6 稳定性

6.1 引言

在前几章中, 基于假设空间 $\mathcal{H}$ 复杂度的不同衡量标准 (如增长函数, VC 维和 Rademacher 复杂度) 分别给出了泛化误差界. 这些泛化误差界均与具体的学习算法无关, 对所有的学习算法都适用. 或许可能会问能否借助关于具体算法的分析给出更好的泛化界? 基于算法的分析的确可以给出更好的理论结果, 但这样的理论结果可能无法推广到作用于同一个假设空间的其他算法. 在这一章将看到, 在保证可将分析结果推广到其他具有相同性质的算法前提下, 学习算法一些一般的性质可以与算法的特性结合起来以得到更好的理论结果.

本章将基于算法的稳定性分析导出依赖于算法的理论结果, 具体组织结构如下: 6.2节介绍算法稳定性的定义, 6.3节给出满足稳定性的算法的泛化误差界, 6.4节说明算法稳定性与假设空间可学习性之间的关联, 6.5.1节指出一大类基于核正则化的算法均满算法的稳定性, 并给出常用分类、回归算法基于稳定性的泛化误差界.

6.2 稳定性

令 $\boldsymbol{z}$ 表示样本 $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$, 给定 $D = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m\}, \boldsymbol{x}_i \in \mathcal{X}$ 独立同分布地来自分布 $\mathcal{D}$, $y_i \in \{-1, +1\}$. 对假设空间 $\mathcal{H} : \mathcal{X} \to \{-1, +1\}$ 和学习算法 $\mathfrak{L}$, 令 $\mathfrak{L}_D \in \mathcal{H}$ 表示基于训练集 D 从假设空间 $\mathcal{H}$ 中学得的假设. 考虑 D 的以下变化:

- $D^{\setminus i}$ 表示移除 D 中第 i 个样例得到的集合

$$D^{\setminus i} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{i-1}, \boldsymbol{z}_{i+1}, \ldots, \boldsymbol{z}_m\},$$

- D^i 表示替换 D 中第 i 个样例得到的集合

$$D^i = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{i-1}, \boldsymbol{z}_i', \boldsymbol{z}_{i+1}, \ldots, \boldsymbol{z}_m\}.$$

  其中 $\boldsymbol{z}_i' = (\boldsymbol{x}_i', y_i')$, $\boldsymbol{x}_i'$ 服从分布 $\mathcal{D}$ 并独立于 D.

损失函数 $\ell(\mathfrak{L}_D(\boldsymbol{x}), y) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ 刻画了假设 $\mathfrak{L}_D$ 的预测标记 $\mathfrak{L}_D(\boldsymbol{x})$ 与真实标记 y 之间的差别, 简记为 $\ell(\mathfrak{L}_D, \boldsymbol{z})$, 其中 $\boldsymbol{z} = (\boldsymbol{x}, y)$. 下面定义关于假设 $\mathfrak{L}_D$ 的几种损失.

- 泛化损失

$$\ell(\mathfrak{L}, D) = \mathbb{E}_{\boldsymbol{x} \in \mathcal{X}, \boldsymbol{z} = (\boldsymbol{x}, y)}[\ell(\mathfrak{L}_D, \boldsymbol{z})] \tag{32}$$

- 经验损失

$$\widehat{\ell}(\mathfrak{L}, D) = \frac{1}{m} \sum_{i=1}^{m} \ell(\mathfrak{L}_D, \boldsymbol{z}_i) \tag{33}$$

- 留一 (leave-one-out) 损失

$$\ell_{loo}(\mathfrak{L}, D) = \frac{1}{m}\sum_{i=1}^{m} \ell(\mathfrak{L}_{D^{\setminus i}}, \boldsymbol{z}_i) \tag{34}$$

接下来定义算法的均匀稳定性 (uniform stability).

定义 6.1. 关于移除样例的稳定性: 对任何 $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{z} = (\boldsymbol{x}, y)$, 若学习算法 $\mathfrak{L}$ 满足

$$\left|\ell(\mathfrak{L}_D, \boldsymbol{z}) - \ell(\mathfrak{L}_{D^{\setminus i}}, \boldsymbol{z})\right| \leq \beta, \quad i = 1, 2, \ldots, m, \tag{35}$$

则称 $\mathfrak{L}$ 关于损失函数 ℓ 满足关于移除样例的 β-均匀稳定性.

同时也可以从替换示例的角度来定义算法关于损失函数的稳定性.

定义 6.2. 关于替换示例的稳定性: 令 D 和 D^i 为仅有一个样本点不同的两个训练集, 如果算法 $\mathfrak{L}$ 基于训练集 D 和 D^i 返回的假设满足

$$|\ell(\mathfrak{L}_D, \boldsymbol{z}) - \ell(\mathfrak{L}_{D^i}, \boldsymbol{z})| \leq \gamma, \quad i = 1, 2, \ldots, m,$$

则称 $\mathfrak{L}$ 关于损失函数 ℓ 满足关于替换样例的 γ-均匀稳定性.

若算法 $\mathfrak{L}$ 关于损失函数 ℓ 满足 β-均匀稳定性, 则有

$$\begin{aligned} |\ell(\mathfrak{L}_D, \boldsymbol{z}) - \ell(\mathfrak{L}_{D^i}, \boldsymbol{z})| &\leq |\ell(\mathfrak{L}_D, \boldsymbol{z}) - \ell(\mathfrak{L}_{D^{\setminus i}}, \boldsymbol{z})| + |\ell(\mathfrak{L}_{D^i}, \boldsymbol{z}) - \ell(\mathfrak{L}_{D^{\setminus i}}, \boldsymbol{z})| \\ &\leq 2\beta \end{aligned}$$

也就是说, 移除示例的稳定性包含替换示例的稳定性, 即有如下关系成立: $\gamma \leq 2\beta$. 本章后续主要使用算法关于替换示例的稳定性, 为了简洁起见, 下文中提到均匀 γ 稳定均是指算法关于替换示例是 γ 稳定的.

当 $\mathfrak{L}$ 在两个数目相同且仅有一个样本不同的样本集上训练时, $\mathfrak{L}$ 返回假设的损失差距应该不大于 γ. 一个 γ-均匀稳定的算法通常也被简称为 γ-稳定的或者稳定的 (对某个未显式给出的 γ). 一般而言, 系数 γ 依赖于样本的大小 m. 在 6.3节中可以看到, 要使得基于稳定性的学习算法具有收敛性, $\gamma = o(1/\sqrt{m})$ 是一个必要条件. 在 6.5.1 节中给出的一类算法, 有更好的条件 $\gamma = O(1/m)$ 成立.

6.3 基于稳定性的泛化误差界

假设 h 为从 $\mathcal{X}$ 到 $\mathcal{Y}'$ 的映射, 这里的 $\mathcal{Y}'$ 有时不同于 $\mathcal{Y}$. 例如, 对于分类问题, 通常令 $\mathcal{Y} = \{-1, +1\}$, 而学得的假设 h 则在 $\mathbb{R}$ 中取值. 因此将损失函数 ℓ 定义在 $\mathcal{Y}' \times \mathcal{Y}$ 上, 虽然在大多数情形下 $\mathcal{Y}' = \mathcal{Y}$. 给定损失函数 $\ell : \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_+$, 定义假设 h 在点 $\boldsymbol{z}$ 的损失为 $\ell(h, \boldsymbol{z}) = \ell(h(\boldsymbol{x}), y)$. 用 $\mathcal{D}$ 表示样本满足的分布, $\mathcal{H}$ 表示假设空间. 某个假设 $h \in \mathcal{H}$ 在样本 $D = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m)$ 上的经验误差和泛化误差分别定义如下:

$$\widehat{\ell}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h, \boldsymbol{z}_i), \quad \ell(h) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\ell(h, \boldsymbol{z})]$$

给定算法 $\mathfrak{L}$, 定义 $\mathfrak{L}_D$ 为算法 $\mathfrak{L}$ 基于样本集 D 返回的假设. 定义损失函数 ℓ 有上界 $M \geq 0$ 是指对于所有的 $h \in \mathcal{H}$ 和 $\boldsymbol{z} \in \mathcal{X} \times \mathcal{Y}$, 均有 $\ell(h, \boldsymbol{z}) \leq M$. 这一节中将给出稳定学习算法返回假设的泛化误差界, 主要结果如下述定理 [Bousquet and Elisseeff, 2002]:

定理 6.1. *假设损失函数 ℓ 有上界 $M \geq 0$. 设 $\mathfrak{L}$ 表示移除示例稳定系数为 β, 替换示例稳定性系数为 γ 的学习算法 (则有 $\gamma \leq 2\beta$), D 表示独立同分布来自 $\mathcal{D}$ 的 m 个样本组成的集合, 则下列不等式以至少 $1-\delta$ 的概率成立:*

$$\ell(\mathfrak{L}, D) \leq \widehat{\ell}(\mathfrak{L}, D) + \gamma + (2m\gamma + M)\sqrt{\frac{\ln(1/\delta)}{2m}} \tag{36}$$

$$\ell(\mathfrak{L}, D) \leq \ell_{loo}(\mathfrak{L}, D) + \beta + (4m\beta + M)\sqrt{\frac{\ln(1/\delta)}{2m}}. \tag{37}$$

证明. 仅对式 (36) 证明. 证明过程中会将 McDiarmid 不等式应用于函数 Φ, 其中 Φ 定义为: $\Phi(D) = \ell(\mathfrak{L}, D) - \widehat{\ell}(\mathfrak{L}, D)$. 令 D' 表示与 D 仅有一个样本不同的集合, 即:

$$D = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{m-1}, \boldsymbol{z}_m), \quad D' = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{m-1}, \boldsymbol{z}'_m).$$

由 Φ 的定义可知下面的不等式成立:

$$|\Phi(D') - \Phi(D)| \leq |\ell(\mathfrak{L}, D') - \ell(\mathfrak{L}, D)| + |\widehat{\ell}(\mathfrak{L}, D') - \widehat{\ell}(\mathfrak{L}, D)|. \tag{38}$$

对上式右边两项分别给出上界. 由算法 $\mathfrak{L}$ 的 γ-稳定性可知

$$|\ell(\mathfrak{L}, D) - \ell(\mathfrak{L}, D')| = |\mathbb{E}_{\boldsymbol{z}}[\ell(\mathfrak{L}_D, \boldsymbol{z})] - \mathbb{E}_{\boldsymbol{z}}[\ell(\mathfrak{L}_{D'}, \boldsymbol{z})]| \leq \mathbb{E}_{\boldsymbol{z}}[|\ell(\mathfrak{L}_D, \boldsymbol{z}) - \ell(\mathfrak{L}_{D'}, \boldsymbol{z})|] \leq \gamma.$$

利用 ℓ 的有界性以及算法 $\mathfrak{L}$ 的 γ-稳定性, 有

$$\begin{aligned}
|\widehat{\ell}(\mathfrak{L}, D) - \widehat{\ell}(\mathfrak{L}, D')| &= \frac{1}{m}\left|\left(\sum_{i=1}^{m-1} \ell(\mathfrak{L}_D, \boldsymbol{z}_i) - \ell(\mathfrak{L}_{D'}, \boldsymbol{z}_i)\right) + \ell(\mathfrak{L}_D, \boldsymbol{z}_m) - \ell(\mathfrak{L}_{D'}, \boldsymbol{z}'_m))\right| \\
&\leq \frac{1}{m}\left[\left(\sum_{i=1}^{m-1} |\ell(\mathfrak{L}_D, \boldsymbol{z}_i) - \ell(\mathfrak{L}_{D'}, \boldsymbol{z}_i)|\right) + |\ell(\mathfrak{L}_D, \boldsymbol{z}_m) - \ell(\mathfrak{L}_{D'}, \boldsymbol{z}'_m)|\right] \\
&\leq \frac{m-1}{m}\gamma + \frac{M}{m} \leq \gamma + \frac{M}{m}.
\end{aligned}$$

因此, 由式 (38) 可知 Φ 满足条件 $|\Phi(D) - \Phi(D')| \leq 2\gamma + \frac{M}{m}$. 将 McDiarmid 不等式应用于 $\Phi(D)$ 可得如下不等式界定 Φ 与其平均值之间的偏差:

$$P\left[\Phi(D) \geq \epsilon + \mathbb{E}_D[\Phi(D)]\right] \leq \exp\left(\frac{-2m\epsilon^2}{(2m\gamma + M)^2}\right).$$

令 $\delta = \exp(\frac{-2m\epsilon^2}{(2m\gamma+M)^2})$, 可知下列不等式以至少 $1-\delta$ 的概率成立:

$$\Phi(D) < \epsilon + \mathbb{E}_D[\Phi(D)]. \tag{39}$$

求解出 ϵ 代入式 (39) 可知:

$$\Phi(D) \leq \mathbb{E}_{D \sim \mathcal{D}^m}[\Phi(D)] + (2m\gamma + M)\sqrt{\frac{\ln(1/\delta)}{2m}}. \tag{40}$$

以至少 $1-\delta$ 的概率成立. 接下来分析期望项, 由期望的线性性可知 $\mathbb{E}_D[\Phi(D)] = \mathbb{E}_D[\ell(\mathfrak{L},D)] - \mathbb{E}_D[\widehat{\ell}(\mathfrak{L},D)]$. 由泛化误差的定义可知

$$\mathbb{E}_{D\sim\mathcal{D}^m}[\ell(\mathfrak{L},D)] = \mathbb{E}_{D\sim\mathcal{D}^m}[\mathbb{E}_{\boldsymbol{z}\sim\mathcal{D}}[\ell(\mathfrak{L}_D,\boldsymbol{z})]] = \mathbb{E}_{D,\boldsymbol{z}\sim\mathcal{D}^{m+1}}[\ell(\mathfrak{L}_D,\boldsymbol{z})]. \tag{41}$$

由期望的线性性可知

$$\mathbb{E}_{D\sim\mathcal{D}^m}[\widehat{\ell}(\mathfrak{L},D)] = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{D\sim\mathcal{D}^m}[\ell(\mathfrak{L}_D,\boldsymbol{z}_i)] = \mathbb{E}_{D\sim\mathcal{D}^m}[\ell(\mathfrak{L}_D,\boldsymbol{z}_1)]. \tag{42}$$

这里第二个等号成立是因为 $\boldsymbol{z}_i$ 是独立同分布的, 因此期望值 $\mathbb{E}_{D\sim\mathcal{D}^m}[\ell(\mathfrak{L}_D,\boldsymbol{z}_i)]$ 全都相等, $i\in[1,m]$. 式 (42) 中的最后一项是假设 $\mathfrak{L}_D$ 在样本上的期望误差, 可以重写为

$$\mathbb{E}_{D\sim\mathcal{D}^m}[\ell(\mathfrak{L}_D,\boldsymbol{z}_i)] = \mathbb{E}_{D,\boldsymbol{z}\sim\mathcal{D}^{m+1}}[\ell(\mathfrak{L}_{D'},\boldsymbol{z})],$$

这里 D' 是由样本 $\boldsymbol{z}$ 和 D 中抽取 $m-1$ 样本组成的集合. 故由式 (41) 和算法 $\mathfrak{L}$ 的 γ-稳定性可知

$$\begin{aligned}|\mathbb{E}_{D\sim\mathcal{D}^m}[\Phi(D)]| &= |\mathbb{E}_{D,\boldsymbol{z}\sim\mathcal{D}^{m+1}}[\ell(\mathfrak{L}_D,\boldsymbol{z})] - \mathbb{E}_{D,\boldsymbol{z}\sim\mathcal{D}^{m+1}}[\ell(\mathfrak{L}_{D'},\boldsymbol{z})]| \\ &\leq \mathbb{E}_{D,\boldsymbol{z}\sim\mathcal{D}^{m+1}}[|\ell(\mathfrak{L}_D,\boldsymbol{z}) - \ell(\mathfrak{L}_{D'},\boldsymbol{z})|] \\ &\leq \mathbb{E}_{D,\boldsymbol{z}\sim\mathcal{D}^{m+1}}[\gamma] = \gamma.\end{aligned}$$

最后在式 (40) 中用 γ 取代 $\mathbb{E}_D[\Phi(D)]$, 定理证毕. ■

定理 6.1 给出了基于稳定性推导出的算法 $\mathfrak{L}$ 学得假设的泛化误差界. 从式 (36) 可以看出, 经验误差与泛化误差之间差别的收敛率为 $\gamma\sqrt{m}$; 若 $\gamma = O(\frac{1}{m})$, 则可保证收敛率为 $O(\frac{1}{\sqrt{m}})$. 与前面章节中的结果比较可知, 这与基于 VC 维或 Rademacher 复杂度得到泛化误差的收敛率一致.

当 $(m\gamma)/\sqrt{m} = o(1)$, 即 $\gamma = o(1/\sqrt{m})$, 上述定理给出的泛化误差界收敛. 特别地, 当稳定性系数 γ 为 $O(1/m)$ 时, 有结论 $\ell(\mathfrak{L},D) - \widehat{\ell}(\mathfrak{L},D) = O(1/\sqrt{m})$ 成立. 在 6.5.1 节将证明基于核正则化的算法在一些通常的假设条件下具有这一性质.

6.4 稳定性与可学习性间的关系

本小节将讨论稳定性与可学习性之间的关系. 学习算法的稳定性分析所关注的是 $|\widehat{\ell}(\mathfrak{L},D) - \ell(\mathfrak{L},\mathcal{D})|$, 而假设空间复杂度分析所关注的是 $\sup_{h\in\mathcal{H}}|\widehat{E}(h) - E(h)|$; 也就是说, 稳定性分析不必考虑假设空间中所有可能的假设, 只需根据算法自身的特性 (稳定性) 来讨论输出假设 $\mathfrak{L}_D$ 的泛化误差界. 那么, 稳定性和可学习性之间有什么关系呢?

首先, 必须假设 $\gamma\sqrt{m}\to 0$, 这样才能保证稳定的学习算法 $\mathfrak{L}$ 具有一定的泛化能力, 即经验损失收敛于泛化损失, 否则可学习性无从谈起. 为便于计算, 假定 $\gamma = \frac{1}{m}$, 代入到式 (36) 中可得:

$$\ell(\mathfrak{L},\mathcal{D}) \leq \widehat{\ell}(\mathfrak{L},D) + \frac{1}{m} + (2+M)\sqrt{\frac{\ln(1/\delta)}{2m}}. \tag{43}$$

对损失函数 ℓ, 若学习算法 $\mathfrak{L}$ 所输出的假设满足经验损失最小化, 则称算法 $\mathfrak{L}$ 满足经验风险最小化 (Empirical Risk Minimization) 原则, 简称算法是 ERM 的. 关于学习算法的稳定性和可学习性, 有如下定理:

定理 6.2. *若学习算法 $\mathfrak{L}$ 是 ERM 且稳定的, 则假设空间 $\mathcal{H}$ 是 PAC 可学习的.*

证明. 令 g 表示 $\mathcal{H}$ 中具有最小泛化损失的假设, 即

$$\ell(g,\mathcal{D}) = \min_{h\in\mathcal{H}} \ell(h,\mathcal{D}).$$

再令

$$\epsilon' = \frac{\epsilon}{2},$$

$$\frac{\delta}{2} = 2\exp\left(-2m(\epsilon')^2\right),$$

由 Hoeffding 不等式可知, 当 $m \geq \frac{2}{\epsilon^2}\ln\frac{4}{\delta}$ 时,

$$|\ell(g,\mathcal{D}) - \widehat{\ell}(g,D)| \leq \frac{\epsilon}{2}$$

以至少 $1-\delta/2$ 的概率成立. 令式 (43) 中

$$\frac{1}{m} + (2+M)\sqrt{\frac{\ln(2/\delta)}{2m}} = \frac{\epsilon}{2},$$

解得 $m = O\left(\frac{1}{\epsilon^2}\ln\frac{1}{\delta}\right)$ 使

$$\begin{aligned}\ell(\mathfrak{L},\mathcal{D}) - \ell(g,\mathcal{D}) &\leq \widehat{\ell}(\mathfrak{L},D) + \frac{\epsilon}{2} - \left(\widehat{\ell}(g,D) - \frac{\epsilon}{2}\right)\\ &\leq \widehat{\ell}(\mathfrak{L},D) - \widehat{\ell}(g,D) + \epsilon\\ &\leq 0 + \epsilon\\ &= \epsilon\end{aligned} \tag{44}$$

以至少 $1-\delta$ 的概率成立, 从而定理 6.2 得证. ■

在这一定理中, 为何学习算法的稳定性能导出假设空间的可学习性? 学习算法和假设空间看起来是两码事. 事实上, 要注意稳定性与假设空间并非无关, 由稳定性的定义可知两者通过损失函数 ℓ 联系起来.

6.5 实例

本节将介绍一大类基于核正则化的算法（分类算法 SVMs 和回归算法 SVR、KRR）的稳定性，并基于前两节的内容分析这些算法的泛化误差界.

6.5.1 核正则化算法

令 K 为正定对称核, $\mathbb{H}$ 对应于 K 的再生核 Hilbert 空间, 并且 $||\cdot||_K$ 是 $\mathbb{H}$ 中由 K 诱导的范数. 一个基于核正则化的算法定义如下: 在空间 $\mathbb{H}$ 中极小化目标函数 F_D, 并输出具有极小值的假设, 其中 $D = (\boldsymbol{z}_1,\ldots,\boldsymbol{z}_m)$ 为训练集, 而 F_D 定义如下: 对所有 $h\in\mathbb{H}$,

$$F_D(h) = \widehat{\ell}_D(h) + \lambda||h||_K^2. \tag{45}$$

在此式中, $\widehat{\ell}_D(h) = \frac{1}{m}\sum_{i=1}^{m}\ell(h, \boldsymbol{z}_i)$ 是损失函数 ℓ 的经验误差, $\lambda \geq 0$ 用来平衡经验误差和正则化项 $||h||_K^2$ 的重要性. 假设空间 $\mathcal{H}$ 是 $\mathbb{H}$ 的一个子集. 在这一通用模型, 将会考虑 KRR, SVR 和 SVMs 等算法.

定义 6.3. *σ-可容许性: 如果存在 $\sigma \in \mathbb{R}_+$, 使得对于任何两个假设 $h, h' \in \mathcal{H}$ 和所有 $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$, 有*

$$|\ell(h'(\boldsymbol{x}), y) - \ell(h(\boldsymbol{x}), y)| \leq \sigma|h'(\boldsymbol{x}) - h(\boldsymbol{x})|, \tag{46}$$

则称损失函数 ℓ 对于假设空间 $\mathcal{H}$ 是 σ-可容许的.

这里的假设对于二次损失函数和大部分其他损失函数都是成立的, 只要假设空间和输出标记空间有上界 $M \in \mathbb{R}_+ : \forall h \in \mathcal{H}, \forall \boldsymbol{x} \in \mathcal{X}, |h(\boldsymbol{x})| \leq M$ 以及 $\forall y \in \mathcal{Y}, |y| \leq M$ 即可.

下面将使用到 *Bregman Divergence*(B_F) 的概念, 其中 $F : \mathbb{H} \to \mathbb{R}$ 为任何凸可微函数, B_F 的定义如下: 对所有 $f, g \in \mathbb{H}$,

$$B_F(f||g) = F(f) - F(g) - \langle f - g, \nabla F(g)\rangle.$$

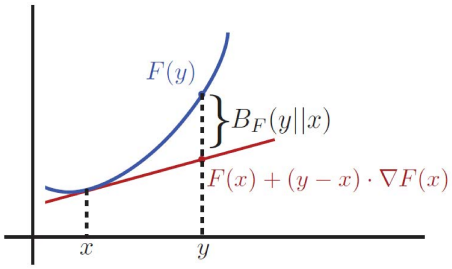

图 6: Bregman divergence 的示意图.

图 6给出了 Bregman Divergence 的几何解释. 通过使用次梯度 (subgradient), 可将这一定义推广到凸的但非可微的损失函数 F. 对一个凸函数 $F : \mathbb{H} \to \mathbb{R}$, F 在 h 的次梯度定义为 $\partial F(h)$, 即

$$\partial F(h) = \{g \in \mathbb{H} : \forall h' \in \mathbb{H}, F(h') - F(h) \geq \langle h' - h, g\rangle\}$$

$\partial F(h)$ 由在 h 处支撑函数 F 的某个超平面的向量 g 组成 (见图 7).

当 F 在点 h 处可微时, $\partial F(h)$ 恰为 $\nabla F(h)$, i.e., $\partial F(h) = \{\nabla F(h)\}$. 注意到, 当 F 在点 h 处极小时, 0 为 $\partial F(h)$ 的一个元素. 次梯度具有可加性, 即对任何两个凸函数 F_1 和 F_2, $\partial(F_1 + F_2)(h) = \{g_1 + g_2 : g_1 \in \partial F_1(h), g_2 \in \partial F_2(h)\}$. 对任何 $h \in \mathbb{H}$, 固定 $\delta F(h)$ 为 $\partial F(h)$ 的一个任意元素, 对任何这样的选择 δF 定义广义的 *Bergman Divergence*为:

$$\forall h', h \in \mathbb{H}, B_F(h'||h) = F(h') - F(h) - \langle h' - h, \delta F(h)\rangle. \tag{47}$$

注意到, 由次导数的定义有 $B_F(h'||h) \geq 0$ 对所有的 h',$h \in \mathbb{H}$. 由 (45) 式可以定义 F_D 的广义 *Bregman divergence*. 令 N 表示凸函数 $h \to ||h||_K^2$. 由于 N 是可微的, 对所有 $h \in \mathcal{H}$ 有 $\delta N(h) = \nabla N(h)$, 因此 δN 和 B_N 均唯一确定. 为了让 F_D 和 $\widehat{\ell}_D$ 的 Bregman divergence 的定义相容, 即满足 $B_{F_D} = B_{\widehat{\ell}_D} + \lambda B_N$,

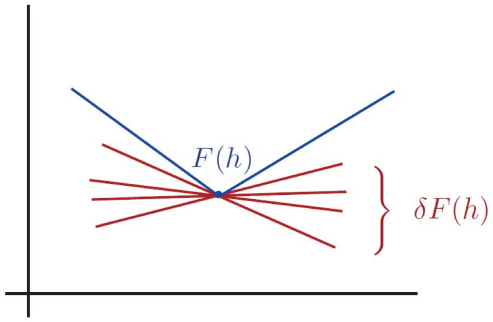

图 7: sub-gradient 概念的示意图. 蓝线表示 F 函数, 红线表示集合 $\partial F(h)$ 中的元素.

可以定义 $\delta\widehat{\ell}_D$ 为: $\delta\widehat{\ell}_D(h) = \delta F_D(h) - \lambda\nabla N(h), \forall h \in \mathbb{H}$. 注意到, 当 F_D 在任何点 h 为最小时, 可以让 $\delta F_D(h)$ 为 0, 同时对所有其他的 $h \in H$, 可令 $\delta F_D(h)$ 为 $\partial F_D(h)$ 中任一个元素. 采用类似的方式, 可以对 $F_{D'}$ 和 $\widehat{\ell}_{D'}$ 定义 Bregman divergence 使得 $B_{F_{D'}} = B_{\widehat{\ell}_{D'}} + \lambda B_N$.

6.5.2 基于核正则化的算法的稳定性

在下面关于基于核正则化的算法稳定性系数上界的证明中, 将用到上述定义的广义 *Bregman Divergence*.

命题 6.1. *令 K 表示一个正定对称核且满足: 存在某个 $r \in \mathbb{R}_+$, 对所有 $\boldsymbol{x} \in \mathcal{X}$, $K(\boldsymbol{x}, \boldsymbol{x}) \leq r^2$, ℓ 为凸的且具有 σ 相容性的损失函数, 则极小化式 (45) 基于核正则化的算法是 γ-稳定的, 且 γ 有上界:*

$$\gamma \leq \frac{\sigma^2 r^2}{m\lambda}.$$

证明. 令 h 为让 F_D 取最小值的点 (minimizer), h' 为让 $F_{D'}$ 取最小值的点, 这里样本集 D 和 D' 仅有一个点不同, 相应地 $\boldsymbol{z}_m$ 在 D 中, 而 $\boldsymbol{z}'_m$ 在 D' 中. 由于广义的 Bregman divergence 是非负的, 且 $B_{F_D} = B_{\widehat{\ell}_D} + \lambda B_N$, $B_{F_{D'}} = B_{\widehat{\ell}_{D'}} + \lambda B_N$, 这样有:

$$B_{F_D}(h'||h) + B_{F_{D'}}(h||h') \geq \lambda(B_N(h'||h) + B_N(h||h').$$

注意到, $B_N(h'||h) + B_N(h||h') = -\langle h' - h, 2h\rangle - \langle h - h', 2h'\rangle = 2||h' - h||_K^2$. 用 Δh 表示 $h' - h$, 则有:

$$\begin{aligned}
2\lambda||\Delta h||_K^2 &\leq B_{F_D}(h'||h) + B_{F_{D'}}(h||h') \\
&= F_D(h') - F_D(h) - \langle h' - h, \delta F_D(h)\rangle + F_{D'}(h) - F_{D'}(h') - \langle h - h', \delta F_{D'}(h')\rangle \\
&= F_D(h') - F_D(h) + F_{D'}(h) - F_{D'}(h') \\
&= \widehat{\ell}_D(h') - \widehat{\ell}_D(h) + \widehat{\ell}_{D'}(h) - \widehat{\ell}_{D'}(h').
\end{aligned}$$

第二个等号成立是因为 h' 和 h 的定义和次梯度在极小值点的选择, 有 $\delta F_{D'}(h') = 0$ 和 $\delta F_D(h) = 0$. 最后一个等号成立来自于 F_D 和 $F_{D'}$ 的定义. 接下来, 将所得不等式写成关于损失函数 ℓ 的表达式, 注意到 D 和 D' 仅有一个样本点不同, 以及 ℓ 的σ 可容许性, 可以得到

$$\begin{aligned}
2\lambda||\Delta h||_K^2 &\leq \frac{1}{m}[\ell(h', \boldsymbol{z}_m) - \ell(h, \boldsymbol{z}_m) + \ell(h, \boldsymbol{z}'_m) - \ell(h', \boldsymbol{z}'_m)] \\
&\leq \frac{\sigma}{m}[|\Delta h(\boldsymbol{x}_m)| + |\Delta h(\boldsymbol{x}'_m)|].
\end{aligned} \tag{48}$$

由再生核性质和 Cauchy-Schwarz 不等式, 对所有 $\boldsymbol{x} \in \mathcal{X}$,

$$\Delta h(\boldsymbol{x}) = \langle \Delta h, K(\boldsymbol{x}, \cdot) \rangle \leq ||\Delta h||_K ||K(\boldsymbol{x}, \cdot)||_K = \sqrt{K(\boldsymbol{x}, \boldsymbol{x})} ||\Delta h||_K \leq r ||\Delta h||_K.$$

由上式和式 (48) 推出 $||\Delta h||_K \leq \frac{\sigma r}{\lambda m}$. 由 ℓ 函数的 σ 可容许性和再生性质可知下列等式成立:

$$\forall \boldsymbol{z} \in \mathcal{X} \times \mathcal{Y}, |\ell(h', \boldsymbol{z}) - \ell(h, \boldsymbol{z})| \leq \sigma |\Delta h(\boldsymbol{x})| \leq r\sigma ||\Delta h||_K.$$

因此

$$\forall \boldsymbol{z} \in \mathcal{X} \times \mathcal{Y}, |\ell(h', \boldsymbol{z}) - \ell(h, \boldsymbol{z})| \leq \frac{\sigma^2 r^2}{m\lambda},$$

证明结束. ■

在这一命题的条件下, 对于固定的 λ, 基于核正则化的算法稳定性系数 γ 是 $O(1/m)$ 的.

分类算法 SVMs

这一节将给出 SVMs 算法基于稳定性的泛化误差界, 其中使用标准的 hinge-loss 函数, 即对所有 $y \in \mathcal{Y} = \{-1, 1\}$ 以及 $y' \in \mathbb{R}$,

$$\ell_{\text{hinge}}(y', y) = \begin{cases} 0 & \text{if } 1 - yy' \leq 0 \\ 1 - yy' & \text{otherwise} \end{cases} \tag{49}$$

推论 6.1. SVMs 算法基于稳定性的泛化误差界: *假设对所有 $\boldsymbol{x} \in \mathcal{X}$ 和某个 $r \geq 0$ 均有 $K(\boldsymbol{x}, \boldsymbol{x}) \leq r^2$ 成立. 令 h_D 为 SVMs 在大小为 m 的独立同分布样本 D 上返回的假设, 则对任何 $\delta > 0$, 下面的不等式以至少 $1 - \delta$ 的概率成立:*

$$\ell(\mathfrak{L}, D) \leq \widehat{\ell}(\mathfrak{L}, D) + \frac{r^2}{m\lambda} + \left(\frac{2r^2}{\lambda} + \frac{r}{\sqrt{\lambda}} + 1 \right) \sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

证明. 显然, 对任何 $y \in \mathcal{Y}$, $\ell_{\text{hinge}}(\cdot, y)$ 是 1-lipschitz 的, 因此它是 σ 可容许的, 其中 $\sigma = 1$. 由命题 (6.1) 可知 SVMs 具有 γ-稳定性, 其中 $\gamma \leq \frac{r^2}{m\lambda}$. 由于对任何 $y \in \mathcal{Y}$, $|\ell_{\text{hinge}}(0, y)| \leq 1$. 由引理 (6.1), $\forall \boldsymbol{x} \in \mathcal{X}$, $|h_D(\boldsymbol{x})| \leq r/\sqrt{\lambda}$. 因此, 对所有样本 D 以及任何 $\boldsymbol{x} \in \mathcal{X}$ 和 $y \in \mathcal{Y}$, 损失函数有上界: $\ell_{\text{hinge}}(h_D(\boldsymbol{x}), y) \leq r/\sqrt{\lambda} + 1$. 将此值作为 M 及上面 γ 的上界代入定理 (6.1) 中, 定理得证. ■

由于 hinge-loss 函数可看做是 binary-loss 函数的上界, 若用分类中常用的 binary-loss 函数来衡量 h_D 的误差损失, 对由此产生的泛化误差, 推论 (6.1) 中的界也是成立的.

回归算法 SVR 和 KRR

这一节具体地分析两个广泛使用的回归算法, SVR(Support Vector Regression) 和 KRR(Kernel Ridge Regression), 这两种算法都是基于核正则化的算法代表.

SVR 是基于 ϵ 敏感的损失函数 ℓ_ϵ, 定义如下: 对所有 $(y, y') \in \mathcal{Y} \times \mathcal{Y}$,

$$\ell_\epsilon(y', y) = \begin{cases} 0 & \text{if } |y' - y| \leq \epsilon; \\ |y' - y| - \epsilon & \text{otherwise} \end{cases} \tag{50}$$

假定 ℓ_ϵ 对于 SVR 算法返回的假设是有界的, 可以得到 SVR 算法的一个基于稳定性的泛化误差界.

推论 6.2. SVR 算法基于稳定性的泛化误差界: 假设 $K(\boldsymbol{x},\boldsymbol{x}) \le r^2$ 对所有 $\boldsymbol{x} \in \mathcal{X}$ 和某个 $r \ge 0$ 成立, 且 ℓ_ϵ 有上界 $M \ge 0$. 令 h_D 为算法 *SVR* 在大小为 m 的独立同分布样本 D 上返回的假设, 则对任何 $\delta > 0$, 下列不等式以至少 $1-\delta$ 的概率成立

$$\ell(\mathfrak{L}, D) \le \widehat{\ell}(\mathfrak{L}, D) + \frac{r^2}{m\lambda} + (\frac{2r^2}{\lambda} + M)\sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

证明. 首先证明对任何 $y \in \mathcal{Y}$, $\ell_\epsilon(\cdot) = \ell_\epsilon(\cdot, y)$ 是 1-Lipschitz 函数. 对任意 y',$y'' \in \mathcal{Y}$, 考虑如下四种情况: 首先, 如果 $|y'-y| \le \epsilon$ 且 $|y''-y| \le \epsilon$, 则 $|\ell_\epsilon(y'') - \ell_\epsilon(y')| = 0$; 第二, 如果 $|y'-y| > \epsilon$ 且 $|y''-y| > \epsilon$, 则由三角不等式有 $|\ell_\epsilon(y'') - \ell_\epsilon(y')| = ||y''-y| - |y'-y|| \le |y''-y'|$; 第三, 如果 $|y'-y| \le \epsilon$ 且 $|y''-y| > \epsilon$, 则 $|\ell_\epsilon(y'') - \ell_\epsilon(y')| = ||y''-y| - \epsilon| = |y''-y| - \epsilon \le |y''-y| - |y'-y| \le |y''-y'|$; 第四, 如果 $|y''-y| \le \epsilon$ 且 $|y'-y| > \epsilon$, 由对称性可得与上相同的不等式.

因此, 对于所有情形均有 $\ell_\epsilon(y'', y) - \ell_\epsilon(y', y)| \le |y''-y|$. 特别地, 这表明 ℓ_ϵ 具有 σ 可容许性, 其中 $\sigma = 1$. 由命题 (6.1)SVR 是 γ 稳定的, $\gamma \le \frac{r^2}{m\lambda}$. 将其代入定理 (6.1) 中的界, 定理得证. ∎

接下来是关于 KRR 算法基于稳定性的界, 基于平方损失 ℓ_2 函数定义如下: 对所有 $y', y \in \mathcal{Y}$,

$$\ell_2(y', y) = (y'-y)^2. \tag{51}$$

正如上面关于 SVR 结论的证明, 同样需要假设关于 KRR 算法返回的假设 ℓ_2 是有界的.

推论 6.3. KRR 算法基于稳定性的泛化误差界: 假设对所有 $\boldsymbol{x} \in \mathcal{X}$ 和某个 $r \ge 0$ 均有 $K(\boldsymbol{x},\boldsymbol{x}) \le r^2$ 成立, 且 ℓ_2 有上界 $M \ge 0$. 令 h_D 为 *KRR* 在大小为 m 的独立同分布样本 D 上返回的假设, 则对任何 $\delta > 0$, 下面的不等式以至少 $1-\delta$ 的概率成立:

$$\ell(\mathfrak{L}, D) \le \widehat{\ell}(\mathfrak{L}, D) + \frac{4Mr^2}{\lambda m} + \left(\frac{8Mr^2}{\lambda} + M\right)\sqrt{\frac{\log \frac{1}{\delta}}{2m}}.$$

证明. 对任何 $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$ 和 $h, h' \in h$,

$$\begin{aligned}
|\ell_2(h'(\boldsymbol{x}), y) - \ell_2(h(\boldsymbol{x}), y)| &= |(h'(\boldsymbol{x}) - y)^2 - (h(\boldsymbol{x}) - y)^2| \\
&= |[h'(\boldsymbol{x}) - h(\boldsymbol{x})][(h'(\boldsymbol{x}) - y) + (h(\boldsymbol{x}) - y)]| \\
&\le (|h'(\boldsymbol{x}) - y| + |h(\boldsymbol{x}) - y|)|h(\boldsymbol{x}) - h'(\boldsymbol{x})| \\
&\le 2\sqrt{M}|h(\boldsymbol{x}) - h'(\boldsymbol{x})|.
\end{aligned}$$

这里利用了损失函数的 M 上界, 因此 ℓ_2 是 σ 可容许的, 其中 $\sigma = 2\sqrt{M}$. 再由命题 (6.1), KRR 是 γ 稳定的, 其中 $\gamma \le \frac{4r^2M}{m\lambda}$. 将其代入定理 (6.1) 中的界, 定理得证. ∎

之前的两个推论都假定损失函数是有界的. 接下来给出的引理说明只要标记空间 $\mathcal{Y}$ 具有某种有界性, 则 SVR 和 KRR 算法使用的损失函数是有界的.

引理 6.1. 假设对所有 $\boldsymbol{x} \in \mathcal{X}$ 和某个 $r \ge 0$ 均有 $K(\boldsymbol{x},\boldsymbol{x}) \le r^2$ 成立, 且存在某个 $B \ge 0$ 对所有 $y \in \mathcal{Y}$, $\ell(0, y) \le B$ 成立, 则核正则化算法基于样本集 D 返回的假设 h_D 有上界:

$$\forall \boldsymbol{x} \in \mathcal{X}, |h_D(\boldsymbol{x})| \le r\sqrt{B/\lambda}.$$

证明. 由再生核性质和 Cauchy-Schwarz 不等式有

$$\forall \boldsymbol{x} \in \mathcal{X}, |h_D(\boldsymbol{x})| \leq \langle h_D, K(\boldsymbol{x}, \cdot)\rangle \leq ||h_D||_K \sqrt{K(\boldsymbol{x}, \boldsymbol{x})} \leq r||h_D||_K. \tag{52}$$

在 $\mathbb{H}$ 上极小化 (45) 式, 由于 $\mathbb{H}$ 包含原点 0, 由 F_D 和 h_D 的定义可知如下不等式成立.

$$F_D(h_D) \leq F_D(0) = \frac{1}{m}\sum_{i=1}^{m} \ell(0, y_i) \leq B.$$

由于损失函数 ℓ 是非负的, 于是 $\lambda||h_D||_K^2 \leq F_D(h_D)$, 进一步可得 $\lambda||h_D||_K^2 \leq B$. 将此不等式与式 (52) 结合, 定理证毕. ■

注意到, 基于核正则化的算法泛化误差界都具有形式: $\ell(\mathfrak{L}, D) - \widehat{\ell}(\mathfrak{L}, D) \leq O(\frac{1}{\lambda\sqrt{m}})$. 因此, 仅当 $\lambda \gg 1/\sqrt{m}$ 时, 上面给出的泛化误差界才是有意义的. 正则化参数 λ 是样本大小 m 的函数: 对于较大的 m, 希望正则化参数 λ 较小, 以减小正则化项的作用. λ 的大小影响着线性假设空间的范数大小, 较大的 λ 会导致较小的假设空间范数. 在这种意义下, λ 也可以看做假设空间复杂程度的一种度量, 而 λ 需要满足的条件也可以解释为: 复杂度更小的假设空间具有更好的泛化性保障. 在这一章关于稳定性的分析中, 假设 λ 为固定值: 训练集中样本点发生变化时, 正则化参数被认为是不变的. 虽然这是一种很弱的假设, 但对于一般情形, 该假设可能并不成立.

另外，正如本章所言，基于均匀稳定性能够导出泛化误差界, 但均匀稳定性并不是算法奏效的必要条件. 一些算法可能并不具有均匀稳定性, 但在实际的监督学习场景中仍具有很好的泛化性, 例如 Lasso 算法. 关于算法稳定性方面的文章可以进一步阅读 [Shalev-Shwartz et al., 2009b].

参考文献

Abbasi-yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pages 2312–2320.

Abernethy, J., Bartlett, P. L., Rakhlin, A., and Tewari, A. (2008a). Optimal stragies and minimax lower bounds for online convex games. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 415–423.

Abernethy, J., Hazan, E., and Rakhlin, A. (2008b). Competing in the dark: An efficient algorithm for bandit linear optimization. In *Proceedings of the 21st Annual Conference on Learning*, pages 263–274.

Agarwal, A., Dekel, O., and Xiao, L. (2010). Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the 23rd Annual Conference on Learning Theory*, pages 28–40.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537.

Bartlett, P. L., Dani, V., Hayes, T. P., Kakade, S. M., Rakhlin, A., and Tewari, A. (2008). High-probability regret bounds for bandit online linear optimization. In *Proceedings of the 21st Annual Conference on Learning*, pages 335–341.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2:499–526.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122.

Cesa-bianchi, N., Conconi, A., and Gentile, C. (2002). On the generalization ability of on-line learning algorithms. In *Advances in Neural Information Processing Systems 14*, pages 359–366.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008a). The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems 20*, pages 345–352.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008b). Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning*, pages 355–366.

Daniely, A., Gonen, A., and Shalev-Shwartz, S. (2015). Strongly adaptive online learning. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1405–1411.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems 23*, pages 586–594.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. (2005). Online convex optimization in the bandit setting: Gradient descent without a gradient. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 385–394.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix computations, 3rd Edition*. Johns Hopkins University Press.

Hazan, E., Agarwal, A., and Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192.

Hazan, E. and Kale, S. (2011). Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 421–436.

Hazan, E. and Kale, S. (2014). Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15:2489–2512.

Hazan, E. and Seshadhri, C. (2007). Adaptive algorithms for online decision problems. *Electronic Colloquium on Computational Complexity*, 88.

Hazan, E. and Seshadhri, C. (2009). Efficient learning algorithms for changing environments. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 393–400.

Hou, B.-J., Zhang, L., and Zhou, Z.-H. (2017). Learning with feature evolvable streams. In *Advances in Neural Information Processing Systems 30*.

Jun, K.-S., Bhargava, A., Nowak, R., and Willett, R. (2017). Scalable generalized linear bandits: Online computation and hashing. In *Advances in Neural Information Processing Systems 30*, pages 99–109.

Kakade, S. M., Shalev-Shwartz, S., and Tewari, A. (2008). Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th International Conference on Machine Learning*, pages 440–447.

Kakade, S. M. and Tewari, A. (2009). On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems 21*, pages 801–808.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer.

Kushner, H. J. and Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, second edition.

Mahdavi, M., Zhang, L., and Jin, R. (2015). Lower and upper bounds on the generalization of stochastic exponentially concave optimization. In *Proceedings of the 28th Annual Conference on Learning Theory*.

Mohri, M., Rostamizadeh, A., and Talwalkar, A., editors (2012). *Foundations of Machine Learning*. MIT Press, Cambridge, MA.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609.

Nesterov, Y. (2004). *Introductory lectures on convex optimization: a basic course*, volume 87 of *Applied optimization*. Kluwer Academic Publishers.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152.

Nesterov, Y. (2007). Gradient methods for minimizing composite objective function. Core discussion papers.

Nesterov, Y. (2011). Random gradient-free minimization of convex functions. Core discussion papers.

Rakhlin, A., Shamir, O., and Sridharan, K. (2012). Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 449–456.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

Saha, A. and Tewari, A. (2011). Improved regret guarantees for online smooth convex optimization with bandit feedback. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 636–642.

Sauer, N. (1972). On the density of families of sets. *Journal of Combinatorial Theory - Series A*, 13(1):145–147.

Shalev-Shwartz, S. (2011). Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2009a). Stochastic convex optimization. In *Proceedings of the 22nd Annual Conference on Learning Theory*.

Shalev-Shwartz, S., Shamir, O., Sridharan, K., and Srebro, N. (2009b). Learnability and stability in the general learning setting.

Shalev-Shwartz, S., Singer, Y., and Srebro, N. (2007). Pegasos: primal estimated sub-gradient solver for SVM. In *Proceedings of the 24th International Conference on Machine Learning*, pages 807–814.

Shelah, S. (1972). A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261.

Srebro, N., Sridharan, K., and Tewari, A. (2010). Optimistic rates for learning with a smooth loss. *ArXiv e-prints*, arXiv:1009.3896.

Sridharan, K., Shalev-shwartz, S., and Srebro, N. (2009). Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems 21*, pages 1545–1552.

Tseng, P. (2008). On acclerated proximal gradient methods for convex-concave optimization. Technical report, University of Washington.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience.

Vapnik, V. N. and Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280.

Yang, T., Zhang, L., Jin, R., and Yi, J. (2016). Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 449–457.

Zhang, L., Yang, T., and Jin, R. (2017). Empirical risk minimization for stochastic convex optimization: $O(1/n)$- and $O(1/n^2)$-type of risk bounds. In *Proceedings of the 30th Annual Conference on Learning Theory*, pages 1954–1979.

Zhang, L., Yang, T., Jin, R., Xiao, Y., and Zhou, Z.-H. (2016). Online stochastic linear optimization under one-bit feedback. In *Proceedings of the 33rd International Conference on Machine Learning*.

Zhang, L., Yang, T., Jin, R., and Zhou, Z.-H. (2018). Dynamic regret of strongly adaptive methods. In *Proceedings of the 35th International Conference on Machine Learning*.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pages 928–936.