arXiv:2303.04444v1 [math.ST] 8 Mar 2023

关于快速增长无界函数的经验最小化点的 L^1收敛

Pierre Bras*

摘要

对于 $V:\ \mathbb{R}^d \to \mathbb{R}$ 强制的情形，我们研究经验最小化点的 L^1距离的收敛速率。该经验最小化点是函数 V 在有限数量的n 个带噪样本上取到的真实最小值，与 V的最小值之间的距离。我们证明，一般而言，对于快速增长的无界函数，收敛速率的上界为 $a_n n^{-1/q}$，其中 q 是潜在随机变量的维数，并且对每个 $\varepsilon > 0$，都有 $a_n = o(n^{\varepsilon})$。随后，我们介绍其在机器学习和蒙特卡罗模拟中产生的优化问题上的应用。

关键词–随机优化、经验测度、经验最小化点

1 引言

随机算法是求解高维复杂优化问题的有力工具。我们考虑一个待最小化的函数 $V:\ \mathbb{R}^d \to (0,\infty)$，并假设存在一个取值于 $\mathbb{R}^q$ 的随机变量 Z 和一个函数 $v:\ \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}^+$，使得

$$\forall x \in \mathbb{R}^d,\ \mathbb{E}[v(x,Z)] = V(x).$$

此外，我们重点考虑无法精确计算 V 或精确计算其代价过高的情形，此时我们需要依赖来自 v 的含噪观测来估计 V。为了最小化 V，我们按如下步骤进行：首先抽取独立同分布样本 $Z_1, \ldots, Z_n$，然后不再最小化 V，而是最小化其经验版本

$$V_n(x) = \frac{1}{n}\sum_{i=1}^{n} v(x, Z_i),$$

通常使用随机梯度下降算法。这类方法常见于机器学习相关问题中，此时原始数据集过大，我们只使用其中一个（仍然很大的）子集；也常见于蒙特卡罗模拟中，此时我们用经验平均来近似期望。

一个自然的问题是证明经验最小值收敛到真实最小值。记 $X_n^\star$ 为 V_n 的一个最小化点，$x^\star$ 为 V 的一个最小化点，我们给出 $\mathbb{E}|V_n(X_n^\star) - V(x^\star)|$ 的收敛速率，其中期望是对抽样 $Z_1, \ldots, Z_n$ 而取的。结果表明，其阶为 $a_n n^{-1/q}$，其中对每个 $\varepsilon > 0$，都有 $a_n = o(n^{\varepsilon})$。当 V 的最小值仅在一个点处取得时，我们还给出 $\mathbb{E}|X_n^\star - x^\star|^2$ 的收敛速率。为此，我们依赖 [FG15]，它给出了经验测度 $(1/n)\sum_{i=1}^n \delta_{Z_i}$ 关于 L^1-Wasserstein 距离收敛到 Z 的分布的收敛速率，并表明一般而言，该收敛速率的阶为 $n^{-1/q}$，且这一界是紧的，因此一般情况下无法期待我们的问题具有更好的收敛速率。

经验风险最小化点的收敛问题在 [VA71] 中针对二分类问题得到了深入研究，其中证明了量 $|V_n(X_n^\star) - V(x^\star)|$ 与$\sup_{x\in\mathbb{R}^d} |V_n(x) -$

*索邦大学，概率、统计与建模实验室，UMR 8001，188号信箱，朱西厄广场4号，法国巴黎邮政编码75252，CEDEX 5。电子邮箱：pierre.bras@sorbonne-universite.fr

$V(x)|$（见引理 4.2）；并且，如果 $\mathcal{V}:=\{v(x,\cdot):\ x\in\mathbb{R}^d\}$ 是一个由 $\{0,1\}$值函数构成的函数类，且 $\mathcal{V}$的 Vapnik-Chervonenkis 维数有限（该维数是一个整数，用于衡量 $\mathcal{V}$的组合性质），那么 $\mathbb{E}\sup_{x\in\mathbb{R}^d}|V_n(x)-V(x)|$ 以$1/\sqrt{n}$ 的速率收敛至零（见 [BM06，引理 2.4]）。 [BM06] 和 [BMP10] 利用集中不等式，将这些结果推广到了有界、Bernstein 型和星形函数类。然而，在蒙特卡罗模拟的背景下，这一问题尚未得到研究。

[BDL08] 和 [LP20]研究了通过 k-均值聚类对随机向量进行最优量化的风险，并证明了一个 $O(\log(n)/n)$上界，然而这些结果不能直接应用于经验最小化解的收敛速率。

在本文中，我们证明了 $n^{-1/q}$ 对于具有快速增长的更一般的无界函数类的收敛速率，并且不假设凸性。

本文的组织结构如下。在第2节中，我们陈述假设并给出主要结果。在第3节中，我们给出机器学习和蒙特卡罗模拟中的经典示例，在这些示例中我们的假设得到验证，且我们的结果可以得到应用。

记号

我们赋予空间 $\mathbb{R}^d$ 以由 $|\cdot|$表示的标准欧几里得范数。对于 $x\in\mathbb{R}^d$ 和$R>0$，我们记$\mathbf{B}(x,R)=\{y\in\mathbb{R}^d:\ |y-x|\le R\}$。

对于 f：满足$\min_{\mathbb{R}^d}(f)$存在的 $\mathbb{R}^d\to\mathbb{R}$，我们记$\operatorname{argmin}(f)=\{x\in\mathbb{R}^d:\ f(x)=\min_{\mathbb{R}^d}(f)\}$。

若 $f:\ \mathbb{R}^d\to\mathbb{R}$ 是 Lipschitz 连续的，我们用 $[f]_{\mathrm{Lip}}$ 表示其 Lipschitz 常数。我们常考虑 f 在某个集合 $\mathcal{E}\subset\mathbb{R}^d$ 上的 Lipschitz 常数，并将其记为 $[f_{|\mathcal{E}}]_{\mathrm{Lip}}$。

若 f满足 $\lim_{|x|\to\infty}f(x)=+\infty$，则称其为强制的。

我们将两个分布 π_1 和 π_2 在 $\mathbb{R}$上的 L^p-Wasserstein 距离记为 ^{d}:

$$\mathcal{W}_p(\pi_1,\pi_2)=\inf\left\{\left(\int_{\mathbb{R}^d}|x-y|^p\pi(dx,dy)\right)^{1/p}:\ \pi\in\mathcal{P}(\pi_1,\pi_2)\right\},$$

其中，$\mathcal{P}(\pi_1,\pi_2)$表示在 $(\mathbb{R}^d\times\mathbb{R}^d,\mathcal{B}or(\mathbb{R}^d)^{\otimes 2})$ 上边缘分布分别为 π_1 和 π_2的概率分布集合。

对于 $x\in\mathbb{R}^d$，我们用 δ_x 表示位于 x处的狄拉克质量。

对于集合 $\mathcal{E}\subset\mathbb{R}^d$，我们定义 $\operatorname{Radius}(\mathcal{E}):=\sup\{|x|:\ x\in\mathcal{E}\}$。

本文中，我们使用记号 C 和 C' 表示正常数，它们可能随行而变化。

2 主要结果

设 $V:\ \mathbb{R}^d\to(0,\infty)$ 是 $\mathcal{C}^0$ 且为强制的，即当 $|x|\to\infty$ 时，$V(x)\to+\infty$。这保证了 $\min_{\mathbb{R}^d}V$ 存在，记作 $V^\star:=\min_{\mathbb{R}^d}V>0$。我们假设存在一个定义在某个概率空间 $(\Omega\ \mathcal{A},\mathbb{P})$ 上的随机变量 Z，其分布为 μ，取值于 $\mathbb{R}^q$，以及某个函数 $v:\ \mathbb{R}^d\times\mathbb{R}^q\to\mathbb{R}^+$，该函数是 $\mathcal{C}^0$，使得

$$\forall x\in\mathbb{R}^d,\ \mathbb{E}[v(x,Z)]=V(x). \tag{2.1}$$

我们观测到 $Z_1,\ldots,Z_n$ 个来自 Z 的独立同分布样本，并令

$$V_n(x):=\frac{1}{n}\sum_{i=1}^n v(x,Z_i),\quad x\in\mathbb{R}^d. \tag{2.2}$$

在以下假设成立的条件下，定理 2.1(ii) 紧接着给出，几乎必然地，对于 n 足够大时，V_n 是强制的，接下来我们定义 $X_n^\star\in\mathbb{R}^d$ 为 V_n 的一个经验最小化点，即

$$X_n^\star\in\operatorname{argmin}(V_n)=\operatorname{argmin}_{x\in\mathbb{R}^d}\left(\frac{1}{n}\sum_{i=1}^n v(x,Z_i)\right). \tag{2.3}$$

我们的目标是证明，最小化 V_n V 的经验观测可以很好地逼近真实最小值 $V^\star$，也就是说，我们旨在证明 $V_n(X_n^\star)$ 收敛到 $V^\star$，并给出 $\mathbb{E}|V_n(X_n^\star) - V^\star|$ 的界。为方便起见，我们还定义了依赖于维数的收敛速率，一般情况下等于 $n^{-1/q}$：

$$\mathcal{R}_q(n) = \begin{cases} n^{-1/2} & \text{if } q = 1, \\ n^{-1/2}\log(1+n) & \text{if } q = 2, \\ n^{-1/q} & \text{if } q > 2. \end{cases}$$

我们定义 μ 的 $r > 0$ 阶矩:

$$\mathscr{M}_r(\mu) := \int_{\mathbb{R}^q} |x|^r \mu(dx) \in [0, \infty].$$

现在陈述我们的主要结果。

定理 2.1。假设:

(i) 分布 μ 满足 $\mathscr{M}_r(\mu) < \infty$，其中 $r > q/(q-1)$（若 $q \neq 1$），且对某个 $r > 2$（若 $q = 1$）。

(ii) 存在一个紧集 $\mathcal{Q} \subset \mathbb{R}^q$，使得 $x \mapsto \inf_{z\in\mathcal{Q}} v(x,z)$ 是强制的，且 $\mathbb{P}(Z \in \mathcal{Q}) > 0$。*(iii)* 对所有 $M > 0$，存在一个常数 C，使得对所有紧集 $\mathcal{K} \subset \mathbb{R}^q$:

$$\sup_{x\in\boldsymbol{B}(0,M)} \mathbb{E}[v(x,Z)\mathbb{1}_{Z\notin\mathcal{K}}] \leq Ce^{-C\text{Radius}(\mathcal{K})^2} \tag{2.4}$$

$$\sup_{x\in\boldsymbol{B}(0,M)} [v(x,\cdot)_{|\mathcal{K}}]_{\text{Lip}} \leq Ce^{C\text{Radius}(\mathcal{K})}. \tag{2.5}$$

(iv) 存在 $x^\star \in$ 极小点集(V)，使得 $\mathbb{E}[v^2(x^\star, Z)] < \infty$。

Then

(a) $V_n(X_n^\star)$ 在 $L^1(\mathbb{P})$ 中收敛于 $V^\star$，并且

$$\mathbb{E}|V_n(X_n^\star) - V^\star| \leq Ca_n\mathcal{R}_q(n), \tag{2.6}$$

(b) 如果进一步假设 $\text{argmin}(V) = \{x^\star\}$ 仅包含一个点，$\nabla^2 V(x^\star)$ 存在且为正定，并且 $(\mathbb{E}|X_n|^2)$ 有界，那么 $X_n^\star$ 在 $L^2(\mathbb{P})$ 中收敛于 $x^\star$，并且

$$\mathbb{E}|X_n^\star - x^\star|^2 \leq Ca_n\mathcal{R}_q(n), \tag{2.7}$$

其中 (a_n) 是某个正数列，使得对每个 $\varepsilon > 0$，都有 $a_n = o(n^\varepsilon)$。

注 2.2。在存在 $A \geq 0$ 和 $B > 0$ 满足 $V(x) \geq A + B|x|^2$ 的假设下，条件 $(\mathbb{E}|X_n|^2)$ 有界得以满足。事实上，此时我们有

$$\mathbb{E}|X_n|^2 \leq B^{-1}(\mathbb{E}V(X_n^\star) - A) = B^{-1}(\mathbb{E}V_n(X_n^\star) - A),$$

根据定理 *2.1* (a) 可知其有界。

3 应用

3.1 蒙特卡罗模拟的重要性抽样

我们来介绍一个来自 [LP10]的问题。设 $\varphi\colon\ \mathbb{R}^d \to \mathbb{R}$ 为 $\mathcal{C}^0$，并设 $Z \sim \mathcal{N}(0,\ I_d)$。为了估计 $\mathbb{E}[\varphi(Z)]$,，我们对变量 $\varphi(Z)$ 采用重要性抽样进行方差缩减。例如，对于布莱克-斯科尔斯模型中的看涨期权和看跌期权，分别有

$$\varphi(Z) = \left(\sum_{i=1}^{d} a_i s_0^i \exp\left[\left(r - \sum_j \frac{\sigma_{i,j}^2}{2} \right) T + \sqrt{T} \sum_j \sigma_{i,j} Z_j \right] - K \right)_+, \tag{3.1}$$

$$\varphi(Z) = \left(K - \sum_{i=1}^{d} a_i s_0^i \exp\left[\left(r - \sum_j \frac{\sigma_{i,j}^2}{2} \right) T + \sqrt{T} \sum_j \sigma_{i,j} Z_j \right] \right)_+, \tag{3.2}$$

其中 $r,\ T,\ K \in \mathbb{R}^+,\ \sigma \in \mathcal{M}_d(\mathbb{R})$ 是对称正定的，且 $s_0,\ a \in \mathbb{R}^d$。让我们通过平移进行重要性抽样。对于 $x \in \mathbb{R}^d$，我们有

$$\begin{aligned} \mathbb{E}[\varphi(Z)] &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \varphi(z) e^{-\frac{|z|^2}{2}} dz = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \varphi(z+x) e^{\frac{-|x|^2}{2} - \langle x, z \rangle} e^{-\frac{z^2}{2}} dz \\ &= \mathbb{E}\left[\varphi(Z+x) e^{\frac{-|x|^2}{2} - \langle x, Z \rangle} \right] \end{aligned} \tag{3.3}$$

并且我们有

$$\begin{aligned} \mathrm{Var}\left[\varphi(Z+x) e^{\frac{-|x|^2}{2} - \langle x, Z \rangle} \right] &= \mathbb{E}\left[\varphi^2(Z+x) e^{-|x|^2 - 2\langle x, Z \rangle} \right] - \mathbb{E}\left[\varphi(Z+x) e^{\frac{-|x|^2}{2} - \langle x, Z \rangle} \right]^2 \\ &= \mathbb{E}\left[\varphi^2(Z+x) e^{-|x|^2 - 2\langle x, Z \rangle} \right] - \mathbb{E}[\varphi(Z)]^2. \end{aligned}$$

因此，目标是

$$\underset{x \in \mathbb{R}^d}{\text{Minimize}}\ V(x) = \mathbb{E}\left[\varphi^2(Z+x) e^{-|x|^2 - 2\langle x, Z \rangle} \right] = \mathbb{E}\left[\varphi^2(Z) e^{-\langle x, Z \rangle + |x|^2/2} \right] =: \mathbb{E}[v(x, Z)]. \tag{3.4}$$

为了 用较小的方差通过蒙特卡罗模拟估计 $\mathbb{E}[\varphi(Z)]$，使用 (3.3) .

命题 3.1。设 $Z \sim \mathcal{N}(0, I_d)$ 且设 $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ 满足以下条件：存在 A、$B \geq 0$，使得

$$\varphi \not\equiv 0, \quad \forall z \in \mathbb{R}^d,\ |\varphi(z)| \leq A e^{B|z|} \quad \textit{and} \quad \forall r > 0,\ [\varphi_{|\boldsymbol{B}(0,r)}]_{\mathrm{Lip}} \leq A e^{Br}. \tag{3.5}$$

则由下式定义的函数 $v\colon\ \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$

$$v(x, z) = \varphi^2(z) e^{-\langle x, Z \rangle + |x|^2/2}$$

满足定理 *2.1* 的假设，特别是：

$$\mathbb{E}|V_n(X_n^\star) - V^\star| \leq C a_n \mathcal{R}_d(n).$$

证明。 • 由于 $\varphi \not\equiv 0$，存在一个内部非空的紧集 $\mathcal{Q} \subset \mathbb{R}^q$，使得 $\inf_{z \in \mathcal{Q}} \varphi(z) > 0$；又由于 $Z \sim \mathcal{N}(0, I_q)$，我们有 $\mathbb{P}(Z \in \mathcal{Q}) > 0$。

• 对于 $\mathcal{K} = \mathbf{B}(0, R)$，我们有

$$\mathbb{E}\left[\sup_{x \in \mathbf{B}(0,M)} v(x, Z) \mathbb{1}_{Z \notin \mathcal{K}} \right] = C \int_{\mathbf{B}(0,R)^c} \sup_{x \in \mathbf{B}(0,M)} \varphi^2(z) e^{-\langle x, z \rangle + |x|^2/2 - |z|^2/2} dz$$

$$\leq C\int_{\mathbf{B}(0,R)^c} e^{2B|z|+M|z|+|M|^2/2-|z|^2/2}dz \leq C\int_{\mathbf{B}(0,R)^c} e^{C|z|-|z|^2/2}dz \leq Ce^{-C'R^2}.$$

• 对于 $\mathcal{K}=\mathbf{B}(0,R)$ 和 $x\in\mathbf{B}(0,M)$，我们有

$$\begin{aligned}[v(x,\cdot)_{|\mathcal{K}}]_{\mathrm{Lip}} &\leq \sup_{z\in\mathcal{K}}\left(\varphi^2(z)|x|e^{|x||z|}+2\varphi(z)[\varphi_{\mathcal{K}}]_{\mathrm{Lip}}e^{|x||z|}\right)\\ &\leq Ce^{2BR+MR}+Ce^{(B+B')R+MR}\leq Ce^{C'R}.\end{aligned}$$

□

注 3.2。 如果 φ 定义为布莱克-斯科尔斯看涨期权 (3.1) 或看跌期权 (3.2)则满足假设(3.5)。

3.2 神经网络

考虑一个使用全连接神经网络、采用二次损失和二次正则化的回归问题。令 $\varphi:\ \mathbb{R}\to\mathbb{R}$ 为 S 形函数 $x\mapsto(1+e^{-x})^{-1}$。令 $K\in\mathbb{N}$ 为层数，对于 $k=1,\ .....,\ K$，令 $d_k\in\mathbb{N}$ 为 k^{th}的大小。对于 $u\in\mathbb{R}^{d_{k-1}}$ 以及对于$x\in\mathcal{M}_{d_k,\ d_{k-1}}(\mathbb{R})$，我们定义 $\varphi_x(u):=[\varphi([x\cdot u]_i)]_{1\leq i\leq d_k}$。神经网络的输出为

$$\begin{aligned}&\psi:\mathbb{R}^{d_1,d_0}\times\cdots\times\mathbb{R}^{d_K,d_{K-1}}\times\mathbb{R}^{d_0}\to\mathbb{R}^{d_K}\\ &\psi(x_1,\ldots,x_K,u)=\psi(x,u)=x_K\cdot\varphi_{x_{K-1}}\circ\ldots\circ\varphi_{x_1}(u).\end{aligned}$$

令 $u_i\in\mathbb{R}^{d_0}$ 和 $y_i\in\mathbb{R}^{d_K}$ 为 $1\leq i\leq N$的数据。目标为

$$\underset{x_1,\ldots,x_K}{\text{minimize}}\quad V(x):=\frac{1}{2N}\sum_{i=1}^{N}|\psi(x_1,\ldots,x_K,u_i)-y_i|^2+\frac{\lambda}{2}|x|^2,$$

其中 $x=(x_1,\ .....,\ x_K)$，且其中 $\lambda>0$ 为正则化参数。我们将 $d:=d_0d_1+\cdots+d_{K-1}d_K$ 记为优化问题的维数。

命题 3.3。设 $\mu:=(1/N)\sum_{i=1}^{N}\delta_{u_i,\ y_i}$ 为与数据 $\mathbb{R}^{d_0}\times\mathbb{R}^{d_K}$ 相关联的概率测度，设 $Z\sim\mu$，并设

$$v:\mathbb{R}^d\times\mathbb{R}^{d_0}\times\mathbb{R}^{d_K}\to\mathbb{R}^+,\quad (x,u,y)\mapsto\frac{1}{2}|\psi(x,u)-y|^2+\frac{\lambda}{2}|x|^2.$$

则 Z 和 v 满足定理 2.1的假设。

证明。 • 取任意包含至少一个数据 $\mathcal{Q}\subset\mathbb{R}^{d_0}\times\mathbb{R}^{d_K}$ $(u_i,\ y_i)$ 的紧集 v，并注意到 $x\ (u,\ y)\ \geq\ (\lambda/2)$ $|x|^2$，即可满足条件（ii）。

•条件（2.4）得到满足，因为 Z 具有紧支撑。

•我们有

$$\partial_y v(x,u,y)=|\psi(x,u)-y|\leq|y|+\|x_K\|$$

and

$$\partial_u v(x,u,y)=|\psi(x,u)-y|\cdot\|\partial_u\psi(x,u)|\leq(|y|+\|x_K\|)\cdot\|x_K\|\cdot|\varphi'_{x_{K-1}}\circ\varphi_{x_{K-2}}\cdots\circ\varphi_{x_1}(u)|\cdots|\varphi'_{x_1}(u)|,$$

并且由于所有函数 φ'_{x_k} 在 $x\in\mathbf{B}(0,M)$上均有界，因此我们得到（2.5）。 □

定理 2.1 的证明

4.1 经验测度关于 L^1-Wasserstein 距离的收敛

设 μ 是 $\mathbb{R}^q$ 上的一个概率分布，并设 $(Z_n)_{n\geq 1}$ 是服从分布 μ的独立同分布随机变量序列。对于 $n \geq 1$, 记经验测度为

$$\mu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i}.$$

[FG15，定理 1] 给出了 L^p-Wasserstein 距离的收敛速率，适用于 μ_n 到 μ。

定理 4.1。设 $p > 0$，并假设存在某个 $r > p$，使得 $\mathscr{M}_r(\mu) < \infty$。那么存在一个常数$C > 0$，它仅取决于 q、p、r，使得对于所有 $n \geq 1$,

$$\mathbb{E}[\mathcal{W}_p(\mu_n, \mu)] \leq C\mathscr{M}_r^{p/r}(\mu) \begin{cases} n^{-1/2} + n^{-(r-p)/r} & \text{if } p > q/2 \text{ and } r \neq 2p, \\ n^{-1/2}\log(1+n) + n^{-(r-p)/r} & \text{if } p = q/2 \text{ and } r \neq 2p, \\ n^{-p/q} + n^{-(r-p)/r} & \text{if } p \in (0, q/2) \text{ and } r \neq q/(q-p). \end{cases}$$

其中，期望是对样本 $X_1, \ldots, X_n$ 取的。

紧接着 [FG15，定理 1] 给出了恰当的例子，证明这些界是紧的，这表明一般而言，我们无法在定理 2.1 中得到更好的界。

现在考虑 $p = 1$ 的情形。根据定理 2.1(i) 中的假设，即 $\mathscr{M}_r(\mu) < \infty$对某个 $r > q/(q-1)$ 若 $q \neq 1$，且对某个 $r > 2$ 若 $q = 1$，则项 $n^{-(r-1)/1}$ 变得可以忽略，因而

$$\mathbb{E}[\mathcal{W}_1(\mu_n, \mu)] \leq C_\mu \mathcal{R}_q(n).$$

回顾一下 L^1-Wasserstein 距离的康托罗维奇-鲁宾斯坦表示 [Vil09，公式 (6.3)：

$$\mathcal{W}_1(\mu_1, \mu_2) = \sup\left\{ \int_{\mathbb{R}^q} f(x)(\mu_1 - \mu_2)(dx) : \ f : \mathbb{R}^d \to \mathbb{R}, \ [f]_{\text{Lip}} = 1 \right\}. \tag{4.1}$$

4.2 定理 2.1(a) 的证明

引理 4.2。假设 $X_n^\star \in \Theta$ 且 $x^\star \in \Theta$ 对某个 $x^\star \in \operatorname{argmin}(V)$ 成立，并且对某个集合$\Theta \subset \mathbb{R}^d$成立。那么我们有

$$|V_n(X_n^\star) - V^\star| \leq \sup_{x\in\Theta} |V_n(x) - V(x)|.$$

证明。若 $V_n(X_n^\star) \geq V^\star$，则

$$|V_n(X_n^\star) - V^\star| = V_n(X_n^\star) - V^\star \leq V_n(x^\star) - V(x^\star) = |V_n(x^\star) - V(x^\star)|.$$

若 $V_n(X_n^\star) \leq V^\star$，则

$$|V_n(X_n^\star) - V^\star| = V(x^\star) - V_n(X_n^\star) \leq V(X_n^\star) - V_n(X_n^\star) = |V(X_n^\star) - V_n(X_n^\star)|.$$

□

从现在起，我们的证明策略如下。如果 $X_n^\star$ 和 $x^\star$ 属于某个集合 $\Theta \subset \mathbb{R}^d$，那么利用（4.1）可得：

$$\mathbb{E}|V_n(X_n^\star) - V^\star| \leq \mathbb{E}\sup_{x\in\Theta} |V_n(x) - V(x)|$$

$$= \mathbb{E} \sup_{x\in\Theta} \left| \int_{\mathbb{R}^q} v(x,z)\mu_n(dz) - \int_{\mathbb{R}^q} v(x,z)\mu(dz) \right|$$
$$\leq \left(\sup_{x\in\Theta} [v(x,\cdot)]_{\mathrm{Lip}} \right) \mathbb{E}[\mathcal{W}_1(\mu_n,\mu)].$$

因此，如果 v（x，$\cdot$）是 Lipschitz 连续的，且关于 $x \in \Theta$一致，那么我们可以直接应用定理4.1，直接得到 $\mathbb{E}|V_n(X_n^\star) - V^\star|$的界。然而，正如第3.1节所示，一般情况下并非如此。

我们首先证明，可以认为 $X_n^\star \in \Theta$，其中 $\Theta = \mathbf{B}(\mathbf{0}, M)$ 对某个 $M > 0$ 成立。

引理 4.3。设 $M_0 > 0$ 使得 $x^\star \in \boldsymbol{B}(0, M_0)$，并且

$$\inf_{x\notin \boldsymbol{B}(0,M_0)} \inf_{z\in\mathcal{Q}} v(x,z) > 4V^\star/\mathbb{P}(Z \in \mathcal{Q}).$$

注意到，利用定理 *2.1(ii)* 的假设，可以得到这样的 M_0 存在。于是

$$\mathbb{E}|(V_n(X_n^\star) - V^\star)\mathbb{1}_{|X_n^\star|>M}| \leq Cn^{-1}. \tag{4.2}$$

证明。我们定义 $\kappa := \mathbb{P}(Z \in \mathcal{Q}) > 0$。我们有

$$\mathbb{E}|(V_n(X_n^\star) - V^\star)\mathbb{1}_{|X_n^\star|>M_0}| \leq \mathbb{E}[V_n(X_n^\star)\mathbb{1}_{|X_n^\star|>M_0}] + V^\star \mathbb{P}\left(|X_n^\star| > M_0\right).$$

但是，如果$|X_n^\star| > M_0$，那么

$$\inf_{|x|>M_0} V_n(x) \leq V_n(X_n^\star) \leq V_n(x^\star),$$

因此必然不可能有

$$V_n(x^\star) < 2V^\star \leq \inf_{|x|>M_0} V_n(x),$$

因此

$$\mathbb{P}\left(|X_n^\star| > M_0\right) \leq \mathbb{P}(V_n(x^\star) \geq 2V^\star) + \mathbb{P}\left(\inf_{|x|>M_0} V_n(x) < 2V^\star \right). \tag{4.3}$$

根据切比雪夫不等式，我们有

$$\mathbb{P}(V_n(x^\star) \geq 2V^\star) \leq \mathbb{P}(|V_n(x^\star) - V^\star| \geq V^\star) \leq \frac{\mathrm{Var}[v(x^\star, Z)]}{V^\star n}.$$

并且如果 $\inf_{|x|>M_0} V_n(x) < 2V^\star$，那么利用 $v \geq 0$，我们有

$$2V^\star > \inf_{|x|>M_0} V_n(x) \geq \inf_{|x|>M_0} \frac{1}{n}\sum_{i=1}^n v(x,Z_i)\mathbb{1}_{Z_i\in\mathcal{Q}} \geq \left(\inf_{|x|>M_0} \inf_{z\in\mathcal{Q}} v(x,z) \right) \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Z_i\in\mathcal{Q}},$$

因此必有 $(1/n)\sum_{i=1}^n \mathbb{1}_{Z_i\in\mathcal{Q}} < \kappa/2$，再次利用切比雪夫不等式：

$$\mathbb{P}\left(\inf_{|x|>M_0} V_n(x) < 2V^\star \right) \leq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Z_i\in\mathcal{Q}} < \frac{\kappa}{2} \right) \leq \mathbb{P}\left(\left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Z_i\in\mathcal{Q}} - \kappa \right| > \frac{\kappa}{2} \right) \leq \frac{4}{n\kappa^2}.$$

因此由（4.3）可得

$$\mathbb{P}(|X_n^\star| > M_0) \leq C/n \tag{4.4}$$

然后

$$\begin{aligned}
\mathbb{E}|(V_n(X_n^\star) - V^\star)\mathbb{1}_{|X_n^\star|>M_0}| &\leq \mathbb{E}|V_n(X_n^\star)\mathbb{1}_{|X_n^\star|>M_0}| + V^\star\mathbb{P}\left(|X_n^\star| > M_0\right) \\
&\leq C\sqrt{\mathbb{E}[V_n^2(X_n^\star)]}n^{-1/2} + Cn^{-1} \\
&\leq C\mathrm{Var}[v(x^\star, Z)]n^{-1} + Cn^{-1}.
\end{aligned}$$

□

现在我们证明定理 2.1。

证明。定义紧集:

$$\Theta = \mathbf{B}(0, M_0) \subset \mathbb{R}^d, \quad \mathcal{K}_n := \mathbf{B}(0, \beta_n) \subset \mathbb{R}^q, \quad \widetilde{\mathcal{K}}_n := \mathbf{B}(0, \beta_n + 1) \subset \mathbb{R}^q,$$

其中，(β_n) 是一个稍后将定义的正递增序列。对于 $n \in \mathbb{N}$，我们还定义连续函数

$$\phi_n : \mathbb{R}^q \to \mathbb{R}, \quad z \mapsto \begin{cases} 1 & \text{if } z \in \mathcal{K}_n, \\ 0 & \text{if } z \notin \widetilde{\mathcal{K}}_n, \\ \beta_n + 1 - |z| & \text{if } z \in \widetilde{\mathcal{K}}_n \setminus \mathcal{K}_n. \end{cases}$$

于是我们有

$$\mathbb{E}|V_n(X_n^\star) - V^\star| = \mathbb{E}|(V_n(X_n^\star) - V^\star)\mathbb{1}_{|X_n^\star| > M_0}| + \mathbb{E}|(V_n(X_n^\star) - V^\star)\mathbb{1}_{|X_n^\star| \le M_0}|. \tag{4.5}$$

利用引理 a 4.3，第一项有界于 Cn^{-1}。现在转向（4.5）中的第二项:

$$\begin{aligned} \sup_{x\in\Theta} |V_n(x) - V(x)| &= \sup_{x\in\Theta} \left| \int_{\mathbb{R}^q} v(x,z)\mu_n(dz) - \int_{\mathbb{R}^q} v(x,z)\mu(dz) \right| \\ &= \sup_{x\in\Theta} \left| \int_{\mathbb{R}^q} v(x,z)\phi_n(z)(\mu_n - \mu)(dz) + \int_{\mathbb{R}^q} v(x,z)(1-\phi_n(z))(\mu_n - \mu)(dz) \right| \\ &\le \left(\sup_{x\in\Theta} [(v(x,\cdot)\phi_n)]_{\mathrm{Lip}} \right) \mathcal{W}_1(\mu_n, \mu) + \sup_{x\in\Theta} \int_{z\notin\mathcal{K}_n} v(x,z)\mu(dz) \\ &\quad + \sup_{x\in\Theta} \int_{z\notin\mathcal{K}_n} v(x,z)\mu_n(dz). \end{aligned} \tag{4.6}$$

利用假设（2.4），我们有

$$\sup_{x\in\Theta} \int_{z\notin\mathcal{K}_n} v(x,z)\mu(dz) = \sup_{x\in\Theta} \mathbb{E}[v(x,Z)\mathbb{1}_{z\notin\mathcal{K}_n}] \le Ce^{-C\beta_n^2}.$$

利用假设（2.5），我们得到

$$\sup_{x\in\Theta} [(v(x,\cdot)\phi_n)]_{\mathrm{Lip}} \le Ce^{C\beta_n}.$$

接着，对 $Z_1, \ldots, Z_n$ 取期望，并注意到

$$\mathbb{E}\left[\sup_{x\in\Theta} \int_{z\notin\mathcal{K}_n} v(x,z)\mu_n(dz) \right] = \sup_{x\in\Theta} \int_{z\notin\mathcal{K}_n} v(x,z)\mu(dz),$$

利用引理 4.2和定理 4.1，我们得到

$$\mathbb{E}|V_n(X_n^\star) - V^\star| \le \mathbb{E} \sup_{x\in\Theta} |V_n(x) - V(x)| \le Ce^{C\beta_n}\mathbb{E}[\mathcal{W}_1(\mu_n,\mu)] + Ce^{-C\beta_n^2} \le Ce^{C\beta_n} n^{-1/q} + Ce^{-C\beta_n^2}.$$

现在我们选取 $\beta_n = D\sqrt{\log(n)}$，其中 $D > 0$，使得

$$\mathbb{E}|V_n(X_n^\star) - V^\star| \le Ce^{C'\sqrt{\log(n)}} n^{-1/q} + Cn^{-CD^2}.$$

取足够大的 $D > 0$ 可得

$$\mathbb{E}|V_n(X_n^\star) - V^\star| \le Ce^{C'\sqrt{\log(n)}} n^{-1/q}.$$

□

4.3 定理2.1(b) 的证明

证明。我们定义 $\Theta = \mathbf{B}(0, M_0) \subset \mathbb{R}^d$。由于 $\nabla^2 V(x^\star)$ 为正定的，存在 $r > 0$，使得对于所有 $x \in \mathbf{B}(x^\star, r)$，$\nabla^2 V(x)$ 均为正定的，并且我们定义

$$\lambda_{\min} := \min\left\{\lambda : \ \lambda \text{ eigenvalue of } \nabla^2 V(x),\ x \in \mathbf{B}(x^\star, r)\right\} > 0.$$

于是我们有

$$\begin{aligned}\mathbb{E}|X_n^\star - x^\star|^2 &= \mathbb{E}|(X_n^\star - x^\star)\mathbb{1}_{X_n^\star \notin \Theta}|^2 + \mathbb{E}|(X_n^\star - x^\star)\mathbb{1}_{X_n^\star \in \Theta \cap \mathbf{B}(x^\star, r)^c}|^2 + \mathbb{E}|(X_n^\star - x^\star)\mathbb{1}_{X_n^\star \in \Theta \cap \mathbf{B}(x^\star, r)}|^2 \\ &=: E_1 + E_2 + E_3.\end{aligned}$$

并利用（4.4），可得

$$E_1 \le (\mathbb{E}|X_n^2| + |x^\star|)\mathbb{P}(X_n^\star \notin \Theta) \le C/n$$

对于第二项，由于 $x^\star$ 是 V 的唯一极小化点，我们有 $\beta := \inf_{x \notin \mathbf{B}(x^\star, r)}(V(x) - V(x^\star)) > 0$，进而

$$\begin{aligned}E_2 &\le (M_0 + |x^\star|)^2\, \mathbb{P}(X_n^\star \notin \mathbf{B}(x^\star, r)) \le \frac{(M_0 + |x^\star|)^2}{\beta}\mathbb{E}[(V(X_n^\star) - V^\star)\mathbb{1}_{X_n^\star \notin \mathbf{B}(x^\star, r)}] \\ &\le C\mathbb{E}[V(X_n^\star) - V(x^\star)] = C\mathbb{E}[V_n(X_n^\star) - V(x^\star)] \le C a_n \mathcal{R}_q(n).\end{aligned}$$

对于第三项，由泰勒展开可知，存在某个 $\xi_n \in (X_n^\star, x^\star)$，使得 t

$$V(X_n^\star) = V(x^\star) + \frac{1}{2}\nabla^2 V(\xi_n) \cdot (X_n^\star - x^\star)^{\otimes 2}.$$

因此

$$\begin{aligned}\nabla^2 V(\xi_n) \cdot (X_n^\star - x^\star)^{\otimes 2} &= 2(V(X_n^\star) - V(x^\star)) \le 2(V(X_n^\star) - V_n(X_n^\star) + V_n(x^\star) - V(x^\star)) \\ &\le 4 \sup_{x \in \Theta} |V_n(x) - V(x)|\end{aligned}$$

然后

$$E_3 \le \mathbb{E}\left[\nabla^2 V(\xi_n) \cdot (X_n^\star - x^\star)^{\otimes 2}\right] \le \frac{4}{\lambda_{\min}}\mathbb{E}\sup_{x \in \Theta}|V_n(x) - V(x)| \le C a_n \mathcal{R}_q(n).$$

□

参考文献

[BDL08] Gérard Biau、Luc Devroye 和 Gábor Lugosi。希尔伯特空间中的聚类性能。*IEEE* 信息论汇刊，54(2):781–790，2008。[BM06] Peter L. Bartlett 和 Shahar Mendelson。经验最小化。概率论及相关领域，135(3):311–334，2006。[BMP10] Peter L. Bartlett、Shahar Mendelson 和 Petra Philips。关于经验最小化解的期望的基于样本的估计的最优性。*ESAIM* 概率与统计，14:315–337，2010。[FG15] Nicolas Fournier 和 Arnaud Guillin。经验测度在 Wasserstein 距离下的收敛速率。概率论及相关领域，162(3-4):707–738，2015。

[LP10] Vincent Lemaire 和 Gilles Pagès。无约束递归重要性抽样。《应用概率年刊》，20(3):1029–1067，2010。[LP20] Yating Liu 和 Gilles Pagès。最优量化的收敛速率及其在经验测度聚类性能中的应用。《机器学习研究杂志》，21:论文编号 86，36，2020。[VA71] V.N. Vapnik 和 A.Ya.Chervonenkis。事件的相对频率一致收敛于其概率。《概率论及其应用》，16(2):264–280，1971。[Vil09] Cédric Villani。最优传输：古老与现代。《数学科学基础丛书》。施普普林格，柏林，2009。