第 8 章 鲁棒优化与机器学习（Machine learning）

覃含章

本章中我们介绍鲁棒优化与机器学习相结合的一些研究的最新进展。这是一个很有趣的研究方向并有诸多实际应用, 其出发点是: 传统统计学习和机器学习模型中，往往数据噪声的假设是随机（stochastic）的。然而，我们却也可以假设数据噪声并不满足一个随机分布（在机器学习中，经常会假设一个正态或者 sub-Gaussian 的分布），而是只是假设其在一个不确定集中。由这种新的假设，我们可以得到许多新的机器学习模型与算法，比如著名的 LASSO 算法虽然并不是最早通过鲁棒优化的模型推出的，但也可以看成一种具有鲁棒性的回归算法。

8.1 从鲁棒优化角度看回归模型（Regression）：正则性（Regularization）和鲁棒性（Robustness）

我们知道，著名的 LASSO 算法实际上是求解带有 L_1 正则项的线性回归模型。即，一般是考虑求解这样一个优化问题

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2 + \lambda\|\boldsymbol{\beta}\|_1,$$

其中 $\boldsymbol{X} \in \mathbb{R}^{M\times N}$ 是描述数据特征（feature）的矩阵，$\boldsymbol{y} \in \mathbb{R}^M$ 是描述数据标签（label）的向量，$\lambda > 0$ 是正则项前的系数。在一些限制条件和假设下，可以证明存在某个自然数 k, 使得 LASSO 等价于求解如下问题（我们使用 $\|\cdot\|_0$ 表示一个向量非零元素的个数，即 $\|\boldsymbol{\beta}\|_0 = \mathrm{card}(\{i : \boldsymbol{\beta}_i \neq 0\})$）:

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2 \\ &\text{s.t. } \|\boldsymbol{\beta}\|_0 \leq k. \end{aligned}$$

也就是说在这种情况下 LASSO 所得到的解是稀疏（sparse）的。这里我们主要考虑这样一种非概率的统计模型，即我们认为我们只能得到 $\boldsymbol{X}$ 的一个带有误差的样本 $\boldsymbol{X}'$。我们利用鲁棒优化的思想，认为 $\boldsymbol{X}' = \boldsymbol{X} + \boldsymbol{\Delta}$，而 $\boldsymbol{\Delta} \in \mathcal{U} \subset \mathbb{R}^{M\times N}$，这里的 $\mathcal{U}$ 就是我们的不确定集合（uncertainty set），注意这个集合是非随机的（non-stochastic）的。我们因此就可以考虑这样一个鲁棒线性回归问题:

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta}\in\mathcal{U}} \|\boldsymbol{y} - (\boldsymbol{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2. \tag{8.1}$$

这个鲁棒线性回归是个什么意思呢，也就是说我们现在优化的时候，所选择的 $\boldsymbol{\beta}$ 是最小化了不确定集里最差的那个 $\boldsymbol{X}'$，即我们要让“最坏情况”下的损失函数值最小。而在传统的 LASSO 或者线性回归中，我们的目标可以看成是要让期望的损失函数值最小。下面先初步解答如下问题：在线性回归模型中，我们什么时候可以将正则性和鲁棒性，这样一个来自统计/机器学习，一个来自优化理论的性质等同看待？这里就以回归模型中最出名的脊回归（Ridge Regression）和 LASSO 为例。

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2+\lambda\|\boldsymbol{\beta}\|_2=\min_{\boldsymbol{\beta}}\max_{\boldsymbol{\Delta}\in\mathcal{U}_{\text{RLS}}}\|\boldsymbol{y}-(\boldsymbol{X}+\boldsymbol{\Delta})\boldsymbol{\beta}\|_2, \tag{8.2}$$

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2+\lambda\|\boldsymbol{\beta}\|_1=\min_{\boldsymbol{\beta}}\max_{\boldsymbol{\Delta}\in\mathcal{U}_{\text{LASSO}}}\|\boldsymbol{y}-(\boldsymbol{X}+\boldsymbol{\Delta})\boldsymbol{\beta}\|_2. \tag{8.3}$$

注意到脊回归和 LASSO 说白了只是正则项不同（选用 L_1 和 L_2 正则）的最小二乘法（least squares method），那么我们就发现上面的结果告诉了我们它们都对应特定的鲁棒线性优化模型，只是对应的不确定集不同罢了！具体来说，我们有 $\mathcal{U}_{\text{RLS}}$ 和 $\mathcal{U}_{\text{LASSO}}$ 对应两个不同的二阶锥（second-order cone）约束集：

$$\mathcal{U}_{\text{RLS}}=\left\{\boldsymbol{\Delta}:\left(\sum_{ij}\Delta_{ij}^2\right)^{1/2}\leq\lambda\right\}, \tag{8.4}$$

$$\mathcal{U}_{\text{LASSO}}=\left\{\boldsymbol{\Delta}:\boldsymbol{\Delta}\text{的每列 }\Delta_i\text{ 都满足: }\|\Delta_i\|_2\leq\lambda\right\}. \tag{8.5}$$

那么这边我们就获得了对脊回归、LASSO 的一种基于鲁棒优化的新认识：这两种带正则项的线性回归其实可以看成一种鲁棒线性回归算法！那么自然，我们接下来应该也会对这两个问题感兴趣：

- 正则性和鲁棒性在线性回归中是否都是一回事？
- 如果不都是一回事的话，在什么条件下是？什么条件下不是？

定理 8.1: 鲁棒线性回归定理 (Bertsimas and Copenhaver, 2018)

（1）存在 $0<\alpha\leq 1$ 使得对任何 $\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta}$,

$$\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_p+\alpha\max_{\boldsymbol{\Delta}\in\mathcal{U}}\|\boldsymbol{\Delta}\boldsymbol{\beta}\|_p\leq\max_{\boldsymbol{\Delta}\in\mathcal{U}}\|\boldsymbol{y}-(\boldsymbol{X}+\boldsymbol{\Delta})\boldsymbol{\beta}\|_p\leq\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_p+\max_{\boldsymbol{\Delta}\in\mathcal{U}}\|\boldsymbol{\Delta}\boldsymbol{\beta}\|_p.$$

（2）我们令 $\mathcal{U}=\{\boldsymbol{\Delta}:\|\boldsymbol{\Delta}\|\leq\lambda\}$，其中的范数 $\|\cdot\|$ 如下表中所示，则有

$\|\cdot\|$	$\|\boldsymbol{\Delta}\|$ 的取值	$\alpha=1$ 的“当且仅当”条件
q-Frobenuis 范数	$\left(\sum_{ij}\lvert\Delta_{ij}\rvert^q\right)^{1/q}$	$p\in\{1,q,\infty\}$
q-谱范数	奇异值的 L_q 范数	$p\in\{1,2,\infty\}$
$(L_q,L_r)-$ 诱导范数	$\max_{\boldsymbol{\beta}}\frac{\|\boldsymbol{\Delta}\boldsymbol{\beta}\|_r}{\|\boldsymbol{\beta}\|_q}$	$p\in\{1,r,\infty\}$

♡

定理8.1是Bertsimas and Copenhaver (2018) 的文章中给出的基于上述两个问题的一般化回答。定理中的（1）表明一般来说我们都能用正则项的形式将鲁棒问题的取值控制住，但一般来说两者并不是完全一致的（文章中也给出了一些详细的例子来佐证）；

而（2）则给出了鲁棒性 = 正则性，即（1）中的 $\alpha = 1$ 的“当且仅当”条件。比如这其中的 q-Frobenuis 范数情形，其中 $p = 1$ 和 $q = 2$ 的时候就对应了我们前面的 (8.3) 和 (8.2)。表中的其他内容则表明类似结论也可以被推广到其它矩阵范数上。

8.2 基于对抗样本（Adversarial samples）的鲁棒学习（Robust learning）

本节我们将前一节仅仅针对回归模型的思路拓展，介绍在更一般的机器学习任务里，如果出现所谓的对抗样本，如何训练我们的模型，和相应的样本复杂度（相比于非对抗情境的机器学习任务）。我们主要考虑如下优化问题：

$$\min_{\boldsymbol{\theta}} \frac{1}{N}\sum_{i=1}^{N} \max_{\boldsymbol{x}_i' \in \mathcal{U}_i} g(\boldsymbol{\theta}, \boldsymbol{x}_i', y_i).$$

其中，$g(\boldsymbol{\theta}, \boldsymbol{x}_i', y_i)$ 是一个损失函数（和前一节不同，这里的 g 不一定要是某个范数了），比如说，现在我们也可以将它看成一个基于人工神经网络（artifical neural network）的损失函数。$\boldsymbol{\theta}$ 就是我们要优化的参数，$(\boldsymbol{x}_i, y_i)$ 是一个数据点，$\mathcal{U}_i$ 则是针对每个数据点定义的一个不确定集。我们注意到，这个最优化问题可以利用常见的交替方向法来求解。具体来说，算法每一步中我们将 $\boldsymbol{\theta}$ 的值固定，然后我们通过如下方式计算 $\boldsymbol{\Delta}_{x_i}$：

$$\boldsymbol{\Delta}_{x_i} = \arg \max_{\boldsymbol{\Delta}: \boldsymbol{x}_i + \boldsymbol{\Delta} \in \mathcal{U}_i} g(\boldsymbol{\theta}, \boldsymbol{x}_i + \boldsymbol{\Delta}, y_i). \tag{8.6}$$

当然这个优化问题 (8.6) 一般来说是难以直接求得的（我们这里没有限制 g），那么如果我们认为 g 是光滑的，就可以求解一个一阶泰勒展开的近似问题：

$$\boldsymbol{\Delta}_{x_i}' = \arg \max_{\boldsymbol{\Delta}: \boldsymbol{x}_i + \boldsymbol{\Delta} \in \mathcal{U}_i} g(\boldsymbol{\theta}, \boldsymbol{x}_i, y_i) + \left\langle \nabla_{\boldsymbol{x}} g(\boldsymbol{\theta}, \boldsymbol{x}, y_i), \boldsymbol{\Delta} \right\rangle. \tag{8.7}$$

根据式 (8.7) 我们就可以在固定 $\boldsymbol{\theta}$ 值的情况下更新 $\boldsymbol{\Delta}_{x_i}'$，也即 $\boldsymbol{x}_i' = \boldsymbol{x}_i + \boldsymbol{\Delta}_{x_i}'$。假设算法每步一共更新了 mb 次，那么在固定数据点集 $\{(\boldsymbol{x}_i', y_i)\}_{i=1}^{|mb|}$ 的情况下，我们就可以对 $\boldsymbol{\theta}$ 采用一步批梯度下降法（mini-batch gradient descent）的迭代。如此，我们就描述了我们的对抗训练（adversarial training）算法，算法的具体实现细节和一些数值实例可见Shaham et al. (2018) 的工作。本节我们接着讨论对抗训练中的一些复杂度问题，我们将仅限于讨论分类（classification）问题。

我们先定义分类错误（classification error）为：对分布 $\mathcal{P}: \mathbb{R}^d \times \{\pm 1\} \to \mathbb{R}$, 分类器（classifier）$f: \mathbb{R} \to \{\pm 1\}$ 的分类错误率 e 为 $e = \mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{P}}[f(\boldsymbol{x}) \neq y]$. 也就是分类错误率其实就是分类器出错的概率。然后我们将这个定义拓展，定义所谓的 $\mathcal{U}$-鲁棒分类错误率：对分布 $\mathcal{P}: \mathbb{R}^d \times \{\pm 1\} \to \mathbb{R}$, 定义 $\mathcal{U}: \mathbb{R}^d \to \mathrm{P}(\mathbb{R}^d)$（$\mathrm{P}(\mathbb{R}^d)$ 表示 $\mathbb{R}^d$ 的支撑集，即所有子集的集合）。分类器（classifier）$f: \mathbb{R} \to \{\pm 1\}$ 的 $\mathcal{U}$-鲁棒分类错误率 e 为 $e = \mathbb{P}_{(\boldsymbol{x},y)\sim\mathcal{P}}[\exists\, \boldsymbol{x}' \in \mathcal{U}(\boldsymbol{x}): f(\boldsymbol{x}') \neq y]$. Schmidt et al. (2018) 的文章在 $\mathcal{P}$ 为高斯分布和伯努利分布的假设下研究了 $\mathcal{U}(\boldsymbol{x}) = \mathcal{U}_\infty^\epsilon(x) = \{x' \in \mathbb{R}^d | \|x' - x\|_\infty \leq \epsilon\}$.

定理 8.2: 鲁棒分类样本复杂度定理 (Schmidt et al., 2018)

(1) 高斯模型：令 $(x_1, y_1), \ldots, (x_n, y_n)$ 是从 $\mathcal{N}(\theta^*, \sigma)$ 中独立同分布抽样得到的样本，其中高斯分布的参数满足 $\|\theta^*\|_2 = \sqrt{d}, \sigma \le c \cdot d^{1/4}$（$c > 0$ 是一个常数）.

- 有高概率 $f_{\hat{w}} = y_1 \cdot x_1$ 的分类错误率不超过 1%。
- 如果取 ϵ 使得 $\frac{1}{4}d^{-1/4} \le \epsilon \le \frac{1}{4}$，那么如果 $n \ge \epsilon^2\sqrt{d}$，有高概率 $f_{\hat{w}} = \frac{1}{n}\sum_{i=1}^{n} y_i x_i$ 的 $\mathcal{U}_\infty^\epsilon$-鲁棒分类错误率不超过 1%。
- 基于样本 $(x_1, y_1), \ldots, (x_n, y_n)$，对任意 0/1 分类器 f_n，如果 $n \le c'\frac{\epsilon^2\sqrt{d}}{\log d}$（$c' > 0$ 是一个常数），期望的 $\mathcal{U}_\infty^\epsilon$-鲁棒分类错误率至少为 $\frac{1}{2}(1 - 1/d)$.

(2) 伯努利模型：存在参数 $\theta^* \in \{\pm 1\}^d, \tau > 0$，使得抽样机制定义为先均匀地抽样 $y \in \{-1, 1\}$，再对 $\boldsymbol{x}$ 的每个坐标独立以 $1/2 + \tau$ 概率抽得 $y \cdot \theta_i^*$，以 $1/2 - \tau$ 概率抽得 $-y \cdot \theta_i^*$。令 $\tau \ge c \cdot d^{-1/4}$，$\epsilon < 3\tau < 1, \gamma < 1/2$。

- 有高概率 $f_{\hat{w}} = y_1 \cdot \boldsymbol{x}_1$ 的分类错误率不超过 1%。
- 如果取 ϵ 使得 $\frac{1}{4}d^{-1/4} \le \epsilon \le \frac{1}{4}$，那么如果 $n \ge \epsilon^2\sqrt{d}$，有高概率 $f_{\hat{w}} = y_1 \cdot T(\boldsymbol{x}_1)$ 的 $\mathcal{U}_\infty^\epsilon$-鲁棒分类错误率不超过 1%。$T$ 是一个非线性算子，使得 $\boldsymbol{x}$ 的每个非负坐标取 1，负坐标取 -1。
- 基于样本 $(x_1, y_1), \ldots, (x_n, y_n)$，对任意 0/1 分类器 f_n，如果 $n \le c'\frac{\epsilon^2\gamma^2 d}{\log d/\gamma}$（$c' > 0$ 是一个常数），期望的 $\mathcal{U}_\infty^\epsilon$-鲁棒分类错误率至少为 $\frac{1}{2} - \gamma$. ♡

定理8.2给我们最大的一个启示就是看起来对于分类问题，鲁棒分类错误是和样本遵循的分布高度相关的。具体来说，在高斯模型中，我们对于常规的分类错误来说只需要一个数据点就可以做到高概率的完美分类，但对于鲁棒分类错误我们至少需要 $\sqrt{d}$ 阶的样本数量才能做到比较好的分类（这是由线性分类器对于鲁棒分类错误的样本复杂度和鲁棒分类错误的样本复杂度下界放在一起说明的）。而作为对比，对于伯努利模型，除了对于常规的分类错误来说同样一个数据点就可以做到高概率的完美分类，对于鲁棒分类错误来说用一个非线性的分类器也只需要一个数据点就可以做到完美分类。而在高斯模型中，下界保证了任何非线性的分类器在理论上也无法突破 $\sqrt{d}$ 样本数量。

这便是Schmidt et al. (2018) 的工作主要要说明的，为此它们利用了两个著名的开放分类数据集，MNIST 和 CIFAR10，利用卷积神经网络和前面提到的对抗训练算法，它们发现基于前者训练出来的分类器，在测试数据集上利用 $\mathcal{U}_\infty^\epsilon$ 的定义扰动数据，可以达到很低的鲁棒分类错误率。而对于 CIFAR，虽然还能保持一般意义上很低的分类错误率，却有很高的鲁棒分类错误率。基于前面的结果，一种可以接受的解释就是 MNIST 这个数据集更接近伯努利模型，而 CIFAR10 更接近高斯模型。另外，我们也可以体会到实际上神经网络模型对于标准意义上的分类错误能达到很高的标准，但往往对于鲁棒分类错误就不那么在行了。

关于鲁棒分类器，Bertsimas et al. (2019) 给出了基于支持向量机（Support Vector Machine），逻辑回归（Logistic Regression）和决策树（Decision Tree）的鲁棒版本。在

大规模数据集上，他们发现这些更加具有鲁棒性的分类器可以提升传统机器学习模型的分类精度。具体来说，样本外的精度（out-of-sample accuracy）对支持向量机提升了5.3%，逻辑回归提升了 4.0%，决策树提升了 1.3%。

8.3 神经网络（Neural network）中的分布鲁棒优化

本节我们讨论上一节所引入的问题的一种更高级的尝试，利用分布鲁棒优化进行对抗训练，具体来说我们主要介绍Sinha et al. (2018) 的工作。注意，和前面不同的是，这里的分布式鲁棒优化里的不确定集不再是“确定性”的了，而是成了一个描述“分布”（测度）的集合。因此，我们相当于要考虑这样一个以对抗训练为目标的优化问题

$$\min_{\boldsymbol{\theta}\in\Theta}\max_{P\in\mathcal{P}}\mathbb{E}_P[g(\boldsymbol{\theta};\boldsymbol{Z})]. \tag{8.8}$$

其中，$\boldsymbol{\theta}$ 仍然是训练模型的参数，$\mathcal{P}$ 就是一个描述分布/测度的模糊集，而 $\boldsymbol{Z}$ 为将所有数据 $Z=[\boldsymbol{X};\boldsymbol{y}]$ 简写起来的形式。这边一个很关键的概念就是近些年机器学习和优化领域都十分热门的 Wasserstein 度量，令我们考虑的数据集 $\boldsymbol{Z}\in\mathcal{Z}$ 且 $\mathcal{Z}$ 是实数域的一个子集，如果存在一个“价格”函数 $c:\mathcal{Z}\times\mathcal{Z}\to\mathbb{R}_+\cup\{\infty\}$，那么对任意两个取值在 $\mathcal{Z}$ 上的概率测度 P,Q，我们有 P,Q 之间的 Wasserstein 距离为

$$W_c(P,Q):=\inf_{\mu\in\Gamma(P,Q)}\int_{\mathcal{Z}\times\mathcal{Z}}c(p,q)d\mu(p,q), \tag{8.9}$$

其中 $\Gamma(P,Q)$ 是所有取值在 $\mathcal{Z}\times\mathcal{Z}$ 上的边缘分布为 P 和 Q 的概率测度的集合。于是，我们考虑我们的模糊集用 Wasserstein 度量定义，即 $\mathcal{P}=\{P:W_c(P,P_0)\le\rho\}$。然后我们可以证明，考虑问题 (8.8) 的拉格朗日松弛形式，我们有如下等价关系（松弛因子 $\gamma>0$）：

$$\min_{\boldsymbol{\theta}\in\Theta}\underbrace{\max_{P}\mathbb{E}_P[g(\boldsymbol{\theta};\boldsymbol{Z})-\gamma W_c(P,P_0)]}_{F(\boldsymbol{\theta})}=\min_{\boldsymbol{\theta}\in\Theta}\mathbb{E}_{P_0}[\underbrace{\max_{\boldsymbol{Z}'\in\mathcal{Z}}(g(\boldsymbol{\theta};\boldsymbol{Z}')-\gamma c(\boldsymbol{Z}',\boldsymbol{Z}))}_{\phi_\gamma(\boldsymbol{\theta};\boldsymbol{Z})}]. \tag{8.10}$$

然后，利用 P_0 的样本得到的经验分布 $\hat{P}_n$ 代替 P_0，我们需要求解优化问题：

$$\min_{\boldsymbol{\theta}\in\Theta}\mathbb{E}_{\hat{P}_n}[\phi_\gamma(\boldsymbol{\theta};\boldsymbol{Z})]. \tag{8.11}$$

对此，Sinha et al. (2018) 给出了基于随机梯度下降法（SGD）的算法：

- 输入：分布 P_0 的样本，$\Theta,\mathcal{Z}$，算法步长 $\{\alpha_t>0\}_{t=0}^{T}$
- **for** $t=0,\dots,T-1$ **do**
  - 抽样 $\boldsymbol{Z}^t\sim P_0$ 且找到 $g(\boldsymbol{\theta}^t;\boldsymbol{Z})-\gamma c(\boldsymbol{Z},\boldsymbol{Z}^t)$ 的一个（局部）ϵ-最优解 $\boldsymbol{Z}_t'$
  - $\boldsymbol{\theta}^{t+1}\leftarrow\mathrm{Proj}_\Theta(\boldsymbol{\theta}^t-\alpha_t\nabla_{\boldsymbol{\theta}})g(\boldsymbol{\theta}^t;\boldsymbol{Z}_t')$

Sinha et al. (2018) 证明，在假设 c 是连续，且 $c(\cdot,\boldsymbol{Z})$ 对任意 $\boldsymbol{Z}\in\mathcal{Z}$ 是 1-强凸，并且 g 对于$\boldsymbol{\theta}$ 和 $\boldsymbol{Z}$都是关于系数 $L_{\boldsymbol{\theta}\boldsymbol{\theta}},L_{\boldsymbol{\theta}\boldsymbol{Z}},L_{\boldsymbol{Z}\boldsymbol{Z}},L_{\boldsymbol{Z}\boldsymbol{\theta}}$ 和 L_2 范数李普希茨（Lipschitz）连续的，算法对于问题 (8.10) 的全局最优值（g 是凸的）/局部最优值的收敛速度。

定理 8.3: 非凸 SGD 的分布鲁棒优化收敛性定理 (Sinha et al., 2018)

假设 $\mathbb{E}[\|\nabla F(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}}\phi_\gamma(\boldsymbol{\theta}; \boldsymbol{Z})\|_2^2] \le \sigma^2$, $\Theta = \mathbb{R}^d$, 取 $\Delta_F \ge F(\boldsymbol{\theta}^0) - \min_{\boldsymbol{\theta}} F(\theta)$, $L_\phi := L_{\boldsymbol{\theta}\boldsymbol{\theta}} + \frac{L_{\boldsymbol{\theta}\boldsymbol{Z}} L_{\boldsymbol{Z}\boldsymbol{\theta}}}{\gamma - L_{\boldsymbol{Z}\boldsymbol{Z}}}$, $\alpha_t \equiv \sqrt{\frac{2\Delta_F}{L_\phi \sigma^2 T}}$。我们的算法保证

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla F(\boldsymbol{\theta}^t)\|_2^2\right] - \frac{2L_{\boldsymbol{\theta}\boldsymbol{Z}}^2}{\gamma - L_{\boldsymbol{Z}\boldsymbol{Z}}}\epsilon \le \sigma\sqrt{\frac{8L_\phi\Delta_F}{T}}.$$

♡

这个收敛性定理成立的关键还是在于我们对于 g 有光滑性假设，也就是说这里的分析对于常见的光滑的神经网络损失函数都是成立的。接下来，我们考虑理论上在鲁棒情形下这个算法的泛化（generalization）能力（对应前一节的鲁棒分类错误率）。事实上，Sinha et al. (2018) 证明了如下结论：对任意 $\theta \in \Theta$，

$$\max_{P: W_c(P, P_0) \le \rho} \mathbb{E}_P[g(\boldsymbol{\theta}; \boldsymbol{Z})] \le \gamma\rho + \mathbb{E}_{\hat{P}_n}[\phi_\gamma(\boldsymbol{\theta}; \boldsymbol{Z})] + O(1/\sqrt{n}). \tag{8.12}$$

具体的分析仍然是基于 Monge 映射 $T_\gamma(\boldsymbol{\theta}; \boldsymbol{Z}_0) := \arg\max_{\boldsymbol{Z} \in \mathcal{Z}}\{g(\boldsymbol{\Theta}; \boldsymbol{Z}) - \gamma c(\boldsymbol{Z}, \boldsymbol{Z}_0)\}$ 的光滑性，这里我们不再展开讨论了，有兴趣的读者可以参阅他们的文章。

本章参考文献

Bertsimas, Dimitris and Martin S Copenhaver, "Characterization of the equivalence of robustification and regularization in linear and matrix regression," *European Journal of Operational Research*, 2018, *270* (3), 931–942.

__ **, Jack Dunn, Colin Pawlowski, and Ying Daisy Zhuo**, "Robust classification," *INFORMS Journal on Optimization*, 2019, *1* (1), 2–34.

Schmidt, Ludwig, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry, "Adversarially robust generalization requires more data," in "Advances in Neural Information Processing Systems" 2018, pp. 5019–5031.

Shaham, Uri, Yutaro Yamada, and Sahand Negahban, "Understanding adversarial training: Increasing local stability of supervised models through robust optimization," *Neurocomputing*, 2018, *307*, 195–204.

Sinha, Aman, Hongseok Namkoong, and John Duchi, "Certifying some distributional robustness with principled adversarial training," in "International Conference on Learning Representations" 2018.