Stochastic Gradient Descent

论文标题: A Stochastic Approximation Method[1].

1 研究问题

已知$M(x)$为一未知函数, α为一个给定常量使得等式

$$M(x) = \alpha \tag{1}$$

有唯一解$x = \theta$. 我们想要找到上述的等式的解.

2 解决方法: 构造一个收敛至θ的数列$\{x_n\}$

通过给定初始值$x_1, \cdots, x_r$迭代生成x_n, x_n由$x_1, \cdots, x_{n-1}, M(x_1), \cdots, M(x_{n-1})$以及可能存在的导数$M'(x_1), \cdots, M'(x_{n-1})$作为自变量的函数生成.
我们将证明通过上述方法生成的数列$\{x_n\}$将*依概率收敛*至θ.(收敛速度取决于迭代方式的选取)
如果

$$\lim_{n \to \infty} x_n = \theta \tag{2}$$

与初始值$x_1, \cdots, x_r$的选取无关, 我们称上述方法对于特定的$M(x)$与α是*有效*的.

3 理论证明

引入

我们可以认为每一个x对应一个随机变量$Y = Y(x), Y(X)$有分布函数$Pr[Y(x) \leq y] = H(y|x)$使得

$$M(x) = \int_{-\infty}^{\infty} y \,\mathrm{d}H(y|x) = E[Y|x] \tag{3}$$

与$M(x)$一样的, 这里的$H(y|x)$也是未知的.
如果(2)初始值依概率收敛到θ, 且$x_1, \cdots, x_r$的选取无关, 我们称这个过程对于给定的$H(y|x)$和α是*一致的*.

Lemma 1.

在下述证明过程中我们会给$M(x)$添加一些重要的限制, 这些限制往往在实际中容易满足.
接下来, 我们认为$H(y|x)$是对于每个x的关于y的分布函数, 并且存在一个正实数C, 使得

$$Pr[|Y(x)| \leq C] = \int_{-C}^{C} \mathrm{d}H(y|x) = 1 \qquad \text{for all } x. \tag{4}$$

认为对每个x, 由(2)定义的$M(x)$存在且有限, 进一步认为存在有限常数α, θ使得

$$M(x) \leq \alpha \quad \text{for} \quad x < \theta, \quad M(x) \geq \alpha \quad \text{for} \quad x > \theta. \tag{5}$$

此处我们暂不关心是否$M(\theta) = \alpha$.
令$\{a_n\}$为给定的正项级数使得

$$0 < \sum_1^\infty a_n^2 = A < \infty. \tag{6}$$

接下来我们定义(非平稳的)马尔科夫链$\{x_n\}$, 令x_1为*任意*常数,

$$x_{n+1} - x_n = a_n(\alpha - y_n), \tag{7}$$

其y_n为随机向量使得

$$Pr[y_n \leq y | x_n] = H(y|x_n). \tag{8}$$

令

$$b_n = E(x_n - \theta)^2. \tag{9}$$

我们将证明

$$\lim_{n\to\infty} b_n = 0 \tag{10}$$

在任意初始值x_1的条件下均成立. 而 (10)可以推出x_n依概率收敛到θ.
由(7)可知

$$\begin{aligned} b_{n+1} = E(x_{n+1} - \theta)^2 = E[E[(x_{n+1} - \theta)^2]|x_n] &= E\left[E\left[\{(x_n - \theta) + a_n(\alpha - y_n)\}^2\right]|x_n\right] \\ &= E\left[\int_{-\infty}^{\infty} \{(x_n - \theta) - a_n(y - \alpha)\}^2 \mathrm{d}H(y|x_n)\right] \\ &= b_n + a_n^2 E\left[\int_{-\infty}^{\infty} (y - \alpha)^2 \mathrm{d}H(y|x_n)\right] - 2a_n E[(x_n - \theta)(M(x_n) - \alpha)] < spanid =''^9e7a9e''class =''blocki \end{aligned}$$

令

$$d_n = E[(x_n - \theta)(M(x) - \alpha)], \tag{12}$$

$$e_n = E\left[\int_{-\infty}^{\infty} (y - \alpha)^2 \mathrm{d}H(y|x_n)\right], \tag{13}$$

则由(11)

$$b_{n+1} - b_n = a_n^2 e_n - 2a_n d_n \tag{14}$$

由(5)可知

$$d_n \geq 0,$$

由(4)可知

$$0 \leq e_n \leq [C + |\alpha|]^2 < \infty.$$

结合(6)可知正项级数$\sum_1^\infty a_n^2 e_n$收敛.
累加(14)式可得

$$b_{n+1} = b_1 + \sum_{j=1}^{n} a_j^2 e_j - 2\sum_{j=1}^{n} a_j d_j \tag{15}$$

因为$b_{n+1} \geq 0$, 从而

$$\sum_{j=1}^{n} a_j d_j \leq \frac{1}{2}\left[b_1 + \sum_{j=1}^{n} a_j^2 e_j\right] < \infty. \tag{16}$$

令$n \to \infty$可知正项级数

$$\sum_1^\infty a_n d_n \tag{17}$$

收敛, 令(15)中$n \to \infty$

$$\lim_{n\to\infty} b_n = b_1 + \sum_1^\infty a_n^2 e_n - 2\sum_i^\infty a_n d_n = b \tag{18}$$

存在; $b \geq 0$.
我们认为存在(稍后证明其存在性)一个非负数列$\{k_n\}$使得

$$d_n \geq k_n b_n, \qquad \sum_1^\infty a_n k_n = \infty \tag{19}$$

由(17)的收敛性及(19)的前半部分可知

$$\sum_1^\infty a_n k_n b_n < \infty \tag{20}$$

由(19)的后半部分及(20)可知, 对任意$\epsilon > 0$, 必定存在无限多项$b_n < \epsilon$. 我们已知数列$\{b_n\}$的极限存在, 故$b = \lim_{n\to\infty} b_n = 0$. 因此我们证明了引理:
Lemma 1. 如果存在非负数列$\{k_n\}$满足式(19), 那么$b = 0$.

Lemma 2.

令

$$A_n = |x_1 - \theta| + [C + |\alpha|](a_1 + a_2 + \cdots + a_{n-1}); \tag{21}$$

那么由(4)和(7)可得

$$Pr[|x_n - \theta| \leq A_n] = 1. \tag{22}$$

设

$$\bar{k}_n = \inf\left[\frac{M(x)-\alpha}{x-\theta}\right] \quad \text{for} \quad 0<|x-\theta|\le A_n \tag{23}$$

由(5)可知$\bar{k}_n \ge 0$. 用$P_n(x)$表示x_n的概率分布函数, 有

$$d_n = \int_{|x-\theta|\le A_n}(x-\theta)(M(x)-\alpha)\mathrm{d}P_n(x) \ge \int_{|x-\theta|\le A_n}\bar{k}_n|x-\theta|^2\mathrm{d}P_n(x) = \bar{k}_n b_n \tag{24}$$

我们证明了由(23)定义的$\bar{k}_n$满足(19)的前半部分, 为了证明后半部分我们作出如下假设

$$\bar{k}_n \ge \frac{K}{A_n} \tag{25}$$

对某个常数K以及足够大的n成立, 并且有(稍后证明这样的K和a_n存在)

$$\sum_{n=2}^{\infty}\frac{a_n}{(a_1+\cdots+a_{n-1})} = \infty. \tag{26}$$

从(26)可知

$$\sum_{1}^{\infty}a_n = \infty, \tag{27}$$

因此对于足够大的n

$$2[C+|\alpha|](a_1+\cdots+a_{n-1}) \ge A_n. \tag{28}$$

由(25)我们可知对于足够大的n

$$a_n\bar{k}_n \ge a_n\frac{K}{A_n} \ge \frac{a_nK}{2[C+|\alpha|](a_1+\cdots+a_{n-1})}, \tag{29}$$

由(26)及(29)可证明(19)的后半部分.

Lemma 2. 如果(25)和(26)成立, 那么$b=0$.

如此，问题变为了寻找满足(25)和(26)的K和a_n.
对于满足(6)和(26)的数列, 我们有一个非常熟悉的数列$a_n = 1/n$, 因为

$$\sum_{1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}, \quad \sum_{n=2}^{\infty}\left[\frac{1}{n\left(1+\frac{1}{2}+\cdots+\frac{1}{n-1}\right)}\right] = \infty$$

更一般的, 对任何数列$\{a_n\}$, 存在两个正常数c', c''满足

$$\frac{c'}{n} \le a_n \le \frac{c''}{n} \tag{30}$$

也会满足(6)和(26). 我们把任何满足条件(6)和(26)的数列称为$1/n$*型数列*, 无论其是否满足(30). 下面将寻找满足(25)的条件.

Theorom 1.

如果$\{a_n\}$是$1/n$型数列, 那么将很容易找到满足(5)和(25)的$M(x)$(但这样的(5)还不足以证明(25)), 对于(5)的强化条件: 对某个$\delta > 0$,

$$M(x) \le \alpha - \delta \quad \text{for} \quad x < \theta, \quad M(x) \ge \alpha + \delta \quad \text{for} \quad x > \theta \tag{5.1}$$

那么对于$0 < |x - \theta| \le A_n$, 有

$$\frac{M(x) - \alpha}{x - \theta} \ge \frac{\delta}{A_n}, \tag{31}$$

从而

$$\bar{k}_n \ge \frac{\delta}{A_n}, \tag{32}$$

即(25)中$K = \delta$, 再由Lemma 2., 我们可以总结出
Theorom 1. 如果$\{a_n\}$是$1/n$型数列, 且(4)成立, 如果$M(x)$满足(5.1), 那么$b = 0$.

Theorom 2.

我们关注$M(x)$可以满足(25)的另一种情况

$$M(x) \text{ is nondecreasing,} \tag{33}$$

$$M(\theta) = \alpha, \tag{34}$$

$$M'(\theta) > 0. \tag{35}$$

我们讲证明(25)在$M(X)$满足以上情况下成立.由(34), 可知

$$M(x) - \alpha = (x - \theta)[M'(\theta) + \epsilon(x - \theta)], \tag{36}$$

其中$\epsilon(t)$是满足如下条件的方程

$$\lim_{t \to \infty} \epsilon(t) = 0. \tag{37}$$

因此存在常数$\delta > 0$使得

$$\epsilon(t) \ge -\frac{1}{2} M'(\theta) \quad \text{for} \quad |t| \le \delta \tag{38}$$

从而

$$\frac{M(x) - \alpha}{x - \theta} \ge \frac{1}{2} M'(\theta) > 0 \quad \text{for} \quad |x - \theta| \le \delta \tag{39}$$

另外, 对于$\theta + \delta \le x \le \theta + A_n$, 因为$M(x)$是非降函数,

$$\frac{M(x)-\alpha}{x-\theta} \geq \frac{M(\theta+\delta)-\alpha}{A_n} \geq \frac{\delta M'(\theta)}{2A_n}, \tag{40}$$

而另一边, 对于$\theta - A_n \leq x \leq \theta - \delta$,

$$\frac{M(x)-\alpha}{x-\theta} = \frac{\alpha - M(x)}{\theta - x} \geq \frac{\alpha - M(\theta-\delta)}{A_n} \geq \frac{\delta M'(\theta)}{2A_n}. \tag{41}$$

不失一般性, 认为$\delta/A_n \leq 1$, 由(39), (40)及(41), 可知

$$\frac{M(x)-\alpha}{x-\theta} \geq \frac{\delta M'(\theta)}{2A_n} > 0 \quad \text{for} \quad 0 < |x-\theta| \leq A_n, \tag{42}$$

即(25)中$K = \delta M'(\theta)/2 > 0$, 我们可以总结出

Theorom 2. 如果$\{a_n\}$是$1/n$型数列, 且(4)成立, 如果$M(x)$满足(33), (34)和(35), 那么$b = 0$.

---

1. Robbins, H. & Monro, S. A Stochastic Approximation Method. *The Annals of Mathematical Statistics* **22**, 400–407 (1951). ↩