arXiv:2304.02064v1 [cs.LG] 4 Apr 2023

多源领域自适应表示学习的算法依赖的界

Qi Chen
拉瓦尔大学

Mario Marchand
拉瓦尔大学

摘要

我们利用信息论工具，从表示学习的视角对多源域适应（MDA）进行全新分析。具体而言，我们研究有监督多源域适应中少量目标标签条件下的联合分布对齐，以及利用伪标签进行的无监督多源域适应；后者相对困难，研究也较少。我们进一步为这两种情形提供依赖算法的泛化界，其中泛化由参数与数据之间的互信息刻画。随后，我们提出一种新颖的深度多源域适应算法，通过联合对齐隐式处理目标偏移。最后，我们将互信息界扩展至该算法，并由此得到非平凡的梯度范数估计。所提出的算法在目标偏移多源域适应基准上的表现可与最先进水平媲美，同时提高了内存效率。

1 引言

通常的机器学习理论假设测试数据与训练数据服从相同的分布，但这一假设在现实应用中经常不成立。这种分布（领域）偏移会降低算法的性能。因此，人们提出了多种方法，通过领域自适应（DA）或迁移学习来解决这一问题（Huang 等人，2007年；Ben-David 等人，2010年（a）；Pan 和 Yang，2009年）。领域自适应旨在为目标域学习一个泛化能力良好的预测器 T，使用从源域采样的数据 S，其中 $S \neq T$。随着深度学习凭借其在计算机视觉等许多复杂任务中的优异表现而日益普及（Csurka，2017年）以及自然语言处理（Blitzer，2008年），一系列关于深度领域自适应的研究（Ganin 等人，2016年；Tzeng 等人，2017年）在实践中取得了巨大成功。

上述工作均聚焦于单源域自适应。然而，在实践中，利用多个源域（$S_{1:N}$）的知识更具吸引力。简单地将所有源域合并为一个源域，然后应用单源域自适应算法，显然不是最优的，因为某些源分布通常会比其他源分布与目标分布更相似。在这种情况下，均匀合并后的源分布可能与目标分布极不相似，从而违背成功进行领域自适应的条件，例如 Ben-David 等人（2010年（b））提出的条件。

因此，为了从最相关的源域中学习，许多现有的多源域适应（Zhao 等人，2018年；Wen 等人，2020年；Peng 等人，2019年；Li 等人，2018年）工作利用每个源域边缘分布与目标域边缘分布之间的散度$d(S_i(X)\ T(X))$ $\forall i \in [N]$来推断领域关系。当标签分布不同时，即使在单源领域自适应中（Zhao 等人，2019年），这类边缘分布对齐方法也已被证明存在问题$T(Y) \neq S(Y)$——我们将其称为**目标偏移**[1]。最近，其他方法（Redko 等人，2019年a；Shui 等人，2021年）引入了额外的算法层，用于估计标签比例，以减轻目标偏移。然而，在某些情况下，这种估计可能不准确，进而导致性能下降。

上述目标偏移校正方法在数学上是联合分布对齐的一种特殊情况。此外，当源域数量 N 非常大时，先前的成对对齐方法在内存和样本复杂度方面都难以估计大量判别器。因此，我们提出在常用于深度学习的隐藏层表示空间中，对源分布的凸组合 $S^{\alpha} = \sum_{i=1}^{N} \alpha_i S_i$，$\forall \boldsymbol{\alpha} \in \Delta_N$与目标分布 T进行联合对齐，其中向量 $\boldsymbol{\alpha}$ 的

《 26th国际人工智能与统计会议（AISTATS）论文集》，2023年，西班牙瓦伦西亚。PMLR：第206卷。版权归作者所有，2023年。

[1]为避免与标签偏移假设混淆，本文其余部分均采用**目标偏移** 这一术语，其中$S(X|Y) = T(X|Y)\ S(Y) \neq T(Y)$。

主权重由学习算法进行优化。对齐是在深度学习中常用的隐藏层表示空间上进行的，而相应的分析却鲜有研究。此外，先前采用拉德马赫复杂度或 VC 维数的理论并不依赖于具体算法。相比之下，我们使用信息论工具，对 MDA 表示学习进行针对特定算法的泛化误差分析。

与以往工作相比，本文的主要贡献如下:

统一方法 我们提出了一种适用于有监督和无监督多源域适应的统一方法，该方法针对表示空间进行联合对齐，适用于目标偏移问题。所提出的算法能够同时学习域权重，并且所提出的非成对对齐比以往方法更加节省内存。此外，在无监督场景中，当目标偏移显著时，所提出的算法优于以往方法，而这种情况在实践中具有现实意义。

依赖算法的泛化界 我们首先为多源域适应提出依赖算法的泛化界，该界取决于模型参数与输入数据之间的互信息：模型参数对数据的依赖越小，算法就越不容易过拟合。我们进一步将该界应用于所提出的深度多源域适应算法，并得到梯度范数估计，将其用作优化域权重的正则化系数。

2 相关工作

由于领域自适应在各种设定下已有大量研究，本节仅讨论与本文最相关的方法。

领域自适应中的分布偏移度量早期关于单源域适应的理论研究基于 $\mathcal{H}$-散度（Ben-David 等人，2006年，2010年a、b）。随后，研究者开始探索不同的度量，例如差异（Mansour 等人，2009年）、Wasserstein 距离（Courty 等人，2014年，2017年）、Jensen-Shannon 散度（Shui 等人，2020年，2022年b），以及更一般的散度定义，如积分概率度量（IPM）（Zhang 等人，2012年）和f-散度（Acuna 等人，2021年）。如需更全面的讨论，请参阅 Redko 等人（2019年b）和 Wang 等人（2023年）。

本文采用 Wasserstein 距离，因为它具有诸多优良性质。在次高斯假设下，通过传输代价不等式，它比 KL 散度更紧，还能够捕捉数据的内在几何结构。Courty 等人（2017）首次从理论上分析了联合对齐.....

了使用 Wasserstein 距离进行的单源领域自适应中的联合对齐。基于求解最优传输问题的算法，其复杂度随样本量呈二次增长，因此难以应用于大型数据集。因此，Damodaran 等人（2018）通过交替更新小批量耦合矩阵和网络参数，在表示空间中近似 Wasserstein 距离，以实证方式解决这一难题。我们从理论上分析了表示空间中关于 Wasserstein 距离的联合对齐，并给出了依赖算法的界。

领域自适应中的信息论学习近期，Xu 和 Raginsky（2017年）以及 Russo 和 Zou（2019年）提出的信息论分析，已被用于深入理解深度学习模型的泛化能力，例如复杂的元学习算法（Chen 等人，2021年；Jose 和 Simeone，2020年）。与传统的 VC 维数界和一致稳定性界相比，该方法的一项显著优势是能够纳入对数据分布、假设空间和学习算法的依赖关系。Wu 等人（2020年、2022年）也使用信息论工具为单源监督领域自适应推导界限。然而，他们的界限包含一个无法优化的项 $D_{\mathrm{KL}}(\mathcal{T},\ \mathcal{S})$。与之相反，我们首次为无监督和有监督多源域适应提供完全依赖算法的界限；这一结论在单源领域自适应的情形下同样成立。

多源域适应Konstantinov 和 Lampert（2019年）以及 Shui 等人（2021年）研究了目标域仅有少量标签时的有监督多源域适应情形。Konstantinov 和 Lampert（2019年）使用成对差异 $d(\mathcal{S}_i,\ \mathcal{T})$估计域间关系。Shui 等人（2021年）通过标签比例对领域间的损失和 Wasserstein 距离重新加权，以缓解目标偏移；该标签比例则根据少量目标标签和源标签所提供的统计信息进行估计。

对于无监督多源域适应，现有的大多数方法根据目标偏移可分为两类。边缘分布对齐方法利用各源边缘分布与目标边缘分布之间的不同散度（即，$d(\mathcal{S}_i(X)\ \mathcal{T}(X))\ \forall i \in [N]$）来估计域间关系。例如，Zhao 等人（2018年）和 Li 等人（2018年）使用 $\mathcal{H}$-散度，该方法仅适用于二分类。Wen 等人（2020年）和 Peng 等人（2019年）采用差异。上述进行无监督边缘分布对齐的方法对目标偏移并不稳定（Zhao 等人，2019年）。

Redko 等人（2019年a）和 Shui 等人（2021年）考虑通过联合分布对齐的特殊情形处理目标偏移。JCPOT（Redko 等人，2019年a）假设$\mathcal{S}_i(X|Y) = \mathcal{T}(X|Y)\ \forall i \in [N]$，并使用迭代 Bregman 投影求解受约束的 Wasserstein 重心-

分类问题，通过优化类别比例来纠正目标偏移。然而，上述标签偏移假设通常并不成立，而且与 Courty 等人（2017）类似，该算法在大型数据集上难以计算。Shui 等人（2021）假设广义标签偏移（GLS）条件（Combes 等人，2020）成立，并使用黑箱偏移估计（Lipton 等人，2018）估计标签比例。此外，他们使用 Wasserstein 距离和预测的伪标签，在表示空间上最小化 $d(\mathcal{S}_i(X|Y)\ \mathcal{T}(X|Y))$。Shui 等人（2021）的定理是针对有监督多源域适应提出的，并未对使用伪标签的无监督多源域适应作出正式论证。MOST（Nguyen 等，2021）不优化域间关系，而是简单地学习一个源域判别器，以获得加权集成专家（教师模型）。学生分类器通过使用预测标签最小化伪源域与伪目标域联合分布之间的 Wasserstein 距离，来模仿教师模型。

我们的方法与上述所有方法都不同。我们不进行成对对齐，而是定义一个组合源分布，并直接最小化组合源分布与目标分布之间的 Wasserstein 距离。我们不显式估计标签比例，而是通过联合表示对齐隐式处理目标偏移。此外，我们针对目标域和源域偏移优化任务权重 α，并采用自适应的信息论正则化系数。

3 问题设定

基本符号 除非特别说明，我们用大写字母表示随机变量，用相应的花体字母表示其定义所在的集合，例如，X，Y在 $\mathcal{X}$，$\mathcal{Y}$上。

令 $\mathcal{X}$ 为输入空间。令 $\mathcal{Y}$ 为标签空间，令 $\mathcal{Z}$ def $= \mathcal{X} \times \mathcal{Y}$ 为样本空间。于是，一组 N源分布 $\mathcal{S}_{1:N}$ def$= \{\mathcal{S}_1, \ldots, \mathcal{S}_N\}$和目标分布 $\mathcal{T}$定义在 $\mathcal{Z}$上。现在我们定义一个表示学习函数 $g\colon \mathcal{U} \times \mathcal{X} \to \mathcal{X}$，用于将输入转换为特征表示，其中 g 由定义在 $\mathcal{U}$上的 u 参数化。因此，原始样本空间上的分布在新空间上诱导出一个分布 $\tilde{\mathcal{Z}}$ def$= \tilde{\mathcal{X}} \times \mathcal{Y}$。给定权重单纯形$\Delta_N$ def$= \{\boldsymbol{\alpha}\colon \alpha_i \geq 0, \sum_{i=1}^N \alpha_i = 1\}$，我们考虑源分布混合 $\mathcal{S}^{\boldsymbol{\alpha}}$ def$= \sum_{i=1}^N \alpha_i \mathcal{S}_i$，将其作为组合源分布~。$\mathcal{Z}$ 上的诱导分布由 u 确定，记为 $\tilde{\mathcal{T}}_u$ 和 $\tilde{\mathcal{S}}_u^{\alpha}$ def $= \sum_{i=1}^N \alpha_i \tilde{\mathcal{S}}_{i,u}$。

给定参数化假设空间 $\mathcal{V}$ 后，我们定义预测器 $h\colon \mathcal{V} \times \mathcal{X} \to \mathcal{Y}$，它根据特征表示预测标签。为了衡量预测器的性能，令 $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ 为取正值的损失函数。

因此，$(u,\ v) \in \mathcal{U} \times \mathcal{V}$ 在目标分布和组合源分布上的**总体（真实）风险** 分别定义为

$$
\begin{aligned}
R_{\mathcal{T}}(u,v) &\stackrel{\text{def}}{=} \mathbb{E}_{Z\sim\mathcal{T}}\ell(h(v,g(u,X)),Y)\\
&= \mathbb{E}_{\tilde{Z}\sim\tilde{\mathcal{T}}_u}\ell(h(v,\tilde{X}),Y)\\
R_{\mathcal{S}^{\boldsymbol{\alpha}}}(u,v) &\stackrel{\text{def}}{=} \sum_{i=1}^N \alpha_i \mathbb{E}_{Z\sim\mathcal{S}_i}\ell(h(v,g(u,X)),Y)\\
&= \mathbb{E}_{\tilde{Z}\sim\tilde{\mathcal{S}}_u^{\boldsymbol{\alpha}}}\ell(h(v,\tilde{X}),Y)\,,
\end{aligned}
$$

其中 Z 定义= $(X,\ Y)$ $\tilde{Z}$ 定义= $(\tilde{X},\ Y)$。

令源数据集 $S_{1:N}$ 和目标数据集 T 分别为从相应分布中采样得到的样本，其中 $S_i = \{Z_{i,j}^s\}_{j=1}^{m_i}$，$Z_{i,\ j}^s = (X_{i,\ j}^s,\ Y_{i,\ j}^s) \sim \mathcal{S}_i$，$\forall i \in [N]$以及 $T = \{Z_j^t\}_{j=1}^{m_t}$，$Z_j^t = (X_j^t,\ Y_j^t) \sim \mathcal{T}$。我们进一步将对应于合并源域分布的数据集记为 $S^{\boldsymbol{\alpha}}$。因此，相应的**经验风险**定义如下：

$$
\begin{aligned}
\hat{R}_{\mathcal{T}}(u,v) &\stackrel{\text{def}}{=} \tfrac{1}{m_t}\textstyle\sum_{j=1}^{m_t}\ell(h(v,g(u,X_j^t)),Y_j^t)\\
\hat{R}_{\mathcal{S}^{\boldsymbol{\alpha}}}(u,v) &\stackrel{\text{def}}{=} \textstyle\sum_{i=1}^N \frac{\alpha_i}{m_i}\sum_{j=1}^{m_i}\ell(h(v,g(u,X_{i,j}^s)),Y_{i,j}^s)\,.
\end{aligned}
$$

在无监督领域自适应中，训练阶段无法访问目标标签。因此，若干方法（Xie 等人，2018年；Courty 等人，2017年）将目标输入的预测标签作为伪标签。我们在多源域适应中形式化相应的设定。令 $f_p\colon\ \mathcal{X} \to \mathcal{Y}$ 为一个将目标输入映射到标签空间的伪标签函数。本文中，我们设定 $f_p(x) = h(v, g(u, x))$，侧重于表示学习。因此，$\forall u,\ v \in \mathcal{U} \times \mathcal{V}$，令 $\mathcal{T}_{u,\ v}$ 为由 f_p诱导的 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$~上的伪目标分布，令 $\mathcal{T}_{u,\ v}$ 为 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$上的相应分布~ ~。于是，从伪分布中得到的随机样本记为 $\hat{}Z$ 定义= $(X, \hat{}Y) = (X, f_p(X))$。最后，将无标签目标数据集定义为 $T'_X = \{X_j^t\}_{j=1}^{m'_t}, X_j^t \sim \mathcal{T}(X)$，其中 $\mathcal{T}(X)$表示在 $\mathcal{X}$上边缘化的目标分布 $\mathcal{T}$。

现在，我们介绍推导本文多源域适应界所需的一些常见假设。

假设 *1. (a)* 对于任意$u \in \mathcal{U}$，表示学习函数g：$\mathcal{U} \times \mathcal{X} \to \tilde{\mathcal{X}}$ 相对于度量 $\rho_{\tilde{x}}$ 和 ρ_x是 K-利普希茨的 [2]，即 $\rho_{\tilde{x}}(g(u,x)\ g\ u\ x' \leq (,)) K\rho_x(x,x')$。*(b)* 预测器 h：$\mathcal{V} \times \mathcal{X} \to \mathcal{Y}$~对于任意 $v \in \mathcal{V}$，相对于度量 ρ_y 和 $\rho_{\tilde{x}}$是 L-利普希茨的，即 $\rho_y(h(v,\tilde{x})\ h(v,\tilde{x}')) \leq L\rho_{\tilde{x}}(\tilde{x},\tilde{x}')$。*(c)* 假设损失函数 ℓ：$\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ 是对称的，相对于第一个参数中的 ρ_y 满足 M-利普希茨条件，并满足三角不等式。

[2]利普希茨连续性的一般定义见附录 A.2。

4 有监督多源域适应

我们首先考虑有监督多源域适应情形。在大多数情况下，只有少量带标签的目标数据可用。

4.1 总体风险界

基于最优传输理论（Villani，2009年；Peyr´e 等，2017年），我们利用 Wasserstein 距离（定义 A.6）得到以下总体风险界。

定理4.1. $\forall(u,v)\in\mathcal{U}\times\mathcal{V}$，且 $\forall\boldsymbol{\alpha}\in\Delta_N$，若假设*1*成立，则

$$|R_{\mathcal{T}}(u,v)-R_{\mathcal{S}^{\boldsymbol{\alpha}}}(u,v)|\le\mathbf{W}_1(\tilde{\mathcal{T}}_u,\tilde{\mathcal{S}}_u^{\boldsymbol{\alpha}})\le\mathbf{W}_1(\mathcal{T},\mathcal{S}^{\boldsymbol{\alpha}}),$$

其中，第一个 *Wasserstein* $_1$ 距离定义在度量空间 $(\mathcal{Z},\rho_{\mathcal{Z}})$ 上，对于 $\rho_{\tilde{\mathcal{Z}}}(\tilde{z},\tilde{z}')=\ell(y,y')+LM\rho_{\mathcal{X}}(\tilde{x},\tilde{x}')$；第二个则定义在度量空间 $(\mathcal{Z},\rho_{\mathcal{Z}})$ 上，对于 $\rho_{\mathcal{Z}}(z,z')=\ell(y,y')+LMK\rho_{\mathcal{X}}(x,x')$。

定理4.1说明了学习变换函数 g的重要性，其中$\mathbf{W}_1(\mathcal{T}_u,\mathcal{S}_u^{\boldsymbol{\alpha}})$始终小于$\mathbf{W}_1(\mathcal{T},\mathcal{S}^{\boldsymbol{\alpha}})$。后者是一个固定项，用于刻画给定 $\boldsymbol{\alpha}$时源域组合分布与目标分布之间的偏移。更具体地说，可能不存在一个 $\boldsymbol{\alpha}$ 能使$\mathbf{W}_1(\mathcal{T},\mathcal{S}^{\boldsymbol{\alpha}})=0$。相反，通过优化变换参数 u，我们可能得到$\mathbf{W}_1(\mathcal{T}_u,\mathcal{S}_u^{\boldsymbol{\alpha}})=0$。详细证明见附录C.1。

4.2 泛化界

根据定理4.1，目标总体风险有如下界：$R_{\mathcal{S}^{\boldsymbol{\alpha}},\mathcal{T}}(u,v)+\mathbf{W}_1(\mathcal{T}_u,\mathcal{S}_u^{\boldsymbol{\alpha}})$。估计 Wasserstein 距离需要目标数据。考虑极端情况 $m_t\to+\infty$，最小化该上界的经验估计，效果永远不会优于直接最小化目标经验风险。因此，我们提出通过考虑以下含参数$0\le\epsilon\le1$的**组合风险** 进行学习：

$$R^{\epsilon}_{\mathcal{S}^{\boldsymbol{\alpha}},\mathcal{T}}(u,v)\overset{\text{def}}{=}(1-\epsilon)R_{\mathcal{T}}(u,v)+\epsilon(R_{\mathcal{S}^{\boldsymbol{\alpha}}}(u,v)+\mathbf{W}_1(\tilde{\mathcal{T}}_u,\tilde{\mathcal{S}}_u^{\boldsymbol{\alpha}}))\,.$$

因此，我们有 $R_{\mathcal{T}}(u,v)\le R^{\epsilon}_{\mathcal{S}^{\boldsymbol{\alpha}},\mathcal{T}}(u,v)$。接下来，令 $\hat{}$ 表示经验组合风险 $R^{\epsilon}_{\mathcal{S}^{\boldsymbol{\alpha}},\mathcal{T}}(u,v)=(1-\epsilon)\hat{R}_{\mathcal{T}}(u,v)+\epsilon\hat{R}_{\mathcal{S}^{\boldsymbol{\alpha}}}(u,v)+\epsilon\hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_u,\tilde{\mathcal{S}}_u^{\boldsymbol{\alpha}})$.

我们考虑一种用于监督式多源域适应（MDA）的随机学习算法 $\mathcal{A}$，它以 N 源数据集和目标数据集作为输入，并输出随机参数 $(U,V)=\mathcal{A}(S^{\boldsymbol{\alpha}},T)\sim P_{U,V|S^{\boldsymbol{\alpha}},T}$，其中 $P_{U,V|S^{\boldsymbol{\alpha}},T}$是由 $\mathcal{A}$在给定 (S^{α},T) 的条件下诱导出的 $\mathcal{U}\times\mathcal{V}$上的分布。期望泛化差距定义为

$$\begin{aligned}gen(\mathcal{T},\mathcal{A})&\overset{\text{def}}{=}\mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T}[R_{\mathcal{T}}(U,V)-\hat{R}^{\epsilon}_{\mathcal{S}^{\boldsymbol{\alpha}},\mathcal{T}}(U,V)]\\&\le\mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T}[R^{\epsilon}_{\mathcal{S}^{\boldsymbol{\alpha}},\mathcal{T}}(U,V)-\hat{R}^{\epsilon}_{\mathcal{S}^{\boldsymbol{\alpha}},\mathcal{T}}(U,V)]\,.\end{aligned}$$

为界定 Wasserstein 距离下的泛化差距，传统方法会应用 Wasserstein 距离的三角不等式，然后分别界定源域和目标域的差距（Courty 等人，2017年；Shui 等人，2021年）。这些证明基于先前关于 Wasserstein 距离的集中不等式结果（Weed 和 Bach，2019年），并额外假设损失有界。为获得非平凡的依赖算法的界，我们采用不同的方法处理这一问题。根据 Kantorovich-Rubinstein 对偶（Villani，2009年），有

$$\mathbf{W}_1(\tilde{\mathcal{T}}_u,\tilde{\mathcal{S}}_u^{\boldsymbol{\alpha}})=\sup_{f:\|f\|_{Lip}\le1}\mathbb{E}_{\tilde{z}\sim\tilde{\mathcal{T}}_u}f(\tilde{z})-\mathbb{E}_{\tilde{z}'\sim\tilde{\mathcal{S}}_u^{\boldsymbol{\alpha}}}f(\tilde{z}')\,.$$

上确界取遍所有 1-Lipschitz 函数 $\mathcal{F}\overset{\text{定义为}}{=}\{f\colon\mathcal{Z}\to\mathbb{R}^+,|f(\tilde{z})-f(\tilde{z}')|\le\rho_{\tilde{\mathcal{Z}}}(\tilde{z},\tilde{z}')\}$。由于学习这一函数类具有不可处理性（Arjovsky 等，2017年），我们转而考虑参数化函数类 $\tilde{\mathcal{F}}\overset{\text{定义为}}{=}\{\tilde{f}\colon\tilde{\mathcal{Z}}\times\mathcal{V}'\to\mathbb{R}^+,\ \|\tilde{f}\|_{Lip}\le1\}$，其参数集为 $\mathcal{V}'$。我们假设上述上确界在 $\mathcal{F}\subset\tilde{\mathcal{F}}$处取得。在此假设下，我们有

$$\mathbf{W}_1(\tilde{\mathcal{T}}_u,\tilde{\mathcal{S}}_u^{\boldsymbol{\alpha}})=\sup_{v'\in\mathcal{V}'}\mathbb{E}_{\tilde{z}\sim\tilde{\mathcal{T}}_u}\tilde{f}(v',\tilde{z})-\mathbb{E}_{\tilde{z}'\sim\tilde{\mathcal{S}}_u^{\boldsymbol{\alpha}}}\tilde{f}(v',\tilde{z}')\,.$$

现在我们提出次高斯（定义 A.4）假设，并为有监督多源域适应给出一个互信息（定义 A.5）上界。

假设 *2.* $\forall(u,v)\in\mathcal{U}\times\mathcal{V}$, $\ell(h(v,g(u,X)),Y)$ 关于 $Z\sim\mathcal{T}$ 和 $Z\sim\mathcal{S}_i,\forall i\in[N]$ 是 σ-次高斯的。此外，对于任意满足上确界可达的适当 $\mathcal{F}$ 选择，$\forall(u,v')\in\mathcal{U}\times\mathcal{V}'$, $f(v',(g(u,X),Y))\in\mathcal{F}$ 关于 $Z\sim\mathcal{T}$ 和 $Z\sim\mathcal{S}_i,\forall i\in[N]$ 是 σ'-次高斯的。

定理 4.2. 若假设 *1* 和 *2* 均成立，则对于任意 $\boldsymbol{\alpha}\in\Delta_N$ 和 $0\le\epsilon\le1$，我们可以得到有监督多源领域自适应算法 $\mathcal{A}$ 的泛化差距上界：

$$\begin{aligned}gen(\mathcal{T},\mathcal{A})\le\sigma&\sqrt{2(\frac{(1-\epsilon)^2}{m_t}+\sum_{i=1}^{N}\frac{\epsilon^2\alpha_i^2}{m_i})I(U,V;S^{\boldsymbol{\alpha}},T)}\\&+\sigma'\sqrt{2\epsilon^2(\sum_{i=1}^{N}\frac{\alpha_i^2}{m_i}+\frac{1}{m_t})I(U;S^{\boldsymbol{\alpha}},T)}\,.\end{aligned}$$

上述针对有监督多源域适应的**泛化差距** 上界由两项组成。第一项由输入数据集与输出假设$S^{\boldsymbol{\alpha}},T$之间的互信息 U,V表征。第二项源于对表示空间上 Wasserstein 距离的估计，取决于表示参数 U 与输入数据集之间的互信息。注意，如果我们考虑一个任意随机的表示学习函数（即 U 不依赖任何数据），那么互信息项等于零。然而，经验风险会非常大。定理 4.2 的详细证明见附录 C.2。

利用目标数据 我们首先关注目标数据较少的情况，即也就是说，m_t 较小。通常，源域样本的数量远多于 m_t。因此，当 $\sum_{i=1}^{N} \alpha_i^2/m_i \ll 1/m_t$时，泛化差距主要由目标数据决定，其速率为 $\mathcal{O}(\sqrt{1/m_t})$。在这种情况下，相较于直接使用目标数据进行训练，使用目标数据来近似 Wasserstein 距离（$\epsilon = 1$）更可取，因为在假设 $\sigma = \sigma'$的情况下，$I(U; \mathcal{S}^{\boldsymbol{\alpha}}, T) \leq I(U, V; \mathcal{S}^{\boldsymbol{\alpha}}, T)$。相反，当$1/m_t \ll \sum_{i=1}^{N} \alpha_i^2/m_i$时，我们更倾向于直接使用目标数据进行训练（$\epsilon = 0$）。对于其他情况，存在一个非平凡的 ϵ 取值，可使相对于该算法的泛化差距最小。

5 无监督多源域适应

对于无监督领域自适应，在训练阶段无法获取目标标签。因此，如果没有任何前提条件，就很难保证迁移成功。Ben-David 等人（2010年（a））指出了无监督领域自适应成功的一项必要条件：理想联合误差必须足够小，否则，没有任何分类器能够同时在源域和目标域上表现良好。因此，文献中通常假设理想联合误差较小。基于同样的假设，我们给出了带伪标签的无监督多源域适应定理。

定义 5.1. （理想联合误差）设理想联合假设为 $(u^*, v^*) \triangleq \arg\min_{u,v} R_{\mathcal{S}^{\boldsymbol{\alpha}}}(u, v) + R_{\mathcal{T}}(u, v)$，则理想联合误差为 $R^* = R_{\mathcal{S}^{\boldsymbol{\alpha}}}(u^*, v^*) + R_{\mathcal{T}}(u^*, v^*)$。

5.1 总体风险界

定义 5.2. *(*联合近似误差*)* 令 $\tilde{x} = g(u, x)$ $\tilde{x}' = g(u, x')$ 以及在伪标签函数 $h(v, g(u, x))$ 下 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ 上的最优 耦合为

$$\gamma^*_{u,v} \stackrel{def}{=} \underset{\gamma \in \Pi(\mathcal{T}_{u,v}, \mathcal{S}^{\boldsymbol{\alpha}})}{\arg\min} \int [LM\rho_{\tilde{x}}(\tilde{x}, \tilde{x}') + \ell(\hat{y}, y')] \, d\gamma(\hat{z}, z')$$

则联合近似误差为

$$R^*_{rep}(u, v) \stackrel{def}{=} LM \int_{\mathcal{Z} \times \mathcal{Z}} [\rho_{\tilde{x}}(g(u^*, x), g(u^*, x')) - \rho_{\tilde{x}}(g(u, x), g(u, x'))] \, d\gamma^*_{u,v}(\hat{z}, z') \,,$$

其中 $R^*_{rep}(u, v)$ 刻画了 u^* 的表示空间中域偏移所导致的近似误差（即理想联合假设在表示空间中的对应形式）。

定理 5.1. 对于任意给定的 $\boldsymbol{\alpha} \in \Delta_N$和 $\forall u, v \in \mathcal{U} \times \mathcal{V}$，若假设 *1* 成立，则目标总体风险可以界定为：

$$R_{\mathcal{T}}(u, v) \leq \mathbf{W}_1(\tilde{\mathcal{T}}_{u,v}, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_u) + R^*_{rep}(u, v) + R^* \,.$$

详细证明见附录 D.1。由上述定理可知，若我们最小化 $\mathbf{W}_1(\mathcal{T}_{u,v}, \mathcal{S}^{\boldsymbol{\alpha}}_u)$，$R^*_{rep}(u, v)$ 可能会增大。因此，对于基于联合表示对齐的成功无监督多源域适应，所提出的理论表明，只需对使 $\mathbf{W}_1(\mathcal{T}_{u,v}, \mathcal{S}^{\boldsymbol{\alpha}}_u)$ 最小化的 (u, v)，令 $R^*_{rep}(u, v)$ 足够小即可。

目标偏移 联合对齐方法的一个重要优势是，当存在目标偏移时，相比边缘对齐方法更有利。因为通过联合最优传输，我们可以通过最小化$\mathbf{W}_1(\mathcal{T}_{u,v}, \mathcal{S}^{\boldsymbol{\alpha}}_u)$ 关于 v 来自动纠正目标偏移。而传统边缘对齐方法最小化 $\mathbf{W}_1(\mathcal{S}_u(X), \mathcal{T}_u(X)) + R_{\mathcal{S}^{\boldsymbol{\alpha}}}(u, v)$，却无法纠正这种偏移，因为目标函数中的两项都与源域和目标域标签分布之间的散度无关。

5.2 泛化界

与上述有监督多源域适应方法类似，我们使用随机算法 $\mathcal{A}^{un}$，它以 N 源数据集和无标签目标数据集作为输入，然后输出随机参数 $(U, V) = \mathcal{A}^{un}(S^{\boldsymbol{\alpha}}, T'_X) \sim P_{U,V|S^{\boldsymbol{\alpha}}, T'_X}$。结合定理 5.1，定义无监督多源域适应的期望泛化差距如下：

$$gen(\mathcal{T}, \mathcal{A}^{un}) \stackrel{\text{def}}{=} \mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T'_X}[R_{\mathcal{T}}(U, V) - \hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_{U,V}, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_U)] \,,$$

其中，$\hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_{U,V}, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_U)$ 是$\tilde{\mathbf{W}}_1(\mathcal{T}_{U,V}, \mathcal{S}^{\boldsymbol{\alpha}}_U)$ 的经验估计，其形式因对 $\mathcal{F}$的具体选择而异，详见第 6.1 节。不指定 $\mathcal{F}$ 的定义见附录。现在，考虑以下次高斯假设。

假设 *3.* 对于任何使上确界可达的适当 $\mathcal{F}$ 选择，$\forall (u, v, v') \in \mathcal{U} \times \mathcal{V} \times \mathcal{V}'$，$f(v', (g(u, X) Y)) \in \mathcal{F}$相对于 $Z \sim \mathcal{T}_{u,v}$是 σ'-次高斯的；且 $\forall (u, v') \in \mathcal{U} \times \mathcal{V}'$，$f(v', (g(u, X) Y)) \in \mathcal{F}$相对于 $Z \sim \mathcal{S}_i$，$\forall i \in [N]$是σ'-次高斯的。

然后，我们得到以下关于使用伪标签进行联合表示对齐时泛化差距的依赖算法的界。

定理 5.2。 若假设 *1* 和假设 *3* 成立，则对于任意 $\boldsymbol{\alpha} \in \Delta_N$，无监督多源域适应算法使用伪标签 $\mathcal{A}^{un}$ 时的期望泛化差距满足以下界：

$$gen(\mathcal{T}, \mathcal{A}^{un}) \leq \sqrt{2\sigma'^2 (\sum_{i=1}^{N} \frac{\alpha_i^2}{m_i} + \frac{1}{m'_t}) I(U, V; S^{\boldsymbol{\alpha}}, T'_X)} + \mathbb{E}_{U,V} R^*_{rep}(U, V) + R^*$$

讨论 </style'>上述上界包含两项，$_V R^*_{rep}(U, V)$ 和 R^*，假设它们足够小

从而能够成功进行无监督表示迁移。然后，联合表示对齐算法的泛化差距以上界为 $(S^{\boldsymbol{\alpha}},\ T'_X)$ 与算法输出 $(U,\ V)$ 之间的互信息。这在一定程度上是因为，联合对齐还会通过伪标签预测影响预测器参数 V 的学习。对于无监督多源域适应，我们通常有较大的 m'_t，其中 $m'_t \gg m_i,\ \forall i \in [N]$。因此，样本复杂度由 $\sum_{i=1}^{N} \frac{\alpha_i^2}{m_i}$ 主导。考虑一种特殊情况：所有源域与目标域的相似性都同样强，因此有$\overline{\alpha_i} = 1/N$ 和 $\sum_{i=1}^{N} \frac{\alpha_i^2}{m_i} = \frac{\sum_{i=1}^{N} 1/m_i}{N^2}$。在这种情况下，可以通过增加领域数量来降低泛化差距，~~即 $N \to \infty$~~。定理 5.2 的详细证明见附录 D.2。

6 DeepMDA 统一方法

6.1 信息论多源域自适应（IMDA）

结合有监督与无监督多源域适应 有监督多源域适应算法没有利用无标签目标数据所传递的信息。与此同时，无监督多源域自适应算法要求理想联合误差较小，而这一条件在实践中未必成立。因此，我们考虑结合这两种都采用联合表示对齐的方法。接着，我们通过权重参数 $0 \leq \tau \leq 1$平衡两种方案的置信度，提出一种通用方法。当 τ 既不为 0 也不为 1 时，所提出的算法便成为一种半监督迁移学习算法。因此，我们定义如下经验风险：

$$\hat{R}^{\boldsymbol{\alpha}}_{\epsilon,\tau}(u,v) \stackrel{\text{def}}{=} \tau(1-\epsilon)\hat{R}_{\mathcal{T}}(u,v) + \tau\epsilon\hat{R}_{\mathcal{S}^{\boldsymbol{\alpha}}}(u,v) + \tau\epsilon\hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_u, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_u) + (1-\tau)\hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_{u,v}, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_u)$$

~选择 $\mathcal{F}$由于 $\ell(h(v,\cdot)\ \cdot)\ \forall v \in \mathcal{V}$满足利普希茨条件，其中 $|\ell(h(v,\tilde{x})\ y)\ -\ \ell(h(v,\tilde{x}')\ y')| \leq \rho_{\tilde{z}}(\tilde{z},\tilde{z}')$适用于两个 Wasserstein 距离中的 $\rho_{\tilde{z}}$，因此~ ~我们可以选择 $\mathcal{F} = \{f = \ell(h(v',\cdot)\ \cdot):\ v' \in \mathcal{V}' = \mathcal{V}\}$。正如第 4.2 节所述，我们假设在所有 1-Lipschitz 函数的空间上取的上确界 $\mathcal{F}$可以达到~ ~某个 $f \in \mathcal{F}$。因此，该模型包含两个预测器 $h(v,.)$和 $h(v',.)$，它们使用相同的网络结构（$\mathcal{V}' = \mathcal{V}$），并根据特征表示 $g(u,x)$进行预测。前者相对于 v最小化经验风险 ^ ^$\tau(1-\epsilon)R_{\mathcal{T}}(u,v)\ + \tau\epsilon R_{\mathcal{S}^{\boldsymbol{\alpha}}}(u,v)$。另一个辅助预测器相对于 v'最大化两个 Wasserstein 距离的经验估计。然后，我们相对于 u,v最小化这些经验估计之和。上述方法不同于边际表示对齐方法，在后者中，W_1 距离定义在 $\mathcal{X}$ 上，而相应的 f 是类似 DANN 的方法所采用的域判别器（Wen 等人，2020年；Shui 等人，2021年）。

优化 Wasserstein 距离 </style}使用上述所述的~f 后，伪目标分布的经验风险为：

$$\hat{R}_{\mathcal{T}_{u,v}}(u,v,v') = \frac{1}{m'_t}\sum_{j=1}^{m'_t} \ell(h(v', g(u, X^t_j)), h(v, g(u, X^t_j)))$$

因此，我们可以对两个 Wasserstein 距离作如下近似：

$$\hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_u, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_u) = \max_{v'}[\hat{R}_{\mathcal{T}}(u,v') - \hat{R}_{\mathcal{S}^{\boldsymbol{\alpha}}}(u,v')]$$

$$\hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_{u,v}, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_u) = \max_{v'}[\hat{R}_{\mathcal{T}_{u,v}}(u,v,v') - \hat{R}_{\mathcal{S}^{\boldsymbol{\alpha}}}(u,v')]$$

然后，我们得到以下两个极小极大优化目标 ^ ~ ~ ^ ~ ~$\min_u \mathbf{W}_1(\mathcal{T}_u, \mathcal{S}^{\boldsymbol{\alpha}}_u)$ 和 $\min_{u,v} \mathbf{W}_1(\mathcal{T}_{u,v}, \mathcal{S}^{\boldsymbol{\alpha}}_u)$，可通过梯度反转层实现（Ganin 等人，2016年）。

使用随机梯度算法进行联合优化Pensia等人（2018年）提出了利用梯度更新为含噪迭代算法中的互信息建立界限的方法。这种方法尤其适用于深度学习，因为随机梯度下降（SGD）及其变体常用于深度学习。我们使用随机梯度朗之万动力学（SGLD）算法（Welling和Teh，2011年）来分析上述多源迁移方案。SGLD是随机梯度下降的一种随机变体，每一步都会独立注入噪声。

在每次迭代 k中，从源数据集 $S_{1:N}$中采样一批源数据 $S^k_{B_{1:N}}$。同时，从相应的目标数据集中分别采样一批有标签目标数据 T^k_B 和一批无标签目标数据$T^k_{X_B}$。随后，我们通过以下更新来更新表示参数 U和预测器参数 V：

$$U_k = U_{k-1} - \eta^k_u G^k_u + \xi^k_u; \quad V_k = V_{k-1} - \eta^k_v G^k_v + \xi^k_v,$$

其中，$\xi^u_k \sim N,\ \sigma^2_k I_{d_u}$和 $\xi^v_k \sim N,\ \sigma^2_k I_{d_v}$是注入的方差为 σ^2_k的各向同性高斯噪声。参数η^k_u和 η^k_v分别是每一步中 U和 V的学习率。 G^k_u和 G^k_v分别是使用每次迭代采样的批次数据集对$\nabla_U R^{\boldsymbol{\alpha}}_{\epsilon,\ \tau}(U,\ V)$和$\nabla_V R^{\boldsymbol{\alpha}}_{\epsilon,\ \tau}(U,\ V)$计算得到的梯度估计。请注意，采样策略需要独立于参数此前的迭代值。

6.2 基于梯度范数界的理论保证

现在，我们给出所提出的 IMDA 算法的梯度范数界。

定理 6.1。 根据定理*4.1*、*5.1*、*4.2*、*5.2*，定义权重参数 τ，用于平衡无监督和~有监督多源域适应。采用上述 $\mathcal{F}$ 选择，并使用上文所述的 $\sigma' = \sigma$ 和随机梯度朗之万动力学更新。于是，我们可以得到期望的梯度范数界

目标风险，$\forall \boldsymbol{\alpha} \in \Delta_N, \tau, \epsilon \in [0,1]$，有：

$$\mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T,T'_X} R_{\mathcal{T}}(U,V) \le \mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T,T'_X} \hat{R}^{\boldsymbol{\alpha}}_{\epsilon,\tau}(U,V) + \tau\sigma\sqrt{2(\frac{(1-\epsilon)^2}{m_t} + \epsilon^2 \sum_{i=1}^{N} \frac{\alpha_i^2}{m_i})(\delta_u + \delta_v)} + \tau\epsilon\sigma\sqrt{2(\sum_{i=1}^{N} \frac{\alpha_i^2}{m_i} + \frac{1}{m_t})\delta_u} + (1-\tau)\mathbb{E}_{U,V} R^*_{rep}(U,V) + (1-\tau)(\sigma\sqrt{2(\sum_{i=1}^{N} \frac{\alpha_i^2}{m_i} + \frac{1}{m'_t})(\delta_u + \delta_v)} + R^*)\,,$$

$\delta_u \overset{\text{定义}}{=} \sum_{k=1}^{K} \frac{(\eta_u^k)^2 \mathbb{E}\|G_u^k\|_2^2}{2\sigma_k^2}$ 和 $\delta_v \overset{\text{定义}}{=} \sum_{k=1}^{K} \frac{(\eta_u^k)^2 \mathbb{E}\|G_v^k\|_2^2}{2\sigma_k^2}$ 分别是 U 和 V 的累积梯度范数。

定理 6.1 相当一般。当 $\tau = 1$时，它涵盖有监督多源域适应。当 $\tau = 0$时，它涵盖无监督设置。在其他情况下，它可用于半监督迁移学习。我们可以看到，对于半监督迁移，关于较小 $R^* + \mathbb{E}_{U,V} R^*_{rep}(U, V)$ 的假设有所弱化。当我们有大量有标签目标数据（m_t较大）时，可以选择较大的 τ。此外，如果我们只有一个源域，$\boldsymbol{\alpha}$ 就会变成一个标量并等于 1。因此，该界也适用于单源领域自适应。证明见附录 E.1。

6.3 实现与讨论

优化域权重 $\boldsymbol{\alpha}$ 根据定理 6.1，对于给定的 u、v、v' 和 ϵ，τ，我们可以针对上界中的相关项优化 $\boldsymbol{\alpha}$。因此，我们得到如下带有常数 C_0、$C_1 > 0$的优化问题：

$$\min_{\boldsymbol{\alpha}} \Big((\epsilon\tau + C_0(1-\tau))\hat{R}_{S^{\boldsymbol{\alpha}}}(u,v) - (\epsilon\tau + 1 - \tau)\hat{R}_{S^{\boldsymbol{\alpha}}}(u,v') + C_1((1-\tau+\tau\epsilon)\sqrt{\delta_u + \delta_v} + \tau\epsilon\sqrt{\delta_u}) R(\boldsymbol{\alpha}) \Big)\,,$$

$$R(\boldsymbol{\alpha}) = \sqrt{\sum_{i=1}^{N} \frac{\alpha_i^2}{m_i}},\ \text{s.t.} \forall i \in [N], \alpha_i \ge 0, \sum_{i=1}^{N} \alpha_i = 1\,,$$

该问题可以使用标准的凸优化工具箱求解。我们在每个训练轮次固定 $\boldsymbol{\alpha}$，通过最小化经验风险 $\hat{R}^{\boldsymbol{\alpha}}_{\epsilon,\tau}(u, v)$ 来优化网络参数 u、v、v'。然后，在每个训练轮次结束时，使用上述凸优化目标，基于更新后的 u、v、v' 优化 $\boldsymbol{\alpha}$。对于无监督多源域适应（$\tau = 0$），在优化 $\boldsymbol{\alpha}$时，我们添加 $C_0 R_{S^{\boldsymbol{\alpha}}}(u, v)$，作为$R^*$ 中 $R_{S^{\boldsymbol{\alpha}}}(u^*, v^*)$ 的近似项；对于 u、v的优化目标，该项并不存在。$\boldsymbol{\alpha}$在 δ_u、δ_v 中的其他高阶项将被忽略。

我们强调，上述目标与Shui等人（2021）不同，后者也对域权重进行优化 $\boldsymbol{\alpha}$。正则化项 $R(\boldsymbol{\alpha})$ 前的系数会随梯度步而调整，这有着有趣的解释。随着梯度步数增加，模型参数和域权重 $\boldsymbol{\alpha}$ 都可能对现有数据集过拟合。因此，$R(\boldsymbol{\alpha})$ 的自适应正则化系数可以逐渐控制对特定源分布过拟合的可能性。

复杂度比较 对于内存复杂度，所提出的方法只需要一个 重复预测器。相比之下，WADN（Shui 等人，2021年）需要 N 个域判别器和 $N|\mathcal{Y}|$ 个类别特征质心。MDAN 和 DARN（Wen 等人，2020年）需要 N 个域判别器。M^{3}SDA（Peng 等人，2019年）和 MDMN（Li 等人，2018年）需要 N^2 个域判别器。对于时间复杂度，所提出的方法需要在每个批次计算 $\mathcal{O}(N)$ 个小批量风险及相应的反向梯度，以优化模型参数，这与其他方法相同。此外，每个训练周期后，我们需要以 $\mathcal{O}(N)$ 的时间复杂度优化域权重；与 WADN 的 $\mathcal{O}(N|\mathcal{Y}|)$ 时间复杂度相比，这是一个改进，其中 $|\mathcal{Y}|$ 源自对标签比例的估计。

7 实验

本节中，我们在 Shui 等人（2021）提出的目标偏移多源域适应基准上测试 IMDA（第 6 节）。实验细节和更多结果见附录 F 和 G。

7.1 数据集与基线方法

亚马逊评论数据集 Blitzer 等人（2007）包含来自四个领域的正面和负面评论：“书籍”“DVD”“电子产品”和“厨房用品”。我们依次将每个领域作为目标域，其余三个作为源域。数据按照 Chen 等人（2012）采用的相同方法进行预处理，使用频率最高的 5000 个一元词/二元词作为词袋特征。原始亚马逊评论数据的标签分布均衡。为了创建目标偏移数据，我们遵循 Shui 等人（2021）的策略，随机删除源域中 50% 的负面评论，同时保持目标域不变。

数字数据集 包含四种不同的数字识别任务。MNIST（LeCun，1998年）和 USPS（Hull，1994年）是标准的手写数字数据集。街景门牌号数据集（SVHN）（Netzer 等人，2011年）是一个由谷歌街景门牌号图像组成的数字数据集。SYNTH（Ganin 和 Lempitsky，2015年）是一个对 SVHN 应用各种变换的合成数据集。与 Shui 等人（2021年）一样，我们也将所有源域中数字 5–9 的数据删去 50%，使其目标标签分布

表 1：U 的准确率(%)　目标偏移数据上的无监督多源域适应（丢弃率 50%）：亚马逊评论（左），数字（右）

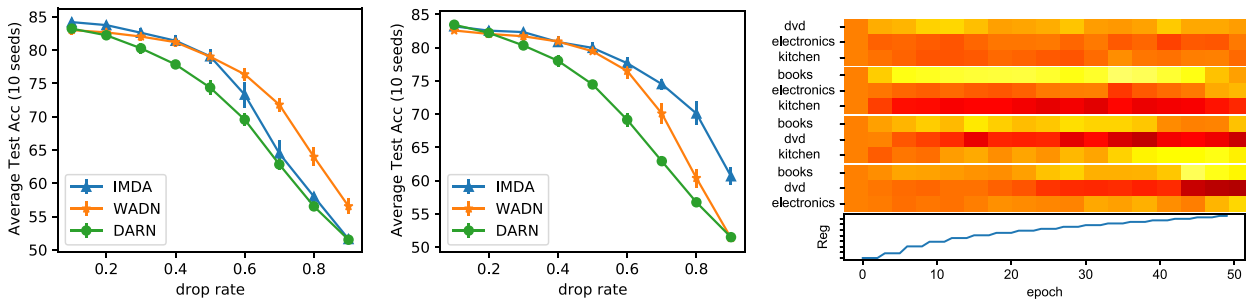
方法	目标域 书籍	DVD	电子产品	厨房用品	平均值
源域	68.15±1.37	69.51±0.74	82.09±0.88	75.30±1.29	73.81
DANN	65.59±1.35	67.23±0.71	80.49±1.11	74.71±1.53	72.00
MDAN	68.77±2.31	67.81±2.46	80.96±0.77	75.67±1.96	73.30
MDMN	70.56±1.05	69.64±0.73	82.71±0.71	77.05±0.78	74.99
M³SDA	69.09±1.26	68.67±1.37	81.34±0.66	76.10±1.47	73.79
DARN	69.88±1.91	69.63±1.09	80.83±1.13	77.47±1.05	74.45
WADN	**75.74**±1.60	**78.46**±1.72	82.10±2.09	81.05±2.25	79.34
IMDA（我们的方法）	75.21±2.3	78.15±1.79	**83.93**±0.24	**81.44**±1.44	**79.68**

方法	目标域 MNIST	SVHN	SYNTH	USPS	平均值
源域	84.93±1.50	67.14±1.40	78.11±1.31	86.02±1.12	79.05
DANN	86.99±1.53	69.56±2.26	78.73±1.30	86.81±1.74	80.52
MDAN	87.86±2.24	69.13±1.56	79.77±1.69	86.50±1.59	80.81
MDMN	87.31±1.88	69.84±1.59	80.27±0.88	86.61±1.41	81.00
M³S领域自适应	87.22±1.70	68.89±1.93	80.01±1.77	86.39±1.68	80.87
DARN	86.96±1.27	68.99±1.76	80.65±1.12	86.89±1.64	80.87
WADN	89.04±0.86	71.68±1.27	**82.06**±0.97	**90.06**±1.12	**83.21**
IMDA（我们的方法）	**89.26**±0.89	**76.6**±0.86	81.76±0.69	84.45±1.2	83.02

表 2：准确率(%) o　f 有监督多源域适应在目标偏移数据上的准确率（丢弃率 50%）：亚马逊评论（左），数字（右）1)

方法	目标域 书籍	DVD	电子产品	厨房用品	平均值
源域 + Tar	72.59±1.89	73.02±1.84	81.59±1.58	77.03±1.73	76.06
DANN	67.35±2.28	66.33±2.42	78.03±1.72	74.31±1.71	71.50
MDAN	68.70±2.99	69.30±2.21	78.78±2.21	74.07±1.89	72.71
MDMN	69.19±2.09	68.71±2.39	81.88±1.46	78.51±1.91	74.57
M³SDA	69.28±1.78	67.40±0.46	76.28±0.81	76.50±1.19	72.36
RLUS	71.83±1.71	69.64±2.39	81.98±1.04	78.69±1.15	75.54
MME	69.66±0.58	71.36±0.96	78.88±1.51	76.64±1.73	74.14
DARN	69.48±2.28	69.59±2.90	80.66±1.38	77.25±0.94	74.25
WADN	74.52±1.45	**77.39**±1.04	81.40±1.61	**81.52**±1.64	78.71
IMDA（我们的方法）	**75.17**±0.94	77.92±1.73	**83.04**±1.34	80.11±1.17	**79.06**

方法	目标域 MNIST	SVHN	SYNTH	USPS	平均值
源域 + Tar	79.63±1.74	56.48±1.90	69.64±1.38	86.29±1.56	73.01
DANN	86.77±1.30	69.13±1.09	78.82±1.35	86.54±1.03	80.32
MDAN	86.93±1.05	68.25±1.53	79.80±1.17	86.23±1.41	80.30
MDMN	77.59±1.36	69.62±1.26	78.93±1.64	87.26±1.13	78.35
M³SDA	85.88±2.06	68.84±1.05	76.29±0.95	87.15±1.10	79.54
RLUS	87.61±1.08	70.50±0.94	79.52±1.30	86.70±1.13	81.08
MME	87.24±0.95	65.20±1.35	80.31±0.60	87.88±0.76	80.16
DARN	86.56±1.48	68.76±1.36	80.4±1.27	86.82±1.09	80.64
WADN	**88.02**±1.27	70.54±1.06	**81.58**±0.96	90.48±1.11	82.66
IMDA（我们的方法）	84.82±1.15	**75.46**±1.2	79.87±1.04	**91.6**±0.96	**82.94**



(a) 有监督多源域适应　(b) 无监督多源域适应　(c) 域权重与正则化

图 1：亚马逊数据集上，不同目标偏移程度下所有领域的平均测试准确率：(a) 少量目标标签条件下的有监督多源域适应；(b) 使用伪标签的无监督多源域适应。(c) 亚马逊数据集上，丢弃率为 0。4 时，有监督多源域适应在不同训练轮次中的域权重和正则化系数可视化。

保持不变，以生成目标偏移 数据。

基线方法 我们首先介绍无监督多源域适应的基线方法。**源域** 使用所有带标签的源域数据集训练预测器，不进行适应。**DANN** （Ganin 等人，2016年）是一种单源领域自适应方法，因此会在合并后的源域数据与目标数据之间进行边缘分布匹配。**MDAN** （Zhao 等人，2018年）、**MDMN** （Li 等人，2018年）、**M^{3}SDA** （Peng 等人，2019年）和**DARN** （Wen 等人，2020年）都是边缘分布匹配方法。**WADN** （Shui 等人，2021年）首先估计标签比例以校正目标偏移，然后对齐条件分布。因此，它也采用了一种（不同的）联合分布匹配方法。对于有监督多源域适应，**源域 + 目标域** 通过合并源数据和少量有标签目标数据进行训练。另有两个基线方法，**RLUS**（Konstantinov 和 Lampert，2019年）以及**MME**（Saito 等，2019年）也纳入了考虑。

7.2 结果与分析

基准数据集上的结果 我们将上述数据集上的实验结果与 Shui 等人（2021）提供的基线方法进行比较。对于有监督多源域适应，我们考虑只能获取少量带标签的目标样本的情况。因此，我们随机抽取 10% 的有标签目标数据作为训练集，其余 90% 作为测试集。对于无监督多源域适应，我们使用上述不带标签的目标数据进行训练，并在另一个未见过的目标数据集上测试结果。我们使用这两篇论文提供的代码复现了 WADN 和 DARN 的结果，并直接采用了其中展示的结果

作为其他基线方法。无监督多源域适应的比较结果见表 1，有监督多源域适应的比较结果见表 2。复现结果和所提出方法（IMDA）的结果均为运行十次的平均值。报告了各方法的均值和标准差，并以粗体标出最佳值。

从表中可以看出，与当前最先进的方法（WADN）相比，对于无监督多源域适应，IMDA 在“电子产品”和“厨房用品”上的表现有所提升，而在“书籍”和“DVD”上的表现略有下降。对于有监督多源域适应，IMDA 在“电子产品”和“书籍”上的表现有所提升。平均准确率相对于 WADN 有 $0.3-0.4\%$ 的提升。对于数字数据集，所提出的方法在“SVHN”上有显著提升，而对于有监督和无监督多源域适应，在“MNIST”或“USPS”上均有下降。就平均准确率而言，与 WADN 相比，我们取得了可比的结果：有监督多源域适应略有提升，而无监督多源域适应略有下降。此外，**源域 + 目标域**的表现逊于所有基于散度的方法，这验证了我们在定理 4.2 中关于利用目标数据的论述。

目标偏移消融研究 我们在亚马逊数据集上进行消融研究，以展示 IMDA 在缓解目标偏移方面的有效性。我们将负面评论的丢弃率从 10% 调整为 -90%。图 1 (a) 和 (b) 分别展示了少量目标标签条件下的有监督多源域适应，以及使用伪标签的无监督多源域适应的结果。

图 1 (b) 表明，IMDA 能够应对无监督多源域适应中的显著目标偏移，并且优于 WADN；此外，WADN 还需要利用伪标签估计标签比例。由于标签比例估计不准确，预测器在显著目标偏移下往往不够准确，进而导致 WADN 的性能下降。

对于有监督设置，仅提供少量已标注的目标样本 。当丢弃率低于 50% 时，IMDA 的表现与 WADN 相当；而随着丢弃率升高，其性能会下降。这一结果与所提出的理论一致。当源域和目标域的联合分布在标签分布方面存在显著差异时，基于真实目标标签利用简单统计量校正目标偏移，比最小化 Wasserstein 距离更有效；后者需要充足的数据。

域关系可视化 我们在图 1(c)的热图中展示了域权重和正则化系数随训练轮次的变化，其中颜色越深表示权重越大。此外，我们还能看到一种避免对特定领域过拟合的自适应效果，例如，在估计“电子产品”时，某些训练轮次中“DVD”的权重被调低了。

8 讨论与结论

局限性 Kumar等（2020）和Wang等（2022）在目标领域持续变化时（渐进式域自适应/连续测试时域自适应）采用自训练进行无监督领域自适应。在这种情形下，误差累积现象往往会出现。由于使用了伪标签，IMDA也面临这一问题。Wang等（2022）提出了两种缓解误差累积的方法：权重平均伪标签和增强平均伪标签。IMDA也可以采用类似的权重平均方法来提高算法的稳定性。

计算机视觉高级任务中的未来应用 语义分割等复杂计算机视觉任务通常需要结构适应。因此，要扩展传统的特征对齐方法，需要针对特征映射进行特定设计。例如，ADVENT（Vu 等，2019）采用加权自信息空间。然而，IMDA 可以通过针对经验风险的极大极小目标函数直接扩展到分割任务，其中结构信息包含在 $\mathcal{Y}$的定义中。

公平性领域的未来应用 近年来，机器学习算法相关的伦理问题日益受到学界关注，尤其是如何避免隐性歧视。由于不公平预测通常与分布偏移有关，因此 Zhao 和 Gordon（2019）以及 Schumann 等（2019）考虑采用领域自适应方法来提升公平性。所提出的方法也可以应用于其他公平学习算法，或与之结合（Hardt 等，2016；Barocas 等，2017；Liu 等，2019；Shui 等人，2022a、c）。

总之，本文对有监督多源域适应和无监督多源域适应中的表示学习进行了信息论分析。我们首先为多源域适应提出了完全依赖于算法且适用于基于深度学习的算法的界限。随后，我们提出了一种新颖的算法（IMDA），以缓解目标偏移，并通过实验证明，该算法在提高内存效率的同时，性能可与以往最先进的方法相媲美。值得注意的是，在更贴近实际的无监督学习场景中，面对显著的目标偏移，所提出的算法优于以往工作。最后，所提出的方法具有基础性，且易于扩展至其他应用，例如复杂的计算机视觉任务和公平性。

致谢

我们感谢匿名审稿人和元审稿人提出的建设性意见。本研究得到加拿大自然科学与工程研究委员会（NSERC）探索基金、SSQ Assurances 与 NSERC 的协作研究与开发基金，以及中国国家留学基金委员会的支持。

参考文献

David Acuna、Guojun Zhang、Marc T Law 和 Sanja Fidler。《f-域对抗学习：理论与算法》。发表于国际机器学习大会，第 66–75 页。PMLR，2021年。

MartinArjovsky、Soumith Chintala 和L´eon Bottou。Wasserstein生成对抗网络。*arXiv*预印本 *arXiv:1701.07875*，2017年。

Solon Barocas、Moritz Hardt 和 ArvindNarayanan。机器学习中的公平性。*NIPS* 教程，1:2017，2017年。

Shai Ben-David、John Blitzer、Koby Crammer 和 Fernando Pereira。领域自适应的表征分析。神经信息处理系统进展，19，2006年。

Shai Ben-David、John Blitzer、Koby Crammer、Alex Kulesza、Fernando Pereira 和 Jennifer Wortman Vaughan。跨不同领域学习理论。《机器学习》，79(1):151–175，2010年（a）。

Shai Ben-David、Tyler Lu、Teresa Luu 和 D´avid P´al。领域自适应的不可能性定理。载于国际人工智能与统计学会议，第 129–136 页，2010年（b）。John Blitzer。自然语言处理系统的领域自适应。博士论文，宾夕法尼亚大学，2008年。John Blitzer、Mark Dredze 和 Fernando Pereira。传记、宝莱坞、音响和搅拌机：面向情感分类的领域自适应。载于计算语言学协会第 *45* 届年会论文集，第 440–447 页，2007年。St´ephane Boucheron、G´abor Lugosi 和 Pascal Massart。集中不等式：独立性的非渐近理论。牛津大学出版社，2013年。Minmin Chen、Zhixiang Xu、Kilian Weinberger 和 Fei Sha。用于领域自适应的边缘化去噪自编码器。*arXiv* 预印本 *arXiv:1206.4683*，2012年。Qi Chen、Changjian Shui 和 Mario Marchand。元学习的泛化界：信息论分析。神经信息处理系统进展，34：25878–25890，2021年。Remi Tachet des Combes、Han Zhao、Yu-Xiang Wang 和 Geoff Gordon。结合条件分布匹配与广义标签偏移的领域自适应。*arXiv* 预印本 *arXiv:2003.04475*，2020年。Nicolas Courty、R´emi Flamary 和 Devis Tuia。结合正则化最优传输的领域自适应。载于数据库中的机器学习与知识发现：第 *14* 届欧洲 *ECML PKDD* 会议，法国南锡，*2014*年*9*月*15*日至*19*日。论文集，第一部分，第 274–289 页。Springer，2014年。

Nicolas Courty、R´emi Flamary、Amaury Habrard 和 Alain Rakotomamonjy。用于领域自适应的联合分布最优传输。神经信息处理系统进展，30，2017年。

Gabriela Csurka。《视觉应用中的领域自适应综述》。计算机视觉应用中的领域自适应，第 1–35 页，2017年。

Bharath Bhushan Damodaran、Benjamin Kellenberger、R´emi Flamary、Devis Tuia 和 Nicolas Courty。Deep-jdot：用于无监督领域自适应的深度联合分布最优传输。载于 欧洲计算机视觉会议（*ECCV*）论文集，第 447–463 页，2018年。

John Duchi。统计学 311/电气工程 377 课程讲义。网址：*https://stanford. edu/class/stats311/Lectures/full notes. pdf*。最后访问时间，2:23，2016年。

Yaroslav Ganin 和 Victor Lempitsky。通过反向传播实现无监督领域自适应。载于国际机器学习会议，第 1180–1189 页。PMLR，2015年。Yaroslav Ganin、Evgeniya Ustinova、Hana Ajakan、Pascal Germain、Hugo Larochelle、Franc¸ois Laviolette、Mario Marchand 和 Victor Lempitsky。神经网络的领域对抗训练。《机器学习研究期刊》，17(1):2096–2030，2016年。Ishaan Gulrajani、Faruk Ahmed、Martin Arjovsky、Vincent Dumoulin 和 Aaron C Courville。改进 Wasserstein 生成对抗网络的训练。神经信息处理系统进展，30，2017年。Moritz Hardt、Eric Price 和 Nati Srebro。监督学习中的机会平等。神经信息处理系统进展，29，2016年。Jiayuan Huang、Arthur Gretton、Karsten Borgwardt、Bern- hard Sch¨olkopf 和 Alex J Smola。利用未标注数据纠正样本选择偏差。载于神经信息处理系统进展，第 601–608 页，2007年。Jonathan J. Hull。用于手写文本识别研究的数据库。*IEEE*模式分析与机器智能汇刊，16(5):550–554，1994年。Sharu Theresa Jose 和 Osvaldo Simeone。元学习的信息论泛化界及其应用。*arXiv* 预印本 *arXiv:2005.04372*，2020年。Nikola Konstantinov 和 Christoph Lampert。从不可信来源进行稳健学习。载于国际机器学习会议，第 3488–3498 页。PMLR，2019年。Ananya Kumar、Tengyu Ma 和 Percy Liang。理解渐进式领域自适应中的自训练。载于国际机器学习会议，第 5468–5479 页。PMLR，2020年。Yann LeCun。手写数字的 MNIST 数据库。*http://yann. lecun. co m/exdb/mnist/*，1998年。

Yitong Li、David E Carlson 等。通过多域匹配提取关系。神经信息处理系统进展，31，2018年。Zachary Lipton、Yu-Xiang Wang 和 Alexander Smola。使用黑箱预测器检测并纠正标签偏移。载于国际机器学习会议，第 3122–3130 页。PMLR，2018年。Lydia T Liu、Max Simchowitz 和 Moritz Hardt。无约束学习的隐式公平性准则。载于国际机器学习会议，第 4051–4060 页。PMLR，2019年。Yishay Mansour、Mehryar Mohri 和 Afshin Ros- tamizadeh。领域自适应：学习界与算法。*arXiv* 预印本 *arXiv:0902.3430*，2009年。Yuval Netzer、Tao Wang、Adam Coates、Alessandro Bis- sacco、Bo Wu 和 Andrew Y Ng。利用无监督特征学习识别自然图像中的数字。2011年。Tuan Nguyen、Trung Le、He Zhao、Quan Hung Tran、Truyen Nguyen 和 Dinh Phung。MOST：通过最优传输实现学生—教师学习的多源领域自适应。载于人工智能中的不确定性，第 225–235 页。PMLR，2021年。Sinno Jialin Pan 和 Qiang Yang。迁移学习综述。*IEEE*知识与数据工程汇刊，22(10):1345–1359，2009年。Xingchao Peng、Qinxun Bai、Xide Xia、Zijun Huang、Kate Saenko 和 Bo Wang。多源领域自适应的矩匹配。载于*IEEE/CVF*国际计算机视觉会议论文集，第 1406–1415 页，2019年。Ankit Pensia、Varun Jog 和 Po-Ling Loh。含噪迭代算法的泛化误差界。载于*2018*年*IEEE*国际信息论研讨会 *(ISIT)*，第 546–550 页。IEEE，2018年。Gabriel Peyr´e、Marco Cuturi 等。计算最优传输。经济与统计研究中心工作论文，(2017-86)，2017年。Ievgen Redko、Nicolas Courty、R´emi Flamary 和 Devis Tuia。目标偏移下多源领域自适应的最优传输。载于第*22*届国际人工智能与统计会议，第 849–858 页。PMLR，2019年a。Ievgen Redko、Emilie Morvant、Amaury Habrard、Marc Sebban 和 Younes Bennani。领域自适应理论进展。Elsevier，2019年b。Daniel Russo 和 James Zou。你的数据探索过拟合了多少？通过信息使用控制偏差。*IEEE*信息论汇刊，66(1):302–323，2019年。Kuniaki Saito、Donghyun Kim、Stan Sclaroff、Trevor Darrell 和 Kate Saenko。通过极小极大熵实现半监督领域自适应。载于论文集

IEEE/CVF 计算机视觉国际会议，第8050–8058页，2019年。Candice Schumann、Xuezhi Wang、Alex Beutel、Jilin Chen、Hai Qian 和 Ed H Chi。跨领域迁移机器学习公平性。*arXiv* 预印本 *arXiv:1906.09688*，2019年。Changjian Shui、Qi Chen、Jun Wen、Fan Zhou、Christian Gagn´e 和 Boyu Wang。超越 H-散度：基于 Jensen-Shannon 散度的领域自适应理论。*arXiv* 预印本 *arXiv:2007.15567*，第6卷，2020年。Changjian Shui、Zijian Li、Jiaqi Li、Christian Gagn´e、Charles X Ling 和 Boyu Wang。从多个目标偏移源域进行聚合。载于国际机器学习会议，第9638–9648页。PMLR，2021年。Changjian Shui、Qi Chen、Jiaqi Li、Boyu Wang 和 Chris- tian Gagn´e。通过隐式路径对齐进行公平表征学习。载于国际机器学习会议，第20156–20175页。PMLR，2022年a。Changjian Shui、Qi Chen、Jun Wen、Fan Zhou、Christian Gagn´e 和 Boyu Wang。基于 Jensen–Shannon 散度的新型领域自适应理论。《知识型系统》，257:109808，2022年b。Changjian Shui、Gezheng Xu、Qi Chen、Jiaqi Li、Charles Ling、Tal Arbel、Boyu Wang 和 Christian Gagn´e。关于在多个子群体上学习公平性与准确率。*arXiv* 预印本 *arXiv:2210.10837*，2022年c。Eric Tzeng、Judy Hoffman、Kate Saenko 和 Trevor Dar- rell。对抗判别式领域自适应。载于*IEEE* 计算机视觉与模式识别会议论文集，第7167–7176页，2017年。C´edric Villani。最优传输：旧与新，第338卷。Springer，2009年。Tuan-Hung Vu、Himalaya Jain、Maxime Bucher、Matthieu Cord 和 Patrick P´erez。ADVENT：用于语义分割领域自适应的对抗熵最小化。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第2517–2526页，2019年。Boyu Wang、Jorge A Mendez、Changjian Shui、Fan Zhou、Gezheng Xu、Christian Gagn´e 和 Eric Eaton。知识共享与迁移的差距最小化。《机器学习研究杂志》，24(33):1–57，2023年。Qin Wang、Olga Fink、Luc Van Gool 和 Dengxin Dai。持续测试时领域自适应。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第7201–7211页，2022年。Jonathan Weed 和 Francis Bach。经验测度在 Wasserstein 距离下收敛的尖锐渐近率与有限样本率。《伯努利》，25(4A):2620–2648，2019年。Max Welling 和 Yee W Teh。通过随机梯度朗之万动力学进行贝叶斯学习。载于论文集

第*28*届国际机器学习会议（*ICML-11*），第681–688页，2011年。Junfeng Wen、Russell Greiner 和 Dale Schuurmans。用于多源领域自适应的领域聚合网络。载于国际机器学习会议，第10214–10224页。PMLR，2020年。Xuetong Wu、Jonathan H Manton、Uwe Aickelin 和 Jingge Zhu。迁移学习的信息论分析。载于*2020 IEEE* 信息论国际研讨会（*ISIT*），第2819–2824页。IEEE，2020年。Xuetong Wu、Jonathan H Manton、Uwe Aickelin 和 Jingge Zhu。迁移学习的信息论分析：误差界与应用。*arXiv* 预印本 *arXiv:2207.05377*，2022年。Shaoan Xie、Zibin Zheng、Liang Chen 和 Chuan Chen。学习用于无监督领域自适应的语义表示。载于国际机器学习会议，第5423–5432页，2018年。Aolin Xu 和 Maxim Raginsky。学习算法泛化能力的信息论分析。载于神经信息处理系统进展，第2524–2533页，2017年。Chao Zhang、Lei Zhang 和 Jieping Ye。领域自适应的泛化界。神经信息处理系统进展，25，2012年。Han Zhao 和 Geoff Gordon。学习公平表示时的内在权衡。神经信息处理系统进展，32，2019年。Han Zhao、Shanghang Zhang、Guanhang Wu、Jos´e MF Moura、Joao P Costeira 和 Geoffrey J Gordon。对抗式多源领域自适应。神经信息处理系统进展，31，2018年。Han Zhao、Remi Tachet Des Combes、Kun Zhang 和 Geoffrey Gordon。关于学习用于领域自适应的不变表示。载于国际机器学习会议，第7523–7532页。PMLR，2019年。

A 定义

定义 A.1.（耦合）。设 $(\mathcal{X}, \mu)$ 和 $(\mathcal{Y}, \nu)$ 是两个概率空间。耦合 μ, ν 意指在某个概率空间 $(\mathcal{Z}, \pi)$上构造两个随机变量 X和 Y，使得 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $(proj_{\mathcal{X}})_{\#}\pi = \mu$和 $(proj_{\mathcal{Y}})_{\#}\pi = \nu$，这意味着 π是定义在 $\mathcal{X} \times \mathcal{Y}$上的联合测度，其在 $\mathcal{X}$和 $\mathcal{Y}$上的边缘分布分别为 μ, ν。称 (X, Y) 为 (μ, ν)的耦合。

定义 *A.2.*（利普希茨连续性）。给定两个度量空间 $(\mathcal{X}, \rho_x)$, $(\mathcal{Y}, \rho_y)$，函数 $f\colon \mathcal{X} \to \mathcal{Y}$关于 $\mathcal{X}$是 L-利普希茨的，当且仅当$\forall x$, $x' \in \mathcal{X}$, $\rho_y(f(x)\ f(x')) \leq L\rho_x(x, x')$。特别地，若 $\mathcal{X} \subset \mathbb{R}^{d_1}, \mathcal{Y} \subset \mathbb{R}^{d_2}$，我们通常取 $\rho_x(x, x') = \|x - x'\|$和$\rho_y(\cdot, \cdot) = \|y - y'\|$，关于范数 $\|\cdot\|$。

定义 A.3。（传输代价不等式）。度量空间 （$\mathcal{X}$， ρ）上的概率测度 μ满足常数为 $c > 0$ 的 L^p 传输不等式，如果对于每个概率 $\nu \ll \mu$，都有：

$$\mathbf{W}_p(\mu, \nu) \leq \sqrt{2cD_{\mathrm{KL}}(\nu \| \mu)}$$

定义 A.4。（次高斯）。将随机变量 X 的累积量生成函数（*CGF*）定义为 $\psi_X(\lambda) \overset{\text{定义为}}{=}$对数$\mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}]$。如果 X 满足 σ- 次高斯，则

$$\psi_X(\lambda) \leq \frac{\lambda^2\sigma^2}{2}, \forall \lambda \in \mathbb{R}\,.$$

定义 A.5。（互信息）。设 X 和 Y为任意随机变量， D_{KL} 表示 *KL* 散度。 X 与 Y之间的互信息定义为：

$$I(X; Y) \overset{def}{=} D_{\mathrm{KL}}(P(X, Y) \| P(X)P(Y))$$

定义 A.6。（*Wasserstein*-p距离）。设两个分布定义在同一个波兰度量空间 （$\mathcal{X}$， ρ）上，其中ρ（$\cdot$，$\cdot$）为度量，而 $p \in [1, +\infty)$， Π（μ， ν）是 μ、 ν的所有耦合（参见定义 *A.1*）所构成的集合。 μ 与 ν 之间的 p 阶 *Wasserstein* 距离定义为：

$$\mathbf{W}_p(\mu, \nu) \overset{def}{=} \inf_{\pi \in \Pi(\mu, \nu)} \left[\int_{\mathcal{X} \times \mathcal{X}} \rho(x, x')^p d\pi(x, x')\right]^{1/p}.$$

如果 $\rho(x,\ x') = \|x - x'\|$ 和 $p = 1$，上述定义即为地球搬运（EM）距离或 Wasserstein-1 距离。

B 技术引理

引理 B.1. ***(Donsker-Varadhan** 表示**)*** 设 P 和 Q 是定义在集合 $\mathcal{X}$上的两个概率测度。设$\mathfrak{g}$： $\mathcal{X} \to \mathbb{R}$ 为可测函数，且 $\mathbb{E}_{x\sim Q}[\exp \mathfrak{g}(x)] \leq \infty$。则有

$$D_{KL}(P||Q) = \sup_{\mathfrak{g}}\{\mathbb{E}_{x\sim P}[\mathfrak{g}(x)] - \log \mathbb{E}_{x\sim Q}[\exp \mathfrak{g}(x)]\}.$$

上述引理的证明见推论 4.15（Boucheron 等，2013年）或定理4.1（Duchi，20 16).

引理 B.2. 设 X_1、 X_2、 $...X_n$ 为相互独立的 σ_i-次高斯随机变量。则

$$\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right)\right] \leq \exp\left(\frac{\lambda^2\sum_{i=1}^{n}\sigma_i^2}{2}\right), \forall\lambda \in \mathbb{R}$$

即， $\sum_{i=1}^{n} X_i$ 是 $\sqrt{\sum_{i=1}^{n}\sigma_i^2}$-次高斯的。

证明。 X_i， $\forall i = \{1, 2, ..., n\}$ 是 σ_i-次高斯的，由定义 A.4 可得 $\mathbb{E}[\exp(\lambda(X_i - \mathbb{E}X_i))] \leq \exp(\frac{\lambda^2\sigma_i^2}{2})$。由于这些变量相互独立，我们有

$$\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right)\right] = \prod_{i=1}^{n}\mathbb{E}[\exp(\lambda(X_i - \mathbb{E}X_i))] \leq \prod_{i=1}^{n}\exp(\frac{\lambda^2\sigma_i^2}{2}) = \exp\left(\frac{\lambda^2\sum_{i=1}^{n}\sigma_i^2}{2}\right)$$

□

引理 B.3. 对于相互独立的变量 X 和 Y，有如下熵不等式：

$$\max\{H(X), H(Y)\} \leq H(X+Y) \leq H(X,Y) = H(X) + H(Y)$$

证明。设 $Z = X + Y$，根据互信息的定义，我们可以得到一个众所周知的结论：“条件化使熵减小”。 $I(Z; X) = H(Z) - H(Z|X) \geq 0$，那么 $H(Z) \geq H(Z|X)$。类似地，我们有 $H(Z) \geq H(Z|Y)$。 X, Y相互独立，因此我们可以得到 $H(Z|X) = H(Y|X) = H(Y)$ 和 $H(Z|Y) = H(X|Y) = H(Y)$。因此，我们证明了 $H(X+Y) \geq \max\{H(X)\ H(Y)\}$。

为证明右侧，我们首先需要证明，随机变量函数的熵小于这些随机变量的联合熵。记 $Z = f(X,\ Y) = X + Y$，我们有

$$I((X,Y); Z) = H(Z) - H(Z|(X,Y)) = H(Z) = H(X,Y) - H((X,Y)|Z),$$

其中，上述等式是根据 $H(Z|(X,\ Y)) = 0$ 这一事实得到的。

根据熵的定义，我们有 $H((X,Y)|Z) \geq 0$。结合 X, Y 的独立性，可得$H(Z) \leq H(X,Y) = H(X) + H(Y)$。

□

C 第 4 节缺失证明

C.1 定理 4.1 的证明

定理。 $\forall(u,v) \in \mathcal{U} \times \mathcal{V}$，且 $\forall \boldsymbol{\alpha} \in \Delta_N$，若满足假设 *1*，则

$$|R_{\mathcal{T}}(u,v) - R_{\mathcal{S}^{\boldsymbol{\alpha}}}(u,v)| \leq \mathbf{W}_1(\tilde{\mathcal{T}}_u, \tilde{\mathcal{S}}_u^{\boldsymbol{\alpha}}) \leq \mathbf{W}_1(\mathcal{T}, \mathcal{S}^{\boldsymbol{\alpha}}),$$

其中，第一个 W_1 距离定义在度量空间 $(\mathcal{Z}, \rho_{\tilde{z}})$ 上，对于 $\rho_{\tilde{z}}(\tilde{z}, \tilde{z}') = \ell(y, y') + LM\rho_{\tilde{x}}(\tilde{x}, \tilde{x}')$；第二个距离定义在度量空间 $(\mathcal{Z}, \rho_z)$ 上，对于 $\rho_z(z, z') = \ell(y, y') + LMK\rho_x(x, x')$。

证明。对于任意 $\pi \in \Pi(\mathcal{T}_u, \mathcal{S}_u^{\alpha})$，我们有

$$\begin{aligned}
|R_{\mathcal{T}}(u,v) - R_{\mathcal{S}^{\boldsymbol{\alpha}}}(u,v)| &= |\mathbb{E}_{\tilde{Z}\sim\tilde{\mathcal{T}}_u}\ell(h(v,\tilde{X}),Y) - \mathbb{E}_{\tilde{Z}'\sim\tilde{\mathcal{S}}_u^{\boldsymbol{\alpha}}}\ell(h(v,\tilde{X}'),Y')| \\
&= |\int_{\tilde{\mathcal{Z}}\times\tilde{\mathcal{Z}}} \ell(h(v,\tilde{x}),y) - \ell(h(v,\tilde{x}'),y')d\pi(\tilde{z},\tilde{z}')| \\
&\leq \int_{\tilde{Z}\times\tilde{Z}} |\ell(h(v,\tilde{x}),y) - \ell(h(v,\tilde{x}'),y')|d\pi(\tilde{z},\tilde{z}') \\
&= \int_{\tilde{Z}\times\tilde{Z}} |\ell(h(v,\tilde{x}),y) - \ell(h(v,\tilde{x}),y') + \ell(h(v,\tilde{x}),y') - \ell(h(v,\tilde{x}'),y')|d\pi(\tilde{z},\tilde{z}') \\
&\leq \int_{\tilde{Z}\times\tilde{Z}} |\ell(h(v,\tilde{x}),y) - \ell(h(v,\tilde{x}),y')| + |\ell(h(v,\tilde{x}),y') - \ell(h(v,\tilde{x}'),y')|d\pi(\tilde{z},\tilde{z}') \\
&\leq \int_{\tilde{Z}\times\tilde{Z}} \left(\ell(y,y') + M\rho_y(h(v,\tilde{x}),h(v,\tilde{x}'))\right) d\pi(\tilde{z},\tilde{z}') \\
&\leq \int_{\tilde{Z}\times\tilde{Z}} \left(\ell(y,y') + LM\rho_{\tilde{x}}(\tilde{x},\tilde{x}')\right) d\pi(\tilde{z},\tilde{z}')
\end{aligned}$$

前两个等式由总体风险的定义和关于耦合的定义 A.1 得出。第一个不等式是对绝对值函数应用詹森不等式得到的。倒数第二个不等式则利用损失函数的三角不等式以及假设 1 中的利普希茨连续性得出。

设 $\mathcal{Z}$ 上的度量为 $\rho_{\tilde{z}}(\tilde{z}, \tilde{z}') = \ell(y, y') + LM\rho_{\tilde{x}}(\tilde{x}, \tilde{x}')$。

$$\begin{aligned}
\int_{\tilde{\mathcal{Z}}\times\tilde{\mathcal{Z}}} \rho_{\tilde{z}}(\tilde{z},\tilde{z}')d\pi(\tilde{z},\tilde{z}') &= \int_{\mathcal{Z}\times\mathcal{Z}} \left(\ell(y,y') + LM\rho_{\tilde{x}}(\tilde{x},\tilde{x}')\right) d\pi(\tilde{z},\tilde{z}') \\
&= \int_{\mathcal{Z}\times\mathcal{Z}} \left(\ell(y,y') + LM\rho_{\tilde{x}}(g(u,x),g(u,x'))\right) d\gamma(z,z') \\
&\leq \int_{\mathcal{Z}\times\mathcal{Z}} \left(\ell(y,y') + LMK\rho_x(x,x')\right) d\gamma(z,z')
\end{aligned}$$

设 $\mathcal{Z}$ 上的度量为 $\rho_z(z, z') = \ell(y, y') + LMK\rho_x(x, x')$。
设表示空间上的最优耦合为 $\pi_u^* = \arg\min_{\tilde{\pi}\in\Pi(\tilde{\mathcal{T}}_u,\tilde{\mathcal{S}}_u^{\alpha})} \int_{\tilde{\mathcal{Z}}\times\tilde{\mathcal{Z}}} \rho_{\tilde{z}}(\tilde{z},\tilde{z}')d\tilde{\pi}(\tilde{z},\tilde{z}')$，相应的$\pi\in\Pi(\mathcal{T}_u,\mathcal{S}_u^{\alpha})$原始样本空间上的耦合为 γ_u^*。

设原始样本空间上的最优耦合为 $\gamma^* = \arg\min_{\gamma\in\Pi(\mathcal{T},\mathcal{S}^{\alpha})} \int_{\mathcal{Z}\times\mathcal{Z}} \rho_z(z,z')d\gamma(z,z')$，相应的表示空间上的耦合为 π_u。
因此我们有：

$$\begin{aligned}
|R_{\mathcal{T}}(u,v) - R_{\mathcal{S}^{\boldsymbol{\alpha}}}(u,v)| &\leq \int_{\tilde{\mathcal{Z}}\times\tilde{\mathcal{Z}}} \rho_{\tilde{z}}(\tilde{z},\tilde{z}')d\pi_u^*(\tilde{z},\tilde{z}') = \mathbf{W}_1(\tilde{\mathcal{T}}_u, \tilde{\mathcal{S}}_u^{\boldsymbol{\alpha}}) \\
&\leq \int_{\tilde{\mathcal{Z}}\times\tilde{\mathcal{Z}}} \rho_{\tilde{z}}(\tilde{z},\tilde{z}')d\pi_u(\tilde{z},\tilde{z}') \leq \int_{\mathcal{Z}\times\mathcal{Z}} \rho_z(z,z')d\gamma^*(z,z') = \mathbf{W}_1(\mathcal{T}, \mathcal{S}^{\boldsymbol{\alpha}})
\end{aligned}$$

最后一个不等式成立，是因为$\pi_u \neq \pi_u^*$ 可能不是表示空间上的最优耦合。

□

C.2 定理4.2的证明

定理。 如果假设 *1* 和假设 *2* 均成立，那么对于任意 $\boldsymbol{\alpha} \in \Delta_N$ 和 $0 \leq \epsilon \leq 1$，我们可以界定有监督多源领域自适应算法 $\mathcal{A}$ 的泛化差距：

$$gen(\mathcal{T}, \mathcal{A}) \leq \sigma \sqrt{2(\frac{(1-\epsilon)^2}{m_t} + \sum_{i=1}^{N} \frac{\epsilon^2 \alpha_i^2}{m_i}) I(U,V; S^{\boldsymbol{\alpha}}, T)} + \sigma' \sqrt{2\epsilon^2 (\sum_{i=1}^{N} \frac{\alpha_i^2}{m_i} + \frac{1}{m_t}) I(U; S^{\boldsymbol{\alpha}}, T)}\,.$$

证明。

$$\begin{aligned} gen(\mathcal{T}, \mathcal{A}) &\leq \mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T}[R^{\epsilon}_{\mathcal{S}^{\boldsymbol{\alpha}},\mathcal{T}}(U,V) - \hat{R}^{\epsilon}_{\mathcal{S}^{\boldsymbol{\alpha}},\mathcal{T}}(U,V)] \\ &= \mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T}[(1-\epsilon)(R_{\mathcal{T}}(U,V) - \hat{R}_{\mathcal{T}}(U,V)) + \epsilon(R_{\mathcal{S}^{\boldsymbol{\alpha}}}(U,V) - \hat{R}_{\mathcal{S}^{\boldsymbol{\alpha}}}(U,V))] \\ &\quad + \epsilon \mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T}[\mathbf{W}_1(\tilde{\mathcal{T}}_U, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_U) - \hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_U, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_U)] \end{aligned}$$

步骤 1：界定经验风险

$$B_r = \mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T}\left[(1-\epsilon)[R_{\mathcal{T}}(U,V) - \hat{R}_{\mathcal{T}}(U,V)] + \epsilon[R_{\mathcal{S}^{\boldsymbol{\alpha}}}(U,V) - \hat{R}_{\mathcal{S}^{\boldsymbol{\alpha}}}(U,V)]\right]$$

$(U,V) \in \mathcal{U} \times \mathcal{V}$是有监督多源域适应算法的输出随机变量，取决于输入数据集 $S_{1:N}$，

$\tilde{}$ $\tilde{}$ $\tilde{}$ T和一个固定的 $\boldsymbol{\alpha}$（非随机）。令 $(\tilde{U},\tilde{V})$ 为 (U,V) 的一个独立副本，使得 $(\tilde{U},\tilde{V}) \perp\!\!\!\perp (S_{1:N}, T)$。接下来定义

$$\begin{aligned} f(U,V,S^{\boldsymbol{\alpha}},T) &\stackrel{\text{def}}{=} (1-\epsilon)\hat{R}_{\mathcal{T}}(U,V) + \epsilon \hat{R}_{S^{\boldsymbol{\alpha}}}(U,V) \\ &= (1-\epsilon)\frac{1}{m_t}\sum_{j=1}^{m_t} \ell(h(V, g(U, X_j^t)), Y_j^t) + \epsilon \sum_{i=1}^{N} \frac{\alpha_i}{m_i} \sum_{j=1}^{m_i} \ell(h(V, g(U, X_{i,j}^s)), Y_{i,j}^s) \end{aligned}$$

$\tilde{}$ $\tilde{}$由于损失函数是 σ-次高斯的（假设 2），应用引理 B.2，可得 $f(\tilde{U},\tilde{V},S^{\boldsymbol{\alpha}},T)$ 是$\sqrt{(1-\epsilon)^2/m_t + \epsilon^2 \sum_{i=1}^{N} \alpha_i^2/m_i}\,\sigma$-次高斯的。

$$\begin{aligned} I(U,V;S^{\boldsymbol{\alpha}},T) &= D_{\mathrm{KL}}(P_{U,V,S^{\boldsymbol{\alpha}},T} \| P_{U,V} P_{S^{\boldsymbol{\alpha}},T}) \\ &= \sup_{\mathfrak{g}} \{\mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T}\mathfrak{g}(U,V,S^{\boldsymbol{\alpha}},T) - \log \mathbb{E}_{\tilde{U},\tilde{V},S^{\boldsymbol{\alpha}},T}[\exp^{\mathfrak{g}(\tilde{U},\tilde{V},S^{\boldsymbol{\alpha}},T)}]\} \\ &\geq \lambda \mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T}[f(U,V,S^{\boldsymbol{\alpha}},T)] - \lambda \mathbb{E}_{\tilde{U},\tilde{V},S^{\boldsymbol{\alpha}},T}[f(\tilde{U},\tilde{V},S^{\boldsymbol{\alpha}},T)] - \psi_{\tilde{U},\tilde{V},S^{\boldsymbol{\alpha}},T}(\lambda), \forall \lambda \in \mathbb{R} \\ &\geq \lambda \mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T}[f(U,V,S^{\boldsymbol{\alpha}},T)] - \lambda \mathbb{E}_{\tilde{U},\tilde{V},S^{\boldsymbol{\alpha}},T}[f(\tilde{U},\tilde{V},S^{\boldsymbol{\alpha}},T)] - \frac{\lambda^2\sigma^2(\frac{(1-\epsilon)^2}{m_t} + \epsilon^2 \sum_{i=1}^{N} \frac{\alpha_i^2}{m_i})}{2} \end{aligned}$$

我们有：

$$\begin{aligned} \mathbb{E}_{\tilde{U},\tilde{V},S^{\boldsymbol{\alpha}},T}[f(\tilde{U},\tilde{V},S^{\boldsymbol{\alpha}},T)] &= \epsilon \mathbb{E}_{\tilde{U},\tilde{V},S^{\boldsymbol{\alpha}}} \hat{R}_{\mathcal{S}^{\boldsymbol{\alpha}}}(\tilde{U},\tilde{V}) + (1-\epsilon)\mathbb{E}_{\tilde{U},\tilde{V},T} \hat{R}_{\mathcal{T}}(\tilde{U},\tilde{V}) \\ &= \epsilon \mathbb{E}_{\tilde{U},} {}^{\wp R_{S^{\boldsymbol{\alpha}}}(} \tilde{U},\tilde{V}) + (1-\epsilon)\mathbb{E}_{\tilde{U}\,,} {}^{\wp R_{\mathcal{T}}(} \tilde{U},\tilde{V}) \\ &= \epsilon \mathbb{E}_{U,\ V,\ S^{\boldsymbol{\alpha}},\ T} R_{S^{\boldsymbol{\alpha}}}(U,\ V) + (1-\epsilon)\mathbb{E}_{U,\ V,\ S^{\boldsymbol{\alpha}},\ T} R_{\mathcal{T}}(U,\ V) \end{aligned}$$

将上述项代入不等式，可得 $-\lambda B_r - \frac{\lambda^2\sigma^2(\frac{(1-\epsilon)^2}{m_t} + \epsilon^2 \sum_{i=1}^{N} \frac{\alpha_i^2}{m_i})}{2} \leq I(U,\ V;\ S^{\boldsymbol{\alpha}},\ T)$

因此，我们有 $|B_r| \leq \sqrt{2(\sigma^2(\frac{(1-\epsilon)^2}{m_t} + \epsilon^2 \sum_{i=1}^{N} \frac{\alpha_i^2}{m_i}))I(U,\ V;\ S^{\boldsymbol{\alpha}},\ T)}$。

步骤 2：界定 Wasserstein 距离

$$B_w = \mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T}\left[\epsilon \mathbf{W}_1(\tilde{\mathcal{T}}_U, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_U) - \epsilon \hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_U, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_U)\right] = \mathbb{E}_{U,S^{\boldsymbol{\alpha}},T}\left[\epsilon \mathbf{W}_1(\tilde{\mathcal{T}}_U, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_U) - \epsilon \hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_U, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_U)\right]$$

Let

$$f'(U, S^{\boldsymbol{\alpha}}, T) = \hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_U, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_U) = \sup_{v'} \left[\frac{1}{m_t}\sum_{j=1}^{m_t} \tilde{f}(v', (g(U, X_j^t), Y_j^t)) - \sum_{i=1}^{N} \frac{\alpha_i}{m_i} \sum_{j=1}^{m_i} \tilde{f}(v', (g(U, X_{i,j}^s), Y_{i,j}^s))\right],$$

$$
\begin{aligned}
I(U; S^{\boldsymbol{\alpha}}, T) &= D_{\mathrm{KL}}(P_{U,S^{\boldsymbol{\alpha}},T} \| P_U P_{S^{\boldsymbol{\alpha}},T}) \\
&= \sup_{\mathfrak{g}} \{\mathbb{E}_{U,S^{\boldsymbol{\alpha}},T} \mathfrak{g}(U, S^{\boldsymbol{\alpha}}, T) - \log \mathbb{E}_{\tilde{U},S^{\boldsymbol{\alpha}},T}[\exp^{\mathfrak{g}(\tilde{U},S^{\boldsymbol{\alpha}},T)}]\} \\
&\geq \lambda \mathbb{E}_{U,S^{\boldsymbol{\alpha}},T}[f'(U, S^{\boldsymbol{\alpha}}, T)] - \lambda \mathbb{E}_{\tilde{U},S^{\boldsymbol{\alpha}},T}[f'(\tilde{U}, S^{\boldsymbol{\alpha}}, T)] - \psi_{\tilde{U},S^{\boldsymbol{\alpha}},T}(\lambda), \forall \lambda \in \mathbb{R}
\end{aligned}
$$

假设 $\mathcal{V}'$ 是不影响上确界的子集。因此，我们有：

$$
\begin{aligned}
\mathbb{E}_{\tilde{U},S^{\boldsymbol{\alpha}},T}[f'(\tilde{U}, S^{\boldsymbol{\alpha}}, T)] &= \mathbb{E}_{\tilde{U}} \sup_{v'} \mathbb{E}_{S^{\boldsymbol{\alpha}},T} \left[\frac{1}{m_t} \sum_{j=1}^{m_t} \tilde{f}(v', (g(\tilde{U}, X_j^t), Y_j^t)) - \sum_{i=1}^{N} \frac{\alpha_i}{m_i} \sum_{j=1}^{m_i} \tilde{f}(v', (g(\tilde{U}, X_{i,j}^s), Y_{i,j}^s)) \right] \\
&= \mathbb{E}_{\tilde{U}} \mathbf{W}_1(\tilde{\mathcal{T}}_{\tilde{U}}, \tilde{\mathcal{S}}_{\tilde{U}}^{\boldsymbol{\alpha}}) = \mathbb{E}_{U,S^{\boldsymbol{\alpha}},T} \mathbf{W}_1(\tilde{\mathcal{T}}_U, \tilde{\mathcal{S}}_U^{\boldsymbol{\alpha}})
\end{aligned}
$$

˜根据假设 2， f 对于任意 u 和 v' 都是 σ'-次高斯的，因此上确界不会影响其次高斯性。应用引理 B.2，可知 ˆ$\mathbf{W}_1(\tilde{\mathcal{T}}_u, \tilde{\mathcal{S}}_u^{\boldsymbol{\alpha}})$ 是 $\sqrt{\frac{1}{m_t} + \sum_{i=1}^{N} \frac{\alpha_i^2}{m_i}} \sigma'$-次高斯的。 —

采用与界定经验风险相同的证明过程，我们得到：

$$
|B_w| \leq \sqrt{2\sigma'^2 \epsilon^2 (\frac{1}{m_t} + \sum_{i=1}^{N} \frac{\alpha_i^2}{m_i}) I(U; S^{\boldsymbol{\alpha}}, T)}
$$

最后，我们有：

$$
\begin{aligned}
gen(\mathcal{T}, \mathcal{A}) &\leq B_r + B_w \\
&\leq \sqrt{2(\sigma^2 (\frac{(1-\epsilon)^2}{m_t} + \epsilon^2 \sum_{i=1}^{N} \frac{\alpha_i^2}{m_i})) I(U, V; S^{\boldsymbol{\alpha}}, T)} + \sqrt{2\sigma'^2 \epsilon^2 (\frac{1}{m_t} + \sum_{i=1}^{N} \frac{\alpha_i^2}{m_i}) I(U; S^{\boldsymbol{\alpha}}, T)}
\end{aligned}
$$

证明完毕。

□

D 第 5 节缺失的证明

D.1 定理 5.1 的证明

定义 D.1。 *(联合近似误差)* 令 $\tilde{x} = g(u, x)$ $\tilde{x}' = g(u, x')$，则令基于伪标签函数 $h(v, g(u, x))$ 在 $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ 上的最优 耦合为

$$\gamma^*_{u,v} \overset{def}{=} \underset{\gamma \in \Pi(\mathcal{T}_{u,v}, \mathcal{S}^{\boldsymbol{\alpha}})}{\arg\min} \int_{\mathcal{Z}\times\mathcal{Z}} [LM\rho_{\tilde{x}}(\tilde{x}, \tilde{x}') + \ell(\hat{y}, y')]\, d\gamma(\hat{z}, z')$$

那么，联合近似误差为

$$R^*_{rep}(u, v) \overset{def}{=} LM \int_{\mathcal{Z}\times\mathcal{Z}} [\rho_{\tilde{x}}(g(u^*, x), g(u^*, x')) - \rho_{\tilde{x}}(g(u, x), g(u, x'))]\, d\gamma^*_{u,v}(\hat{z}, z')\,,$$

其中 $R^*_{rep}(u, v)$ 刻画了 u^* 的表示空间中域偏移的近似误差（即理想联合假设在表示空间中的对应形式）。

定理。 对于任意给定的 $\boldsymbol{\alpha} \in \Delta_N$ 和 $\forall u, v \in \mathcal{U} \times \mathcal{V}$，若满足假设 *1*，则目标总体风险可界定为:

$$R_{\mathcal{T}}(u, v) \le \mathbf{W}_1(\tilde{\mathcal{T}}_{u,v}, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_u) + R^*_{rep}(u, v) + R^*\,.$$

证明。 令 $u^*, v^* = \arg\min_{u,v} R_{\mathcal{S}^{\boldsymbol{\alpha}}}(u, v) + R_{\mathcal{T}}(u, v)$ 为理想联合假设，则有:

$$\begin{aligned}
R_{\mathcal{T}}(u, v) &= \mathbb{E}_{Z\sim\mathcal{T}} \ell(h(v, g(u, X)), Y) \\
&\le \mathbb{E}_{Z\sim\mathcal{T}} \ell(h(v, g(u, X)), h(v^*, g(u^*, X))) + \mathbb{E}_{Z\sim\mathcal{T}} \ell(h(v^*, g(u^*, X)), Y) \\
&= \mathbb{E}_{\hat{Z}\sim\mathcal{T}_{u,v}} \ell(h(v^*, g(u^*, X)), \hat{Y}) + R_{\mathcal{T}}(u^*, v^*) \\
&= \mathbb{E}_{Z\sim\mathcal{T}_{u,v}} \ell(h(v^*, g(u^*, X)), Y) - \mathbb{E}_{Z'\sim\mathcal{S}^{\boldsymbol{\alpha}}} \ell(h(v^*, g(u^*, X')), Y') \\
&\quad + \mathbb{E}_{Z'\sim\mathcal{S}^{\boldsymbol{\alpha}}} \ell(h(v^*, g(u^*, X')), Y') + R_{\mathcal{T}}(u^*, v^*) \\
&\le |\mathbb{E}_{\hat{Z}\sim\mathcal{T}_{u,v}} \ell(h(v^*, g(u^*, X)), \hat{Y}) - \mathbb{E}_{Z'\sim\mathcal{S}^{\boldsymbol{\alpha}}} \ell(h(v^*, g(u^*, X')), Y')| \\
&\quad + R_{\mathcal{S}^{\boldsymbol{\alpha}}}(u^*, v^*) + R_{\mathcal{T}}(u^*, v^*)\,.
\end{aligned}$$

对于表示空间上的对齐，我们有 $\forall \gamma \in \Pi(\mathcal{T}_{u,v}, \mathcal{S}^{\boldsymbol{\alpha}})$:

$$\begin{aligned}
&|\mathbb{E}_{Z\sim\mathcal{T}_{u,v}} \ell(h(v^*, g(u^*, X)), \hat{Y}) - \mathbb{E}_{Z'\sim\mathcal{S}^{\boldsymbol{\alpha}}} \ell(h(v^*, g(u^*, X')), Y')| \\
&\quad \le \int_{\mathcal{Z}\times\mathcal{Z}} |\ell(h(v^*, g(u^*, x)), \hat{y}) - \ell(h(v^*, g(u^*, x')), y')| d\gamma(\hat{z}, z') \\
&\quad \le \int_{\mathcal{Z}\times\mathcal{Z}} |\ell(h(v^*, g(u^*, x)), y') - \ell(h(v^*, g(u^*, x')), y')| \\
&\qquad\quad + |\ell(h(v^*, g(u^*, x)), \hat{y}) - \ell(h(v^*, g(u^*, x)), y')| d\gamma(\hat{z}, z') \\
&\quad \le \int_{\mathcal{Z}\times\mathcal{Z}} LM\rho_{\tilde{x}}(g(u^*, x), g(u^*, x')) + \ell(\hat{y}, y') d\gamma(\hat{z}, z') \\
&\quad = \int_{\mathcal{Z}\times\mathcal{Z}} LM\rho_{\tilde{x}}(g(u^*, x), g(u^*, x')) - LM\rho_{\tilde{x}}(g(u, x), g(u, x')) d\gamma(\hat{z}, z') \\
&\qquad + \int_{\mathcal{Z}\times\mathcal{Z}} LM\rho_{\tilde{x}}(g(u, x), g(u, x')) + \ell(\hat{y}, y') d\gamma(\hat{z}, z')
\end{aligned}$$

设 $\tilde{\mathcal{Z}}$ 上的度量为 $\rho_{\tilde{z}}(\tilde{z}, \tilde{z}') = LM\rho_{\tilde{x}}(\tilde{x}, \tilde{x}') + \ell(y, y')$。设变换后的联合空间 $\tilde{\mathcal{Z}}$ 上的最优耦合为 $\pi^*_{u,v} = \arg\min_{\tilde{\pi}} \int_{\tilde{\mathcal{Z}}\times\tilde{\mathcal{Z}}} \rho_{\tilde{z}}(\tilde{z}, \tilde{z}') d\pi(\tilde{z}, \tilde{z}')$，则原始样本空间上的对应耦合为: $\pi \in \Pi(\mathcal{T}_{u,v}, \mathcal{S}^{\boldsymbol{\alpha}}_u)$

$$\gamma^*_{u,v} = \underset{\gamma \in \Pi(\mathcal{T}_{u,v}, \mathcal{S}^{\boldsymbol{\alpha}})}{\arg\min} \int_{\mathcal{Z}\times\mathcal{Z}} LM\rho_{\tilde{x}}(g(u, x), g(u, x')) + \ell(h(v, g(u, x')), y') d\gamma(\hat{z}, z')\,.$$

因此，

$$R_{\mathcal{T}}(u,v) \le \mathbf{W}_1(\tilde{\mathcal{T}}_{u,v}, \tilde{\mathcal{S}}_u^{\boldsymbol{\alpha}}) + R^*_{rep}(u,v) + R^* .$$

其中 $R^*_{rep}(u,\ v) = \int_{\mathcal{Z}\times\mathcal{Z}} LM\rho_{\tilde{x}}(g(u^*,\ x)\ g(u^*,\ x')) - LM\rho_{\tilde{x}}(g(u,\ x)\ g(u,\ x'))d\gamma^*_{u,\ v}(\hat{}z,\ z')$ 是联合近似误差（定义 D.1），该误差由表示学习函数 $g(u,\ x)$ 和最优耦合$\gamma^*_{u,\ v}(\hat{}z,\ z')$ 控制。

□

D.2 定理 5.2 的证明

定理。 如果假设 *1* 和假设 *3* 成立，则对于任意 $\boldsymbol{\alpha} \in \Delta_N$，带伪标签 $\mathcal{A}^{un}$ 的无监督多源域适应算法的期望泛化差距满足以下界:

$$gen(\mathcal{T}, \mathcal{A}^{un}) \le \sqrt{2\sigma'^2(\sum_{i=1}^N \frac{\alpha_i^2}{m_i} + \frac{1}{m'_t})I(U,V;S^{\boldsymbol{\alpha}},T'_X)} + \mathbb{E}_{U,V} R^*_{rep}(U,V) + R^*$$

证明。 根据定义:

$$gen(\mathcal{T}, \mathcal{A}^{un}) \stackrel{\text{def}}{=} \mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T'_X}[R_{\mathcal{T}}(U,V) - \hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_{U,V}, \tilde{\mathcal{S}}_U^{\boldsymbol{\alpha}})]$$

结合定理 5.1，可得

$$gen(\mathcal{T}, \mathcal{A}^{un}) \le \mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T'_X}[\mathbf{W}_1(\tilde{\mathcal{T}}_{U,V}, \tilde{\mathcal{S}}_U^{\boldsymbol{\alpha}}) - \hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_{U,V}, \tilde{\mathcal{S}}_U^{\boldsymbol{\alpha}})] + \mathbb{E}_{U,V} R^*_{rep}(U,V) + R^*$$

Let

$$\begin{aligned} f'(U,V,S^{\boldsymbol{\alpha}},T'_X) &= \hat{W}_1(\tilde{\mathcal{T}}_{U,V}, \tilde{\mathcal{S}}_U^{\boldsymbol{\alpha}}) \\ &= \sup_{v'} \left[\frac{1}{m'_t}\sum_{j=1}^{m'_t} \tilde{f}(v', (g(U,X_j^t), h(V,g(U,X_j^t)))) - \sum_{i=1}^N \frac{\alpha_i}{m_i}\sum_{j=1}^{m_i} \tilde{f}(v', (g(U,X_{i,j}^s), Y_{i,j}^s)) \right] \end{aligned}$$

$$\begin{aligned} I(U,V;S^{\boldsymbol{\alpha}},T'_X) &= D_{\mathrm{KL}}(P_{U,V,S^{\boldsymbol{\alpha}},T} \| P_{U,V}P_{S^{\boldsymbol{\alpha}},T'_X}) \\ &= \sup_{\mathfrak{g}} \{\mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T'_X}\mathfrak{g}(U,V,S^{\boldsymbol{\alpha}},T'_X) - \log \mathbb{E}_{\tilde{U},\tilde{V},S^{\boldsymbol{\alpha}},T'_X}[\exp^{\mathfrak{g}(\tilde{U},\tilde{V},S^{\boldsymbol{\alpha}},T'_X)}]\} \\ &\ge \lambda \mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T'_X}[f(U,V,S^{\boldsymbol{\alpha}},T'_X)] - \lambda \mathbb{E}_{\tilde{U},\tilde{V},S^{\boldsymbol{\alpha}},T'_X}[f(\tilde{U},\tilde{V},S^{\boldsymbol{\alpha}},T'_X)] - \psi_{\tilde{U},\tilde{V},S^{\boldsymbol{\alpha}},T'_X}(\lambda), \forall \lambda \in \mathbb{R} \end{aligned}$$

$\mathcal{V}'$ 被假设为不影响上确界的集合，因此我们有:

$$\begin{aligned} \mathbb{E}_{\tilde{U},\tilde{V},S^{\boldsymbol{\alpha}},T'_X}[f'(\tilde{U},\tilde{V},S^{\boldsymbol{\alpha}},T'_X)] &= \mathbb{E}_{\tilde{U},\tilde{V}} \sup_{v'} \mathbb{E}_{\hat{Z}\sim\mathcal{T}_{\tilde{U},\tilde{V}}} \tilde{f}(v', (g(\tilde{U},X),\hat{Y})) - \mathbb{E}_{Z'\sim\mathcal{S}^{\boldsymbol{\alpha}}} \tilde{f}(v', (g(\tilde{U},X'),Y')) \\ &= \mathbb{E}_{\tilde{U},\tilde{V}} \mathbf{W}_1(\tilde{\mathcal{T}}_{\tilde{U},\tilde{V}}, \tilde{\mathcal{S}}_{\tilde{U}}^{\boldsymbol{\alpha}}) = \mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T'_X} \mathbf{W}_1(\tilde{\mathcal{T}}_{U,V}, \tilde{\mathcal{S}}_U^{\boldsymbol{\alpha}}) \end{aligned}$$

根据假设 3， $\tilde{f}$在$\hat{}Z \sim \mathcal{T}_{u,v}$下对任意 u,v,v' 均为 σ'-次高斯；在$Z \sim \mathcal{S}_i, \forall i \in [N]$下对任意 u,v' 均为 σ'-次高斯。应用引理 B.2，可得$\hat{}$ $W_1(\tilde{\mathcal{T}}_{U,V}, \tilde{\mathcal{S}}_U^{\boldsymbol{\alpha}})$为 $\sqrt{\frac{1}{m'_t} + \sum_{i=1}^N \frac{\alpha_i^2}{m_i}}\sigma'$-次高斯。因此，我们可以得到$-\lambda \mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T'_X}\left(\mathbf{W}_1(\tilde{\mathcal{T}}_{U,V}, \tilde{\mathcal{S}}_U^{\boldsymbol{\alpha}}) - \hat{}\mathbf{W}_1(\tilde{\mathcal{T}}_{U,V}, \tilde{\mathcal{S}}_U^{\boldsymbol{\alpha}})\right) - \lambda^2\sigma'^2(\frac{1}{m'_t} + \sum_{i=1}^N \frac{\alpha_i^2}{m_i})/2 \le I(U,V;S^{\boldsymbol{\alpha}},T'_X)$。

因此，我们有:

$$|\mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T'_X}\left(\mathbf{W}_1(\tilde{\mathcal{T}}_{U,V}, \tilde{\mathcal{S}}_U^{\boldsymbol{\alpha}}) - \hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_{U,V}, \tilde{\mathcal{S}}_U^{\boldsymbol{\alpha}})\right)| \le \sqrt{2\sigma'^2(\frac{1}{m'_t} + \sum_{i=1}^N \frac{\alpha_i^2}{m_i})I(U,V;S^{\boldsymbol{\alpha}},T'_X)}$$

因此，

$$\begin{aligned} gen(\mathcal{T}, \mathcal{A}^{un}) &\le \mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T'_X}[\mathbf{W}_1(\tilde{\mathcal{T}}_{U,V}, \tilde{\mathcal{S}}_U^{\boldsymbol{\alpha}}) - \hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_{U,V}, \tilde{\mathcal{S}}_U^{\boldsymbol{\alpha}})] + \mathbb{E}_{U,V} R^*_{rep}(U,V) + R^* \\ &\le \sqrt{2\sigma'^2(\sum_{i=1}^N \frac{\alpha_i^2}{m_i} + \frac{1}{m'_t})I(U,V;S^{\boldsymbol{\alpha}},T'_X)} + \mathbb{E}_{U,V} R^*_{rep}(U,V) + R^* \end{aligned}$$

证明结束。 □

E 第6节中遗漏的证明

E.1 定理6.1的证明

定理。 根据定理*4.1*、*5.1*、*4.2*和*5.2*，定义权重参数 τ，以平衡无监督和有监督的~多源域适应。采用上述 $\mathcal{F}$ 的选择，其中 $\sigma' = \sigma$，并使用上文所述的随机梯度朗之万动力学更新。于是，我们可以得到期望目标风险的梯度范数界：$\forall \boldsymbol{\alpha} \in \Delta_N$、$\tau$、$\epsilon \in [0,1]$，有：

$$\begin{aligned}\mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T,T'_X} R_{\mathcal{T}}(U,V) \leq \mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T,T'_X} \hat{R}^{\boldsymbol{\alpha}}_{\epsilon,\tau}(U,V) + \tau\sigma\sqrt{2(\frac{(1-\epsilon)^2}{m_t} + \epsilon^2\sum_{i=1}^N \frac{\alpha_i^2}{m_i})(\delta_u+\delta_v)} \\ + \tau\epsilon\sigma\sqrt{2(\sum_{i=1}^N \frac{\alpha_i^2}{m_i} + \frac{1}{m_t})\delta_u} + (1-\tau)\mathbb{E}_{U,V} R^*_{rep}(U,V) \\ + (1-\tau)(\sigma\sqrt{2(\sum_{i=1}^N \frac{\alpha_i^2}{m_i} + \frac{1}{m'_t})(\delta_u+\delta_v)} + R^*)\,,\end{aligned}$$

$\delta_u \overset{\text{定义}}{=} \sum_{k=1}^K \frac{(\eta_u^k)^2 \mathbb{E}\|G_u^k\|_2^2}{2\sigma_k^2}$ 和 $\delta_v \overset{\text{定义}}{=} \sum_{k=1}^K \frac{(\eta_v^k)^2 \mathbb{E}\|G_v^k\|_2^2}{2\sigma_k^2}$ 分别是 U 和 V 的累积梯度范数。

证明。 结合定理4.1、4.2、5.1和5.2，我们得到：

$$\begin{aligned}\mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T,T_X} R_{\mathcal{T}}(U,V) \leq \mathbb{E}_{U,V,S^{\boldsymbol{\alpha}},T,T_X} \hat{R}^{\boldsymbol{\alpha}}_{\epsilon,\tau}(U,V) + \tau\sqrt{2\epsilon^2\sigma^2(\sum_{i=1}^N \frac{\alpha_i^2}{m_i} + \frac{1}{m_t})I(U;S^{\boldsymbol{\alpha}},T)} \\ + \tau\sqrt{2\sigma^2(\frac{(1-\epsilon)^2}{m_t} + \epsilon^2\sum_{i=1}^N \frac{\alpha_i^2}{m_i})I(U,V;S^{\boldsymbol{\alpha}},T)} + (1-\tau)(R^* + \mathbb{E}_{U,V} R^*_{rep}(U,V)) \\ + (1-\tau)\sqrt{2\sigma^2(\sum_{i=1}^N \frac{\alpha_i^2}{m_i} + \frac{1}{m'_t})I(U,V;S^{\boldsymbol{\alpha}},T'_X)}\,,\end{aligned}$$

我们使用随机梯度朗之万动力学来优化经验风险：

$$\begin{aligned}\hat{R}^{\boldsymbol{\alpha}}_{\epsilon,\tau}(u,v) \overset{\text{def}}{=} \tau(1-\epsilon)\hat{R}_{\mathcal{T}}(u,v) + \tau\epsilon\hat{R}_{\mathcal{S}^{\boldsymbol{\alpha}}}(u,v) \\ + \tau\epsilon\hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_u, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_u)) + (1-\tau)\hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_{u,v}, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_u)\end{aligned}$$

根据 Wasserstein 距离的定义，我们联合最小化目标函数（关于 u，v）并最大化（关于 v'），因此，在给定前面步骤更新后的固定 v' 的情况下，我们分析 U，V。给定批量大小 $|B|$，令经验风险的批量估计为：

$\hat{R}_{T_B^k}(U,V) = \frac{1}{|B|}\sum_{j=1}^{|B|} \ell(h(V, g(U, X_j^t)), Y_j^t)$

$\hat{R}^{\boldsymbol{\alpha}}_{S^k_{B_{1:N}}}(U,V) = \sum_{i=1}^N \frac{\alpha_i}{|B|}\sum_{j=1}^{|B|} \ell(h(V,(g(U, X_{i,j}^s)), Y_{i,j}^s))$

$\hat{R}_{T^k_{X_B}}(U,V,v') = \frac{1}{|B|}\sum_{j=1}^{|B|} \ell(h(v', g(U, X_j^t)), h(V, g(U, X_j^t)))$

$\hat{\mathbf{W}}_1^k(\tilde{\mathcal{T}}_U, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_U)) = \hat{R}_{T_B^k}(U,v') - \hat{R}^{\boldsymbol{\alpha}}_{S^k_{B_{1:N}}}(U,v')$

$\hat{\mathbf{W}}_1^k(\tilde{\mathcal{T}}_{U,V}, \tilde{\mathcal{S}}^{\boldsymbol{\alpha}}_U) = \hat{R}_{T^k_{X_B}}(U,V,v') - \hat{R}^{\boldsymbol{\alpha}}_{S^k_{B_{1:N}}}(U,v')$

更新 U 时，注意三个数据批次对应的梯度如下：

$G_u(U_{k-1}, V_{k-1}, T_B^k) = \tau(1-\epsilon)\nabla_U \hat{R}_{T_B^k}(U_{k-1}, V_{k-1}) + \tau\epsilon\nabla_U \hat{R}_{T_B^k}(U,v')$

$G_u(U_{k-1},\ V_{k-1},\ T^k_{X_B}) = (1-\tau)\nabla_U \hat{R}_{T^k_{X_B}}(U_{k-1},\ V_{k-1},\ v')$

$G_u(U_{k-1}, V_{k-1}, S^k_{B_{1:N}}) = \tau\epsilon\nabla_U \hat{R}^{\boldsymbol{\alpha}}_{S^k_{B_{1:N}}}(U_{k-1}, V_{k-1}) - (1-\tau+\epsilon\tau)\nabla_U \hat{R}^{\boldsymbol{\alpha}}_{S^k_{B_{1:N}}}(U_{k-1}, v')$

同样，对于更新 V，我们有:

$G_v(U_{k-1}, V_{k-1}, T^k_B) = \tau(1-\epsilon)\nabla_V \hat{R}_{T^k_B}(U_{k-1}, V_{k-1})$

$G_v(U_{k-1}, V_{k-1}, T^k_{X_B}) = (1-\tau)\nabla_V \hat{R}_{T^k_{X_B}}(U_{k-1}, V_{k-1}, v')$

$G_v(U_{k-1}, V_{k-1}, S^k_{B_{1:N}}) = \tau\epsilon\nabla_V \hat{R}^{\boldsymbol{\alpha}}_{S^k_{B_{1:N}}}(U_{k-1}, V_{k-1})$

分别将 U 和 V 的整体梯度更新记为:

$$G^k_u = G_u(U_{k-1}, V_{k-1}, T^k_B) + G_u(U_{k-1}, V_{k-1}, T^k_{X_B}) + G_u(U_{k-1}, V_{k-1}, S^k_{B_{1:N}})$$

$$G^k_v = G_v(U_{k-1}, V_{k-1}, T^k_B) + G_v(U_{k-1}, V_{k-1}, T^k_{X_B}) + G_v(U_{k-1}, V_{k-1}, S^k_{B_{1:N}})$$

那么，注入噪声后的更新为:

$$\begin{aligned} U_k &= U_{k-1} - \eta^u_k G^k_u + \xi^u_k \\ V_k &= V_{k-1} - \eta^v_k G^k_v + \xi^v_k \,, \end{aligned}$$

其中 $\xi^u_k \sim N(0, \sigma^2_k I_{d_u})$ 和 $\xi^v_k \sim N(0, \sigma^2_k I_{d_v})$。

定义经过 K 次迭代的表示参数与预测器参数序列为 $(U,V)^{[K]} = (U^{[K]}, V^{[K]}) = ((U_1,V_1),(U_2,V_2)...,(U_K,V_K))$。

每个数据集对应的采样序列定义为 $T^{[K]}_B = (T^1_B,...,T^K_B)$、$S^{[K]}_{B_{1:N}} = (S^1_{B_{1:N}},...,S^K_{B_{1:N}})$ 和 $T^{[K]}_{X_B} = (T^1_{X_B},...,T^K_{X_B})$。

联合优化算法的最终输出 $(U,V) = f((U,V)^{[K]})$。

$$\begin{array}{ccccc} & & T \to T^{[K]}_B & & \\ & & \downarrow & & \\ S_{1:N} \to & S^{[K]}_{B_{1:N}} & \to (U,V)^{[K]} & \to (U,V) \\ & & \uparrow & & \\ & & T'_X \to T^{[K]}_{X_B} & & \end{array}$$

给定特定的 $\boldsymbol{\alpha}$，应用数据处理不等式和链式法则，可得

$$I(U,V;S^{\boldsymbol{\alpha}},T) = I(U,V;S_{1:N},T) \le I((U^{[K]},V^{[K]});S^{[K]}_{B_{1:N}},T^{[K]}_B) = \sum_{k=1}^{K} I(U_k,V_k;S^{[K]}_{B_{1:N}},T^{[K]}_B|U^{[k-1]},V^{[k-1]})$$

此外，由于采样策略与参数的先前迭代值及先前的采样无关，因此可得:

$$\begin{aligned} I(U_k,V_k;S^{[K]}_{B_{1:N}},T^{[K]}_B|U^{[k-1]},V^{[k-1]}) &= I(U_k,V_k;S^k_{B_{1:N}},T^k_B|U^{k-1},V^{k-1}) \\ &= H(U_k,V_k|U_{k-1},V_{k-1}) - H(U_k,V_k|U_{k-1},V_{k-1},S^k_{B_{1:N}},T^k_B) \end{aligned}$$

对于任意的 $(U_{k-1},V_{k-1}) = (u_{k-1},v_{k-1})$，由于一组变量的联合熵小于或等于该组中各变量个体熵之和，因此我们有

$$H(U_k,V_k|U_{k-1}=u_{k-1},V_{k-1}=v_{k-1}) \le H(-\eta^u_k G^k_u + \xi^k_u) + H(-\eta^v_k G^k_v + \xi^k_v))$$

应用引理 B.3，可得

$$\begin{aligned}
H(U_k, V_k | U_{k-1} = u_{k-1}, V_{k-1} = v_{k-1}, & S^k_{B_{1:N}}, T^k_B) \\
&= H(-\eta^u_k G_u(u_{k-1}, v_{k-1}, T^k_{X_B}) + \xi^k_u, -\eta^v_k G_v(u_{k-1}, v_{k-1}, T^k_{X_B}) + \xi^k_v) \\
&\geq H(-\eta^u_k G_u(u_{k-1}, v_{k-1}, T^k_{X_B}) - \eta^v_k G_v(u_{k-1}, v_{k-1}, T^k_{X_B}) + \xi^k_u + \xi^k_v) \\
&\geq H(\xi^k_u) + H(\xi^k_v) = \frac{d_u}{2}\log(2\pi e\sigma_k^2) + \frac{d_v}{2}\log(2\pi e\sigma_k^2)\,.
\end{aligned}$$

第一个不等式成立，是因为随机变量的函数的熵小于随机变量本身的熵（证明见引理 B.3）。第二个不等式则源于梯度与注入的噪声相互独立。

此外，利用独立性，我们可以得到 $\mathbb{E}(\|-\eta^u_k G^k_u + \xi^k_u\|_2^2) = \mathbb{E}(\|\eta^k_u G^k_u\|_2^2 + \|\xi^k_u\|_2^2) = (\eta^k_u)^2\mathbb{E}\|G^k_u\|_2^2 + d_u\sigma_k^2$ 以及 $\mathbb{E}(\|-\eta^v_k G^k_v + \xi^k_v\|_2^2) = \mathbb{E}(\|\eta^k_v G^k_v\|_2^2 + \|\xi^k_v\|_2^2) = (\eta^k_v)^2\mathbb{E}\|G^k_v\|_2^2 + d_v\sigma_k^2$。由于在具有相同二阶矩的变量中，高斯分布的熵最大，因此我们还可以进一步得到：

$$\begin{aligned}
H(U_k, V_k | U_{k-1} = u_{k-1}, V_{k-1} = v_{k-1}) - & H(U_k, V_k | U_{k-1} = u_{k-1}, V_{k-1} = v_{k-1}, S^K_{B_{1:N}}, T^k_B) \\
&\leq H(-\eta^u_k G^k_u + \xi^k_u) + H(-\eta^v_k G^k_v + \xi^k_v)) - (H(\xi^k_u) + H(\xi^k_v)) \\
&\leq \frac{d_u}{2}\log(2\pi e\frac{(\eta^k_u)^2\mathbb{E}\|G^k_u\|_2^2 + d_u\sigma_k^2}{d_u}) + \frac{d_v}{2}\log(2\pi e\frac{(\eta^k_v)^2\mathbb{E}\|G^k_v\|_2^2 + d_v\sigma_k^2}{d_v}) \\
&\quad - \frac{d_u}{2}\log(2\pi e\sigma_k^2) - \frac{d_v}{2}\log(2\pi e\sigma_k^2) \\
&\leq \frac{(\eta^k_u)^2\mathbb{E}\|G^k_u\|_2^2 + (\eta^k_v)^2\mathbb{E}\|G^k_v\|^2}{2\sigma_k^2}
\end{aligned}$$

上述界对所有 u_{k-1}, v_{k-1} 都成立，因此我们可以对该界积分，得到

$$H(U_k, V_k | U_{k-1}, V_{k-1}) - H(U_k, V_k | U_{k-1}, V_{k-1}, S^K_{B_{1:N}}, T^k_B) \leq \frac{(\eta^k_u)^2\mathbb{E}\|G^k_u\|_2^2 + (\eta^k_v)^2\mathbb{E}\|G^k_v\|^2}{2\sigma_k^2}\,,$$

因此得到 $I(U, V; S^{\boldsymbol{\alpha}}, T) \leq \sum_{k=1}^K \frac{(\eta^k_u)^2\mathbb{E}\|G^k_u\|_2^2 + (\eta^k_v)^2\mathbb{E}\|G^k_v\|^2}{2\sigma_k^2} = \delta_u + \delta_v$

同理，采用相同的证明过程可以得到以下结果：

$$I(U; S^{\boldsymbol{\alpha}}, T) \leq \sum_{k=1}^K \frac{(\eta^k_u)^2\mathbb{E}\|G^k_u\|_2^2}{2\sigma_k^2} = \delta_u$$

$$I(U, V; S^{\boldsymbol{\alpha}}, T'_X) \leq \sum_{k=1}^K \frac{(\eta^k_u)^2\mathbb{E}\|G^k_u\|_2^2 + (\eta^k_v)^2\mathbb{E}\|G^k_v\|^2}{2\sigma_k^2} = \delta_u + \delta_v$$

将上述内容放置 将这些项代入本节开头给出的不等式，即可得出证明。 . □

F 实验详情

本节提供实验细节和一些补充结果。所提出的 IMDA 算法的伪代码 [3] 见 H 节。

F.1 亚马逊评论

如主论文所述，原始数据集按照 Chen 等人（2012年）的方法预处理为 5000 维词袋特征。目标偏移数据则通过随机丢弃 50% 的负面评论生成。

模型结构 表征学习器： $[5000, 1000, 500, 100]$ 使用丢弃率为 0.7 的多层感知机网络，并在每个隐藏层后添加 ReLU 激活函数，最终输出 100 维特征表示。
预测器和重复预测器： $[100, 2]$ 线性变换后接一个对数 Softmax 层，将 100 维特征转换为 2 类对数概率。

损失函数：我们选择“负对数似然损失”作为损失函数。

计算 **资源** 实验在一台配备 6 个中央处理器、1 个 32GB 内存图形处理器的服务器上运行。 .

实验设置 亚马逊评论数据集包含 6465 个“书籍”样本、5586 个“DVD”样本、7681 个“电子产品”样本和 7945 个“厨房用品”样本。我们从每个领域中随机抽取 2000 个样本。在无监督多源域适应中，训练阶段使用这 2000 个没有标签的目标样本，其余样本用作测试集。对于只有少量目标标签的有监督多源域适应，我们从 2000 个目标样本中随机抽取 10% 作为训练数据，其余 90% 用作测试集。

我们提供了有监督和无监督设置的主要超参数，这些参数根据交叉验证选定。其他超参数设为代码中提供的默认值。

在有监督多源域适应中，我们将L2 缩放系数$= 0.5$ （ $\boldsymbol{\alpha}$的凸优化目标函数中的 C_1 常数）设为$= 0.5$，Adadelta 优化器的学习率设为 $\eta = 0.5$。$\mathbf{W}_1$ 距离的惩罚项设为 W1 sup coef$= 0.01$。最后，网络训练 40 轮，批量大小为 20。

在主论文所述的无监督多源域适应中，伪目标分布的经验风险为：$\hat{R}_{\mathcal{T}_u,v}(u, v, v') = \frac{1}{m'_t} \sum_{j=1}^{m'_t} \ell(h(v', g(u, X_j^t)) \; h(v, g(u, X_j^t))$. 上述损失无法直接实现，因此我们通过 $\ell(h(v', g(U, X_j^t)), \hat{Y}_j^t) + \ell(h(V, g(U, X_j^t))), \hat{Y}_j'^t)$ 来实现，其中 $\hat{Y}_j^t$和 $\hat{Y}_j'^t$是相应预测器的预测标签。随后，我们分别为两项损失应用不同的惩罚权重 W1 discri coef1 和 W1 discri coef2。我们将 L2 缩放系数设为$= 1$，学习率设为 $\eta = 0.8$，W1 discri coef1$= 0.06$设为 C_0=，W1 discri coef2$= 1.2$。最后，网络训练 50 轮，小批量大小为 20。

为复现 WADN（Shui 等人，2021年）和 DARN（Wen 等人，2020年）的结果，我们使用了代码仓库：https://github.com/cjshui/WADN 和 https://github.com/junfengwen/DARN。超参数设置与各自论文中的说明一致。

F.2 数字

模型结构 表征学习器 由 3 个堆叠模块组成：’卷积层’：[3, 3, 64], ’ReLU’，’最大池- 化：’[2,2,0]；’卷积层’：[3, 3, 128], ’ReLU’，’最大池化：’[2,2,0]；’卷积层’：[3, 3, 256], ’ReLU’，’最大池化：’[2,2,0]。
预测器和重复预测器： $[2304, 512, 100, 10]$ 每个隐藏层后接一个使用“ReLU”激活函数的多层感知机网络，随后连接一个对数 Softmax 层。预测器将 2304 维特征转换为 10 类对数概率。没有使用随机失活。

损失函数：我们选择“负对数似然损失”作为损失函数。

计算 **资源** 实验在一台配备 6 个中央处理器和 1 个 32GB 内存图形处理器的服务器上运行 .

[3] 代码已发布于 https://github.com/livreQ/IMDA

实验设置 数字数据集包含四个领域：“MNIST”、“USPS”、“SVHN”和“SYNTH”。“MNIST”和“USPS”是非常相近的手写数据集。“SYNTH”由“SVHN”转换而来，因此两者也更为相似。下载时，每个领域采用不同的训练-测试划分。各领域的训练样本量分别为 60000、7219、73257、479400，测试样本量分别为 10000、2017、26032、9553。按照 Shui 等人（2021）的相同流程，我们从每个领域随机选取 7000 个样本，然后为每个源域构建一个目标偏移分布，使每个源域包含 5300 个样本，目标域包含 7000 个样本。

在无监督多源域适应中，训练阶段使用 7000 个没有标签的目标样本，测试集则是包含 2000 个样本的未见数据部分。对于只有少量目标标签的有监督多源域适应，我们从 7000 个目标样本中随机抽取 10% 作为训练数据，其余 90% 用作测试集。

我们给出了有监督和无监督设置下的主要超参数，这些参数是通过交叉验证选定的。其他超参数均采用代码中提供的默认值。

在有监督多源域适应中，我们将 L2 缩放系数设为$=1.5$,1，并将 Adadelta 优化器的学习率设为 $\eta=0.5,0.2$。$\mathbf{W}_1$ 距离的惩罚项设为 W1 sup coef$=0.01$,0.03。最后，网络训练 70 个训练轮次，小批量大小为 128。在无监督多源域适应中，我们将 L2 缩放系数设为$=1.5$，学习率设为 $\eta=0.4$，W1 discri coef1$=0.002$,C_0=W1 discri coef2$=0.9$。最后，网络训练 15 个训练轮次，小批量大小为 128。 -

F.3 梯度惩罚

~我们在代码中实现了两种梯度惩罚项。第一种用于确保 Wasserstein~ ~距离下 f 的利普希茨连续性。我们采用了 Gulrajani 等人（2017）提出的方法。例如，对于 $(X,\ Y)\sim\mathcal{T}_u$ 和 $(X',\ Y')\sim\mathcal{S}_u^{\alpha}$，~ ~ ~ 我们生成插值特征 $X_{int}=\lambda X+(1-\lambda)X'$，$\lambda\sim\text{Unif}[0,1]$。梯度惩罚通过添加正则化项 $\|\nabla_{\tilde{X}_{int}}h(v,\ \tilde{X}_{int})\|_2^2$ 来实现。第二种梯度惩罚来自定理 6.1 中的界。我们将每个批次中的梯度范数视为正则化项。请注意，这里的梯度是相对于模型参数而言的，而不是相对于特征或输入。

G 附加结果

G.1 亚马逊评论

现在我们展示一些补充结果。首先，我们展示了无监督多源域适应中各个领域在亚马逊数据集上的不同丢弃率消融研究。如图 2 所示，所提出的 IMDA 算法在每个领域上都始终优于此前的最先进方法。

我们进一步在图 3 中提供了亚马逊数据的 t-SNE 可视化，其中还展示了丢弃率为 50% 时的目标偏移示例。从图中可以观察到，使用 Wasserstein 距离能够保留数据的几何信息，并且适应后分布偏移有所减小。

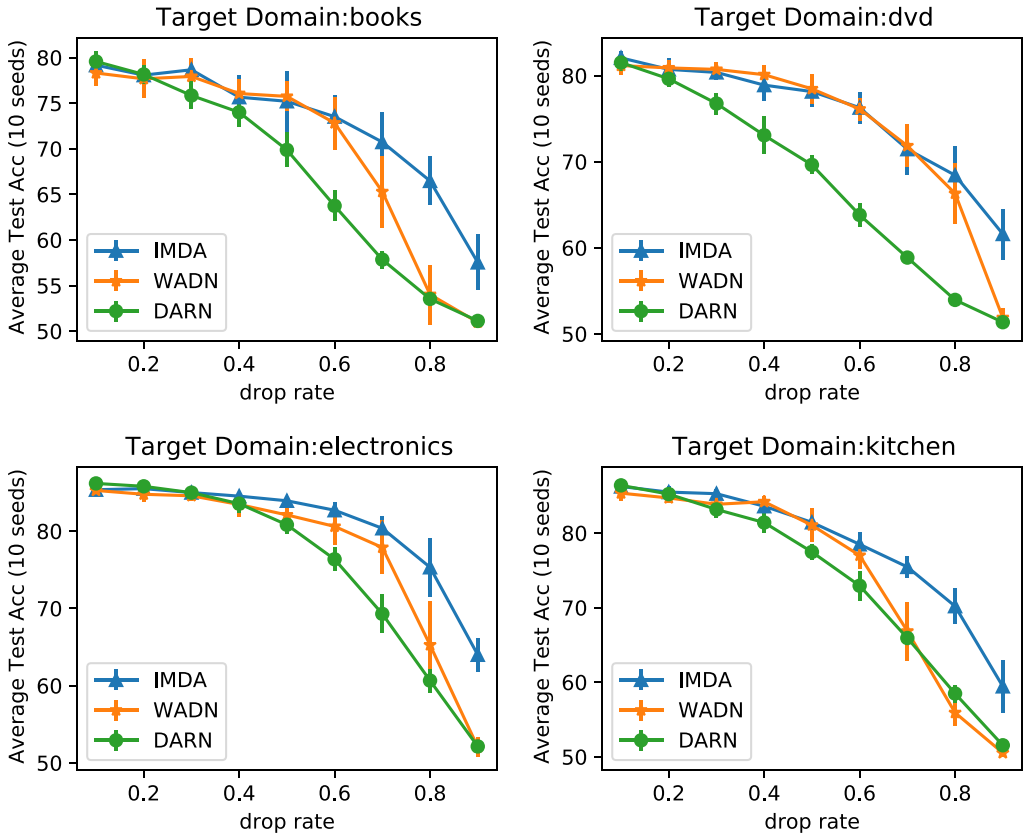


图 2：无监督多源域适应中亚马逊评论数据集在不同目标偏移程度下各领域的测试准确率

G.2 数字

亚马逊评论数据集的四个领域属于难度相近的同质任务。相比之下，四个数字领域属于异质任务，其中两项手写数字识别任务远比两项家庭号码识别任务简单。为了学习更可靠的关系，我们在训练过程开始时（前 5 个训练轮次）设置统一权重，以确保复杂任务得到充分学习，避免因结果较差而被赋予不公平的低权重。

根据数据集描述，我们知道 MNIST 和 USPS 是手写数字数据集，彼此比与其他数据集更加相似。此外，SVHN 和 SYNTH 的相关性更高，因为 SYNTH 是通过对 SVHN 进行变换得到的。我们在图 4 (a) 中观察到了明显的相似性。该算法在学习 MNIST 时赋予 USPS 较高的权重，在学习 SYNTH 时赋予 SVHN 较高的权重，反之亦然。图 4 展示了针对 Wasserstein 距离的极小极大优化过程。在无监督多源域适应中，我们不直接优化合并后的经验风险 $R_{S^\alpha}(u,\ v)$，其估计值持续下降的 $\tilde{\ }$ $\tilde{\ }$演变过程表明，最小化 $\mathbf{W}_1(\mathcal{T}_{u,\ v},\ \mathcal{S}_u)$ 是有效的。

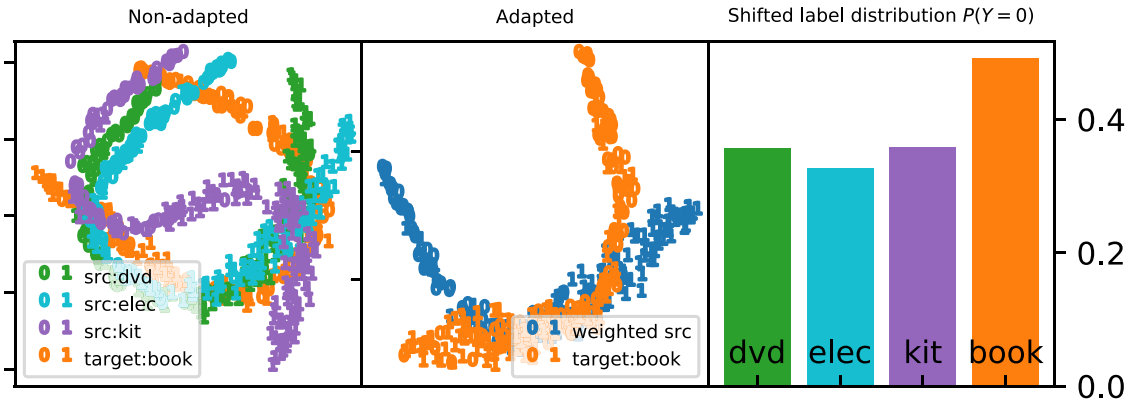
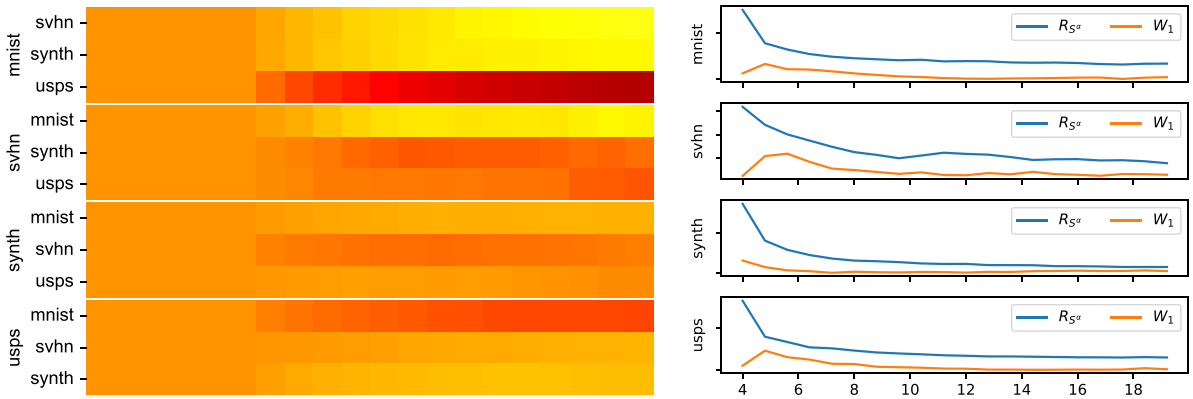


图3：亚马逊评论数据集上的 t-SNE 与目标偏移

我们还在目标偏移数字数据上测试了单源领域自适应方法 DeepJDOT（Damodaran 等，2018）（将多个源域合并为一个）以及 MOST（Nguyen 等，2021）。目标域 MNIST 上的结果见表3。出于以下原因，我们不再对这两种方法进行更详细的比较。首先，我们已经比较了其他采用合并后的源域数据的类似单源领域自适应方法。其次，原始 MOST 算法由多个优化目标组成，但理论结果仅与其中一个目标相关。因此，针对这些目标组合调整超参数对本文的理论研究性质而言毫无意义。

表3：丢弃率为 50% 时数字数据集上的准确率(%)

方法	目标域：MNIST
MOST	88.23
DeepJDOT	87.5
IMDA	89.26



(a) 域权重演化 $\boldsymbol{\alpha}$　　(b) 演化 $\hat{R}_{\mathcal{S}^{\boldsymbol{\alpha}}}(u,\ v)$ 和 $\hat{\mathbf{W}}_1(\tilde{\mathcal{T}}_{u,v},\ \tilde{\mathcal{S}}_u)$

图4：无监督MDA在丢弃率为 0.5 时，数字数据集上不同项随训练轮次演化的可视化

H 算法

算法 1： 信息论多源域适应（IMDA）

输入： 有标签源样本 $S_{1:N}$、少量带标签的目标样本 T、无标签目标样本 T'_X、权重 ϵ、 τ、学习率 η、方差 σ、批量大小 B、常数 C_0， C_1、移动平均权重 $0 < C < 1$;

输出： 表示参数 U、预测器参数 V、重复预测器参数 v'、域权重 $\boldsymbol{\alpha}$;

对于 $t \leftarrow 1$ **至** T，**执行**

如果 $t == 1$，**则**

初始化域权重 $\boldsymbol{\alpha} = \{1/N, ..., 1/N\}$，随机初始化 U_0、 V_0、 v'_0;

$\delta_u = 0$， $\delta_v = 0$;

else

使用上一轮的轮次输出初始化 U_0、 V_0、 v'_0 ;

end

对于 $k \leftarrow 1$ **至** K，**执行**

采样批次数据 T^k_B， $S^k_{B_{1:N}}$， $T^k_{X_B}$来自相应的数据集;

计算梯度:

$$
\begin{aligned}
G^k_u =& \tau(1-\epsilon)\nabla_U \hat{R}_{T^k_B}(U_{k-1},\ V_{k-1}) \\
&+ \tau\epsilon\nabla_U \hat{R}_{T^k_B}(U, v'_{k-1}) + (1-\tau)\nabla_U \hat{R}_{T^k_{X_B}}(U_{k-1},\ V_{k-1},\ v'_{k-1}) \\
&+ \tau\epsilon\nabla_U \hat{R}^{\boldsymbol{\alpha}}_{S^k_{B_{1:N}}}(U_{k-1},\ V_{k-1}) - (1-\tau+\epsilon\tau)\nabla_U \hat{R}^{\boldsymbol{\alpha}}_{S^k_{B_{1:N}}}(U_{k-1},\ v'_{k-1}) \\
G^k_v =& \tau(1-\epsilon)\nabla_V \hat{R}_{T^k_B}(U_{k-1},\ V_{k-1}) \\
&+ (1-\tau)\nabla_V \hat{R}_{T^k_{X_B}}(U_{k-1}, V_{k-1}, v'_{k-1}) + \tau\epsilon\nabla_V \hat{R}^{\boldsymbol{\alpha}}_{S^k_{B_{1:N}}}(U_{k-1},\ V_{k-1}) \\
G^k_{v'} =& \tau\epsilon[\nabla_{v'} \hat{R}_{T^k_B}(U,\ v'_{k-1}) - \nabla_{v'} \hat{R}_{S^k_{B_{1:N}}}(U, v'_{k-1})] \\
&+ (1-\tau)[\nabla_{v'} \hat{R}_{T^k_{X_B}}(U, V, v'_{k-1}) - \nabla_{v'} \hat{R}_{S^k_{B_{1:N}}}(U,\ v'_{k-1})]
\end{aligned}
$$

更新:

$$U_k = U_{k-1} - \eta G^k_u + \xi^u_k,\quad V_k = V_{k-1} - \eta G^k_v + \xi^v_k,\quad v'_k = v'_{k-1} + \eta G^k_{v'};$$

$\delta_u = \delta_u + \frac{\eta^2 \mathbb{E}\|G^k_u\|_2^2}{2\sigma^2}$， $\delta_v = \delta_v + \frac{\eta^2 \mathbb{E}\|G^k_v\|_2^2}{2\sigma^2}$;

end

$U = U_K$， $V = V_K$， $v' = v'_K$;

求解

$$
\begin{aligned}
\boldsymbol{\alpha}' = \text{最小}_{\boldsymbol{\alpha}} \Big(& (\epsilon\tau + C_0(1-\tau)) \hat{R}^{\boldsymbol{\alpha}}_{S^k_B}(U, V) - (\epsilon\tau + 1 - \tau) \hat{R}^{\boldsymbol{\alpha}}_{S^k_B}(U,\ v') \\
& + C_1((1-\tau+\tau\epsilon)\sqrt{\delta_u + \delta_v} + \tau\epsilon\sqrt{\delta_u}) R(\boldsymbol{\alpha}) \Big),
\end{aligned}
$$

$$R(\boldsymbol{\alpha}) = \sqrt{\sum_{i=1}^{N} \frac{\alpha_i^2}{m_i}},\ \text{满足} \forall i \in [N], \alpha_i \geq 0,\ \sum_{i=1}^{N} \alpha_i = 1,$$

更新 $\boldsymbol{\alpha} = C\boldsymbol{\alpha} + (1-C)\boldsymbol{\alpha}'$;

end