効果となってしまっているのではないかと考えられる．また GL+Lap が SCAR と SAR の両方において Lap を下回った原因については，GL+Lap の収束のスピードが極めて遅く，場合によっては収束し切らない・オーバーフローを起こすということがあったため，パラメータを小さくしていたり収束の条件を緩めたりしている点が正則化項の働きを弱めてしまっているのではないかと考えられる．

5.　まとめ

本稿では PU 学習における滑らかさの仮定に着目し，ラプラス正則化項を用いて輸送前後の距離関係に基づいたマッピングを行うことを提案した．

今後の課題は二つ存在する．まず第一に，グループラッソ正則化を併用することでの精度の改善について引き続き検討する必要がある．次に，[2] で用いられていた Partial Gromov-Wasserstein をラプラス正則化に適用することができていないため，ドメイン適応のように異なる次元への学習データの適応は現状不可能となっている．そのため，Partial Gromov-Wasserstein のラプラス正則化への適用について検討していきたい．また，最適輸送を用いず k-NN のみを用いた PU 学習の手法についても現在検討中のため，並行して研究を進めていきたい．

参考文献

[1] Bekker, J. and Davis, J.: Learning from positive and unlabeled data: a survey, *Machine Learning*, Vol. 109, No. 4, pp. 719–760 (online), DOI: 10.1007/s10994-020-05877-5 (2020).
[2] Chapel, L., Alaya, M. Z. and Gasso, G.: Partial Optimal Transport with Applications on Positive-Unlabeled Learning (2020).
[3] Flamary, R., Courty, N., Rakotomamonjy, A. and Tuia, D.: Optimal transport with Laplacian regularization, *NIPS 2014* (2014).
[4] Bao, H., Sakai, T., Sato, I. and Sugiyama, M.: Convex Formulation of Multiple Instance Learning from Positive and Unlabeled Bags (2018).
[5] Bekker, J., Robberechts, P. and Davis, J.: Beyond the Selected Completely At Random Assumption for Learning from Positive and Unlabeled Data (2019).
[6] Neelakantan, A., Roth, B. and McCallum, A.: Compositional Vector Space Models for Knowledge Base Completion (2015).
[7] Caffarelli, L. A. and McCann, R. J.: Free boundaries in optimal transport and Monge-Ampère obstacle problems, *Annals of Mathematics*, Vol. 171, No. 2, pp. 673–730 (online), available from ⟨http://www.jstor.org/stable/20752228⟩ (2010).
[8] Figalli, A.: The Optimal Partial Transport Problem, *Archive for Rational Mechanics and Analysis*, Vol. 195, pp. 533–560 (online), DOI: 10.1007/s00205-008-0212-7 (2010).
[9] Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L. and Peyré, G.: Iterative Bregman Projections for Regularized Transportation Problems (2014).
[10] Chizat, L., Peyré, G., Schmitzer, B. and Vialard, F.-X.: Scaling Algorithms for Unbalanced Transport Problems (2017).
[11] Arjovsky, M., Bottou, L., Gulrajani, I. and Lopez-Paz, D.: Invariant Risk Minimization (2020).