深度学习中的泛化理论与实践：训练轨迹对泛化的影响

汪子乔

同济大学计算机科学与技术学院

2025 年 3 月 4 日

Mixup 方法背景

C-分类问题设定

- 输入空间：$\mathcal{X} \subseteq \mathbb{R}^{d_0}$；标签空间：$\mathcal{Y} = \{1, 2, \ldots, C\}$
- 训练集：$S = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$，其中 $\mathbf{y}_i$ 为独热编码向量
- 预测器：$f_\theta : \mathcal{X} \to [0, 1]^C$；损失函数：$\ell(\theta, \mathbf{x}, \mathbf{y})$；
  经验风险：$\hat{R}_S(\theta) := \frac{1}{n} \sum_{i=1}^n \ell(\theta, \mathbf{x}_i, \mathbf{y}_i)$

Mixup 方法背景

C-分类问题设定

- 输入空间：$\mathcal{X} \subseteq \mathbb{R}^{d_0}$；标签空间：$\mathcal{Y} = \{1, 2, \ldots, C\}$
- 训练集：$S = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$，其中 $\mathbf{y}_i$ 为独热编码向量
- 预测器：$f_\theta : \mathcal{X} \to [0, 1]^C$；损失函数：$\ell(\theta, \mathbf{x}, \mathbf{y})$；
  经验风险：$\hat{R}_S(\theta) := \frac{1}{n} \sum_{i=1}^n \ell(\theta, \mathbf{x}_i, \mathbf{y}_i)$
- Mixup 合成数据集：

$$\widetilde{S}_\lambda := \{(\lambda\mathbf{x} + (1-\lambda)\mathbf{x}', \lambda\mathbf{y} + (1-\lambda)\mathbf{y}') : (\mathbf{x}, \mathbf{y}) \in S, (\mathbf{x}', \mathbf{y}') \in S\}$$

  其中 $\lambda \in [0, 1]$ 服从预设分布，且对所有样本对独立采样
- “Mixup 经验风险/训练损失”定义为：

$$\mathbb{E}_\lambda \hat{R}_{\widetilde{S}_\lambda}(\theta) := \mathbb{E}_\lambda \frac{1}{|\widetilde{S}_\lambda|} \sum_{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \in \widetilde{S}_\lambda} \ell(\theta, \tilde{\mathbf{x}}, \tilde{\mathbf{y}})$$

Mixup 方法可视化示例

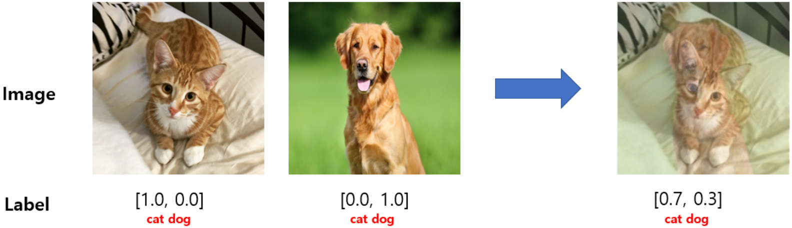
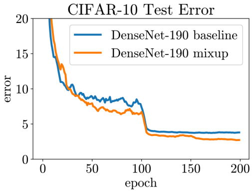


Figure 1: Mixup 数据增强示例（$\lambda = 0.7$），图片源自 *https://www.kaggle.com/code/kaushal2896/data-augmentation-tutorial-basic-cutout-mixup*

Mixup 方法有效提升模型性能

Dataset	Model	ERM	*mixup*
CIFAR-10	PreAct ResNet-18	5.6	**4.2**
	WideResNet-28-10	3.8	**2.7**
	DenseNet-BC-190	3.7	**2.7**
CIFAR-100	PreAct ResNet-18	25.6	**21.1**
	WideResNet-28-10	19.4	**17.5**
	DenseNet-BC-190	19.0	**16.8**

(a) Test errors for the CIFAR experiments.



(b) Test error evolution for the best ERM and *mixup* models.

Figure 2: ERM 与 Mixup 在 CIFAR 数据集上的测试误差对比，数据来自 *Zhang, Hongyi, et al. “mixup: Beyond Empirical Risk Minimization.” ICLR 2018.*

Mixup 损失下界

*Zixuan Liu**, **Ziqiao Wang***, *Hongyu Guo, and Yongyi Mao.*
"Over-Training with Mixup May Hurt Generalization." ICLR 2023.

引理 1

设 $\ell(\cdot)$ 为交叉熵损失函数，且 λ 独立同分布于 $\mathrm{Beta}(1,1)$ 分布（即 $[0,1]$ 上的均匀分布）。则对于所有 $\theta \in \Theta$ 和任意给定的平衡训练集 S，有

$$\mathbb{E}_\lambda \hat{R}_{\widetilde{S}_\lambda}(\theta) \geq \frac{C-1}{2C},$$

当且仅当对每个合成样本 $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \in \widetilde{S}_\lambda$ 满足 $f_\theta(\tilde{\boldsymbol{x}}) = \tilde{\boldsymbol{y}}$ 时等号成立。

例如，在 10 分类任务中，该下界值为 0.45。

观察：训练损失持续下降时（左图），测试误差先降后升（右图）

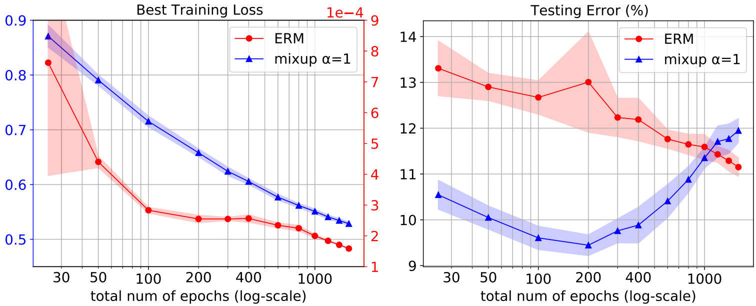


Figure 3: ResNet18 在 CIFAR10 数据集上的表现

实验观察

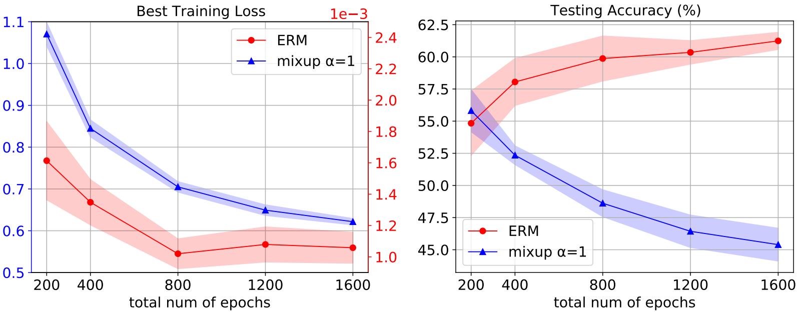


Figure 4: ResNet34 在 CIFAR100 数据集上的表现

普适性验证

该现象同样适用于：

- 不同网络架构（如 VGG16、ResNet34）
- 不同损失函数（如均方误差损失 MSE）
- 其他数据增强方法（在减少样本量情况下），例如" 随机裁剪"（random crop）和" 水平翻转"（horizontal flip）
- 协变量偏移场景（如 CIFAR10.1、CIFAR10.2 数据集）

Mixup 引入标签噪声

- 设 $P(Y|X)$ 为真实条件分布，$f : \mathcal{X} \to [0,1]^C$ 满足 $f_j(\mathbf{x}) \triangleq P(Y = j|X = \mathbf{x})$
  例如，$\mathbf{y} = \arg\max_{j\in\mathcal{Y}} f_j(\mathbf{x})$ 表示预测类别

Mixup 引入标签噪声

- 设 $P(Y|X)$ 为真实条件分布，$f:\mathcal{X}\to[0,1]^C$ 满足 $f_j(\mathbf{x})\triangleq P(Y=j|X=\mathbf{x})$
  例如，$\mathbf{y}=\arg\max_{j\in\mathcal{Y}} f_j(\mathbf{x})$ 表示预测类别
- 对于混合样本 $\widetilde{X}\triangleq\lambda X+(1-\lambda)X'$，存在两种标签分配方式：
  - 真实标签：$\widetilde{Y}_{\mathrm{h}}^*\triangleq\arg\max_{j\in\mathcal{Y}} f_j(\widetilde{X})$

Mixup 引入标签噪声

- 设 $P(Y|X)$ 为真实条件分布，$f:\mathcal{X}\to[0,1]^C$ 满足 $f_j(\mathbf{x})\triangleq P(Y=j|X=\mathbf{x})$
  例如，$\mathbf{y}=\arg\max_{j\in\mathcal{Y}} f_j(\mathbf{x})$ 表示预测类别
- 对于混合样本 $\widetilde{X}\triangleq\lambda X+(1-\lambda)X'$，存在两种标签分配方式：
  - 真实标签：$\widetilde{Y}^*_{\mathrm{h}}\triangleq\arg\max_{j\in\mathcal{Y}} f_j(\widetilde{X})$
  - Mixup 标签：$\widetilde{Y}_{\mathrm{h}}\triangleq\arg\max_{j\in\mathcal{Y}} P(\widetilde{Y}=j|\widetilde{X})$
    其中 $P(\widetilde{Y}=j|\widetilde{X})=\lambda f_j(X)+(1-\lambda)f_j(X')$

Mixup 引入标签噪声

- 设 $P(Y|X)$ 为真实条件分布，$f : \mathcal{X} \to [0,1]^C$ 满足 $f_j(\mathbf{x}) \triangleq P(Y = j|X = \mathbf{x})$
  例如，$\mathbf{y} = \arg\max_{j\in\mathcal{Y}} f_j(\mathbf{x})$ 表示预测类别
- 对于混合样本 $\widetilde{X} \triangleq \lambda X + (1-\lambda)X'$，存在两种标签分配方式：
  - 真实标签：$\widetilde{Y}_{\mathrm{h}}^* \triangleq \arg\max_{j\in\mathcal{Y}} f_j(\widetilde{X})$
  - Mixup 标签：$\widetilde{Y}_{\mathrm{h}} \triangleq \arg\max_{j\in\mathcal{Y}} P(\widetilde{Y} = j|\widetilde{X})$
    其中 $P(\widetilde{Y} = j|\widetilde{X}) = \lambda f_j(X) + (1-\lambda) f_j(X')$
- 当两种分配方式不一致时（$\widetilde{Y}_{\mathrm{h}} \neq \widetilde{Y}_{\mathrm{h}}^*$），Mixup 分配的标签 $\widetilde{Y}_{\mathrm{h}}$ 即为噪声标签

Mixup 引入标签噪声

定理 1

对于任意固定的 X、X' 和 $\widetilde{X} = \lambda X + (1-\lambda)X'$ ($\lambda \in [0,1]$)，分配噪声标签的概率存在下界：

$$
\begin{aligned}
P(\widetilde{Y}_{\mathrm{h}} \neq \widetilde{Y}^*_{\mathrm{h}} | \widetilde{X}) \geq & \mathrm{TV}(P(\widetilde{Y}|\widetilde{X}), P(Y|X)) \\
\geq & \frac{1}{2} \sup_{j \in \mathcal{Y}} \left| f_j(\widetilde{X}) - [(1-\lambda) f_j(X) + \lambda f_j(X')] \right|,
\end{aligned}
$$

其中 $\mathrm{TV}(\cdot,\cdot)$ 表示总变差距离。

带噪声标签的训练过程

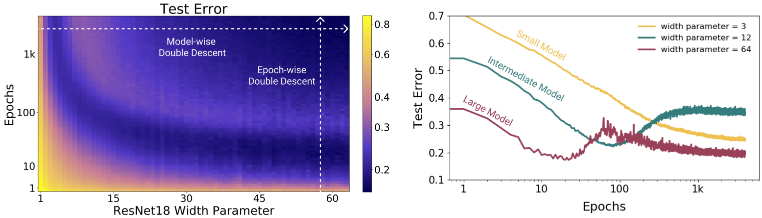


Figure 5: 双下降现象图示（源自 *Nakkiran, Preetum, et al. "Deep Double Descent: Where Bigger Models and More Data Hurt." ICLR 2020*）

U 型曲线的成因分析

- 深度神经网络不再过参数化（满足 $d < m$）
- Mixup 方法引入了噪声标签

神经网络优先学习干净数据

当使用含随机标签的部分数据训练神经网络时：

- ▷ 网络会优先学习干净数据
- ▷ 随后逐渐过拟合噪声标签数据

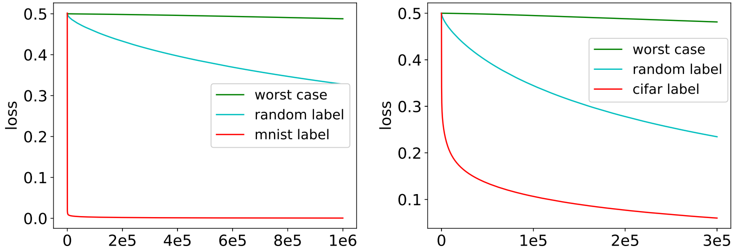


Figure 6: 干净数据与噪声数据的收敛过程（源自 *Arora, Sanjeev, et al. "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks." ICML 2019*）

案例研究：随机特征模型的回归场景

- 设 $\mathcal{Y} = \mathbb{R}$，则 $f : \mathcal{X} \to \mathbb{R}$ 为回归函数
- 定义 $\widetilde{Y}^* = f(\widetilde{X})$，$Z \triangleq \widetilde{Y} - \widetilde{Y}^*$
  此时 Z 表示 Mixup 方法引入的数据相关噪声
  例如，当 f 是 $\rho > 0$ 强凸函数时，$Z \geq \frac{\rho}{2}\lambda(1-\lambda)||X - X'||_2^2$
- 给定 $\widetilde{S} = \{(\widetilde{X}_i, \widetilde{Y}_i)\}_{i=1}^m$ 和模型 $\theta^T\phi(X)$，其中 $\phi : \mathcal{X} \to \mathbb{R}^d$ 为固定特征映射，$\theta \in \mathbb{R}^d$。使用均方误差损失：

$$\hat{R}_{\widetilde{S}}(\theta) \triangleq \frac{1}{2m}\left|\left|\theta^T\widetilde{\Phi} - \widetilde{\mathrm{Y}}^T\right|\right|_2^2,$$

其中 $\widetilde{\Phi} = [\phi(\widetilde{X}_1), \phi(\widetilde{X}_2), \ldots, \phi(\widetilde{X}_m)] \in \mathbb{R}^{d\times m}, \widetilde{\mathrm{Y}} = [\widetilde{Y}_1, \widetilde{Y}_2, \ldots, \widetilde{Y}_m] \in \mathbb{R}^m$。

案例研究：随机特征模型的回归场景

给定 $\widetilde{S}$，期望总体风险为：

$$R_t \triangleq \mathbb{E}_{\theta_t, X, Y} \left|\left|\theta_t^T \phi(X) - Y\right|\right|_2^2 .$$

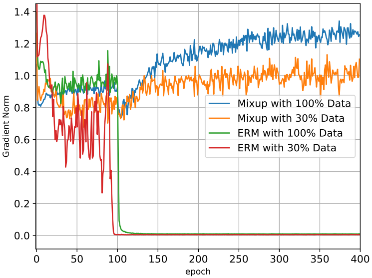
定理 2 (总体风险的动态演化)

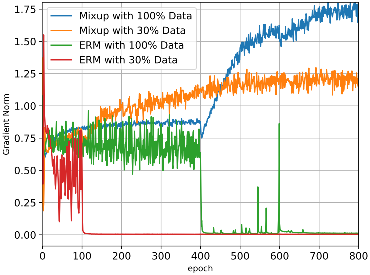
给定合成数据集 $\widetilde{S}$，假设 $\theta_0 \sim \mathcal{N}(0, \xi^2 \mathrm{I}_d)$，$||\phi(X)||^2 \leq C_1/2$ （$C_1 > 0$ 为常数），且 $|Z| \leq \sqrt{C_2}$ （$C_2 > 0$ 为常数），则存在：

$$R_t - R^* \leq C_1 \sum_{k=1}^{d} \left[\underbrace{\left(\xi_k^2 + \theta_k^{*2}\right) e^{-2\eta\mu_k t}}_{\text{递减项}} + \underbrace{\frac{C_2}{\mu_k} \left(1 - e^{-\eta\mu_k t}\right)^2}_{\text{递增项}} \right] + 2\sqrt{C_1 R^* \zeta},$$

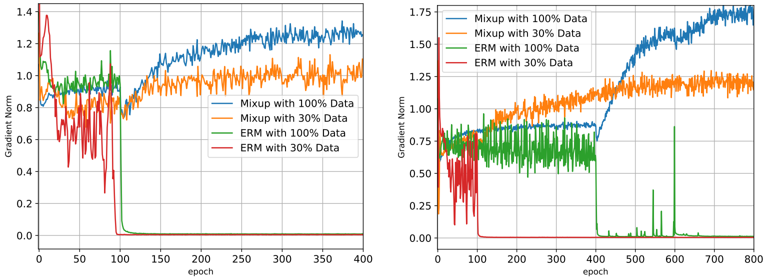
其中 $R^* = \mathbb{E}_{X,Y} \left|\left|Y - \theta^{*T} \phi(X)\right|\right|_2^2$，$\zeta = \sum_{k=1}^{d} \max\{\xi_k^2 + \theta_k^{*2}, \frac{C_2}{\mu_k}\}$，$\mu_k$ 为矩阵 $\frac{1}{m} \widetilde{\Phi} \widetilde{\Phi}^T$ 的第 k 个特征值。

Mixup 训练中梯度范数不消失





Mixup 训练中梯度范数不消失



核心结论：

- ▷ 错误的目标函数仍具效用，关键在于优化轨迹/动态过程
- ▷ 分析泛化性要与数据和算法特性相结合

基础信息度量

- 熵：$H(X) = \mathbb{E}_{P_X}\left[\log \frac{1}{P(X)}\right]$，$H(X,Y) = \mathbb{E}_{P_{X,Y}}\left[\log \frac{1}{P(X,Y)}\right]$，$H(X|Y) = \mathbb{E}_{P_{X,Y}}\left[\log \frac{1}{P(X|Y)}\right]$
  - 离散型 X 满足 $H(X) \geq 0$
  - 链式法则：$H(X,Y) = H(X|Y) + H(Y)$
  - 条件作用降低熵：$H(X|Y) \leq H(X)$
  - 最大值定理：离散型 $H(X) \leq \log |\mathcal{X}|$
- 相对熵（KL 散度）：$\mathrm{D_{KL}}(Q\|P) = \mathbb{E}_Q\left[\log \frac{Q(X)}{P(X)}\right]$
  - 非负性：$\mathrm{D_{KL}}(Q\|P) \geq 0$，当且仅当 $Q = P$ 时取等
  - 非对称性：$\mathrm{D_{KL}}(Q\|P) \neq \mathrm{D_{KL}}(P\|Q)$

基础信息度量

- 互信息：$I(X;Y) = \mathbb{E}_{P_{X,Y}}\left[\log \frac{P(X,Y)}{P(X)P(Y)}\right] = \mathrm{D}_{\mathrm{KL}}\left(P_{X,Y} \| P_X P_Y\right)$
  - 非负性：$I(X;Y) \geq 0$，当且仅当 $X \perp\!\!\!\perp Y$ 时取等
  - 等价形式：$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$
  - 对称性：$I(X;Y) = I(Y;X)$
  - 条件 KL 表达：$I(X;Y) = \mathbb{E}_{P_{X,Y}}\left[\log \frac{P(X|Y)}{P(X)}\right] = \mathbb{E}_{P_Y}\left[\mathrm{D}_{\mathrm{KL}}\left(P_{X|Y} \| P_X\right)\right]$
- 条件互信息：$I(X;Y|Z) = \mathbb{E}_{P_{X,Y,Z}}\left[\log \frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)}\right] = H(X|Z) - H(X|Y,Z)$

基础信息度量

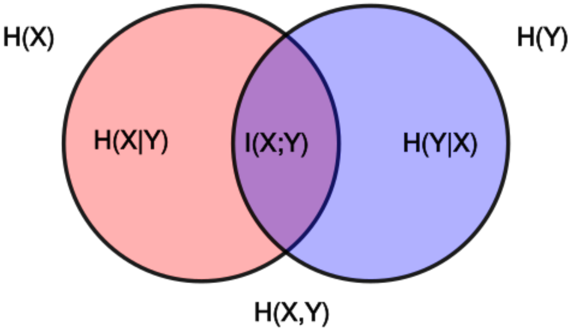


Figure 7: 维恩图（信息量关系示意）。

重要性质

- 链式法则：
  - $H(X_1, X_2, \dots, X_n) = \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1)$
  - $I(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^n I(X_i; Y | X_{i-1}, \dots, X_1)$
  - $\mathrm{D_{KL}}\left(Q_{X,Y} || P_{X,Y}\right) = \mathrm{D_{KL}}\left(Q_X || P_X\right) + \mathrm{D_{KL}}\left(Q_{Y|X} || P_{Y|X}\right)$
- 数据处理不等式（**DPI**）：
  若 $X - Y - Z$ 形成马尔可夫链（即 $P_{X,Z|Y} = P_{X|Y} P_{Z|Y}$），则

  $$I(X;Y) \geq I(X;Z)$$

  例： $f(X,Y) - (X,Y) - Z$ 为马尔可夫链时，
  $I(X,Y;Z) \geq I(f(X,Y);Z)$
- 其他重要工具或性质：费诺不等式，高斯分布的 **KL** 散度，方差相同时高斯分布具有最大熵，等等
- 推荐教材： *Thomas M. Cover, Joy A. Thomas.* 信息论基础（*Elements of Information Theory*）， *Wiley-Interscience*， *2006*

机器学习中的泛化

- 机器学习中真正关心的是模型的**测试性能**

机器学习中的泛化

- 机器学习中真正关心的是模型的**测试性能**
- 泛化误差 = 测试误差 - 训练误差

机器学习中的泛化

- 机器学习中真正关心的是模型的**测试性能**
- 泛化误差 = 测试误差 - 训练误差
  - 理想情况下，希望训练误差 ≈ 0 且泛化误差 ≈ 0
    $\Longleftrightarrow$ 测试误差 ≈ 0

机器学习中的泛化

- 机器学习中真正关心的是模型的**测试性能**
- 泛化误差 = 测试误差 - 训练误差
  - 理想情况下，希望训练误差 ≈ 0 且泛化误差 ≈ 0 $\Longleftrightarrow$ 测试误差 ≈ 0
  - 在实践中，无法获悉数据真实分布

机器学习中的泛化

- 机器学习中真正关心的是模型的**测试性能**
- 泛化误差 = 测试误差 - 训练误差
  - 理想情况下，希望训练误差 ≈ 0 且泛化误差 ≈ 0
    $\Longleftrightarrow$ 测试误差 ≈ 0
  - 在实践中，无法获悉数据真实分布
    $\Longrightarrow$ 小的训练损失和泛化界限/保证 可以带来小的测试误差。

机器学习中的泛化

- 机器学习中真正关心的是模型的**测试性能**
- 泛化误差 = 测试误差 - 训练误差
  - 理想情况下，希望训练误差 ≈ 0 且泛化误差 ≈ 0
    $\Longleftrightarrow$ 测试误差 ≈ 0
  - 在实践中，无法获悉数据真实分布
    $\Longrightarrow$ 小的训练损失和泛化界限/保证 可以带来小的测试误差。
- 高概率意义下泛化边界，

$$\mathrm{P}(ts\_error - tr\_error \geq \epsilon) \leq \delta.$$

即，以不低于 $1-\delta$ 的概率, 泛化误差不超过

$$ts\_error - tr\_error \leq \epsilon.$$

经典统计学原理中的泛化理论（如 Rademacher Complexity）：

$$\epsilon = O\left(\frac{\text{Complexity Measure}}{n}\right).$$

传统泛化理论对机器学习的启示

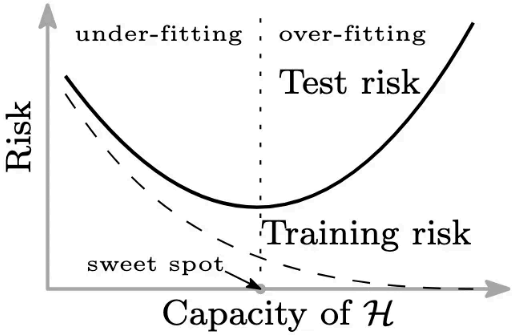


Figure 8: 传统 U-型泛化曲线

传统泛化理论对机器学习的启示

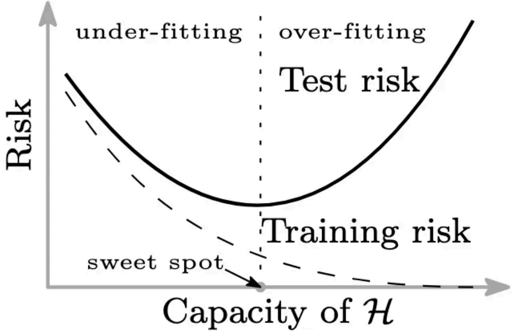


Figure 8: 传统 U-型泛化曲线

$\Longrightarrow$传统统计学习原理认为模型越复杂泛化越糟糕

现代深度学习泛化

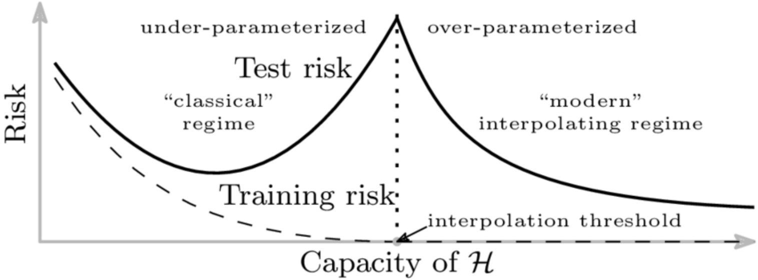


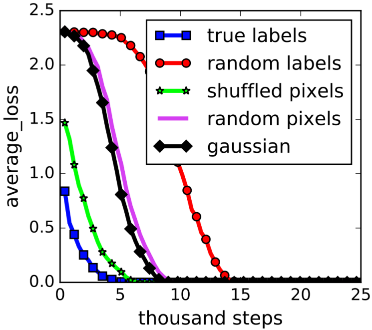
Figure 9: 双下降泛化曲线

$\Longrightarrow$ 过参数化网络中模型越复杂泛化性能会提升

深度学习中的泛化

ICLR 2017 最佳论文奖: Zhang, Chiyuan, et al. "Understanding deep learning requires rethinking generalization."

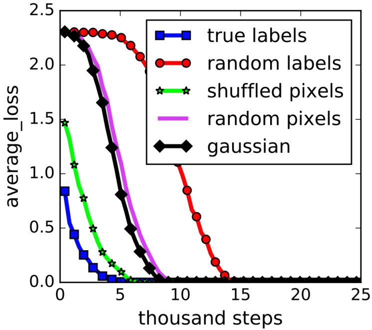
深度神经网络可完美拟合为训练数据随机生成的标签



深度学习中的泛化

ICLR 2017 最佳论文奖: Zhang, Chiyuan, et al. "Understanding deep learning requires rethinking generalization."

深度神经网络可完美拟合为训练数据随机生成的标签

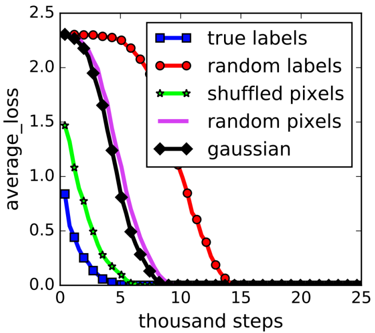


- 深度神经网络参数数量 > 训练数据数量，是个高度复杂的模型 $\Longrightarrow$ 在从未见过的数据上仍然表现良好

深度学习中的泛化

ICLR 2017 最佳论文奖: Zhang, Chiyuan, et al. "Understanding deep learning requires rethinking generalization."

深度神经网络可完美拟合为训练数据随机生成的标签



- 深度神经网络参数数量 > 训练数据数量，是个高度复杂的模型 $\Longrightarrow$ 在从未见过的数据上仍然表现良好
- 传统泛化边界 $ts\_error - tr\_error \leq \mathcal{O}(\frac{\hat{\mathfrak{R}}_n(\mathcal{F})}{n}) \Longrightarrow$ 是空虚的 (*vacuous*)!

深度学习中的泛化

- **挑战：现代深度学习需要新的泛化理论**
  - 如何让泛化理论解释深度学习中的泛化
  - 如何用泛化边界提升深度学习中的表现
  - 如何将泛化分析拓展到更广的学习设置
- **研究思路：模型泛化能力不仅与模型复杂度有关，还与算法特性（如 SGD）相关$\Longrightarrow$ 基于信息论的泛化分析**

相关数学符号说明

- 训练数据集：$S=\{Z_i\}_{i=1}^n \in \mathcal{Z}$ 从未知分布 μ 中独立同分布采样
- 假设集空间：$\mathcal{W} \subseteq \mathbb{R}^d$；
- 机器学习算法：$\mathcal{A}:\mathcal{Z}^n \to \mathcal{W}$；由 $P_{W|S}$ 刻画
- 损失函数：$\ell:\mathcal{W}\times\mathcal{Z}\to\mathbb{R}_0^+$

相关数学符号说明

- 训练数据集：$S=\{Z_i\}_{i=1}^n \in \mathcal{Z}$ 从未知分布 μ 中独立同分布采样
- 假设集空间：$\mathcal{W} \subseteq \mathbb{R}^d$；
- 机器学习算法：$\mathcal{A}: \mathcal{Z}^n \to \mathcal{W}$；由 $P_{W|S}$ 刻画
- 损失函数：$\ell: \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_0^+$
- 泛化误差：
  - 测试误差：$L_\mu(w) \triangleq \mathbb{E}_{Z\sim\mu}[\ell(w,Z)]$；期望测试误差：$L_\mu = \mathbb{E}_W\left[L_\mu(W)\right]$
  - 经验误差：$L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(w,Z_i)$；期望经验误差：$L_n = \mathbb{E}_{W,S}\left[L_S(W)\right]$
  - 期望泛化误差：$\mathcal{E}_\mu(\mathcal{A}) \triangleq L_\mu - L_n = \mathbb{E}_{W,S}[L_\mu(W) - L_S(W)]$

第一个基于互信息的泛化边界

引理 2 (Xu and Raginsky [2017])

假设损失 $\ell(w, Z)$ 对任意 $w \in \mathcal{W}$ 都是 R-次高斯的[1]。算法 $\mathcal{A}$ 的泛化误差界限为

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \sqrt{\frac{2R^2}{n} I(W; S)}.$$

[1] 一个随机变量 X 是 R-次高斯的，即对于任意的 ρ，$\log \mathbb{E} \exp(\rho(X - \mathbb{E}X)) \leq \rho^2 R^2 / 2$。

第一个基于互信息的泛化边界

引理 2 (Xu and Raginsky [2017])

假设损失 $\ell(w,Z)$ 对任意 $w \in \mathcal{W}$ 都是 R-次高斯的[1]。算法 $\mathcal{A}$ 的泛化误差界限为

$$|\mathcal{E}_\mu(\mathcal{A})| \leq \sqrt{\frac{2R^2}{n} I(W;S)}.$$

互信息 $I(W;S) \triangleq \mathrm{D_{KL}}(P_{W,S} || P_W \otimes P_S)$。

$\Longrightarrow$ 依赖于分布和算法

[1] 一个随机变量 X 是 R-次高斯的，即对于任意的 ρ，$\log \mathbb{E} \exp(\rho(X - \mathbb{E}X)) \leq \rho^2 R^2/2$。

随机梯度朗之万动力学（SGLD）

SGLD 更新规则：

$$W_t \triangleq W_{t-1} - \lambda_t g(W_{t-1}, B_t) + N_t$$

其中

$$g(w, B_t) \triangleq \frac{1}{b} \sum_{z \in B_t} \nabla_w \ell(w, z)$$

- λ_t：学习率
- b：批量大小
- B_t：第 t 次更新使用的批量数据
- $N_t \sim \mathcal{N}(0, \sigma_t^2 \mathrm{I}_d)$：高斯噪声项

假设 SGLD 输出 W_T 作为训练得到的模型参数。

SGLD 的信息论界

$$
\begin{aligned}
I(W_T;S) =& I\left(W_{T-1} - \lambda_T g(W_{T-1}, B_T) + N_T; S\right) \\
\leq& I\left(W_{T-1}, -\lambda_T g(W_{T-1}, B_T) + N_T; S\right) \qquad (1)\\
=& I(W_{T-1}; S) + I\left(-\lambda_T g(W_{T-1}, B_T) + N_T; S|W_{T-1}\right) \qquad (2)\\
&\vdots \\
\leq& \sum_{t=1}^{T} I\left(-\lambda_t g(W_{t-1}, B_t) + N_t; S|W_{t-1}\right)
\end{aligned}
$$

$$
\begin{aligned}
&I\left(-\lambda_t g(W_{t-1}, B_t) + N_t; S|W_{t-1}\right) \\
=& \mathbb{E}_{S,W_{t-1}}\left[\mathrm{D_{KL}}\left(Q_{-\lambda_t g(W_{t-1},B_t)+N_t|S,W_{t-1}} || P_{-\lambda_t g(W_{t-1},B_t')+N_t|W_{t-1}}\right)\right] \\
\leq& \frac{d}{2}\mathbb{E}_{W_{t-1}} \log\left(1 + \frac{\lambda_t^2 \mathbb{E}_S^{W_{t-1}} \|g - \mathbb{E}g\|^2}{d\sigma_t^2}\right).
\end{aligned}
$$

SGLD 的信息论误差界

定理 3

SGLD 的泛化误差上界为：

$$\mathcal{E}_\mu\left(\mathcal{A}_{SGLD}\right) \lesssim \sqrt{\frac{d}{n}\sum_{t=1}^{T}\mathbb{E}\log\left(1+\frac{\lambda_t^2\mathbb{E}\|g-\mathbb{E}g\|^2}{d\sigma_t^2}\right)}.$$

随机梯度下降 (SGD)

mini-SGD 更新法则：

$$W_t \triangleq W_{t-1} - \lambda_t g(W_{t-1}, B_t),$$

其中

$$g(w, B_t) \triangleq \frac{1}{b} \sum_{z \in B_t} \nabla_w \ell(w, z),$$

- λ_t：学习率
- b：批量大小
- B_t：第 t 次更新使用的批量。

假设 SGD 输出 W_T 作为学习到的模型参数。**基于互信息界的应用难点：** $I(W_T; S)$ 在 SGD 中过大

构造辅助动力学过程（仅存在于分析中）

Ziqiao Wang, *and Yongyi Mao. "On the Generalization of Models Trained with SGD: Information-Theoretic Bounds and Implications." ICLR 2022.*

定义 $\{\sigma_t\}_{t=1}^{T}$ 是一系列正实数。

令 $\widetilde{W}_0 \triangleq W_0$, $\widetilde{W}_t \triangleq \widetilde{W}_{t-1} - \lambda_t g(W_{t-1}, B_t) + N_t$, for $t > 0$, 其中 $N_t \sim \mathcal{N}(0, \sigma_t^2 \mathrm{I}_d)$ 是高斯噪声。

$$
\begin{array}{ccccccccccc}
 & & N_1 & & N_2 & & \cdots & & N_{T-1} & & N_T \\
 & & \downarrow & & \downarrow & & & & \downarrow & & \downarrow \\
\widetilde{W}_0 & \rightarrow & \widetilde{W}_1 & \rightarrow & \widetilde{W}_2 & \rightarrow & \cdots & \rightarrow & \widetilde{W}_{T-1} & \rightarrow & \widetilde{W}_T \\
\| & \nearrow & & \nearrow & & \nearrow & & & & \nearrow & \\
W_0 & \rightarrow & W_1 & \rightarrow & W_2 & \rightarrow & \cdots & \rightarrow & W_{T-1} & \rightarrow & W_T
\end{array}
$$

令 $\Delta_t = \sum_{\tau=1}^{t} N_\tau \Longrightarrow \widetilde{W}_t = W_t + \Delta_t$.

通过辅助动力学过程进行界限

将这个辅助权重过程表示为 $\mathcal{A}_{AWP}$，并令 $\mathcal{A}_{SGD}$ 为 SGD 的原始算法，

$$
\begin{aligned}
\mathcal{E}_\mu\left(\mathcal{A}_{SGD}\right) =& \mathcal{E}_\mu\left(\mathcal{A}_{SGD}\right) + \mathcal{E}_\mu\left(\mathcal{A}_{AWP}\right) - \mathcal{E}_\mu\left(\mathcal{A}_{AWP}\right) \\
\leq & \underbrace{\mathcal{O}\left(\sqrt{\frac{I(\widetilde{W}_T;S)}{n}}\right)}_{\text{引理 2}} + \underbrace{\left|\mathcal{E}_\mu\left(\mathcal{A}_{SGD}\right) - \mathcal{E}_\mu\left(\mathcal{A}_{AWP}\right)\right|}_{\text{残差项}} \qquad (3) \\
\lesssim & \text{SGD 轨迹中梯度的分散度} + \text{SGD 解的平坦度}.
\end{aligned}
$$

主定理

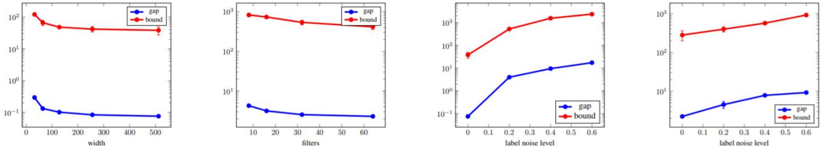
定理 4 (Wang and Mao [2022])

随机梯度下降（SGD）的泛化误差的上界限制为

$$|\mathcal{E}_\mu(\mathcal{A}_{SGD})| \lesssim \sqrt{\sum_{t=1}^{T} \frac{\mathbb{E}\left[\mathbb{V}_t(W_{t-1})\right]}{n\sigma_t^2}} + \mathbb{E}\left[L_S(W_T+\Delta_T) - L_S(W_T)\right]. \tag{4}$$

- 梯度分散 (Gradient Dispersion)：
  $\mathbb{V}_t(w) \triangleq \mathbb{E}_S\left[\|g(w,B_t) - \mathbb{E}_{W,Z}\left[\nabla_w \ell(W,Z)\right]\|_2^2\right]$

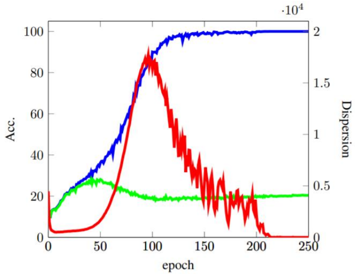
实验：主定理边界的验证



(a) MLP on MNIST (b) AlexNet on CIFAR10 (c) MLP on MNIST (d) AlexNet on CIFAR10

Figure 10: 估计的上界和经验泛化差距（“gap”）。横轴为网络宽度（(a) 和 (b)）。横轴为标签噪声水平（(c) 和 (d)）的函数。

实验：梯度分散的逐时期双下降



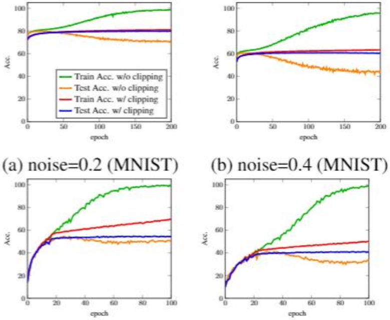
- $\mathbb{V}$ 迅速下降；训练准确度和测试准确度都提高；$\Longrightarrow$ “泛化（Generalization）” 阶段
- $\mathbb{V}$ 开始增加，直到达到峰值；训练准确度和测试准确度逐渐发散；$\Longrightarrow$ “记忆（Memorization）” 阶段
- $\mathbb{V}$ 再次下降；训练和测试曲线分别达到其最大值和最小值；$\Longrightarrow$ “收敛” 阶段

算法：动态梯度裁剪

Algorithm 1 Dynamic Gradient Clipping

Require: Training set S, Batch size b, Loss function ℓ, Ir
λ, Initial minimum gradient norm $\mathcal{G}$, Number of itera
step T_c

for $t \leftarrow 1$ to T **do**
  Sample $\mathcal{B} = \{\boldsymbol{z}_i\}_{i=1}^{b}$ from training set S
  Compute gradient:
  $g_{\mathcal{B}} \leftarrow \sum_{i=1}^{b} \nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}_{t-1}, \boldsymbol{z}_i)/b$
  if $t > T_c$ **then**
    if $\|g_{\mathcal{B}}\|_2 > \mathcal{G}$ **then**
      $g_{\mathcal{B}} \leftarrow \alpha \cdot \mathcal{G} \cdot g_{\mathcal{B}}/\|g_{\mathcal{B}}\|_2$
    else
      $\mathcal{G} \leftarrow \|g_{\mathcal{B}}\|_2$
    end if
  end if
  Update parameter: $\boldsymbol{w}_t \leftarrow \boldsymbol{w}_{t-1} - \lambda \cdot g_{\mathcal{B}}$
end for



(a) noise=0.2 (MNIST) (b) noise=0.4 (MNIST)

(c) noise=0.2 (CIFAR10) (d) noise=0.4 (CIFAR10)

算法：高斯模型扰动 (GMP)

- 主定理启发：当 w^* 处的经验误差曲面（loss landscape）是平坦的，即对 w^* 的小扰动不敏感$\Longrightarrow$ 泛化性能好：

$$\min_{w} L_s(w) + \rho\mathbb{E}_{\Delta\sim\mathcal{N}(0,\sigma^2\mathrm{I}_d)}\left[L_s(w+\Delta) - L_s(w)\right],$$

  其中 ρ 是一个超参数。

- 用 k 次高斯采样平均来估计上述期望，形成以下优化问题：

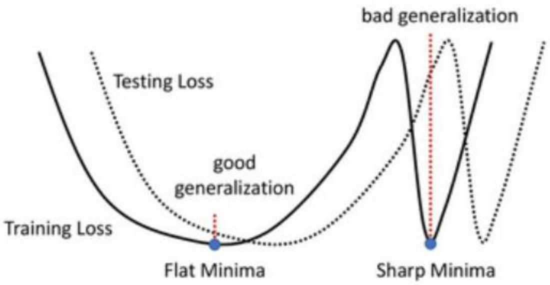
$$\min_{w} \frac{1}{b}\sum_{z\in B}\left((1-\rho)\ell(w,z) + \rho\frac{1}{k}\sum_{i=1}^{k}(\ell(w+\delta_i,z))\right).$$

GMP 算法实验结果

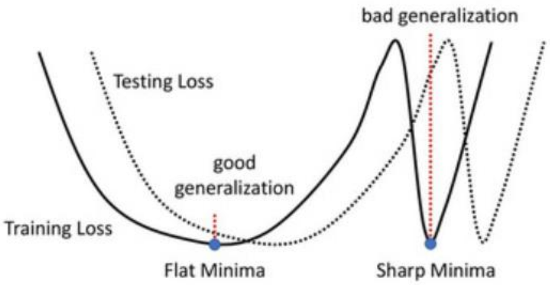
方法	SVHN	CIFAR-10	CIFAR-100
ERM	96.86±0.060	93.68±0.193	72.16±0.297
Dropout	97.04±0.049	93.78±0.147	72.28±0.337
L. S.	96.93±0.070	93.71±0.158	72.51±0.179
Flooding	96.85±0.085	93.74±0.145	72.07±0.271
MixUp	96.91±0.057	**94.52±0.112**	73.19±0.254
Adv. Tr.	97.06±0.091	93.51±0.130	70.88±0.145
GMP^{3}	**97.18±0.057**	94.33±0.094	74.45±0.256
GMP^{10}	97.09±0.068	94.45±0.158	**75.09±0.285**

Table 1: VGG16 的 Top-1 分类准确率（%）。报告结果基于 10 次实验运行。上标表示 k 的取值。

▷ SGD 的解需要平坦（flat minima）:



- SGD 的解需要平坦（flat minima）：



- 后续拓展：用随机微分方程（SDE）建模 SGD 参数更新过程
Two Facets of SDE Under an Information-Theoretic Lens: Generalization of SGD via Training Trajectories and via Terminal States (*with Yongyi Mao, UAI'24*)

1. 数据混合增强带来的启示：训练轨迹 v.s. 最优解
2. 基于信息论的泛化理论
3. 分布外的泛化: 领域自适应
4. 参考文献

领域自适应

问题设置

- 给定源领域数据：$\{X_i, Y_i\} \overset{i.i.d.}{\sim} \mu$
- 目标：获取目标领域模型 $\{X, Y\} \sim \nu$
- **实际目标**：在低成本的前提下，高效地在相关数据分布之间迁移机器学习模型。

额外符号说明

- 源域数据 $Z=(X,Y)\sim\mu$ 与目标域数据 $Z'=(X',Y')\sim\mu'$
- 已标记源样本：$S=\{Z_i\}_{i=1}^n \overset{\text{i.i.d}}{\sim} \mu^{\otimes n}$ 未标记目标样本：$S'_{X'}=\{X'_j\}_{j=1}^m \overset{\text{i.i.d}}{\sim} P_{X'}^{\otimes m}$
- 泛化误差 = 目标域测试误差 - 源域训练误差：

$$\begin{aligned}\mathcal{E} &\triangleq \mathbb{E}_{W,S,S'_{X'}}[R_{\mu'}(W)-R_S(W)]\\ &=\mathbb{E}_{W,S,S'_{X'}}[L_{\mu'}(W){\color{red}-L_\mu(W)+L_\mu(W)}-L_S(W)]\end{aligned}$$

梯度惩罚作为通用正则化器

Ziqiao Wang, *and Yongyi Mao. "Information-Theoretic Analysis of Unsupervised Domain Adaptation." ICLR 2023.*

考虑随机梯度朗之万动力学 (SGLD)。每个时间步 t:

- 已标记源域小批量数据：Z_{B_t}；未标记目标域小批量数据：X'_{B_t}
- 梯度计算：$G_t = g(W_{t-1}, Z_{B_t}, X'_{B_t})$
- 更新规则：$W_t = W_{t-1} - \eta_t G_t + N_t$，其中 $N_t \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_d)$

定理 5

设总迭代次数为 T，则

$$|\mathcal{E}| \leq \sqrt{\frac{R^2}{n}\sum_{t=1}^{T}\frac{\eta_t^2}{\sigma_t^2}\mathbb{E}_{S'_{X'}, W_{t-1}, S}\left[\left\|G_t - \mathbb{E}_{Z_{B_t}}[G_t]\right\|^2\right]} + \sqrt{2R^2 \mathrm{D}_{\mathrm{KL}}(\mu \| \mu')}$$

限制梯度范数 $\Longrightarrow$ 降低泛化误差 $|\mathcal{E}|$

KL 引导的域自适应

- *Nguyen, A. Tuan, et al. "KL Guided Domain Adaptation." ICLR 2022.*

KL 引导的域自适应

- *Nguyen, A. Tuan, et al. "KL Guided Domain Adaptation." ICLR 2022.*
  - **表征网络**:
    - 输入：数据
    - 输出：均值向量 $\hat{\mu} \in \mathbb{R}^d$ 和方差向量 $\hat{\sigma}^2 \in \mathbb{R}^d$
    - 源域高斯分布 $\mathcal{N}(\hat{\mu}_\mathrm{s}, \hat{\sigma}_\mathrm{s}^2 \mathrm{I}_d)$；目标域高斯分布 $\mathcal{N}(\hat{\mu}_\mathrm{t}, \hat{\sigma}_\mathrm{t}^2 \mathrm{I}_d\})$
    - 最小化两个高斯分布之间的 KL 散度

KL 引导的域自适应

- *Nguyen, A. Tuan, et al. "KL Guided Domain Adaptation." ICLR 2022.*
  - **表征网络**：
    - 输入：数据
    - 输出：均值向量 $\hat{\mu} \in \mathbb{R}^d$ 和方差向量 $\hat{\sigma}^2 \in \mathbb{R}^d$
    - 源域高斯分布 $\mathcal{N}(\hat{\mu}_{\mathrm{s}}, \hat{\sigma}_{\mathrm{s}}^2 \mathrm{I}_d)$；目标域高斯分布 $\mathcal{N}(\hat{\mu}_{\mathrm{t}}, \hat{\sigma}_{\mathrm{t}}^2 \mathrm{I}_d\})$
    - 最小化两个高斯分布之间的 KL 散度
  - **分类器**：
    - 从源域高斯分布 $\mathcal{N}(\hat{\mu}_{\mathrm{s}}, \hat{\sigma}_{\mathrm{s}}^2 \mathrm{I}_d)$ 中采样
    - 最小化交叉熵损失

实验结果：RotatedMNIST

RotatedMNIST 基于 MNIST 数据集构建，包含六个域，分别对应旋转角度为 $0°, 15°, 30°, 45°, 60°$ 和 $75°$ 的 MNIST 图像。

Table 2: RotatedMNIST 实验结果

	以 0° 为源域的 RotatedMNIST					
方法	15°	30°	45°	60°	75°	**平均**
ERM	97.5±0.2	84.1±0.8	53.9±0.7	34.2±0.4	22.3±0.5	58.4
DANN	97.3±0.4	90.6±1.1	68.7±4.2	30.8±0.6	19.0±0.6	61.3
MMD	97.5±0.1	95.3±0.4	73.6±2.1	44.2±1.8	32.1±2.1	68.6
CORAL	97.1±0.3	82.3±0.3	56.0±2.4	30.8±0.2	27.1±1.7	58.7
WD	96.7±0.3	93.1±1.2	64.1±3.3	41.4±7.6	27.6±2.0	64.6
KL	97.8±0.1	97.1±0.2	93.4±0.8	75.5±2.4	68.1±1.8	86.4
ERM-GP	97.5±0.1	86.2±0.5	62.0±1.9	34.8±2.1	26.1±1.2	61.2
KL-GP	98.2±0.2	96.9±0.1	95.0±0.6	**88.0±8.1**	**78.1±2.5**	**91.2**

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 2017.

Ziqiao Wang and Yongyi Mao. On the generalization of models trained with SGD: Information-theoretic bounds and implications. In *International Conference on Learning Representations*, 2022.

Thanks!