arXiv:2303.04198v2 [cs.LG] 5 May 2023

线性等变可转向网络的隐式偏置

Ziyu Chen
马萨诸塞大学阿默斯特分校数学与统计学系，阿默斯特，马萨诸塞州01003，ziyuchen@umass.edu

Wei Zhu
马萨诸塞大学阿默斯特分校数学与统计学系，阿默斯特，马萨诸塞州01003，weizhu@umass.edu

摘要

我们研究群不变二分类中，线性等变可转向网络的梯度流所产生的隐式偏置。研究发现，参数化预测器在方向上收敛至由输入群作用定义的、具有最大间隔的唯一群不变分类器。在输入表示满足酉性假设的条件下，我们证明了可转向网络与数据增强之间的等价性。此外，我们还证明，相较于其非不变对应网络，可转向网络具有更大的间隔和更优的泛化界。

1 引言

尽管深度学习近期取得了理论突破，但对于拥有无限多个近乎零训练误差解的过参数化深度神经网络（DNNs），为何能够在新数据上有效泛化，我们仍知之甚少。然而，使用一阶优化方法（如梯度下降（GD））训练的 DNNs 持续取得令人瞩目的成果，这表明训练算法会隐式地引导模型趋向具有强泛化性能的解。

事实上，近期研究表明，基于梯度的训练方法能够通过隐式最小化模型的某种复杂度度量，有效地对解进行正则化 [Vardi, 2022]。例如，Gunasekar 等人 [2018b] 表明，在可分二元分类中，由线性全连接网络参数化并通过 GD 训练的线性预测器，其方向会收敛至最大间隔支持向量机（SVM）；而线性卷积网络则会受到傅里叶域中一种深度相关桥惩罚的隐式正则化。Yun 等人 [2021] 将这一发现推广至线性张量网络。Lyu 和 Li [2020] 以及 Ji 和 Telgarsky [2020] 则在参数空间中证明了（非线性）齐次深度神经网络的隐式最大间隔正则化。

另一方面，另一条研究路线旨在通过架构设计对深度神经网络进行显式正则化，以利用学习问题的内在结构。近年来，鉴于群对称性在科学和工程领域都十分普遍，人们对利用群对称性实现这一目标的兴趣日益浓厚。大量文献致力于设计群等变深度神经网络，以确保输出随输入的对称变换协变地变化。群等变可转向网络是一类广义的保对称模型，它们能够针对任意一对输入—输出群作用实现等变性 [Cohen 等人，2019年，Weiler 和 Cesa，2019年，Cohen 和 Welling, 2017]。经验证据表明，对于具有群对称性的学习任务，等变可转向网络能够显著提升泛化性能，尤其是在数据量有限时。

近期有多项研究尝试通过为等变可转向网络建立更紧的测试风险上界，来解释这类模型在实践中的成功。这通常通过在网络参数满足相同范数约束的条件下，评估等变模型与非等变模型的复杂度度量来实现——

预印本。审稿中。

型的复杂度度量 [Sokolic等人，2017年，Sannai等人，2021]。然而，目前尚不清楚梯度下降训练的可转向网络能否达到参数范数与其非等变对应模型相当的极小化解，也不清楚其原因。因此，这类基于复杂度度量的论证是否能直接用于解释可转向网络在群对称学习任务中的泛化增强，仍有待确定。

鉴于上述问题，本文旨在全面刻画群不变二分类中线性等变可操纵网络训练算法的隐式偏置。我们的结果表明，在梯度流（GF）下训练时（即采用无穷小步长的梯度下降），由可转向网络参数化的预测器在方向上收敛至唯一的群不变分类器，该分类器相对于由输入群表示定义的范数实现最大间隔。这一结果有三项重要含义：在酉输入群作用下，

- 在原始数据集上训练的线性可转向网络，与在群增强数据集上训练的线性全连接网络收敛至相同方向。这表明，使用线性可转向网络进行训练与数据增强之间存在等价性；
- 在相同的原始数据集上训练时，与全连接网络相比，线性可转向网络在群增强数据集上始终能实现更宽的间隔；
- 当底层分布具有群不变性时，经过梯度流训练的线性可转向网络相比其非等变对应模型能够获得更紧的泛化界。这种泛化能力的提升不一定取决于群的大小，而是取决于不变分布的支撑集。

在结束本节之前，我们指出，Lawrence等人[2021] 近期在线性群卷积神经网络（G-CNN）这一背景下探讨了类似主题；群卷积神经网络是本文所研究的等变可转向网络的一种特殊情况。不过，我们指出，他们研究的模型并非真正具有群不变性，因此其关于隐式偏置的结果无法解释G-CNN泛化能力的提升。我们将在第2节进一步阐述本研究与 [Lawrence等人 2021] 工作的比较。

2 相关工作

隐式偏置：近期研究表明，对于线性可分数据上的逻辑损失或指数损失线性回归，GD/随机梯度下降下的线性预测器会在方向上收敛至最大-L^2-间隔支持向量机（SVM） [Soudry 等，2018、Nacson等人，2019、Gunasekar 等人，2018a]。Gunasekar 等人 [2018b] 在假设方向收敛且网络参数对齐的条件下，将这些结果推广至线性全连接网络和线性卷积神经网络（CNN）；随后，Ji 和 Telgarsky [2019a,b], Lyu 和 Li [2020],Ji 和 Telgarsky [2020]证明了这些条件。梯度流（GF）的隐式正则化进一步被Yun 等人推广至线性张量网络 [2021]。对于无限宽度状态下的过参数化非线性网络，人们也从神经切核视角 [Jacot等人，2018、Du等人，2019、Allen-Zhu等人， 2019] 和平均场视角 [Mei等人，2019、Chizat和Bach， 2018]对深度神经网络（DNN）的优化进行了严格分析。然而，这些模型并非专门设计为具有群不变性/等变性，因此，在具有内在群对称性的学习任务中，基于梯度的方法的隐式偏置可能会引导模型收敛至次优解。

等变神经网络。自 Cohen 和 Welling [2016], 首次提出以来，群等变网络设计已发展成为一个蓬勃发展的领域，并广泛应用于计算机视觉和科学计算等领域。与训练算法所诱导的隐式偏置不同，等变网络通过群卷积 [（Cheng等人，2019；Weiler等人，2018；Sosnovik等人，2020；Zhu等人， 2022] ）或更一般的可转向卷积 [（Weiler 和 Cesa，2019；Cohen 等人，2019；Worrall等人， 2017]）将对称性先验显式纳入模型设计。尽管等变网络在实证中表现出色，但对于采用基于梯度的方法训练的等变网络是否确实会收敛到在群对称学习任务中测试风险更小的解，以及为何会如此，目前仍知之甚少 [（Sokolic等人，2017；Sannai等人， 2021]）。Elesedy和Zaidi [2021] 近期的一项研究证明，在线性回归中，等变网络具有严格的泛化收益。然而，该研究所考察的网络并非等变网络，只是在训练后才进行对称化；这种测试时数据增强方式与等变网络的实际使用方式不同。

与 Lawrence等人 **[2021]**的比较。Lawrence等人 [2021] 近期的一项研究也分析了线性等变 G-CNN 的隐式偏置，而这类网络是本文所考察的可转向网络的一个特例。然而，该研究中的网络实际上并非等变/不变网络。这是因为他们所考虑的输入是定义在群上的函数，且群作用为正则表示。在他们的情形下，由于正则表示具有传递性，唯一的群不变线性分类器（相差一个标量乘法）就是对输入取群平均。因此，Lawrence等人 [2021], 采用了一个覆盖整个群的全连接末层，导致分类器不具备群不变性。因此，他们关于隐式偏置的结果并不能解释 G-CNN 泛化能力的提升。相比之下，我们假设输入为欧几里得输入，且群作用非传递，从而使线性可转向网络能够学习非平凡的群不变模型。

3 背景与问题设定

我们简要介绍群论和群等变可转向网络的背景，并说明群不变二分类学习问题的设定。

3.1 群与群等变性

群是一个集合 G，配备了一个二元运算，即群乘法，并满足结合律、单位元和可逆性公理。本文始终假设 G 是有限的，即 $|G| < \infty$，其中 $|G|$ 表示其基数。

给定向量空间 $\mathcal{X}$，令 $\mathrm{GL}(\mathcal{X})$ 表示 $\mathcal{X}$ 的一般线性群，由 $\mathcal{X}$ 上所有可逆线性变换构成。映射 $\rho\colon G \to \mathrm{GL}(\mathcal{X})$ 称为 G 在 $\mathcal{X}$ 上的群表示（或线性群作用），当且仅当 ρ 是从 G 到 $\mathrm{GL}(\mathcal{X})$ 的群同态，即

$$\rho(gh) = \rho(g)\rho(h) \in \mathrm{GL}(\mathcal{X}), \quad \forall g, h \in G. \tag{1}$$

当表示 ρ 在上下文中明确时，我们也将群作用 $\rho(g)\mathbf{x}$ 简写为 $g\mathbf{x}$。

给定一对向量空间 $\mathcal{X}$、$\mathcal{Y}$及其各自的 G 表示 $\rho_{\mathcal{X}}$ 和 $\rho_{\mathcal{Y}}$，若线性映射$\Psi\colon \mathcal{X} \to \mathcal{Y}$与 G 表示 $\rho_{\mathcal{X}}$ 和 $\rho_{\mathcal{Y}}$ 可交换，则称其为 G-等变映射，即:

$$\Psi \circ \rho_{\mathcal{X}}(g) = \rho_{\mathcal{Y}}(g) \circ \Psi, \quad \forall g \in G. \tag{2}$$

线性等变映射也称为交织算子，我们用 $\mathrm{Hom}_G(\rho_{\mathcal{X}}, \rho_{\mathcal{Y}})$ 表示满足（2）的所有交织算子构成的空间。当 $\rho_{\mathcal{Y}} \equiv \mathrm{Id}$ 是平凡表示时，则对所有 $\mathbf{x} \in \mathcal{X}$，有 $\Psi[\rho_{\mathcal{X}}(g)\mathbf{x}] = \Psi[\mathbf{x}]$；也就是说，$\Psi$ 成为一个 G-不变线性映射。

3.2 等变可转向网络和 G-CNN

设 $\mathcal{X}_0 = \mathbb{R}^{d_0}$ 是一个 d_0 维输入欧几里得空间，并配备通常内积。设 $\rho_0\colon G \to \mathrm{GL}(\mathcal{X}_0)$ 是定义在 $\mathcal{X}_0$ 上的 G 表示。假设我们有一个未知的目标函数 $f^*\colon \mathcal{X}_0 = \mathbb{R}^{d_0} \to \mathbb{R}$，它在 ρ_0 下是 G- 不变的，即，对于所有$g \in G$ 和 $\mathbf{x} \in \mathcal{X}_0$，$f^*(g\mathbf{x}) := f^*(\rho_0(g)\mathbf{x}) = f^*(\mathbf{x})$。等变可转向网络的目标是使用一个$L$ 层神经网络 $f = \Psi_L \circ \Psi_{L-1} \circ \cdots \circ \Psi_1$ 来逼近 f^*，并保证该网络同样是 G-不变的。

很容易根据公式（2）验证，等变映射的复合仍然是等变的。因此，要使 $f = \Psi_L \circ \Psi_{L-1} \circ \cdots \circ \Psi_1$ 对 G 不变，只需指定一组 G 表示空间 $\{(\mathcal{X}_l, \rho_l)\}_{l=1}^{L}$，其中 $(\mathcal{X}_L, \rho_L) = (\mathbb{R}, \mathrm{Id})$ 是平凡表示，并使每一层 $\Psi_l \in \mathrm{Hom}_G(\rho_{l-1}, \rho_L)\colon \mathcal{X}_{l-1} \to \mathcal{X}_l$ 都是 G-等变的。等价地，我们希望下图是交换图，

$$\begin{array}{ccccccccc} \mathcal{X}_0 & \xrightarrow{\Psi_1} & \mathcal{X}_1 & \xrightarrow{\Psi_2} & \mathcal{X}_2 & \xrightarrow{\Psi_3} & \cdots & \xrightarrow{\Psi_{L-1}} & \mathcal{X}_{L-1} & \xrightarrow{\Psi_L} & \mathcal{X}_L \\ \downarrow{\scriptstyle \rho_0(g)} & & \downarrow{\scriptstyle \rho_1(g)} & & \downarrow{\scriptstyle \rho_2(g)} & & & & \downarrow{\scriptstyle \rho_{L-1}(g)} & & \downarrow{\scriptstyle \rho_L(g)} \\ \mathcal{X}_0 & \xrightarrow{\Psi_1} & \mathcal{X}_1 & \xrightarrow{\Psi_2} & \mathcal{X}_2 & \xrightarrow{\Psi_3} & \cdots & \xrightarrow{\Psi_{L-1}} & \mathcal{X}_{L-1} & \xrightarrow{\Psi_L} & \mathcal{X}_L \end{array} \quad , \quad \forall g \in G. \tag{3}$$

等变可转向网络。给定表示 $\{(\mathcal{X}_l, \rho_l)\}_{l=0}^{L}$，线性可转向网络按如下方式构建。对于每个 $l \in [L] := \{1, \cdots, L\}$，我们选择一组有限的 N_l（预先计算的）交织算子 $\{\psi_l^j\}_{j=1}^{N_l} \subset \mathrm{Hom}_G(\rho_{l-1}, \rho_l)$。通常，$\{\psi_l^j\}_{j=1}^{N_l}$ 是

$\mathrm{Hom}_G(\rho_{l-1}, \rho_l)$ 的一组基，但在我们的设定中这并非必要。可转向网络第 l 层的等变映射Ψ_l^{steer}: $\mathcal{X}_{l-1} \to \mathcal{X}_l$ 随后由以下参数化

$$\Psi_l^{\text{steer}}(\mathbf{x}; \mathbf{w}_l) = \sum_{j \in [N_l]} w_l^j \psi_l^j(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}_{l-1}, \tag{4}$$

其中，系数 $\mathbf{w}_l = [w_l^j]_{j\in[N_l]}^\top \in \mathbb{R}^{N_l}$ 是第 l 层的可训练参数。L层线性可转向网络 $f_{\text{steer}}(\mathbf{x};\mathbf{W})$ 定义为以下复合函数

$$f_{\text{steer}}(\mathbf{x}; \mathbf{W}) = \Psi_L^{\text{steer}}(\cdots \Psi_2^{\text{steer}}(\Psi_1^{\text{steer}}(\mathbf{x}; \mathbf{w}_1); \mathbf{w}_2) \cdots ; \mathbf{w}_L), \tag{5}$$

其中，$\mathbf{W} = [\mathbf{w}_l]_{l=1}^L \in \prod_{l=1}^L \mathbb{R}^{N_l} =: \mathcal{W}_{\text{steer}}$ 是所有可训练参数的集合。式（5）中定义的网络 $f_{\text{steer}}(\mathbf{x}; \mathbf{W})$ 称为可转向网络，因为它能够引导各层的输出按照任意指定的表示进行变换。

G-CNN。群卷积神经网络（G-CNN）是可转向网络的一种特殊情况，其通过群卷积实现等变性。更具体地说，对于每个$l \in [L-1]$，隐藏表示空间（$\mathcal{X}_l, \rho_l$）设为

$$\mathcal{X}_l = (\mathbb{R}^{d_l})^G = \{\mathbf{x}_l : G \to \mathbb{R}^{d_l}\}, \quad \rho_l(g)\mathbf{x}_l(h) \coloneqq \mathbf{x}_l(g^{-1}h) \in \mathbb{R}^{d_1}, \forall g, h \in G. \tag{6}$$

式（6）中的表示 $\rho_l \in \mathrm{GL}(\mathcal{X}_l)$称为 G 的正则表示。直观地说，$\mathbf{x}_l \in \mathcal{X}_l$ 可以视为一个大小为 $d_l \times |G|$ 的矩阵，而 $\rho_l(g)$ 则是对 $\mathbf{x}_l$ 的列进行的置换。采用这种 $\{(\mathcal{X}_l, \rho_l)\}_{l=0}^L$ 选择后，可按如下方式构造（线性）G-CNN。

第一层：G-卷积神经网络的第一层等变映射 $\Psi_1^{\text{G-CNN}}: \mathcal{X}_0 \to \mathcal{X}_1$ 定义为

$$\Psi_1^{\text{G-CNN}}(\mathbf{x}; \mathbf{w}_1)(g) = \mathbf{w}_1^\top g^{-1}\mathbf{x} \in \mathbb{R}^{d_1}, \quad \forall \mathbf{x} \in \mathbb{R}^{d_0}, \tag{7}$$

其中，$\mathbf{w}_1 = (w^j{}_{,1}{}^k)_{j,k} \in \mathbb{R}^{d_0 \times d_1}$ 是第一层的可训练参数。式 (7) 称为 G-提升映射，因为它将欧几里得信号 $\mathbf{x} \in \mathbb{R}^{d_0}$ 提升为定义在 G 上的函数 $\Psi_1^{\text{G-CNN}}(\mathbf{x};\mathbf{w}_1)$。

隐藏层：对于 $l \in \{2, \cdots, L-1\}$，将 $\Psi_l^{\text{G-CNN}}: \mathcal{X}_{l-1} \to \mathcal{X}_l$ 定义为群卷积，

$$\Psi_l^{\text{G-CNN}}(\mathbf{x}; \mathbf{w}_l)(g) = \sum_{h \in G} \mathbf{w}_l^\top (h^{-1}g)\mathbf{x}(h) \in \mathbb{R}^{d_l}, \quad \forall \mathbf{x} \in \mathcal{X}_{l-1}, \tag{8}$$

其中 $\mathbf{w}_l(g) \in \mathbb{R}^{d_{l-1} \times d_l}, \forall g \in G$。等价地，$\mathbf{w}_l \in \mathbb{R}^{d_{l-1} \times d_l \times |G|}$ 可视为一个三维张量。最后一层：将 $\Psi_L^{\text{G-CNN}}: \mathcal{X}_{L-1} \to \mathcal{X}_L$ 定义为群池化后接一个全连接层，

$$\Psi_L^{\text{G-CNN}}(\mathbf{x}; \mathbf{w}_L) = \mathbf{w}_L^\top \frac{1}{|G|} \sum_{g \in G} \mathbf{x}(g), \quad \forall \mathbf{x} \in \mathcal{X}_{L-1}, \tag{9}$$

whe其中，$\mathbf{w}_L \in \mathbb{R}^{d_{L-1}}$ 是最后一层的权重。一个 L 层线性G-CNN是以下复合 n

$$f_{\text{G-CNN}}(\mathbf{x}; \mathbf{W}) = \Psi_L^{\text{G-CNN}}(\cdots \Psi_2^{\text{G-CNN}}(\Psi_1^{\text{G-CNN}}(\mathbf{x}; \mathbf{w}_1); \mathbf{w}_2) \cdots ; \mathbf{w}_L), \quad \mathbf{W} = [\mathbf{w}_l]_{l=1}^L, \tag{10}$$

其中，$\mathbf{W} \in \mathbb{R}^{d_0 \times d_1} \times \left[\prod_{l=2}^{L-1} \mathbb{R}^{d_{l-1} \times d_l \times |G|}\right] \times \mathbb{R}^{d_{L-1}} =: \mathcal{W}_{\text{G-CNN}}$ 是可训练权重。

关于表示的假设。假设输入表示 $\rho_0 \in \mathrm{GL}(\mathcal{X}_0 = \mathbb{R}^{d_0})$已知，并将 $\rho_L \equiv 1 \in \mathrm{GL}(\mathcal{X}_L = \mathbb{R})$设为平凡表示，以获得群不变输出。在本文中，我们对第一层表示（ρ_1，$\mathcal{X}_1$）以及可转向网络中的等变映射 $\Psi_1^{\text{steer}}(\mathbf{x},\mathbf{w}_1)$作如下特殊选择，而其余表示（$\rho_l$，$\mathcal{X}_l$）和可转向映射 $\Psi_l^{\text{steer}}(\mathbf{x}, \mathbf{w}_l)$、$l \in \{2, \cdots, L-1\}$则可以任意选取。

假设 3.1。我们在第一层采用正则表示 (6)，并将第一层等变映射 $\Psi_1^{\text{steer}}(\mathbf{x};\mathbf{w}_1)$ 设为 G-提升映射 (7)。

假设 3.1 的依据是使网络具备足够的容量，从而能够参数化任意 G 不变的线性分类器；命题 3.6 将对此作进一步说明。在假设 3.1 下，我们可以通过以下命题刻画线性可转向网络 $f_{\text{steer}}(\cdot;\mathbf{W})$。

命题 3.2. 设 $f_{\text{steer}}(\mathbf{x};\mathbf{W})$ 是满足假设 3.1 的线性可转向网络，其中 $\mathbf{W} = [\mathbf{w}_l]_{l=1}^{L} \in \mathcal{W}_{\text{steer}}$ 是所有模型参数的集合。存在一个多线性映射 $M\colon (\mathbf{w}_2,\cdots,\mathbf{w}_L) \mapsto M(\mathbf{w}_2,\cdots,\mathbf{w}_L) \in \mathbb{R}^{d_1}$，使得对于所有 $\mathbf{x} \in \mathbb{R}^{d_0}$ 和 $\mathbf{W} \in \mathcal{W}_{\text{steer}}$，

$$f_{\text{steer}}(\mathbf{x};\mathbf{W}) = f_{\text{steer}}(\overline{\mathbf{x}};\mathbf{W}) = \langle \overline{\mathbf{x}}, \mathbf{w}_1 M(\mathbf{w}_2,\cdots,\mathbf{w}_L)\rangle\,, \tag{11}$$

其中 $\underline{\mathbf{x}} := \frac{1}{|G|}\sum_{g\in G} g\mathbf{x}$ 是 $\mathbf{x}$ 的群轨道 $\in \mathbb{R}^{d_0}$ 上所有元素的平均值。

命题 3.2 的证明以及其他结果均见附录。若定义

$$\overline{\mathcal{P}}_{\text{steer}}(\mathbf{W}) \coloneqq \mathbf{w}_1 M(\mathbf{w}_2,\cdots,\mathbf{w}_L), \quad \mathcal{P}_{\text{steer}}(\mathbf{W}) \coloneqq \left(\frac{1}{|G|}\sum_{g\in G} g^{\top}\right)\overline{\mathcal{P}}_{\text{steer}}(\mathbf{W}), \tag{12}$$

其中 $g^{\top} := \rho_0(g)^{\top}$，则

$$f_{\text{steer}}(\mathbf{x};\mathbf{W}) = \langle \mathbf{x}, \mathcal{P}_{\text{steer}}(\mathbf{W})\rangle = \left\langle \overline{\mathbf{x}}, \overline{\mathcal{P}}_{\text{steer}}(\mathbf{W})\right\rangle. \tag{13}$$

根据 $M(\mathbf{w}_2,\cdots,\mathbf{w}_L)$ 的多重线性，可以验证 $f_{\text{steer}}(\mathbf{x};\mathbf{W})$、$\mathcal{P}_{\text{steer}}(\mathbf{W})$ 和 $\overline{\mathcal{P}}_{\text{steer}}(\mathbf{W})$ 关于 $\mathbf{W}$ 都是 L 次齐次的；也就是说，对于所有 $\nu > 0$ 和 $\mathbf{W} \in \mathcal{W}_{\text{steer}}$，

$$f_{\text{steer}}(\mathbf{x};\nu\mathbf{W}) = \nu^L f_{\text{steer}}(\mathbf{x};\mathbf{W}),\ \mathcal{P}_{\text{steer}}(\nu\mathbf{W}) = \nu^L \mathcal{P}_{\text{steer}}(\mathbf{W}),\ \overline{\mathcal{P}}_{\text{steer}}(\nu\mathbf{W}) = \nu^L \overline{\mathcal{P}}_{\text{steer}}(\mathbf{W}). \tag{14}$$

备注 3.3。作为线性可转向网络的一个特例，线性G-CNN $f_{\text{G-CNN}}(\mathbf{x};\mathbf{W})$ 可以写为 (10)

$$f_{\text{G-CNN}}(\mathbf{x};\mathbf{W}) = \langle \mathbf{x}, \mathcal{P}_{\text{G-CNN}}(\mathbf{W})\rangle = \left\langle \overline{\mathbf{x}}, \overline{\mathcal{P}}_{\text{G-CNN}}(\mathbf{W})\right\rangle, \tag{15}$$

其中

$$\overline{\mathcal{P}}_{\text{G-CNN}}(\mathbf{W}) \coloneqq \mathbf{w}_1 \left[\prod_{l=2}^{L-1}\left(\sum_{g\in G}\mathbf{w}_l(g)\right)\right]\mathbf{w}_L,\ \mathcal{P}_{\text{G-CNN}}(\mathbf{W}) \coloneqq \left[\frac{1}{|G|}\sum_{g\in G} g^{\top}\right]\overline{\mathcal{P}}_{\text{G-CNN}}(\mathbf{W}) \tag{16}$$

备注 3.4。相比之下，一个 L 层线性全连接网络 $f_{\text{fc}}(\mathbf{x};\mathbf{W})$ 定义为

$$f_{\text{fc}}(\mathbf{x};\mathbf{W}) = \mathbf{w}_L^{\top}\mathbf{w}_{L-1}^{\top}\cdots\mathbf{w}_1^{\top}\mathbf{x} = \langle \mathbf{x}, \mathcal{P}_{\text{fc}}(\mathbf{W})\rangle\,, \quad \mathcal{P}_{\text{fc}}(\mathbf{W}) \coloneqq \mathbf{w}_1\cdots\mathbf{w}_L. \tag{17}$$

其中 $\mathbf{W} = [\mathbf{w}_l]_{l=1}^{L} \in \mathcal{W}_{\text{fc}} \coloneqq \left(\prod_{l=1}^{L-1}\mathbb{R}^{d_{l-1}\times d_l}\right)\times\mathbb{R}^{d_{L-1}}$。对比式 (16) 和式 (17) 可知，线性全连接网络 $f_{\text{fc}}(\mathbf{x};\mathbf{W})$ 是线性 G-CNN $f_{\text{G-CNN}}(x;\mathbf{W})$ (16) 的一个更特殊的情况，其中 $G = \{e\}$ 是平凡群。

3.3 群不变二分类

考虑一个二分类数据集 $S = \{(\mathbf{x}_i, y_i) :\ i \in [n]\}$，其中 $\mathbf{x}_i \in \mathbb{R}^{d_0}$ 且 $y_i \in \{\pm 1\}, \forall i \in [n]$。我们假设 S 是来自下文所定义的 G-不变分布 $\mathcal{D}$ 的独立同分布样本。

定义 3.5。若分布 $\mathcal{D}$ 在 $\mathbb{R}^{d_0}\times\{\pm 1\}$ 上相对于表示 $\rho_0 \in \text{GL}(\mathbb{R}^{d_0})$ 是 G-不变的，则称其为

$$(\rho_0(g)\otimes \text{Id})_*\,\mathcal{D} = \mathcal{D},\ \forall g \in G, \tag{18}$$

其中，$(\rho_0(g)\otimes\text{Id})(\mathbf{x}\ y) \coloneqq (\rho_0(g)\mathbf{x}\ y)$，而 $(\rho_0(g)\otimes\text{Id})_*\,\mathcal{D} \coloneqq \mathcal{D}\circ(\rho_0(g)\otimes\text{Id})^{-1}$ 是 $\mathcal{D}$ 在 $\rho_0(g)\otimes\text{Id}$ 下的推前测度。

容易验证，对于具有 G 不变性的分布 $\mathcal{D}$，贝叶斯最优分类器 $f^*\colon \mathbb{R}^{d_0} \to \{\pm 1\}$（即实现最小总体风险的分类器）必然是一个 G 不变函数，即 $f^*(g\mathbf{x}) = f^*(\mathbf{x})\ \forall g \in G$。因此，要使用（线性）神经网络学习 f^*，自然可以使用等变可转向网络来近似 f^*

$$f^*(\mathbf{x}) \approx \text{sign}\left(f_{\text{steer}}(\mathbf{x};\mathbf{W})\right) = \text{sign}\left(\langle \mathbf{x}, \mathcal{P}_{\text{steer}}(\mathbf{W})\rangle\right). \tag{19}$$

选择指数损失 $\ell_{\exp}\colon \mathbb{R}\times\{\pm 1\}\to\mathbb{R}_+$, $\ell_{\exp}(\hat{y},y)\coloneqq \exp(-\hat{y}y)$ 作为替代损失函数后，对于可转向网络 $f_{\text{steer}}(\mathbf{x};\mathbf{W})$ 在 S 上进行经验风险最小化可表示为

$$\min_{\mathbf{W}\in\mathcal{W}_{\text{steer}}} \mathcal{L}_{\mathcal{P}_{\text{steer}}}(\mathbf{W};S)\coloneqq \sum_{i=1}^{n}\ell_{\exp}\left(\langle \mathbf{x}_i,\mathcal{P}_{\text{steer}}(\mathbf{W})\rangle, y_i\right)=\sum_{i=1}^{n}\ell_{\exp}\left(\left\langle \overline{\mathbf{x}}_i,\overline{\mathcal{P}}_{\text{steer}}(\mathbf{W})\right\rangle, y_i\right). \tag{20}$$

另一方面，由于 $\mathcal{P}_{\text{steer}}(\mathbf{W})=\boldsymbol{\beta}\in\mathbb{R}^{d_0}$ 始终对应一个 G-不变线性预测器——这里我们稍微滥用了记号，将 $\boldsymbol{\beta}\in\mathbb{R}^{d_0}$ 与映射 $\mathbf{x}\mapsto\langle \mathbf{x}\ \boldsymbol{\beta}\rangle$等同起来——因此也可以直接在不变线性预测器 $\boldsymbol{\beta}$上进行经验风险最小化：

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{d_0}_G} \mathcal{L}(\boldsymbol{\beta};S)\coloneqq \sum_{i=1}^{n}\ell_{\exp}\left(\langle \mathbf{x}_i,\boldsymbol{\beta}\rangle, y_i\right), \tag{21}$$

其中，$\mathbb{R}^{d_0}_G\subset\mathbb{R}^{d_0}$ 是所有 G-不变线性预测器构成的子空间，其由以下命题刻画。

命题 3.6. 设 $\mathbb{R}^{d_0}_G\subset\mathbb{R}^{d_0}$ 为 G-不变线性预测器构成的子空间，即 $\mathbb{R}^{d_0}_G=\left\{\boldsymbol{\beta}\in\mathbb{R}^{d_0}:\ \boldsymbol{\beta}^\top\mathbf{x}=\boldsymbol{\beta}^\top g\mathbf{x}\ \forall\mathbf{x}\in\mathbb{R}^{d_0},\forall g\in G\right\}$。则

(a) $\mathbb{R}^{d_0}_G$由以下条件刻画：

$$\mathbb{R}^{d_0}_G=\bigcap_{g\in G}\ker(I-g^\top)=\text{Range}\left(\frac{1}{|G|}\sum_{g\in G}g^\top\right). \tag{22}$$

(b) 设 $\mathcal{A}:\ \mathbb{R}^{d_0}\to\mathbb{R}^{d_0}$ 为群平均映射，

$$\mathcal{A}(\boldsymbol{\beta})\coloneqq\overline{\boldsymbol{\beta}}=\frac{1}{|G|}\sum_{g\in G}g\boldsymbol{\beta}. \tag{23}$$

则其伴随算子 $\mathcal{A}^\top:\boldsymbol{\beta}\mapsto\frac{1}{|G|}\sum_{g\in G}g^\top\boldsymbol{\beta}$ 是从 $\mathbb{R}^{d_0}$到 $\mathbb{R}^{d_0}_G$的投影算子。换言之，值域$(\mathcal{A}^\top)=\mathbb{R}^{d_0}_G$且 $\mathcal{A}^\top\circ\mathcal{A}^\top=\mathcal{A}^\top$。

(c) 如果 G 在 $\mathcal{X}_0$上酉地作用，即 $\rho_0(g^{-1})=\rho_0(g)^\top$，那么 $\mathcal{A}=\mathcal{A}^\top$ 是自伴的。这意味着 $\mathcal{A}:\ \boldsymbol{\beta}\mapsto\overline{\boldsymbol{\beta}}$ 是从 $\mathbb{R}^{d_0}$ 到$\mathbb{R}^{d_0}_G$的正交投影。特别地，我们有

$$\overline{\boldsymbol{\beta}}=\boldsymbol{\beta}\iff\boldsymbol{\beta}\in\mathbb{R}^{d_0}_G,\quad \textit{and}\ \|\overline{\boldsymbol{\beta}}\|\le\|\boldsymbol{\beta}\|,\forall\boldsymbol{\beta}\in\mathbb{R}^{d_0}. \tag{24}$$

命题 3.6 与 (12) 结合表明，线性可转向网络 $f_{\text{steer}}(\cdot;\mathbf{W})=\langle\cdot,\mathcal{P}_{\text{steer}}(\mathbf{W})\rangle$ 可以实现任意 G-不变线性预测器 $\boldsymbol{\beta}\in\mathbb{R}^{d_0}_G$；也就是说，$\{\mathcal{P}_{\text{steer}}(\mathbf{W}):\mathbf{W}\in\mathcal{W}_{\text{steer}}\}=\mathbb{R}^{d_0}_G$。因此，(20) 和 (21) 是以不同方式参数化的等价优化问题。然而，使用基于梯度的方法最小化 (20)，可能会得到与直接优化 (21) 所得结果不同的分类器。

梯度流。给定初始化 $\mathbf{W}(0)\in\mathcal{W}_{\text{steer}}$，(20) 的梯度流 $\{\mathbf{W}(t)\}_{t\ge0}$ 是以下常微分方程（ODE）的解，

$$\frac{\mathrm{d}\mathbf{W}}{\mathrm{d}t}=-\nabla_{\mathbf{W}}\mathcal{L}_{\mathcal{P}_{\text{steer}}}(\mathbf{W};S)=-\nabla_{\mathbf{W}}\left[\sum_{i=1}^{n}\ell_{\exp}\left(\langle \mathbf{x}_i,\mathcal{P}_{\text{steer}}(\mathbf{W})\rangle, y_i\right)\right]. \tag{25}$$

本研究旨在考察在梯度流 (25) 下训练的线性可转向网络所参数化的 G-不变线性预测器 $\boldsymbol{\beta}_{\text{steer}}(t)=\mathcal{P}_{\text{steer}}(\mathbf{W}(t))$ 的渐近行为。具体而言，我们旨在分析当 $t\to\infty$ 时 $\boldsymbol{\beta}_{\text{steer}}(t)$ 的方向极限，即 $\lim_{t\to\infty}\frac{\boldsymbol{\beta}_{\text{steer}}(t)}{\|\boldsymbol{\beta}_{\text{steer}}(t)\|}$。在结束本节之前，我们对梯度流 $\mathbf{W}(t)$ 作出如下假设；该假设也出现在 Ji 和 Telgarsky [2020], Lyu 和 Li [2020], Yun 等人[2021]的先前研究中。

假设 3.7。梯度流 $\mathbf{W}(t)$ 满足 $\mathcal{L}_{\mathcal{P}_{\text{steer}}}(\mathbf{W}(t_0);S)<1$，其中某个 $t_0>0$ 存在。

这一假设意味着，数据集 $S=\{(\mathbf{x}_i,y_i):\ i\in[n]\}$ 可以由一个 G-不变线性预测器 $\mathcal{P}_{\text{steer}}(\mathbf{W}(t_0))$ 分离，而我们的分析聚焦于梯度流训练的“后期阶段”，其中 $t\to\infty$。

4 线性可转向网络的隐式偏置

我们关于二元分类中梯度流在线性可转向网络上的隐式偏置的主要结果总结如下定理。

定理 4.1。在假设 3.1 和假设 3.7 下，令 $\boldsymbol{\beta}_{\text{steer}}(t) = \mathcal{P}_{\text{steer}}(\mathbf{W}(t))$ 表示由线性可转向网络参数化、通过梯度流在数据集 $S = \{(\mathbf{x}_i, y_i) : i \in [n]\}$ 上训练得到的 G-不变线性预测器的时间演化；参见公式 (25)。则

(a) 方向极限 $\boldsymbol{\beta}^{\infty}_{\text{steer}} = \lim_{t\to\infty} \frac{\boldsymbol{\beta}_{\text{steer}}(t)}{\|\boldsymbol{\beta}_{\text{steer}}(t)\|}$ 存在，且 $\boldsymbol{\beta}^{\infty}_{\text{steer}} \propto \frac{1}{|G|}\sum_{g\in G} g^{\top}\boldsymbol{\gamma}^*$，其中 $\boldsymbol{\gamma}^*$ 是针对变换后的数据 $\overline{S} = \{(\mathbf{x}_i, y_i) : i \in [n]\}$ 的最大 L^2-间隔支持向量机解：

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^{d_0}} \|\boldsymbol{\gamma}\|^2, \quad \text{s.t. } y_i \langle \overline{\mathbf{x}}_i, \boldsymbol{\gamma}\rangle \geq 1, \forall i \in [n]. \tag{26}$$

此外，如果 G 在输入空间 $\mathcal{X}_0$ 上酉地作用，即 $g^{-1} = g^{\top}$，那么 $\boldsymbol{\beta}^{\infty}_{\text{steer}} \propto \boldsymbol{\gamma}^*$。

(b) 等价地，$\boldsymbol{\beta}^{\infty}_{\text{steer}}$ 与问题的唯一极小化解 $\boldsymbol{\beta}^*$ 成比例

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{d_0}} \|\text{Proj}_{\text{Range}(\mathcal{A})}\boldsymbol{\beta}\|^2, \quad \text{s.t. } \boldsymbol{\beta} \in \mathbb{R}^{d_0}_G, \textit{ and } y_i \langle \mathbf{x}_i, \boldsymbol{\beta}\rangle \geq 1, \forall i \in [n], \tag{27}$$

其中，$\text{Proj}^{\text{Range}(\mathcal{A})}$ 是从 $\mathbb{R}^{d_0}$ 到 $\text{Range}(\mathcal{A}) = \text{Range}\left(\frac{1}{|G|}\sum_{g\in G} g\right)$的投影。此外，若 G 在 $\mathcal{X}_0$上酉作用，则

$$\boldsymbol{\beta}^{\infty}_{\text{steer}} \propto \boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{d_0}} \|\boldsymbol{\beta}\|^2, \quad \text{s.t. } \boldsymbol{\beta} \in \mathbb{R}^{d_0}_G, \textit{ and } y_i \langle \mathbf{x}_i, \boldsymbol{\beta}\rangle \geq 1, \forall i \in [n]. \tag{28}$$

也就是说，在所有 G-不变线性预测器中，$\boldsymbol{\beta}^{\infty}_{\text{steer}}$ 实现了最大的 L^2-间隔。

定理 4.1 表明，梯度流会隐式地引导线性可转向网络趋向唯一的 G-不变分类器，其最大间隔由输入表示 ρ_0定义。

注释4.2。根据备注 3.4，若 $G = \{e\}$ 是平凡群，则线性 G-卷积神经网络$f_{\text{G-CNN}}(\mathbf{x};\mathbf{W})$ (10)（这是线性可转向网络的一种特殊情况）会退化为全连接网络 $f_{\text{fc}}(\mathbf{x};\mathbf{W})$。由于平凡群的表示总是酉的，因此我们得到以下推论，该推论也曾见于 Ji 和 Telgarsky *[2020]* 以及 Yun 等人 *[2021]*。推论 4.3。设 $\{\mathbf{W}(t)\}_{t\geq 0} \subset \mathcal{W}_{\text{fc}}$ 为在数据集 $S = \{(\mathbf{x}_i, y_i)\ i \in [n]\}$上训练线性全连接网络时参数的梯度流，即

$$\frac{\mathrm{d}\mathbf{W}}{\mathrm{d}t} = -\nabla_{\mathbf{W}}\mathcal{L}_{\mathcal{P}_{\text{fc}}}(\mathbf{W}; S) := -\nabla_{\mathbf{W}}\left[\sum_{i=1}^{n} \ell_{\exp}(\langle \mathbf{x}_i, \mathcal{P}_{\text{fc}}(\mathbf{W})\rangle, y_i)\right]. \tag{29}$$

于是，分类器 $\boldsymbol{\beta}_{\text{fc}}(t) = \mathcal{P}_{\text{fc}}(\mathbf{W}(t))$ 收敛于一个方向，该方向与原始数据集 $S = \{(\mathbf{x}_i, y_i) : i \in [n]\}$ 的最大-L^2-间隔支持向量机解 $\boldsymbol{\gamma}^*$ 对齐，

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^{d_0}} \|\boldsymbol{\gamma}\|^2, \quad \text{s.t. } y_i \langle \mathbf{x}_i, \boldsymbol{\gamma}\rangle \geq 1, \forall i \in [n]. \tag{30}$$

5 可转向网络与数据增强之间的等价性

与通过等变可转向网络将对称性先验硬编码到模型架构中相比，将对称性纳入学习过程的另一种方法是借助数据增强来训练非等变模型。本节将证明，在关于 ρ_0 的幺正假设下，这两种方法对于二分类是等价的。

推论 5.1. 设 $\boldsymbol{\beta}^{\infty}_{\text{steer}} = \lim_{t\to\infty} \frac{\boldsymbol{\beta}_{\text{steer}}(t)}{\|\boldsymbol{\beta}_{\text{steer}}(t)\|}$ 为使用梯度流在原始数据集 $S = \{(\mathbf{x}_i, y_i)\ i \in [n]\}$ 上训练的线性可转向网络所参数化的线性预测器 $\boldsymbol{\beta}_{\text{steer}}(t) = \mathcal{P}_{\text{steer}}(\mathbf{W}(t))$ 的方向极限。相应地，设 $\boldsymbol{\beta}^{\infty}_{\text{fc}} = \lim_{t\to\infty} \frac{\boldsymbol{\beta}_{\text{fc}}(t)}{\|\boldsymbol{\beta}_{\text{fc}}(t)\|}$,$\boldsymbol{\beta}_{\text{fc}}(t) = \mathcal{P}_{\text{fc}}(\mathbf{W}(t))$ (17) 为使用增强数据集 $S_{\text{aug}} = \{(g\ \mathbf{x}_i, y_i)\ i \in [n], g \in G\}$ 训练的线性全连接网络的方向极限。若 G 在 $\mathcal{X}_0$ 上酉作用，则

$$\boldsymbol{\beta}^{\infty}_{\text{steer}} = \boldsymbol{\beta}^{\infty}_{\text{fc}}. \tag{31}$$

换言之，对于群不变二分类，使用线性可转向网络的效果与对非不变模型进行数据增强完全相同。

注 5.2. 要使推论 5.1 成立，$\rho_0 \in \mathrm{GL}(\mathcal{X}_0)$ 为酉矩阵至关重要；否则，在增强数据集上训练的线性全连接网络的极限方向 $\boldsymbol{\beta}_{\mathrm{fc}}^{\infty}$ 通常不是 G-不变的（因此不可能等于 $\boldsymbol{\beta}_{\mathrm{steer}}^{\infty} \in \mathbb{R}_G^{d_0}$）；参见例 5.3。

例 5.3。令 $G = \mathbb{Z}_2 = \{\overline{0}, \overline{1}\}$。考虑定义在 $\mathcal{X}_0 = \mathbb{R}^2$ 上的一个（非酉）G-表示 ρ_0，

$$\rho_0\left(\overline{0}\right) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \rho_0\left(\overline{1}\right) = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 2 & 1 \end{bmatrix}. \tag{32}$$

令 $S = \{(\mathbf{x}\ y)\} = \{((1, 2)^\top, +1)\}$ 为仅含一个点的训练集。根据定理 4.1，线性可转向网络参数化线性预测器的极限方向 $\boldsymbol{\beta}_{\mathrm{steer}}^{\infty}$ 满足

$$\boldsymbol{\beta}_{\mathrm{steer}}^{\infty} \propto \frac{1}{2} \sum_{g \in G} g^\top \boldsymbol{\gamma}^* = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \boldsymbol{\gamma}^* = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ \frac{1}{3} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}, \tag{33}$$

其中，$\boldsymbol{\gamma}^* = (0, \frac{1}{3})^\top$ 是变换后的数据集 $\overline{S} = \{(\mathbf{x}\ y)\} = \{((0,3)^\top, +1)\}$ 的最大间隔支持向量机解。相反，根据推论 4.3，在增强数据集 $S_{\mathrm{aug}} = \{((1,2)^\top, +1), ((-1,4)^\top, +1)\}$ 上训练的线性全连接网络的极限方向 $\boldsymbol{\beta}_{\mathrm{fc}}^{\infty}$ 与 S_{aug} 的最大间隔支持向量机解 $\boldsymbol{\gamma}_{\mathrm{aug}}^*$ 对齐，

$$\boldsymbol{\beta}_{\mathrm{fc}}^{\infty} \propto \boldsymbol{\gamma}_{\mathrm{aug}}^* = \arg\min_{\boldsymbol{\gamma} \in \mathbb{R}^2} \|\boldsymbol{\gamma}\|^2, \quad \text{s.t. } y \langle \mathbf{x}, \boldsymbol{\gamma} \rangle \geq 1, \forall (\mathbf{x}, y) \in S_{\mathrm{aug}} \tag{34}$$

$$= (0.2, 0.4)^\top \not\propto \boldsymbol{\beta}_{\mathrm{steer}}^{\infty}. \tag{35}$$

然而，我们将在下文证明，通过在 $\mathbb{R}^{d_0}$ 上定义新的内积，可以重新建立非酉 ρ_0 情况下线性可转向网络与数据增强之间的等价性，

$$\langle \mathbf{v}, \mathbf{w} \rangle_{\rho_0} := \mathbf{v}^\top \left(\frac{1}{|G|} \sum_{g \in G} \rho_0(g)^\top \rho_0(g) \right)^{1/2} \mathbf{w}, \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{X}_0 = \mathbb{R}^{d_0}. \tag{36}$$

当 ρ_0 为酉时，$\langle \mathbf{v}, \mathbf{w} \rangle_{\rho_0} = \langle \mathbf{v}, \mathbf{w} \rangle$ 是通常的欧几里得内积。借助这一新的内积，我们对增强数据集 $S_{\mathrm{aug}} = \{(g\mathbf{x}_i, y_i) : g \in G, i \in [n]\}$ 上的线性全连接网络及其经验损失进行修改，得到

$$f^{\rho_0}{}_{\text{全连接层}}(\mathbf{x}; \mathbf{W}) := \langle \mathbf{x}\ \mathcal{P}_{\mathrm{fc}}(\mathbf{W}) \rangle_{\rho_0}, \tag{37}$$

$$\mathcal{L}_{\mathcal{P}_{\mathrm{fc}}}^{\rho_0}(\mathbf{W}; S_{\mathrm{aug}}) := \sum_{g \in G} \sum_{i=1}^{n} \ell_{\exp}\left(f_{\mathrm{fc}}^{\rho_0}(g\mathbf{x}_i; \mathbf{W}), y_i\right) = \sum_{g \in G} \sum_{i=1}^{n} \ell_{\exp}\left(\langle g\mathbf{x}_i, \mathcal{P}_{\mathrm{fc}}(\mathbf{W}) \rangle_{\rho_0}, y_i\right). \tag{38}$$

下面的推论表明，对于 $\mathcal{X}_0$ 上的非酉 ρ_0，线性可转向网络的隐式偏置再次与经过数据增强训练的修正线性全连接网络 $f^{\rho_0}\mathrm{fc}\ (\mathbf{x}; \mathbf{W})$ 的隐式偏置相同。

推论 5.4. 令 $\boldsymbol{\beta}_{\mathrm{steer}}^{\infty}$ 与推论 5.1 中的相同。令 $\boldsymbol{\beta}^{\rho_0,\ \infty\ \text{全连接}} = \lim_{t \to \infty} \frac{\boldsymbol{\beta}_{\mathrm{fc}}^{\rho_0}(t)}{\|\boldsymbol{\beta}_{\mathrm{fc}}^{\rho_0}(t)\|}$ 为 $\boldsymbol{\beta}^{\rho_0}$ 全连接 $(t) = \mathcal{P}_{\mathrm{fc}}(\mathbf{W}(t))$ 在修正经验损失 $\mathcal{L}_{\mathcal{P}_{\mathrm{fc}}}^{\rho_0}(\mathbf{W};\ S_{\mathrm{aug}})$ (38) 的梯度流下的极限方向，其中该损失对应于增强数据集 S_{aug} 上的线性全连接网络。那么

$$\boldsymbol{\beta}_{\mathrm{steer}}^{\infty} \propto \left(\frac{1}{|G|} \sum_{g \in G} \rho_0(g)^\top \rho_0(g) \right)^{1/2} \boldsymbol{\beta}_{\mathrm{fc}}^{\rho_0, \infty}. \tag{39}$$

因此，对于所有的 $\mathbf{x}$，有 $\langle \mathbf{x}\ \boldsymbol{\beta}_{\mathrm{steer}}^{\infty} \rangle \propto \langle \mathbf{x}\ \boldsymbol{\beta}^{\rho_0,\ \infty\ \mathrm{fc}} \rangle_{\rho_0} \in \mathbb{R}^{d_0}$。

6 改进的间隔与泛化

本节将展示线性可转向网络相比其非不变对应网络，在间隔与泛化方面的改进。下文中，我们假设 ρ_0 为酉矩阵。

下面的定理表明，在增强数据集 S_{aug} 上，线性可转向网络参数化预测器 $\boldsymbol{\beta}_{\mathrm{steer}}^{\infty}$ 的间隔始终大于线性全连接网络 $\boldsymbol{\beta}_{\mathrm{fc}}^{\infty}$ 的间隔，这表明可转向网络参数化分类器具有更强的 L^2-鲁棒性。

定理 6.1. 设 $\boldsymbol{\beta}_{\text{steer}}^{\infty}$ 为在原始数据集 $S=\{(\mathbf{x}_i, y_i)\ i\in[n]\}$ 上训练的线性可转向网络参数化预测器的方向极限；设 $\boldsymbol{\beta}_{\text{fc}}^{\infty}$ 为在同一数据集 S 上训练的线性全连接网络的方向极限。分别设 M_{steer} 和 M_{fc} 为 $\boldsymbol{\beta}_{\text{steer}}^{\infty}$ 和 $\boldsymbol{\beta}_{\text{fc}}^{\infty}$ 在增强数据集 $S_{\text{aug}}=\{(g\mathbf{x}_i, y_i):\ i\in[n], g\in G\}$ 上的（带符号）间隔，即，

$$M_{\text{steer}} \coloneqq \min_{i\in[n], g\in G} y_i \left\langle \boldsymbol{\beta}_{\text{steer}}^{\infty}, g\mathbf{x}_i \right\rangle, \quad M_{\text{fc}} \coloneqq \min_{i\in[n], g\in G} y_i \left\langle \boldsymbol{\beta}_{\text{fc}}^{\infty}, g\mathbf{x}_i \right\rangle. \tag{40}$$

那么，我们总有 $M_{\text{steer}} \geq M_{\text{fc}}$。

最后，我们旨在量化在线性可分且具有群不变性的分布上进行二分类时，线性可转向网络相较于全连接网络所带来的泛化提升。

定义 6.2. 若分布 $\mathcal{D}$ 在 $\mathbb{R}^{d_0}\times\{\pm 1\}$ 上满足以下条件，则称其为线性可分：

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y\left\langle \mathbf{x}, \boldsymbol{\beta}\right\rangle \geq 1\right] = 1. \tag{41}$$

容易验证（由引理 E.1 可知），如果 $\mathcal{D}$ 具有 G不变性且线性可分，那么 $\mathcal{D}$ 可以由一个具有 G不变性的线性分类器分开。下述定理给出了在线性可分且具有群不变性的二分类问题中，线性可转向网络的泛化界。

定理 6.3. 设 $\mathcal{D}$ 是 $\mathbb{R}^{d_0}\times\{\pm 1\}$ 上的一个具有 G不变性的分布，且可由不变分类器 $\boldsymbol{\beta}_0 \in \mathbb{R}_G^{d_0}$线性分开。定义

$$\overline{R} = \inf\left\{r>0 : \|\overline{\mathbf{x}}\| \leq r \text{ with probability } 1\right\}. \tag{42}$$

设 $S=\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ 是来自 $\mathcal{D}$的独立同分布样本，且 $\boldsymbol{\beta}_{\text{steer}}^{\infty}$ 是通过梯度流在 S上训练的、由可转向网络参数化的线性预测器的极限方向。那么，对于任意 $\delta>0$，以至少 $1-\delta$ 的概率（针对随机样本 $S\sim\mathcal{D}^n$），有

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y \neq \operatorname{sign}\left(\langle \mathbf{x}, \boldsymbol{\beta}_{\text{steer}}^{\infty}\rangle\right)\right] \leq \frac{2\overline{R}\|\boldsymbol{\beta}_0\|}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{43}$$

注释 6.4。相比之下，设 $\boldsymbol{\beta}_{\text{fc}}^{\infty}$ 是通过全连接网络参数化、在 S上训练的线性预测器的极限方向。那么，以至少 $1-\delta$的概率，我们有

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y \neq \operatorname{sign}\left(\langle \mathbf{x}, \boldsymbol{\beta}_{\text{fc}}^{\infty}\rangle\right)\right] \leq \frac{2R\|\boldsymbol{\beta}_0\|}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}, \tag{44}$$

其中 $R=\inf\{r>0:\ \|\mathbf{x}\|\leq r$，概率为 $1\}$。这是最大间隔支持向量机的经典泛化结果（例如，参见 Shalev-Shwartz 和 Ben-David *[2014]*。）式 (44) 也可以视为式 (43) 的一个特例，因为全连接网络是一个具有 $G=\{e\}$ 的 G-卷积神经网络（参见注释4.2），因此 $\mathbf{x}=\mathbf{x}$ 和 $R=\overline{R}$。

根据命题 3.6，映射 $\mathbf{x}\to\mathbf{x}$ 是一个正交投影，因此我们始终有 $\overline{R}\leq R$。因此，式 (43) 中可转向网络的泛化界始终小于式 (44) 中全连接网络的泛化界。

注释 6.5。比较式 (43) 和式 (44) 可知，线性可转向网络的泛化能力提升并不一定取决于群的大小 $|G|$。相反，它取决于分布 $\mathcal{D}$ 的支撑集与子空间 $\mathbb{R}_G^{d_0}$ 的距离，因此 $\overline{R}$ 可能远小于 R。事实上，如果 $\mathcal{D}$ 的支撑集包含于 $R_G^{d_0}$，那么 $\overline{R}=R$，且可转向网络不会获得任何泛化增益。这与定理 4.1 一致，因为在这种情况下，变换后的数据集 $\overline{S}=\{(\mathbf{x}_i, y_i):\ i\in[n]\}$ 与原始数据集 S 相同。

7 结论与未来工作

本文研究了群不变二分类中一般线性群等变可转向网络上的梯度流的隐式偏置。我们的研究结果表明，参数化预测器会沿一个方向收敛，该方向与唯一的群不变分类器一致；该分类器具有一个取决于输入表示的最大间隔。作为主要结果的推论，

我们在所研究的设定中证明了数据增强与使用可转向网络进行学习之间的等价性。最后，我们证明，在线性可转向网络的改进间隔与泛化界方面，其表现优于非不变对应网络。

我们结果的一个局限是，本文研究的梯度流的隐式偏置仅在渐近意义下成立，而且收敛至方向极限的速度可能极其缓慢。这与例如 [Soudry 等，2018、Yun 等人， 2021]的研究结果一致。理解非渐近情形下梯度流的行为，是未来一项重要的研究方向。此外，在当前设定中，我们假设第一层等变映射由 G-提升映射给出，从而使线性可转向网络具有足够的容量，能够参数化所有 G-不变线性分类器。去除这一假设后，分析可转向网络的隐式偏置将很有意义。

致谢

本研究部分由美国国家科学基金会通过 DMS-2052525 和 DMS-2140982 资助。

参考文献

Z. Allen-Zhu、Y. Li 和 Z. Song。通过过参数化实现深度学习的收敛理论
《国际机器学习会议论文集》，第242–252页。PMLR，2019年。

X. Cheng、Q. Qiu、R. Calderbank 和 G. Sapiro。RotDCF：用于旋转等变深度网络的卷积滤波器分解。载于《国际学习表征会议》，2019年。网址：https://openreview.net/forum?id=H1gTEj09FX。

L. Chizat 和 F. Bach。关于使用最优传输的过参数化模型中梯度下降的全局收敛。《神经信息处理系统进展》，第31卷，2018年。

T. Cohen 和 M. Welling。群等变卷积网络。载于 M. F. Balcan 和 K. Q. Weinberger（编），《第33届国际机器学习会议论文集》，《机器学习研究会议论文集》第48卷，第2990–2999页，纽约，美国纽约，2016年6月20日至22日。PMLR。网址：https://proceedings.mlr.press/v48/cohenc16.html。T. S. Cohen 和 M. Welling。可转向卷积神经网络。载于《国际学习表征会议》，2017年。网址：https://openreview.net/forum?id=rJQKYt5ll。T. S. Cohen、M. Geiger 和 M. Weiler。齐性空间上等变卷积神经网络的一般理论。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alché-Buc、E. Fox 和 R. Garnett（编），《神经信息处理系统进展》，第32卷。Curran Associates, Inc.，2019年。网址：https://proceedings.neurips.cc/paper/2019/file/b9cfe8b6042cf759dc4c0cccb27a67 37-Paper.pdf ·

S. Du、J. Lee、H. Li、L. Wang 和 X. Zhai。梯度下降找到深度神经网络的全局极小值。载于《国际机器学习会议》，第 1675–1685 页。PMLR，2019年。B. Elesedy 和 S. Zaidi。等变模型可证明具有严格的泛化收益。载于《国际机器学习会议》，第 2959–2969 页。PMLR，2021年。S. Gunasekar、J. Lee、D. Soudry 和 N. Srebro。以优化几何刻画隐式偏置。载于《国际机器学习会议》，第 1832–1841 页。PMLR，2018a。S. Gunasekar、J. D. Lee、D. Soudry 和 N. Srebro。线性卷积网络中梯度下降的隐式偏置。载于《神经信息处理系统进展》，第 31 卷，2018b。A. Jacot、F. Gabriel 和 C. Hongler。神经切核：神经网络中的收敛与泛化。载于《神经信息处理系统进展》，第 31 卷，2018年。Z. Ji 和 M. Telgarsky。梯度下降使深度线性网络的各层对齐。载于《国际学习表征会议》，2019a。网址：https://openreview.net/forum?id=HJflg30qKX。

Z. Ji 和 M. Telgarsky。梯度下降在不可分数据上的隐式偏置。载于《学习理论会议》，第 1772–1798 页。PMLR，2019b。Z. Ji 和 M. Telgarsky。深度学习中的方向收敛与对齐。载于《神经信息处理系统进展》，第 33 卷：17176–17186 页，2020年。

H. Lawrence、K. Georgiev、A. Dienes 和 B. T. Kiani。线性等变网络的隐式偏置。arXiv 预印本 arXiv:2110.06084，2021年。

K. Lyu 和 J. Li。梯度下降最大化齐次神经网络的间隔。载于《国际学习表征会议》，2020年。网址：https://openreview.net/forum?id=SJeLIgBKPS。

S. Mei、T. Misiakiewicz 和 A. Montanari。双层神经网络的平均场理论：与维度无关的界和核极限。载于《学习理论会议》，第 2388–2464 页。PMLR，2019年。

M. Mohri 和 A. M. Medina。带保留价的第二价格拍卖中收益优化的学习理论与算法。载于《国际机器学习会议》，第 262–270 页。PMLR，2014年。

M. S. Nacson、N. Srebro 和 D. Soudry。《可分数据上的随机梯度下降：固定学习率下的精确收敛》。载于第22届人工智能与统计国际会议，第3051–3059页。PMLR，2019年。

A. Sannai、M. Imaizumi 和 M. Kawano。《通过商特征空间改进群不变/等变深度网络的泛化界》。载于人工智能中的不确定性，第771–780页。PMLR，2021年。

S. Shalev-Shwartz 和 S. Ben-David。《理解机器学习：从理论到算法》。剑桥大学出版社，2014年。

J. Sokolic、R. Giryes、G. Sapiro 和 M. Rodrigues。《不变分类器的泛化误差》。载于人工智能与统计，第1094–1103页。PMLR，2017年。

I. Sosnovik、M. Szmaja 和 A. Smeulders。尺度等变可转向网络。载于《国际学习表征会议》，2020年。网址：https://openreview.net/forum?id=HJgpugrKPS。

D. Soudry、E. Hoffer、M. S. Nacson、S. Gunasekar 和 N. Srebro。可分数据上梯度下降的隐式偏置。《机器学习研究杂志》，19(1)：2822–2878，2018年。G. Vardi。深度学习算法中的隐式偏置。arXiv 预印本 arXiv:2208.12591，2022年。M. Weiler 和 G. Cesa。通用 E(2)等变可转向卷积神经网络。载于 H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alché-Buc、E. Fox 和 R. Garnett 编，《神经信息处理系统进展》，第32卷。Curran Associates, Inc.，2019年。网址：https://proceedings.neurips.cc/paper/2019/file/45d6637b718d0f24a237069fe41b0db4-Paper.pdf.

M. Weiler、F. A. Hamprecht 和 M. Storath。《学习用于旋转等变卷积神经网络的可转向滤波器》。载于《IEEE计算机视觉与模式识别会议论文集》，第 849–858 页，2018年。D. E. Worrall、S. J. Garbin、D. Turmukhambetov 和 G. J. Brostow。《调和网络：深度平移与旋转等变性》。载于《IEEE计算机视觉与模式识别会议论文集》，第 5028–5037 页，2017年。C. Yun、S. Krishnan 和 H. Mobahi。《对训练线性神经网络中的隐式偏置的统一观点》。载于《国际学习表征会议》，2021年。网址：https://openreview.net/forum?id=ZsZM-4iMQkH。W. Zhu、Q. Qiu、R. Calderbank、G. Sapiro 和 X. Cheng。《采用分解卷积滤波器的尺度-平移等变网络》。机器学习研究期刊，23(68)：1–45，2022年。

第3节中的证明

命题 3.2。设 $f_{\text{steer}}(\mathbf{x};\mathbf{W})$ 为满足假设 3.1 的线性可转向网络，其中 $\mathbf{W} = [\mathbf{w}_l]_{l=1}^{L} \in \mathcal{W}_{\text{steer}}$ 是所有模型参数的集合。存在一个多线性映射 $M\colon (\mathbf{w}_2,\cdots,\mathbf{w}_L) \mapsto M(\mathbf{w}_2,\cdots,\mathbf{w}_L) \in \mathbb{R}^{d_1}$，使得对所有 $\mathbf{x}\in\mathbb{R}^{d_0}$ 和 $\mathbf{W} \in \mathcal{W}_{\text{steer}}$，

$$f_{\text{steer}}(\mathbf{x};\mathbf{W}) = f_{\text{steer}}(\overline{\mathbf{x}};\mathbf{W}) = \langle \overline{\mathbf{x}}, \mathbf{w}_1 M(\mathbf{w}_2,\cdots,\mathbf{w}_L)\rangle\,, \tag{45}$$

其中 $\underline{\mathbf{x}} := \frac{1}{|G|}\sum_{g\in G} g\mathbf{x}$ 是 $\mathbf{x}$ 的群轨道 $\in \mathbb{R}^{d_0}$ 上所有元素的平均值。

证明。由于线性可转向网络 $f_{\text{steer}}(\mathbf{x};\mathbf{W})$ 是 G-不变的，因此对于所有 $g\in G$，都有 $f_{\text{steer}}(\mathbf{x};\mathbf{w}) = f_{\text{steer}}(g\mathbf{x};\mathbf{w})$。因此，

$$f_{\text{steer}}(\mathbf{x};\mathbf{W}) = \frac{1}{|G|}\sum_{g\in G} f_{\text{steer}}(g\mathbf{x};\mathbf{W}) = f_{\text{steer}}\left(\frac{1}{|G|}\sum_{g\in G} g\mathbf{x};\mathbf{w}\right) = f_{\text{steer}}(\overline{\mathbf{x}};\mathbf{W}), \tag{46}$$

其中，第二个等式是由于 $f_{\text{steer}}(\mathbf{x};\mathbf{w})$ 关于 $\mathbf{x}$ 的线性。由此可得

$$\begin{aligned} f_{\text{steer}}(\mathbf{x};\mathbf{W}) = f_{\text{steer}}(\overline{\mathbf{x}};\mathbf{W}) &= \Psi_L^{\text{steer}}(\cdots\Psi_2^{\text{steer}}(\Psi_1^{\text{steer}}(\overline{\mathbf{x}};\mathbf{w}_1);\mathbf{w}_2)\cdots;\mathbf{w}_L) &(47)\\ &= \left\langle \Psi_1^{\text{steer}}(\overline{\mathbf{x}};\mathbf{w}), \Phi(\mathbf{w}_2,\cdots,\mathbf{w}_L)\right\rangle, &(48)\end{aligned}$$

其中 $\Phi(\mathbf{w}_2,\cdots,\mathbf{w}_L) \in \mathcal{X}_1 = \left(\mathbb{R}^{d_1}\right)^G$ 关于 $(\mathbf{w}_1,\cdots,\mathbf{w}_L)$ 是多线性的。根据假设 3.1，有

$$\begin{aligned} f_{\text{steer}}(\mathbf{x};\mathbf{W}) &= \sum_{g\in G}\left\langle \Psi_1^{\text{steer}}(\overline{\mathbf{x}};\mathbf{w})(g), \Phi(\mathbf{w}_2,\cdots,\mathbf{w}_L)(g)\right\rangle &(49)\\ &= \sum_{g\in G}\left\langle \mathbf{w}_1^\top g^{-1}\overline{\mathbf{x}}, \Phi(\mathbf{w}_2,\cdots,\mathbf{w}_L)(g)\right\rangle &(50)\\ &= \sum_{g\in G}\left\langle \mathbf{w}_1^\top \overline{\mathbf{x}}, \Phi(\mathbf{w}_2,\cdots,\mathbf{w}_L)(g)\right\rangle &(51)\\ &= \left\langle \overline{\mathbf{x}}, \mathbf{w}_1\left[\sum_{g\in G}\Phi(\mathbf{w}_2,\cdots,\mathbf{w}_L)(g)\right]\right\rangle &(52)\\ &= \langle \overline{\mathbf{x}}, \mathbf{w}_1 M(\mathbf{w}_2,\cdots,\mathbf{w}_L)\rangle, &(53)\end{aligned}$$

其中 $M(\mathbf{w}_2,\cdots,\mathbf{w}_L) := \sum_{g\in G}\Phi(\mathbf{w}_2,\cdots,\mathbf{w}_L)(g)$ 关于 $(\mathbf{w}_2,\cdots,\mathbf{w}_L$ 也是多线性的). □

命题 3.6. 设 $\mathbb{R}^{d_0}_G \subset \mathbb{R}^{d_0}$ 是 G-不变线性预测器的子空间，即 $\mathbb{R}^{d_0}_G = \{\boldsymbol{\beta}\in\mathbb{R}^{d_0}:\ \boldsymbol{\beta}^\top\mathbf{x} = \boldsymbol{\beta}^\top g\mathbf{x}\ \forall\mathbf{x}\in\mathbb{R}^{d_0},\ \forall g\in G\}$。则

(a) $\mathbb{R}^{d_0}_G$ 可由下式刻画

$$\mathbb{R}^{d_0}_G = \bigcap_{g\in G}\ker(I - g^\top) = \text{Range}\left(\frac{1}{|G|}\sum_{g\in G} g^\top\right). \tag{54}$$

(b) 设 $\mathcal{A}:\ \mathbb{R}^{d_0}\to\mathbb{R}^{d_0}$ 为群平均映射，

$$\mathcal{A}(\boldsymbol{\beta}) := \overline{\boldsymbol{\beta}} = \frac{1}{|G|}\sum_{g\in G} g\boldsymbol{\beta}. \tag{55}$$

则其伴随算子 $\mathcal{A}^\top:\ \boldsymbol{\beta}\mapsto\frac{1}{|G|}\sum_{g\in G} g^\top\boldsymbol{\beta}$ 是从 $\mathbb{R}^{d_0}$ 到 $\mathbb{R}^{d_0}_G$ 的投影算子。换言之，$\text{Range}(\mathcal{A}^\top) = \mathbb{R}^{d_0}_G$ 且 $\mathcal{A}^\top\circ\mathcal{A}^\top = \mathcal{A}^\top$。

(c) 如果 G 在 $\mathcal{X}_0$上酉地作用，即 $\rho_0(g^{-1}) = \rho_0(g)^\top$，那么 $\mathcal{A} = \mathcal{A}^\top$ 是自伴的。这意味着 $\mathcal{A}:\ \boldsymbol{\beta}\mapsto\overline{\boldsymbol{\beta}}$ 是从 $\mathbb{R}^{d_0}$ 到 $\mathbb{R}^{d_0}_G$的正交投影。特别地，我们有

$$\overline{\boldsymbol{\beta}} = \boldsymbol{\beta} \iff \boldsymbol{\beta}\in\mathbb{R}^{d_0}_G,\quad \textit{and}\ \|\overline{\boldsymbol{\beta}}\| \le \|\boldsymbol{\beta}\|, \forall\boldsymbol{\beta}\in\mathbb{R}^{d_0}. \tag{56}$$

证明。为证明 (a)，对于第一个等式，向量 $\boldsymbol{\beta} \in \mathbb{R}_G^{d_0}$ 当且仅当 $0 = \boldsymbol{\beta}^\top(I-g)\mathbf{x} = \langle (I-g)^\top\boldsymbol{\beta}, \mathbf{x}\rangle, \forall \mathbf{x} \in \mathbb{R}^{d_0}, \forall g \in G$。这等价于 $\boldsymbol{\beta} \in \bigcap_{g\in G} \ker(I-g^\top)$，因此$\mathbb{R}_G^{d_0} = \bigcap_{g\in G}\ker(I-g^\top)$。

对于第二个等式，若 $\boldsymbol{\beta} \in \bigcap_{g\in G} \ker(I-g^\top)$，则 $(I-g^\top)\boldsymbol{\beta} = 0, \forall g \in G$。因此

$$0 = \frac{1}{|G|}\sum_{g\in G}(I-g^\top)\boldsymbol{\beta} = \boldsymbol{\beta} - \frac{1}{|G|}\sum_{g\in G} g^\top\boldsymbol{\beta} \tag{57}$$

$$\Longrightarrow \boldsymbol{\beta} = \frac{1}{|G|}\sum_{g\in G} g^\top\boldsymbol{\beta} \in \text{Range}\left(\frac{1}{|G|}\sum_{g\in G} g^\top\right) \tag{58}$$

另一方面，若对于某个 $\mathbf{y}\in\mathbb{R}^{d_0}$，$\boldsymbol{\beta} = \frac{1}{|G|}\sum_{h\in G} h^\top\mathbf{y}$，则对所有 $g \in G$，

$$\left(I-g^\top\right)\boldsymbol{\beta} = \left(I-g^\top\right)\frac{1}{|G|}\sum_{h\in G} h^\top\mathbf{y} = \frac{1}{|G|}\sum_{h\in G} h^\top\mathbf{y} - \frac{1}{|G|}\sum_{h\in G}(hg)^\top\mathbf{y} = 0. \tag{59}$$

因此 $\boldsymbol{\beta} \in \bigcap_{g\in G}\ker(I-g^\top)$。

点 (b) 可由 (54) 和 (57) 轻易推出。

为证明 (c)，注意到

$$\mathcal{A}^\top = \frac{1}{|G|}\sum_{g\in G} g^\top = \frac{1}{|G|}\sum_{g\in G} g^{-1} = \frac{1}{|G|}\sum_{g\in G} g = \mathcal{A}. \tag{60}$$

因此 $\mathcal{A} = \mathcal{A}^\top$ 是自伴的。结合点 (b)，可知 $\mathcal{A} = \mathcal{A}^\top:\ \boldsymbol{\beta} \mapsto \overline{\boldsymbol{\beta}}$ 是从 $\mathbb{R}^{d_0}$ 到 $\mathbb{R}_G^{d_0}$的正交投影。 □

B 定理 4.1 的证明

定理 4.1. 在假设 3.1 和假设 3.7 下，设 $\boldsymbol{\beta}_{\text{steer}}(t) = \mathcal{P}_{\text{steer}}(\mathbf{W}(t))$ 是由线性可转向网络参数化的不变线性预测器在数据集 $S = \{(\mathbf{x}_i, y_i):\ i\in[n]\}$ 上通过梯度流训练得到的时间演化；参见公式 (25)。那么

(a) 方向极限 $\boldsymbol{\beta}_{\text{steer}}^\infty = \lim_{t\to\infty} \frac{\boldsymbol{\beta}_{\text{steer}}(t)}{\|\boldsymbol{\beta}_{\text{steer}}(t)\|}$ 存在，且 $\boldsymbol{\beta}_{\text{steer}}^\infty \propto \frac{1}{|G|}\sum_{g\in G} g^\top\boldsymbol{\gamma}^*$，其中 $\boldsymbol{\gamma}^*$ 是针对变换后的数据 $\overline{S} = \{(\mathbf{x}_i, y_i):\ i\in[n]\}$的最大-$L^2$-间隔支持向量机解：

$$\boldsymbol{\gamma}^* = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^{d_0}} \|\boldsymbol{\gamma}\|^2, \quad \text{s.t. } y_i\langle\overline{\mathbf{x}}_i, \boldsymbol{\gamma}\rangle \geq 1, \forall i\in[n]. \tag{61}$$

此外，若 G 在输入空间 $\mathcal{X}_0$上酉地作用，即 $g^{-1} = g^\top$，则 $\boldsymbol{\beta}_{\text{steer}}^\infty \propto \boldsymbol{\gamma}^*$。

（b）等价地，$\boldsymbol{\beta}_{\text{steer}}^\infty$ 与问题的唯一极小化解 $\boldsymbol{\beta}^*$ 成比例

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{d_0}} \|\text{Proj}_{\text{Range}(\mathcal{A})}\boldsymbol{\beta}\|^2, \quad \text{s.t. } \boldsymbol{\beta}\in\mathbb{R}_G^{d_0},\ \textit{and}\ y_i\langle\mathbf{x}_i, \boldsymbol{\beta}\rangle \geq 1, \forall i\in[n], \tag{62}$$

其中，$\text{Proj}^{\text{Range}(\mathcal{A})}$ 是从 $\mathbb{R}^{d_0}$ 到值域（$\text{Range}(\mathcal{A}) = \text{Range}\left(\frac{1}{|G|}\sum_{g\in G} g\right)$）的投影。

此外，如果 G 在 $\mathcal{X}_0$上酉地作用，那么

$$\boldsymbol{\beta}_{\text{steer}}^\infty \propto \boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{d_0}} \|\boldsymbol{\beta}\|^2, \quad \text{s.t. } \boldsymbol{\beta}\in\mathbb{R}_G^{d_0},\ \textit{and}\ y_i\langle\mathbf{x}_i, \boldsymbol{\beta}\rangle \geq 1, \forall i\in[n]. \tag{63}$$

也就是说，在所有 G-不变线性预测器中，$\boldsymbol{\beta}_{\text{steer}}^\infty$ 实现了最大的 L^2-间隔。

在证明定理 4.1之前，我们需要以下引理，该引理适用于一般的 L-齐次网络，线性可转向网络是其特例。请注意，在下文中，$\|\cdot\|$ 始终表示将张量视为一维向量时的欧几里得范数。

引理 B.1（改述自 Lyu 和 Li [2020] 以及 Ji 和 Telgarsky [2020]）。在假设 3.7 下，我们得到以下关于方向收敛和对齐的结果。

(a) $\mathcal{L}_{\mathcal{P}_{\text{steer}}}(\mathrm{W}(t);\, S) \to 0$，当 $t \to \infty$ 时。因此，$\|\mathrm{W}(t)\| \to \infty$ 以及 $\|\mathcal{P}_{\text{steer}}(\mathrm{W}(t))\| \to \infty$。

(b) 参数 $\mathrm{W}(t)$ 和梯度 $\nabla_{\mathbf{W}}\mathcal{L}_{\mathcal{P}_{\text{steer}}}(\mathrm{W}(t))$ 的方向收敛与对齐，

$$\lim_{t\to\infty} \frac{\mathbf{W}(t)}{\|\mathbf{W}(t)\|} = \mathbf{W}^\infty = -\lim_{t\to\infty} \frac{\nabla_{\mathbf{W}}\mathcal{L}_{\mathcal{P}_{\text{steer}}}(\mathbf{W}(t);S)}{\|\nabla_{\mathbf{W}}\mathcal{L}_{\mathcal{P}_{\text{steer}}}(\mathbf{W}(t);S)\|}, \tag{64}$$

对于某个 $\mathrm{W}^\infty \in \mathcal{W}_{\text{steer}}$，使得 $\|\mathrm{W}^\infty\| = 1$。

(c) 极限 W^∞ 沿着约束优化问题的一阶驻点方向

$$\min_{\mathbf{W}} \|\mathbf{W}\|^2, \quad \text{s.t. } y_i \langle \mathbf{x}_i, \mathcal{P}_{\text{steer}}(\mathbf{W})\rangle \geq 1,\ \forall i \in [n]. \tag{65}$$

换句话说，存在一个缩放因子 $\tau > 0$，使得 $\tau\mathrm{W}^\infty$ 满足式 (65) 的 Karush-Kuhn-Tucker (KKT) 条件。

定理 4.1 的证明。证明的第一部分受 Gunasekar 等人 [2018b] 的启发。为证明 (a)，令 $\mathrm{W}^\infty = [\mathrm{w}_l^\infty]_{l=1}^L = \lim_{t\to\infty} \mathrm{w}_{(t)}/\|\mathbf{w}_{(t)}\|$ 为 ~~$\mathbf{W}(t)$~~ 的极限方向，并根据引理 B.1，令 $\widetilde{\mathrm{W}}^\infty = [\widetilde{\mathrm{w}}_l^\infty]_{l=1}^L = \tau\mathrm{W}^\infty, \tau > 0$ 为问题 (65) 的驻点。对于 W^∞，KKT 条件意味着存在对偶变量 $\alpha_i \geq 0, i \in [n]$，使得

- 原始可行性：

$$y_i \left\langle \mathcal{P}_{\text{steer}}(\widetilde{\mathbf{W}}^\infty), \mathbf{x}_i \right\rangle \geq 1, \quad \forall i \in [n]. \tag{66}$$

- 驻定性：

$$\widetilde{\mathbf{W}}^\infty = \nabla_{\mathbf{W}}\mathcal{P}_{\text{steer}}(\widetilde{\mathbf{W}}^\infty) \cdot \left(\sum_{i=1}^n \alpha_i y_i \mathbf{x}_i\right) \tag{67}$$

- 互补松弛：

$$y_i \left\langle \mathcal{P}_{\text{steer}}(\widetilde{\mathbf{W}}^\infty), \mathbf{x}_i \right\rangle > 1 \implies \alpha_i = 0. \tag{68}$$

我们声称，$\boldsymbol{\gamma}^* = \overline{\mathcal{P}}_{\text{steer}}(\mathrm{W}^\infty)$ 是以下问题的一个驻点

$$\min_{\boldsymbol{\gamma}\in\mathbb{R}^{d_0}} \|\boldsymbol{\gamma}\|^2, \quad \text{s.t. } y_i \langle \overline{\mathbf{x}}_i, \boldsymbol{\gamma}\rangle \geq 1,\ \forall i \in [n]. \tag{69}$$

也就是说，存在 $\tilde{\alpha}_i \geq 0, i \in [n]$，使得以下条件成立。

- 原始可行性：

$$y_i \left\langle \overline{\mathcal{P}}_{\text{steer}}(\widetilde{\mathbf{W}}^\infty), \overline{\mathbf{x}}_i \right\rangle \geq 1, \quad \forall i \in [n]. \tag{70}$$

- 驻定性：

$$\overline{\mathcal{P}}_{\text{steer}}(\widetilde{\mathbf{W}}^\infty) = \sum_{i=1}^n \tilde{\alpha}_i y_i \overline{\mathbf{x}}_i \tag{71}$$

- 互补松弛：

$$y_i \left\langle \overline{\mathcal{P}}_{\text{steer}}(\widetilde{\mathbf{W}}^\infty), \overline{\mathbf{x}}_i \right\rangle > 1 \implies \tilde{\alpha}_i = 0. \tag{72}$$

事实上，由式 (66) 和 $\left\langle \mathcal{P}_{\text{steer}}(\widetilde{\mathbf{W}}^{\infty}),\ \mathbf{x}_i \right\rangle = \left\langle \overline{\mathcal{P}}_{\text{steer}}(\widetilde{\mathbf{W}}^{\infty}),\ \mathbf{x}_i \right\rangle$ 可知，式 (70) 成立。此外，根据式 (12) 中 $\mathcal{P}_{\text{steer}}(\mathbf{W})$ 和 $\overline{\mathcal{P}}_{\text{steer}}(\mathbf{W})$ 的定义，对于任意 $\mathbf{z} \in \mathbb{R}^{d_0}$,

$$\nabla_{\mathbf{w}_1} \overline{\mathcal{P}}_{\text{steer}}(\mathbf{W}) \cdot \mathbf{z} = \mathbf{z}[M(\mathbf{w}_2, \cdots, \mathbf{w}_L)]^{\top} \tag{73}$$

$$\nabla_{\mathbf{w}_1} \mathcal{P}_{\text{steer}}(\mathbf{W}) \cdot \mathbf{z} = \left(\frac{1}{|G|} \sum_{g \in G} g \right) \mathbf{z}[M(\mathbf{w}_2, \cdots, \mathbf{w}_L)]^{\top} = \overline{\mathbf{z}}[M(\mathbf{w}_2, \cdots, \mathbf{w}_L)]^{\top}. \tag{74}$$

因此，式 (67) 意味着

$$\widetilde{\mathbf{w}}_1^{\infty} = \nabla_{\mathbf{w}_1} \mathcal{P}_{\text{steer}}(\widetilde{\mathbf{W}}^{\infty}) \cdot \left(\sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \right) = \left(\sum_{i=1}^{n} \alpha_i y_i \overline{\mathbf{x}}_i \right) \cdot [M(\widetilde{\mathbf{w}}_2^{\infty}, \cdots, \widetilde{\mathbf{w}}_L^{\infty})]^{\top} \tag{75}$$

因此

$$\begin{aligned} \overline{\mathcal{P}}_{\text{steer}}(\widetilde{\mathbf{W}}^{\infty}) = \widetilde{\mathbf{w}}_1^{\infty} M(\widetilde{\mathbf{w}}_2^{\infty}, \cdots, \widetilde{\mathbf{w}}_L^{\infty}) &= \left(\sum_{i=1}^{n} \alpha_i y_i \overline{\mathbf{x}}_i \right) \|M(\widetilde{\mathbf{w}}_2^{\infty}, \cdots, \widetilde{\mathbf{w}}_L^{\infty})\|^2 \\ &= c \sum_{i=1}^{n} \alpha_i y_i \overline{\mathbf{x}}_i, \end{aligned} \tag{76}$$

其中 $c = \|M(\widetilde{\mathbf{w}}_2^{\infty}, \cdots, \widetilde{\mathbf{w}}_L^{\infty})\|^2 \geq 0$。由 (70) 可知，$\overline{\mathcal{P}}_{\text{steer}}(\widetilde{\mathbf{W}}^{\infty}) \neq 0$，因此 $c > 0$。令 $\tilde{\alpha}_i = c\alpha_i, \forall i \in [n]$。于是 $\tilde{\alpha}_i \geq 0$，并且由于式 (76)，驻定性 (71) 得以满足；此外，由于 $\tilde{\alpha}_i$ 对所有 $i \in [n]$ 都是 α_i 的正比例缩放，互补松弛条件也成立。因此，$\boldsymbol{\gamma}^* = \overline{\mathcal{P}}_{\text{steer}}(\widetilde{\mathbf{W}}^{\infty})$ 是 (69) 的一个驻点。由于问题 (69) 是强凸的，$\boldsymbol{\gamma}^* = \overline{\mathcal{P}}_{\text{steer}}(\widetilde{\mathbf{W}}^{\infty})$ 实际上是 (69) 的唯一极小化解。因此，预测器 $\mathcal{P}_{\text{steer}}(\mathbf{W}(t))$ 的极限方向为

$$\boldsymbol{\beta}_{\text{steer}}^{\infty} = \lim_{t \to \infty} \frac{\mathcal{P}_{\text{steer}}(\mathbf{W}(t))}{\|\mathcal{P}_{\text{steer}}(\mathbf{W}(t))\|} = \lim_{t \to \infty} \frac{\left(\frac{1}{|G|} \sum_{g \in G} g^{\top} \right) \overline{\mathcal{P}}_{\text{steer}}(\mathbf{W}(t))}{\left\| \left(\frac{1}{|G|} \sum_{g \in G} g^{\top} \right) \overline{\mathcal{P}}_{\text{steer}}(\mathbf{W}(t)) \right\|} \tag{77}$$

$$= \lim_{t \to \infty} \frac{\left(\frac{1}{|G|} \sum_{g \in G} g^{\top} \right) \overline{\mathcal{P}}_{\text{steer}} \left(\frac{\tau \mathbf{W}(t)}{\|\mathbf{W}(t)\|} \right)}{\left\| \left(\frac{1}{|G|} \sum_{g \in G} g^{\top} \right) \overline{\mathcal{P}}_{\text{steer}} \left(\frac{\tau \mathbf{W}(t)}{\|\mathbf{W}(t)\|} \right) \right\|} = \frac{\left(\frac{1}{|G|} \sum_{g \in G} g^{\top} \right) \overline{\mathcal{P}}_{\text{steer}} \left(\widetilde{\mathbf{W}}^{\infty} \right)}{\left\| \left(\frac{1}{|G|} \sum_{g \in G} g^{\top} \right) \overline{\mathcal{P}}_{\text{steer}} \left(\widetilde{\mathbf{W}}^{\infty} \right) \right\|} \tag{78}$$

$$\propto \frac{1}{|G|} \sum_{g \in G} g^{\top} \boldsymbol{\gamma}^*, \tag{79}$$

其中，第三个等号是因为 $\overline{\mathcal{P}}_{\text{steer}}$ 是 L-齐次的 (14)，第四个等号则源于 $\overline{\mathcal{P}}_{\text{steer}}$ 的连续性以及 $\mathcal{P}_{\text{steer}}(\widetilde{\mathbf{W}}^{\infty}) = \left(\frac{1}{|G|} \sum_{g \in G} g^{\top} \right) \overline{\mathcal{P}}_{\text{steer}} \left(\widetilde{\mathbf{W}}^{\infty} \right) \neq 0$（否则 (66) 将无法成立）。

最后，若 G 在 $\mathcal{X}_0$上酉地作用，则由式 (71) 与命题 3.6 可得 $\boldsymbol{\gamma}^* = \overline{\mathcal{P}}_{\text{steer}}(\widetilde{\mathbf{W}}^{\infty}) \in \mathbb{R}_G^{d_0}$。因此

$$\boldsymbol{\beta}_{\text{steer}}^{\infty} \propto \frac{1}{|G|} \sum_{g \in G} g^{\top} \boldsymbol{\gamma}^* = \frac{1}{|G|} \sum_{g \in G} g \boldsymbol{\gamma}^* = \overline{\boldsymbol{\gamma}^*} = \boldsymbol{\gamma}^*, \tag{80}$$

最后一个等式同样由命题 3.6 得出。至此，(a) 的证明完成。

为证明 (b)，我们首先证明问题 (62) 存在唯一极小化解。为此，注意到 $\text{Proj}^{\text{Range}(\mathcal{A})}: \mathbb{R}_G^{d_0} \to \text{Range}(\mathcal{A})$ 是单射；事实上，对于任意 $\boldsymbol{\beta} \in \mathbb{R}_G^{d_0}$,

$$\text{Proj}_{\text{Range}(\mathcal{A})} \boldsymbol{\beta} = 0 \implies \boldsymbol{\beta} \in \text{Range}(\mathcal{A})^{\perp} = \ker(\mathcal{A}^{\top}). \tag{81}$$

因此 $0 = \mathcal{A}^{\top} \boldsymbol{\beta} = \boldsymbol{\beta}$，其中第二个等式是因为 $\mathcal{A}^{\top}$ 是到 $\text{Range}(\mathcal{A}^{\top})$ 上的投影 $= \mathbb{R}_G^{d_0}$(命题 3.6)，且 $\boldsymbol{\beta} \in \mathbb{R}_G^{d_0}$。因此，目标函数 (62) 在 $\mathbb{R}_G^{d_0}$ 上是强凸的，且存在唯一极小化解。

要证明 $\boldsymbol{\beta}_{\text{steer}}^{\infty} \propto \boldsymbol{\beta}^*$，只需验证 $\frac{1}{|G|} \sum_{g \in G} g^{\top} \boldsymbol{\gamma}^* = \boldsymbol{\beta}^*$ 是式 (62) 的极小化解。为此，我们注意到

- $\frac{1}{|G|}\sum_{g\in G} g^\top \boldsymbol{\gamma}^* = \mathcal{A}^\top \boldsymbol{\gamma}^* \in \mathbb{R}^{d_0}_G$根据命题 3.6;

- 对所有 $i \in [n]$

$$y_i \left\langle \mathbf{x}_i, \frac{1}{|G|}\sum_{g\in G} g^\top \boldsymbol{\gamma}^* \right\rangle = y_i \left\langle \overline{\mathbf{x}}_i, \boldsymbol{\gamma}^* \right\rangle \geq 1. \tag{82}$$

因此， $\frac{1}{|G|}\sum_{g\in G} g^\top \boldsymbol{\gamma}^*$ 满足 (62) 中的约束。为反证起见，假设$\frac{1}{|G|}\sum_{g\in G} g^\top \boldsymbol{\gamma}^* \neq \boldsymbol{\beta}^*$，那么根据 (62) 中极小化解的唯一性，我们有

$$\|\widetilde{\boldsymbol{\gamma}}\| \coloneqq \|\text{Proj}_{\text{Range}(\mathcal{A})}\boldsymbol{\beta}^*\| < \left\| \text{Proj}_{\text{Range}(\mathcal{A})} \frac{1}{|G|}\sum_{g\in G} g^\top \boldsymbol{\gamma}^* \right\|, \tag{83}$$

其中 $\widetilde{\boldsymbol{\gamma}} \coloneqq \text{Proj}_{\text{Range}(\mathcal{A})}\boldsymbol{\beta}^*$。我们提出以下两个稍后将予以证明的断言：

- **Claim 1:** $y_i \left\langle \widetilde{\boldsymbol{\gamma}}, \overline{\mathbf{x}}_i \right\rangle = y_i \left\langle \boldsymbol{\beta}^*, \mathbf{x}_i \right\rangle \geq 1, \forall i \in [n]$.
- **Claim 2:** $\text{Proj}_{\text{Range}(\mathcal{A})} \frac{1}{|G|}\sum_{g\in G} g^\top \boldsymbol{\gamma}^* = \boldsymbol{\gamma}^*$.

这两个断言结合 (83) 表明， $\widetilde{\boldsymbol{\gamma}}$ 满足 (61) 中的约束，且其范数小于 $\boldsymbol{\gamma}^*$, $\|\widetilde{\boldsymbol{\gamma}}\| < \|\boldsymbol{\gamma}^*\|$。这与 $\boldsymbol{\gamma}^*$ 是问题 (61) 的极小化解这一事实矛盾。因此，我们有$\frac{1}{|G|}\sum_{g\in G} g^\top \boldsymbol{\gamma}^* = \boldsymbol{\beta}^*$。

如果进一步假设 ρ_0 在 $\mathcal{X}_0$上是酉的，那么 $\mathcal{A} = \mathcal{A}^\top$: $\boldsymbol{\beta} \mapsto \overline{\boldsymbol{\beta}}$是从 $\mathbb{R}^{d_0}$ 到 $\mathbb{R}^{d_0}_G = \text{Range}(\mathcal{A})$ 的正交投影。因此，对于 $\boldsymbol{\beta} \in \mathbb{R}^{d_0}_G$，有 $\text{Proj}_{\text{Range}(\mathcal{A})}\boldsymbol{\beta} = \boldsymbol{\beta}$，故问题 (62) 和 (63) 等价。

最后，我们证明上述两个断言。
断言 1 的证明： $y_i \left\langle \widetilde{\boldsymbol{\gamma}}, \overline{\mathbf{x}_i} \right\rangle = y_i \left\langle \boldsymbol{\beta}^*, \mathbf{x}_i \right\rangle \geq 1, \quad \forall i \in [n]$。
由于 $\widetilde{\boldsymbol{\gamma}} = \text{Proj}_{\text{Range}(\mathcal{A})}\boldsymbol{\beta}^*$，我们有

$$\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\beta}^* \in \text{Range}(\mathcal{A})^\perp = \ker(\mathcal{A}^\top). \tag{84}$$

因此

$$\mathcal{A}^\top \widetilde{\boldsymbol{\gamma}} = \mathcal{A}^\top \boldsymbol{\beta}^* = \boldsymbol{\beta}^*, \tag{85}$$

其中，第二个等式是由于 $\boldsymbol{\beta}^* \in \mathbb{R}^{d_0}_G =$值域$(\mathcal{A}^\top)$ 且 $\mathcal{A}^\top = \mathcal{A}^\top \circ \mathcal{A}^\top$ 是一个投影；参见命题 3.6。因此，对于所有 $i \in [n]$,

$$y_i \left\langle \widetilde{\boldsymbol{\gamma}}, \overline{\mathbf{x}}_i \right\rangle = y_i \left\langle \mathcal{A}^\top \widetilde{\boldsymbol{\gamma}}, \mathbf{x}_i \right\rangle = y_i \left\langle \boldsymbol{\beta}^*, \mathbf{x}_i \right\rangle \geq 1. \tag{86}$$

断言 2 的证明：$\text{Proj}_{\text{Range}(\mathcal{A})}\mathcal{A}^\top \boldsymbol{\gamma}^* = \boldsymbol{\gamma}^*$。
We 首先注意到，$\text{Proj}_{\text{Range}(\mathcal{A})}\mathcal{A}^\top \boldsymbol{\gamma}^*$ 和 $\boldsymbol{\gamma}^*$ 都位于仿射空间 $\mathcal{A}^\top \boldsymbol{\gamma}^* + \ker(\mathcal{A}^\top)$中。事实上 ,

$$\left(\mathcal{A}^\top \boldsymbol{\gamma}^* - \text{Proj}_{\text{Range}(\mathcal{A})}\mathcal{A}^\top \boldsymbol{\gamma}^*\right) \in \text{Range}(\mathcal{A})^\perp = \ker(\mathcal{A}^\top), \tag{87}$$

$$\mathcal{A}^\top(\boldsymbol{\gamma}^* - \mathcal{A}^\top \boldsymbol{\gamma}^*) = \mathcal{A}^\top \boldsymbol{\gamma}^* - \mathcal{A}^\top \mathcal{A}^\top \boldsymbol{\gamma}^* = 0. \tag{88}$$

这也意味着

$$\mathcal{A}^\top \left(\text{Proj}_{\text{Range}(\mathcal{A})}\mathcal{A}^\top \boldsymbol{\gamma}^*\right) = \mathcal{A}^\top \boldsymbol{\gamma}^*. \tag{89}$$

此外，对于任意的 $i \in [n]$,

$$y_i \left\langle \text{Proj}_{\text{Range}(\mathcal{A})}\mathcal{A}^\top \boldsymbol{\gamma}^*, \overline{\mathbf{x}}_i \right\rangle = y_i \left\langle \mathcal{A}^\top \left(\text{Proj}_{\text{Range}(\mathcal{A})}\mathcal{A}^\top \boldsymbol{\gamma}^*\right), \mathbf{x}_i \right\rangle \tag{90}$$

$$= y_i \left\langle \mathcal{A}^\top \boldsymbol{\gamma}^*, \mathbf{x}_i \right\rangle = y_i \left\langle \boldsymbol{\gamma}^*, \overline{\mathbf{x}}_i \right\rangle \geq 1. \tag{91}$$

因此，$\text{Proj}_{\text{Range}(\mathcal{A})}\mathcal{A}^\top\boldsymbol{\gamma}^*$ 也满足式 (61) 中的约束。由于正交投影 $\text{Proj}_{\text{Range}(\mathcal{A})}\mathcal{A}^\top\boldsymbol{\gamma}^*$ 在包含 $\boldsymbol{\gamma}^*$ 的仿射空间 $\mathcal{A}^\top\boldsymbol{\gamma}^*+\ker(\mathcal{A}^\top)$ 中的所有向量中实现了最小范数，只有当 $\text{Proj}_{\text{Range}(\mathcal{A})}\mathcal{A}^\top\boldsymbol{\gamma}^* = \boldsymbol{\gamma}^*$ 是 (61) 的唯一极小值点时，才可能发生这种情况。

定理 4.1 的证明至此结束。 □

C 第 5 节中的证明

推论 5.1. 设 $\boldsymbol{\beta}_{\text{steer}}^\infty = \lim_{t\to\infty} \frac{\boldsymbol{\beta}_{\text{steer}}(t)}{\|\boldsymbol{\beta}_{\text{steer}}(t)\|}$ 是通过对原始数据集 $S = \{(\mathbf{x}_i, y_i)\ i \in [n]\}$ 进行梯度流训练的线性可转向网络所参数化的线性预测器$\boldsymbol{\beta}_{\text{steer}}(t) = \mathcal{P}_{\text{steer}}(\mathbf{W}(t))$的方向极限。相应地，设 $\boldsymbol{\beta}_{\text{fc}}^\infty = \lim_{t\to\infty} \frac{\boldsymbol{\beta}_{\text{fc}}(t)}{\|\boldsymbol{\beta}_{\text{fc}}(t)\|}$,$\boldsymbol{\beta}_{\text{fc}}(t) = \mathcal{P}_{\text{fc}}(\mathbf{W}(t))$ (17) 是通过增强数据集 $S_{\text{aug}} = \{(g\mathbf{x}_i, y_i)\ i \in [n], g \in G\}$训练的线性全连接网络的方向极限。若 G 在 $\mathcal{X}_0$上酉地作用，则

$$\boldsymbol{\beta}_{\text{steer}}^\infty = \boldsymbol{\beta}_{\text{fc}}^\infty. \tag{92}$$

换言之，使用线性可转向网络进行群不变二分类，与对非不变模型进行数据增强的效果完全相同。

证明。根据定理 4.1，$\boldsymbol{\gamma}^* = \tau\boldsymbol{\beta}_{\text{steer}}^\infty$ $\boldsymbol{\beta}_{\text{steer}}^\infty$ 的一个正比例缩放满足 (26) 的 KKT 条件。也就是说，存在 $\alpha_i \geq 0$、 $i \in [n]$，使得

- 原始可行性：

$$y_i \langle \boldsymbol{\gamma}^*, \overline{\mathbf{x}}_i \rangle \geq 1, \quad \forall i \in [n]. \tag{93}$$

- 驻定性：

$$\boldsymbol{\gamma}^* = \sum_{i=1}^n \alpha_i y_i \overline{\mathbf{x}}_i \tag{94}$$

- 互补松弛：

$$y_i \langle \boldsymbol{\gamma}^*, \overline{\mathbf{x}}_i \rangle > 1 \implies \alpha_i = 0. \tag{95}$$

利用推论 4.3，我们只需证明 $\boldsymbol{\gamma}^*$ 也是下式的解

$$\arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^{d_0}} \|\boldsymbol{\gamma}\|^2, \quad \text{s.t. } y_i \langle g\mathbf{x}_i, \boldsymbol{\gamma} \rangle \geq 1, \forall i \in [n], \forall g \in G. \tag{96}$$

也就是说，存在 $\tilde{\alpha}_{i,\ g} \geq 0$、 $\forall i \in [n]$、 $\forall g \in G$，使得

- 原始可行性：

$$y_i \langle \boldsymbol{\gamma}^*, g\mathbf{x}_i \rangle \geq 1, \quad \forall i \in [n], \forall g \in G. \tag{97}$$

- 驻定性：

$$\boldsymbol{\gamma}^* = \sum_{i=1}^n \sum_{g\in G} \tilde{\alpha}_{i,g} y_i g\mathbf{x}_i \tag{98}$$

- 互补松弛：

$$y_i \langle \boldsymbol{\gamma}^*, g\mathbf{x}_i \rangle > 1 \implies \tilde{\alpha}_{i,g} = 0. \tag{99}$$

事实上，我们令 $\tilde{\alpha}_{i,g} = \frac{1}{|G|}\alpha_i \geq 0$。由于 $\boldsymbol{\gamma}^* \in \mathbb{R}_G^{d_0}$若 ρ_0 是酉的（从式 (94) 中也可以看出这一点），我们有 $y_i \langle \boldsymbol{\gamma}^*, g\mathbf{x}_i \rangle = y_i \langle \boldsymbol{\gamma}^*, \mathbf{x}_i \rangle, \forall g \in G, \forall i \in [n]$。因此，由式 (93) 可得原始可行性 (97)，

$$y_i \langle \boldsymbol{\gamma}^*, g\mathbf{x}_i \rangle = y_i \left\langle \boldsymbol{\gamma}^*, \frac{1}{|G|}\sum_{h\in G} hg\mathbf{x}_i \right\rangle = y_i \langle \boldsymbol{\gamma}^*, \overline{\mathbf{x}}_i \rangle \geq 1, \quad \forall i \in [n], \forall g \in G. \tag{100}$$

驻定性 (98) 成立，因为

$$\sum_{i=1}^{n}\sum_{g\in G}\tilde{\alpha}_{i,g}y_i g\mathbf{x}_i=\sum_{i=1}^{n}\sum_{g\in G}\frac{1}{|G|}\alpha_i y_i g\mathbf{x}_i=\sum_{i=1}^{n}\alpha_i y_i\overline{\mathbf{x}}_i=\boldsymbol{\gamma}^*, \tag{101}$$

whe最后一个等式由 (94) 得出。最后，若 $y_i\langle\boldsymbol{\gamma}^*, g\mathbf{x}_i\rangle>1$，则 (100) 和 (95) 推出 tha t

$$\tilde{\alpha}_{i,g}=\frac{1}{|G|}\alpha_i=0. \tag{102}$$

这证明了互补松弛条件 (99)。 □

推论 5.4. 令 $\boldsymbol{\beta}^{\infty}_{\text{steer}}$ 与推论 5.1 中的相同。令 $\boldsymbol{\beta}^{\rho_0,\infty}{}_{\text{fc}}=\lim_{t\to\infty}\frac{\boldsymbol{\beta}^{\rho_0}_{\text{fc}}(t)}{\|\boldsymbol{\beta}^{\rho_0}_{\text{fc}}(t)\|}$ 为线性全连接网络在增强数据集 S_{aug} 上，沿修正经验损失 $\mathcal{L}^{\rho_0}_{\mathcal{P}_{\text{fc}}}(\mathrm{W};\,S_{\text{aug}})$ (38) 的梯度流演化所得到的 $\boldsymbol{\beta}^{\rho_0}{}_{\text{fc}}(t)=\mathcal{P}_{\text{fc}}(\mathrm{W}(t))$ 的极限方向。那么

$$\boldsymbol{\beta}^{\infty}_{\text{steer}}\propto\left(\frac{1}{|G|}\sum_{g\in G}\rho_0(g)^{\top}\rho_0(g)\right)^{1/2}\boldsymbol{\beta}^{\rho_0,\infty}_{\text{fc}}. \tag{103}$$

因此，对于所有 $\mathbf{x}\in\mathbb{R}^{d_0}$，我们有 $\langle\mathbf{x}\ \boldsymbol{\beta}^{\infty}_{\text{steer}}\rangle\propto\langle\mathbf{x}\ \boldsymbol{\beta}^{\rho_0,\infty}{}_{\text{fc}}\rangle_{\rho_0}$。

证明。证明与推论 5.1 的证明类似。根据定理 4.1，极限方向 $\boldsymbol{\beta}^{\infty}_{\text{steer}}$ 与 $\frac{1}{|G|}\sum_{g\in G}g^{\top}\boldsymbol{\gamma}^*$ 成比例，其中 $\boldsymbol{\gamma}^*$ 满足与推论 5.1 证明中相同的问题 (26) 的 KKT 条件 ((93)、(94) 和 (95))。另一方面，注意到增强数据集 S_{aug} 上修正全连接网络的损失函数 $\mathcal{L}^{\rho_0}_{\mathcal{P}_{\text{fc}}}(\mathrm{W};\,S_{\text{aug}})$ (38) 等价于

$$\mathcal{L}^{\rho_0}_{\mathcal{P}_{\text{fc}}}(\mathbf{W};S_{\text{aug}})=\sum_{g\in G}\sum_{i=1}^{n}\ell_{\exp}\left(\langle g\mathbf{x}_i,\mathcal{P}_{\text{fc}}(\mathbf{W})\rangle_{\rho_0},y_i\right) \tag{104}$$

$$=\sum_{g\in G}\sum_{i=1}^{n}\ell_{\exp}\left(\langle Ag\mathbf{x}_i,\mathcal{P}_{\text{fc}}(\mathbf{W})\rangle,y_i\right) \tag{105}$$

$$=\mathcal{L}_{\mathcal{P}_{\text{fc}}}(\mathbf{W};\widetilde{S}_{\text{aug}}), \tag{106}$$

其中

$$A=\left(\frac{1}{|G|}\sum_{g\in G}\rho_0(g)^{\top}\rho_0(g)\right)^{1/2},\quad \widetilde{S}_{\text{aug}}=\{(Ag\mathbf{x}_i,y_i):i\in[n],g\in G\}. \tag{107}$$

因此，根据推论 4.3，$\boldsymbol{\beta}^{\rho_0,\ \infty}$ fc 与解 $\widetilde{\boldsymbol{\gamma}}^*_{\text{aug}}$ 成正比

$$\widetilde{\boldsymbol{\gamma}}^*_{\text{aug}}=\arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^{d_0}}\|\boldsymbol{\gamma}\|^2,\quad \text{s.t. } y_i\langle Ag\mathbf{x}_i,\boldsymbol{\gamma}\rangle\geq 1,\forall i\in[n],\forall g\in G. \tag{108}$$

我们声称 $\widetilde{\boldsymbol{\gamma}}^*_{\text{aug}}=A^{-1}\left(\frac{1}{|G|}\sum_{g\in G}g^{\top}\boldsymbol{\gamma}^*\right)$。为此，令 $\widetilde{\alpha}_{i,\ g}=\alpha_i,\forall i\in[n]$，其中 $\alpha_i\geq 0$是问题 (26) 中 $\boldsymbol{\gamma}^*$ 的对偶变量。下面我们验证，原始-对偶对 $A^{-1}\left(\frac{1}{|G|}\sum_{g\in G}g^{\top}\boldsymbol{\gamma}^*\right)$ 和 $(\widetilde{\alpha}_{i,\ g})_{i\in[n],\ g\in G}$满足问题 (108) 的 KKT 条件。

- 原始可行性：对于任意 $i\in[n]$ 和 $h\in G$，

$$y_i\left\langle A^{-1}\left(\frac{1}{|G|}\sum_{g\in G}g^{\top}\boldsymbol{\gamma}^*\right),Ah\mathbf{x}_i\right\rangle=y_i\left\langle\frac{1}{|G|}\sum_{g\in G}g^{\top}\boldsymbol{\gamma}^*,h\mathbf{x}_i\right\rangle \tag{109}$$

$$=y_i\left\langle\boldsymbol{\gamma}^*,\left(\frac{1}{|G|}\sum_{g\in G}g\right)h\mathbf{x}_i\right\rangle \tag{110}$$

$$=y_i\langle\boldsymbol{\gamma}^*,\overline{\mathbf{x}}_i\rangle\geq 1, \tag{111}$$

其中，第一个等式是由于 $A^{\top}=A$ 具有对称性，最后一个不等式则由 (93) 得出。

- 驻定性：

$$A^{-1}\left(\frac{1}{|G|}\sum_{g\in G} g^\top \boldsymbol{\gamma}^*\right) = A^{-1}\left(\frac{1}{|G|}\sum_{g\in G} g^\top\right)\sum_{i=1}^n \alpha_i y_i \overline{\mathbf{x}}_i \tag{112}$$

$$= A^{-1}\frac{1}{|G|^2}\left(\sum_{g\in G} g^\top\right)\sum_{i=1}^n \alpha_i y_i \sum_{h\in G} h\mathbf{x}_i \tag{113}$$

$$= A\cdot A^{-2}\frac{1}{|G|^2}\sum_{i=1}^n \alpha_i y_i \sum_{g\in G}\sum_{h\in G} g^\top h\mathbf{x}_i \tag{114}$$

$$= A\cdot A^{-2}\frac{1}{|G|}\sum_{i=1}^n \alpha_i y_i \left(\frac{1}{|G|}\sum_{g\in G} g^\top g\right)\sum_{h\in G} h\mathbf{x}_i \tag{115}$$

$$= \sum_{i=1}^n \frac{\alpha_i}{|G|} y_i \sum_{h\in G} Ah\mathbf{x}_i \tag{116}$$

$$= \sum_{i=1}^n \sum_{g\in G} \widetilde{\alpha}_{i,g} y_i Ag\mathbf{x}_i, \tag{117}$$

其中，第一个等号由 (94) 得出，最后一个等号则来自对偶变量 $\widetilde{\alpha}_{i,g} = \frac{1}{|G|}\alpha_i, \forall i \in [n]$, $\forall g \in G$ 的定义。

- 互补松弛：若 $y_i \left\langle A^{-1}\left(\frac{1}{|G|}\sum_{g\in G} g^\top \boldsymbol{\gamma}^*\right), Ag\mathbf{x}_i\right\rangle > 1$ 对某些 $i \in [n]$ 和 $g \in G$ 成立，则 (111) 和 (95) 意味着 $\widetilde{\alpha}_{i,g} = \frac{1}{|G|}\alpha_i = 0$。

因此，$\widetilde{\boldsymbol{\gamma}}^*_{\text{aug}} = A^{-1}\left(\frac{1}{|G|}\sum_{g\in G} g^\top \boldsymbol{\gamma}^*\right)$ 确实是 (108) 的解。因此，

$$\boldsymbol{\beta}^\infty_{\text{steer}} \propto \frac{1}{|G|}\sum_{g\in G} g^\top \boldsymbol{\gamma}^* = A\boldsymbol{\gamma}^*_{\text{aug}} \propto A\boldsymbol{\beta}^{\rho_0,\infty}_{\text{fc}}. \tag{118}$$

证明至此完成。 □

D 定理 6.1 的证明

定理 6.1. 设 $\boldsymbol{\beta}^\infty_{\text{steer}}$ 是一个由线性可转向网络参数化的预测器在原始数据集 $S = \{(\mathbf{x}_i, y_i)\ i \in [n]\}$ 上训练后得到的方向极限；设 $\boldsymbol{\beta}^\infty_{\text{fc}}$ 是一个线性全连接网络在同一数据集 S 上训练后得到的方向极限。设 M_{steer} 和 M_{fc} 分别为 $\boldsymbol{\beta}^\infty_{\text{steer}}$ 和 $\boldsymbol{\beta}^\infty_{\text{fc}}$ 在增强数据集 $S_{\text{aug}} = \{(g\mathbf{x}_i, y_i) : i \in [n], g \in G\}$ 上的（带符号）间隔，即，

$$M_{\text{steer}} \coloneqq \min_{i\in[n],g\in G} y_i \left\langle \boldsymbol{\beta}^\infty_{\text{steer}}, g\mathbf{x}_i\right\rangle, \quad M_{\text{fc}} \coloneqq \min_{i\in[n],g\in G} y_i \left\langle \boldsymbol{\beta}^\infty_{\text{fc}}, g\mathbf{x}_i\right\rangle. \tag{119}$$

那么我们总有 $M_{\text{steer}} \geq M_{\text{fc}}$。

证明。根据推论 4.3 和推论 5.1，我们有

$$\boldsymbol{\beta}^\infty_{\text{steer}} \propto \boldsymbol{\gamma}^*_{\text{steer}} = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^{d_0}} \|\boldsymbol{\gamma}\|^2, \quad \text{s.t. } y_i \left\langle \boldsymbol{\gamma}, g\mathbf{x}_i\right\rangle \geq 1, \forall g \in G, \forall i \in [n], \tag{120}$$

$$\boldsymbol{\beta}^\infty_{\text{fc}} \propto \boldsymbol{\gamma}^*_{\text{fc}} = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^{d_0}} \|\boldsymbol{\gamma}\|^2, \quad \text{s.t. } y_i \left\langle \boldsymbol{\gamma}, \mathbf{x}_i\right\rangle \geq 1, \forall i \in [n]. \tag{121}$$

此外，可转向网络 $\boldsymbol{\beta}^\infty_{\text{steer}}$ 在增强数据集 S_{aug} 上的间隔 M_{steer} 为 $M_{\text{steer}} = \frac{1}{\|\boldsymbol{\gamma}^*_{\text{steer}}\|}$。考虑以下三种情况。

- 情形1：$\min_{i\in[n],g\in G} y_i \left\langle \boldsymbol{\gamma}^*\text{fc}\ g\mathbf{x}_i\right\rangle \geq 1$。在这种情况下，$\boldsymbol{\gamma}^*\text{fc}$ 也满足 (120) 中更严格的约束。因此$\boldsymbol{\gamma}^*\text{fc} = \boldsymbol{\gamma}^*$具有 steer 性质，这意味着

$$\boldsymbol{\beta}^\infty_{\text{steer}} = \boldsymbol{\beta}^\infty_{\text{fc}}, \quad \text{and} \quad M_{\text{steer}} = M_{\text{fc}}. \tag{122}$$

• 情形2：$\min_{i\in[n],g\in G} y_i \langle \boldsymbol{\gamma}^*\text{fc}\ g\mathbf{x}_i\rangle \le 0$。这意味着 $\boldsymbol{\gamma}^*\text{fc}$，因此 $\boldsymbol{\beta}^{\infty}\text{fc}$ 在 $D\text{aug}$ 上的间隔为非正。因此

$$M_{\text{fc}} \le 0 < M_{\text{steer}}. \tag{123}$$

• 情形3：$0 < \min_{i\in[n],g\in G} y_i \langle \boldsymbol{\gamma}^*\text{fc}\ g\mathbf{x}_i\rangle < 1$。在此情形下，定义

$$\widetilde{\boldsymbol{\gamma}}^*_{\text{fc}} \coloneqq \frac{\boldsymbol{\gamma}^*_{\text{fc}}}{\min_{i\in[n],g\in G} y_i \langle \boldsymbol{\gamma}^*_{\text{fc}}, g\mathbf{x}_i\rangle}. \tag{124}$$

则 $\widetilde{\boldsymbol{\gamma}}^*_{\text{fc}}$ 满足式 (120) 中的约束，因此

$$\|\widetilde{\boldsymbol{\gamma}}^*_{\text{fc}}\| \ge \|\boldsymbol{\gamma}^*_{\text{steer}}\|. \tag{125}$$

因此

$$M_{\text{fc}} = \min_{i\in[n],g\in G} y_i \langle \boldsymbol{\beta}^{\infty}_{\text{fc}}, g\mathbf{x}_i\rangle = \frac{\min_{i\in[n],g\in G} y_i \langle \boldsymbol{\gamma}^*_{\text{fc}}, g\mathbf{x}_i\rangle}{\|\boldsymbol{\gamma}^*_{\text{fc}}\|} = \frac{1}{\|\widetilde{\boldsymbol{\gamma}}^*_{\text{fc}}\|} \tag{126}$$

$$\le \frac{1}{\|\boldsymbol{\gamma}^*_{\text{steer}}\|} = M_{\text{steer}}. \tag{127}$$

证明至此完成。 □

E 定理 6.3 的证明

为了证明定理 6.3，我们首先需要以下预备知识。

引理 E.1. 如果 $\mathcal{D}$ 是 G-不变的且线性可分，那么 $\mathcal{D}$ 可以由一个 G-不变分类器进行线性分离。也就是说，存在 $\boldsymbol{\beta} \in \mathbb{R}^{d_0}_G$ 使得

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y \langle \mathbf{x}, \boldsymbol{\beta}\rangle \ge 1\right] = 1. \tag{128}$$

证明。由于 $\mathcal{D}$ 是线性可分的，存在 $\boldsymbol{\beta}_0 \in \mathbb{R}^{d_0}$ 使得 $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y \langle \mathbf{x}\ \boldsymbol{\beta}_0\rangle \ge 1\right] = 1$。令 $\boldsymbol{\beta} = \frac{1}{|G|}\sum_{g\in G} g^{\top}\boldsymbol{\beta}_0 \in \mathbb{R}^{d_0}_G$，我们旨在证明 $\mathcal{D}$ 也可以由 $\boldsymbol{\beta}$ 进行线性分离。事实上，对于所有 $(\mathbf{x}\ y) \in \mathbb{R}^{d_0}\times\{\pm 1\}$,

$$y \langle \mathbf{x}, \boldsymbol{\beta}\rangle < 1 \implies y \left\langle \mathbf{x}, \frac{1}{|G|}\sum_{g\in G} g^{\top}\boldsymbol{\beta}_0 \right\rangle < 1 \implies y \left\langle \frac{1}{|G|}\sum_{g\in G} g\mathbf{x}, \boldsymbol{\beta}_0 \right\rangle < 1 \tag{129}$$

$$\implies \exists g \in G,\ \text{s.t. } y \langle g\mathbf{x}, \boldsymbol{\beta}_0\rangle < 1. \tag{130}$$

令 $E = \{(\mathbf{x}\ y)\colon\ y \langle \mathbf{x}\ \boldsymbol{\beta}_0\rangle < 1\} \subset \mathbb{R}^{d_0}\times\{\pm 1\}$，则

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y \langle \mathbf{x}, \boldsymbol{\beta}\rangle < 1\right] \le \sum_{g\in G} \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y \langle g\mathbf{x}, \boldsymbol{\beta}_0\rangle < 1\right] \tag{131}$$

$$= \sum_{g\in G} \mathcal{D}\left\{(\mathbf{x},y) : (\rho_0(g)\otimes \text{Id})(\mathbf{x},y) \in E\right\} \tag{132}$$

$$= \sum_{g\in G} (\rho_0(g)\otimes \text{Id})_* \mathcal{D}\left(E\right) = \sum_{g\in G} \mathcal{D}\left(E\right) \tag{133}$$

$$= \sum_{g\in G} \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y \langle \mathbf{x}, \boldsymbol{\beta}_0\rangle < 1\right] = 0, \tag{134}$$

其中，式 (133) 中的第二个等式是由于 $\mathcal{D}$ 是 G-不变的。因此，$\mathcal{D}$ 可以由 $\boldsymbol{\beta} \in \mathbb{R}^{d_0}_G\cdot$ 分离 □

定义 E.2（经验 Rademacher 复杂度）。设 $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$ 是定义在 $\mathcal{Z}$ 上的一类实值函数，且 $S = (\mathbf{z}_1, \cdots, \mathbf{z}_n)$ 是来自 $\mathcal{Z}$ 的一组 n 个样本。$\mathcal{F}$ 在 S 上的（经验）Rademacher^复杂度 $\mathcal{R}_n(\mathcal{F}, S)$ 定义为

$$\widehat{\mathcal{R}}_n(\mathcal{F}, S) \coloneqq \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}\sim\{\pm 1\}^n}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \sigma_i f(\mathbf{z}_i)\right], \tag{135}$$

其中 $\boldsymbol{\sigma} = (\sigma_1, \cdots, \sigma_n)$ 是满足 $\mathbb{P}[\sigma_i = 1] = \mathbb{P}[\sigma_i = -1] = 1/2$ 的独立同分布 Rademacher变量。

引理 E.3（Talagrand 收缩引理，Mohri 和 Medina [2014] 中的引理 8）。设 $\Phi_i : \mathbb{R} \to \mathbb{R}, i \in [n]$ 是 1-Lipschitz 函数，$\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$ 是一个函数类，且 $S = (\mathbf{z}_1, \ldots, \mathbf{z}_n) \in \mathcal{Z}^n$ 是从 $\mathcal{Z}$ 中抽取的一组 n 样本。那么

$$\widehat{\mathcal{R}}_n(\mathcal{F}, S) \geq \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}\sim\{\pm1\}^n}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i\Phi_i \circ f(\mathbf{z}_i)\right]. \tag{136}$$

引理 E.4. 设 $\mathcal{H} = \left\{\mathbf{x} \mapsto \langle \boldsymbol{\beta}, \mathbf{x}\rangle : \boldsymbol{\beta} \in \mathbb{R}_G^{d_0}, \|\boldsymbol{\beta}\| \leq B\right\}$ 为有界范数的 G-不变线性函数所构成的函数空间，并设 $S_{\mathbf{x}} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\} \subset \mathbb{R}^{d_0}$ 为来自 $\mathbb{R}^{d_0}$ 的一组 n 样本。则

$$\widehat{\mathcal{R}}_n(\mathcal{H}, S_{\mathbf{x}}) \leq \frac{B \max_{i\in[n]} \|\overline{\mathbf{x}}_i\|}{\sqrt{n}} \tag{137}$$

证明. 根据经验 Rademacher复杂度的定义 (135)，我们有

$$n\widehat{\mathcal{R}}_n(\mathcal{H}, S_{\mathbf{x}}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{\beta}\in\mathbb{R}_G^{d_0}, \|\boldsymbol{\beta}\|\leq B}\sum_{i=1}^{n}\sigma_i\langle\boldsymbol{\beta}, \mathbf{x}_i\rangle\right] \tag{138}$$

$$= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{\beta}\in\mathbb{R}_G^{d_0}, \|\boldsymbol{\beta}\|\leq B}\sum_{i=1}^{n}\sigma_i\langle\boldsymbol{\beta}, \overline{\mathbf{x}}_i\rangle\right] \tag{139}$$

$$= \mathbb{E}_{\boldsymbol{\sigma}}\sup_{\boldsymbol{\beta}\in\mathbb{R}_G^{d_0}, \|\boldsymbol{\beta}\|\leq B}\left\langle\boldsymbol{\beta}, \sum_{i=1}^{n}\sigma_i\overline{\mathbf{x}}_i\right\rangle \tag{140}$$

$$= B\mathbb{E}_{\boldsymbol{\sigma}}\left\|\sum_{i=1}^{n}\sigma_i\overline{\mathbf{x}}_i\right\| \tag{141}$$

$$\leq B\left(\mathbb{E}_{\boldsymbol{\sigma}}\left\|\sum_{i=1}^{n}\sigma_i\overline{\mathbf{x}}_i\right\|^2\right)^{1/2}, \tag{142}$$

其中 (139) 由 $\boldsymbol{\beta} \in \mathbb{R}_G^{d_0}$ 得出。由于 $\sigma_1, \cdots, \sigma_n$ 是独立同分布的，我们有

$$\mathbb{E}_{\boldsymbol{\sigma}}\left\|\sum_{i=1}^{n}\sigma_i\overline{\mathbf{x}}_i\right\|^2 = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{i,j}\sigma_i\sigma_j\langle\overline{\mathbf{x}}_i, \overline{\mathbf{x}}_j\rangle\right] \tag{143}$$

$$= \sum_{i\neq j}\langle\overline{\mathbf{x}}_i, \overline{\mathbf{x}}_j\rangle\,\mathbb{E}_{\boldsymbol{\sigma}}\sigma_i\sigma_j + \sum_{i=1}^{n}\|\overline{\mathbf{x}}_i\|^2\mathbb{E}_{\boldsymbol{\sigma}}\sigma_i^2 \tag{144}$$

$$= \sum_{i=1}^{n}\|\overline{\mathbf{x}}_i\|^2 \leq n\max_{i\in[n]}\|\overline{\mathbf{x}}_i\|^2. \tag{145}$$

结合 (145)，等式 (142) 即完成证明。 □

我们还需要以下关于基于 Rademacher 复杂度的泛化界的标准结果 [Shalev-Shwartz 和 Ben-David, 2014]。

引理 E.5. 设 $\mathcal{H}$ 为一组假设，$\mathcal{Z} = \mathbb{R}^{d_0} \times \{\pm1\}$ 为带标签样本的集合，并且

$$\ell : \mathcal{H} \times \mathcal{Z} \to [0, \infty), \quad (h, \mathbf{z}) \mapsto \ell(h, \mathbf{z}), \tag{146}$$

设损失函数满足 $0 \leq \ell(h, \mathbf{z}) \leq c$ 对所有 $\mathbf{z} \in \mathcal{Z}$ 以及 $h \in \mathcal{H}$ 均成立。定义函数类 $\mathcal{F} = \ell(\mathcal{H}, \cdot) := \{\mathbf{z} \mapsto \ell(h, \mathbf{z}) : h \in \mathcal{H}\}$。那么，对于任意 $\delta > 0$ 以及定义在 $\mathcal{Z}$ 上的任意分布 $\mathcal{D}$，以至少 $1 - \delta$ 的概率，在独立同分布样本 $S = (\mathbf{z}_1, \cdots, \mathbf{z}_n) \sim \mathcal{D}^n$ 上，有

$$\sup_{h\in\mathcal{H}}\mathcal{L}_{\mathcal{D}}(h) - \mathcal{L}_S(h) \leq 2\mathcal{R}_n(\mathcal{F}) + c\sqrt{\frac{\log(1/\delta)}{2n}}, \tag{147}$$

其中， $\mathcal{L}_{\mathcal{D}}(h) \coloneqq \mathbb{E}_{\mathbf{z}\sim\mathcal{D}}\ell(h,\mathbf{z})$ 和 $\mathcal{L}_S(h) \coloneqq \frac{1}{n}\sum_{i=1}^n \ell(h,\mathbf{z}_i)$ 分别是假设 $h \in \mathcal{H}$ 的总体和经验损失，而 $\mathcal{R}_n(\mathcal{F}) \coloneqq \mathbb{E}_{S\sim\mathcal{D}^n}\mathcal{R}_n(\mathcal{F},S)$ 是 $\mathcal{F}$的 Rademacher 复杂度。

最后，我们可以证明下文重述的定理 6.3

定理 6.3. 设 $\mathcal{D}$ 是 G 上的一个 $\mathbb{R}^{d_0}\times\{\pm1\}$ -不变分布，该分布可由不变分类器 $\boldsymbol{\beta}_0 \in \mathbb{R}^{d_0}_G$ 线性分离。定义

$$\overline{R} = \inf\left\{r>0 : \|\overline{\mathbf{x}}\| \le r \text{ with probability } 1\right\}. \tag{148}$$

设 $S=\{(\mathbf{x}_i,y_i)\}_{i=1}^n$ 是来自 $\mathcal{D}$ 的独立同分布样本，并设 $\boldsymbol{\beta}^\infty_{\text{steer}}$ 是通过对 S 使用梯度流训练的、由可转向网络参数化的线性预测器的极限方向。那么，对于任意 $\delta>0$，以至少 $1-\delta$ 的概率（针对随机样本 $S\sim\mathcal{D}^n$），有

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y \ne \text{sign}\left(\langle \mathbf{x}, \boldsymbol{\beta}^\infty_{\text{steer}}\rangle\right)\right] \le \frac{2\overline{R}\|\boldsymbol{\beta}_0\|}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{149}$$

定理 6.3 的证明。设 $\mathcal{H} = \{\boldsymbol{\beta}\in\mathbb{R}^{d_0}_G : \ \|\boldsymbol{\beta}\| \le \|\boldsymbol{\beta}_0\|\}$，并稍微滥用一下记号，将 $\boldsymbol{\beta}\in\mathcal{H}$ 视为不变线性映射 $\mathbf{x} \mapsto \langle \mathbf{x}\ \boldsymbol{\beta}\rangle$。根据定理 4.1，令 $\boldsymbol{\beta}^* = \tau\boldsymbol{\beta}^\infty_{\text{steer}}$、 $\tau>0$ 为 (28) 的极小化解。由于根据假设， $\boldsymbol{\beta}_0$ 也满足 (28) 中的约束，因此有 $\|\boldsymbol{\beta}^*\| \le \|\boldsymbol{\beta}_0\|$，进而得到 $\boldsymbol{\beta}^*\in\mathcal{H}$。

考虑定义为 ℓ 的斜坡损失

$$\ell : \mathcal{H}\times(\mathbb{R}^{d_0}\times\{\pm1\}) \to [0,1], \quad (\boldsymbol{\beta},(\mathbf{x},y)) \mapsto \min\{1,\max\{0,1-y\langle\mathbf{x},\boldsymbol{\beta}\rangle\}\}. \tag{150}$$

容易验证， ℓ 是 0-1 损失的上界：对于所有 $\boldsymbol{\beta}\in\mathcal{H}$ 和 $(\mathbf{x}\ y)\ \in\mathbb{R}^{d_0}\times\{\pm1\}$ ，

$$\ell(\boldsymbol{\beta},(\mathbf{x},y)) \ge {}_{\mathbb{I}\{y\ne\text{sign}(\langle\mathbf{x},\boldsymbol{\beta}\rangle)\}}. \tag{151}$$

这意味着

$$\begin{aligned}
\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y \ne \text{sign}\left(\langle \mathbf{x}, \boldsymbol{\beta}^\infty_{\text{steer}}\rangle\right)\right] &= \mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y \ne \text{sign}\left(\langle \mathbf{x}, \boldsymbol{\beta}^*\rangle\right)\right] &(152)\\
&= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}{}_{\mathbb{I}\{y\ne\text{sign}(\langle\mathbf{x},\boldsymbol{\beta}^*\rangle)\}} &(153)\\
&\le \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\ell(\boldsymbol{\beta}^*,(\mathbf{x},y)) = \mathcal{L}_{\mathcal{D}}(\boldsymbol{\beta}^*). &(154)
\end{aligned}$$

由于根据定义 (28)， $\mathcal{L}_S(\boldsymbol{\beta}^*)$ 始终为 0，且 $\boldsymbol{\beta}^*\in\mathcal{H}$，由引理 E.5 可得

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\left[y \ne \text{符号}\left(\langle \mathbf{x}\, \boldsymbol{\beta}^\infty_{\text{steer}}\rangle\right)\right] \le \mathcal{L}_{\mathcal{D}}(\boldsymbol{\beta}^*) = \mathcal{L}_{\mathcal{D}}(\boldsymbol{\beta}^*) - \mathcal{L}_S(\boldsymbol{\beta}^*) \le \text{上确界}_{\boldsymbol{\beta}\in\mathcal{H}}\ \mathcal{L}_{\mathcal{D}}(\boldsymbol{\beta}) - \mathcal{L}_S(\boldsymbol{\beta}) \tag{155}$$

$$\le 2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}, \tag{156}$$

其中 $\mathcal{F} = \ell(\mathcal{H},\cdot) = \{(\mathbf{x}\ y) \mapsto \ell(\boldsymbol{\beta},(\mathbf{x}\ y)) : \ \boldsymbol{\beta}\in\mathcal{H}\}$。因此，要证明定理 6.3，只需证明 $\mathcal{R}_n(\mathcal{F}) \le \frac{\overline{R}\|\boldsymbol{\beta}_0\|}{\sqrt{n}}$。为此，令 $S=\{(\mathbf{x}_i,y_i) : \ i\in[n]\}$ 为一组 n 带标签样本，并令 $S_{\mathbf{x}} = \{\mathbf{x}_i : \ i\in[n]\}$ 为相应的一组无标签输入，则

$$\begin{aligned}
\widehat{\mathcal{R}}_n(\mathcal{F},S) &= \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{f\in\mathcal{F}}\sum_{i=1}^n \sigma_i f(\mathbf{x}_i,y_i) &(157)\\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{\boldsymbol{\beta}\in\mathcal{H}}\sum_{i=1}^n \sigma_i \ell\left(\boldsymbol{\beta},(\mathbf{x}_i,y_i)\right) &(158)\\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{\boldsymbol{\beta}\in\mathcal{H}}\sum_{i=1}^n \sigma_i \min\left\{1,\max\{0,1-y_i\langle\mathbf{x}_i,\boldsymbol{\beta}\rangle\}\right\} &(159)\\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{\boldsymbol{\beta}\in\mathcal{H}}\sum_{i=1}^n \sigma_i \Phi_i(\langle\mathbf{x}_i,\boldsymbol{\beta}\rangle), &(160)
\end{aligned}$$

其中 $\Phi_i : \ \mathbb{R}\to\mathbb{R}$， $a\mapsto$min $\{1,$max$\{0,1-y_i a\}\}$，是 1-Lipschitz 的。因此，引理 E.3 表明 th at

$$\widehat{\mathcal{R}}_n(\mathcal{F},S) \le \widehat{\mathcal{R}}_n(\mathcal{H},S_{\mathbf{x}}). \tag{161}$$

利用引理 E.4 以及 $\mathbf{x} \le \overline{R}$ 以概率 1 关于 $(\mathbf{x}, y) \sim \mathcal{D}$ 的事实，我们得到

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{S\sim\mathcal{D}^n}\widehat{\mathcal{R}}_n(\mathcal{F}, S) \le \mathbb{E}_{S\sim\mathcal{D}^n}\widehat{\mathcal{R}}_n(\mathcal{H}, S_{\mathbf{x}}) \le \frac{\overline{R}\|\boldsymbol{\beta}_0\|}{\sqrt{n}}. \tag{162}$$

定理 6.3 的证明至此完成。 □