arXiv:2301.00631v2 [cs.LG] 8 Mar 2023

用于非凸复合优化的方差缩减随机变度量近端梯度法

Gersende Fort[1*]和 Eric Moulines[2]

[1*]图卢兹数学研究所，法国国家科学研究中心与图卢兹大学，法国图卢兹纳博讷路 118 号，邮编 31400。[2]CMAP，巴黎综合理工学院，萨克雷路，帕莱索，邮编 91128 Cedex，法国。

通讯作者：电子邮箱：gersende.fort@math.univ-toulouse.fr；贡献作者：电子邮箱：eric.moulines@polytechnique.edu；

摘要

本文提出了一种新颖的算法——扰动近端预条件 SPIDER 算法（3P-SPIDER），旨在解决有限和非凸复合优化问题。该算法是一种随机可变度量前向-后向算法，允许对预条件前向算子进行近似，并使用变度量邻近算子作为后向算子；此外，它还提出了一种带方差缩减的小批量策略，以应对有限和情形。我们证明，3P-SPIDER 将一些基于随机预条件梯度下降的算法和一些增量期望最大化算法推广到了复合优化问题，以及前向算子无法以闭式计算的情形。我们还明确控制了 3P-SPIDER 的期望意义下的收敛，并研究了满足近似 ε-驻点条件所需的复杂度。我们的结果首次将非凸复合优化情形、利用小批量策略处理有限和情形的方差缩减技术，以及允许对预条件前向算子进行确定性或随机近似这三者结合起来。最后，我们将该方法应用于含随机效应的逻辑回归模型推断，通过数值实验将 3P-SPIDER 与其他随机前向-后向算法进行比较，并讨论了 3P-SPIDER 若干设计参数的作用。

关键词：随机优化、变度量前向-后向分裂、预条件随机梯度下降、增量式期望最大化、近端方法、方差缩减、非渐近收敛界。

1 引言

从大数据集中高效学习，需要专为适应大数据和复杂模型时代而设计的新型优化算法。在统计学和机器学习中，我们经常需要求解

如下形式的问题

$$\operatorname{argmin}_{s\in\mathbb{R}^q}\left(\frac{1}{n}\sum_{i=1}^{n}W_i(s)+g(s)\right),$$

其中，n 是训练数据集中的样本数量，s 是需要从样本中学习的未知量，W_i 是与样本#i 相关的损失函数，而 g 是正则化

编码了关于 s 的先验知识和约束的项； g 也可以防止过拟合。正则化项 $g: \mathbb{R}^q \to (0, +\infty]$ 通常不可微，而数据保真项$n^{-1}\sum_{i=1}^n W_i$ 在 g 的定义域上是光滑的。

本文研究非凸有限和复合函数的随机优化；更准确地说，本文探讨微分包含问题

$$0 \in \frac{1}{n}\sum_{i=1}^{n} G_i(s) + \partial g(s), \qquad s \in \mathbb{R}^q, \qquad (1)$$

其中 $g: \mathbb{R}^q \to (-\infty, +\infty]$ 是下半连续凸函数，其定义域 $\mathcal{S}$ 非空，并且对于所有 i， $G_i: \mathcal{S} \to \mathbb{R}^q$ 在 $\mathcal{S}$ 上全局利普希茨连续。

本文的首要目标是提出一种新颖的算法。受统计学和机器学习中的应用启发，我们要求该算法满足以下三项约束。(c1) 该算法使用可能经过预条件处理的算子 h_i ，而非前向算子 G_i:

$$\forall s \in \mathcal{S}, \qquad \mathsf{h}_i(s, B) \stackrel{\text{def}}{=} -B^{-1} G_i(s), \qquad (2)$$

其中 B 是一个 $q \times q$ 正定矩阵。此类条件涵盖了例如预条件梯度法，其中也包括采用自适应步长的梯度方法。它还涵盖了为大规模学习设计的期望最大化（EM）算法（Dempster等人(1977)）。(c2) 算法可能只能访问 $\mathsf{h}_i(s, B)$ 的近似值。此类条件适用于 $\mathsf{h}_i(s, B)$ 未显式给出的情况，例如它由难以处理的积分定义时。每个 EM 的 E 步都会出现这种情况，此时无法精确计算后验分布下的条件期望。(c3) 该算法应对有限和挑战，同时尽可能减小算法解所引起的变异性。例如，当求解依赖于随机选择一个由样本构成的小批量时，算法必须提出一种方差缩减方案。

第一类形式为 (1) 的问题，是通过基于梯度的算法最小化正则化损失函数。在这种情况下，$G_i \stackrel{\text{def}}{=} \nabla W_i$. (2) 允许使用预条件梯度；已知这种变度量能够加速收敛。变度量前向-后向算法（VMFB）旨在求解满足 1)-(2) 的问题，其中 G是梯度。然而，正如第 2.1 节所讨论的，据我们所知，VMFB 的各种变体都无法同时满足三个约束 c1、c2 和c3 。

（1）-（2）的另一个应用是 EM 算法。该算法最初旨在计算潜变量模型中未知参数 θ 的最大似然估计量。当完整数据模型属于曲指数族时，EM 等价于所谓的统计量空间中的一种算法（例如参见Delyon 等人（1999））。这一发现是随机 EM 算法（包括为大数据集增量处理而设计的增量 EM 算法）的基石。EM 仅提供预条件前向算子 $-\mathsf{B}(s)^{-1}G_i(s)$。因此，随机 EM 算法自然适用于情形（1）-（2）（参见Fort 等人（2020））。同样，如第2.2节所述，文献中的 EM 变体都没有同时满足约束(c2) 和(c3) 。

我们的首项贡献是设计了一种新颖的迭代算法，称为扰动近端预条件 SPIDER（3P-SPIDER）。该算法结合了：(i) 与光滑部分 $n^{-1}\sum_{i=1}^n G_i(s)$ 相关联的预条件前向算子；(ii) 针对非光滑部分 g 的变度量邻近算子；(iii) 一种随机方法，用于处理由 $n^{-1}\sum_{i=1}^n G_i(s)$ 引起的有限和情形，并结合基于 SPIDER 算法（Nguyen 等人（2017）；Fang 等人（2018））的方差缩减技术；此外，它还允许在预条件前向算子没有闭式时进行 (iv) 数值近似。该算法将在第 3 节介绍，并讨论实现问题。我们还设计了一种随机 VMFB 算法，它满足约束 c1 和 c2，但不包含 c3 所要求的方差缩减步骤。

第二项贡献是针对 3P-SPIDER 提供非渐近的期望收敛性分析，其中变度量 B在第 # 次迭代t 依赖于当前迭代点的取值，并且 G_i是梯度算子；参见第 4节。证明依赖于一个采用创新构造的 Lyapunov不等式，这

是非凸优化问题设定以及算法使用预条件化前向算子和变度量邻近算子的结果（参见第 7.4节）。

定理4.1 给出了 3P-SPIDER期望意义下收敛的控制界，明确揭示了非精确预条件前向算子以及初始化策略的影响。首先，我们证明，当预条件前向算子是精确的或由无偏随机预言机替代时，3P-SPIDER的学习率可以在各次迭代中保持恒定；当其由有偏预言机（确定性或随机的）替代时，学习率则随迭代逐渐减小。其次，我们首次给出了针对非凸有限和复合优化的随机 VMFB 算法的收敛结果，并处理了c1、c2 和 c3 。例如，这是针对带有非光滑惩罚项（$g \neq 0$）且 E 步的蒙特卡罗近似可能有偏的增量 EM 的首个结果。

当前向算子 $\mathsf{h}_i(s,\mathsf{B}(s))$ 精确时，我们研究了 3P-SPIDER 的复杂度（见推论 4.3）：为了满足近似ϵ-平稳条件，对算子 $\mathsf{h}_i(s,\mathsf{B}(s))$ 之一的调用次数为 $O(\sqrt{n}/\epsilon)$，对后向算子的调用次数为$O(1/\epsilon)$，并且学习率可以选取得与准确率 ϵ 无关。将这些明确的收敛控制界应用于梯度法，以及在无约束条件（$g = 0$）下应用于 EM 方法，可以得到文献中已有的结果（例如，Wang 等人（2019）；Fort 等人（2020））， 这些结果被认为达到了当前最先进水平。

最后，我们证明，当对前向算子进行近似时，这一复杂度分析仍然成立。在较为困难的情形下，近似是基于蒙特卡罗和的有偏随机预言机；我们证明，通过在蒙特卡罗和中选取足够多的项， $\mathcal{K}_{\bar{\mathsf{h}}}$ 和 $\mathcal{K}_{\mathrm{prox}}$不受近似影响，与使用精确算子 $\mathsf{h}_i(\cdot,\mathsf{B}(\cdot))$时相同。为此付出的代价是蒙特卡罗复杂度达到 $O(\sqrt{n}/\epsilon^2)$ 阶。

—

在第5节中，3P-SPIDER 被应用于带随机效应的逻辑回归模型中的推断。我们展示了该问题如何具有（1）-（2）的形式。在此示例中，预条件前向算子通过马尔可夫链蒙特卡洛采样器计算得到的蒙特卡罗和进行近似。通过在 3P-SPIDER 作为统计量空间中的随机期望最大化算法这一情形下进行数值分析，我们讨论了设计参数的选择。我们还展示了如何通过利用前向算子的蒙特卡洛近似来增强 SPIDER 方差缩减效应。

证明见第 6 节和第 7节；附录中也提供了技术细节。

记号。 我们用 $\langle\cdot,\cdot\rangle$ 表示 $\mathbb{R}^q$上的点积，用 $\|\cdot\|$ 表示相应的范数。对于一个 $q \times q$ 正定矩阵 B，令$\langle\cdot,\cdot\rangle_B \overset{\text{def}}{=} \langle B\cdot,\cdot\rangle$，并以 $\|\cdot\|_B$ 表示相应的范数。I_q 表示单位矩阵。对于矩阵B，$B^\top$ 表示其转置。 $\mathcal{P}_+^q$ 表示$q \times q$ 正定矩阵的集合。

$\mathbb{N}$ （分别）是非负（分别为正）整数的集合。对于 $n \in \mathbb{N}^\star$，我们令 $[n]$ 定义为$=\{0,\cdots,n\}$，并令 $[n]^\star$ 定义为$=$ $\{1,\cdots,n\}$。 $\mathbb{R}_+$是正实数的集合。对于 $q \in \mathbb{R}$， $\lceil q \rceil$ 表示向上取整。

I 是恒等函数。对于适当函数g ： $\mathbb{R}^q \to (-\infty, +\infty]$， $\partial g(s)$ 表示在 s 处的次微分。对于在 $s \in \mathbb{R}^q$ 处连续可微的函数 W， $\nabla W(s)$ 是 W 在 s处的梯度。

所有随机变量（r.v.）都定义在 $(\Omega\ \mathcal{A},\ \mathbb{P})$ 上；对于随机变量 U， $\sigma(U)$ 表示由 U 生成的 σ-代数。

2 动机示例

本节将说明，变度量近端梯度算法和期望最大化算法都是式（1）和式（2）所描述的一般框架的样本。预条件矩阵 B 在第一种情况下由用户选择，而在第二种情况下则由算法提供。

2.1 变度量近端梯度算法

考虑具有有限和结构的非凸复合问题

$$\text{find } s \in \mathbb{R}^q: \quad 0 \in \frac{1}{n}\sum_{i=1}^{n} \nabla W_i(s) + \partial g(s)\,, \quad (3)$$

其中， g 是定义域为 $\mathcal{S}$的正常下半连续凸函数；对于所有 $i \in [n]^\star$， W_i 在 $\mathcal{S}$上连续可微。它具有形式 (1)，其中 G_i def= ∇W_i 是一个梯度。(3) 是更一般问题的一个例子：在 $\mathbb{R}^q$ 上求两个（集合值）算子 $0 \in \mathrm{A}s + \mathrm{C}s$之和的零点。这里，A def $= n^{-1}\sum_{i=1}^{n} \nabla W_i$和 C def $= \partial g$ 是极大单调算子（参见 Bauschke 和 Combettes, 2011，定理 20.25）。(3) 可以通过前向-后向分裂算法求解（例如参见 Combettes 和 Wajs (2005)；Beck (2017)）：前向步骤在每次迭代中使用部分或全部函数W_i的梯度；后向步骤使用与 g 相关联的邻近算子。这就得到了基于近端梯度的算法。

在情形 $g = 0$ 下（其中包括无约束优化问题），当

$$n^{-1}\sum_{i=1}^{n} G_i(s) = \mathbb{E}\left[\mathcal{G}(Z, s)\right] \tag{4}$$

且可使用随机预言机 $\mathcal{G}(Z, s)$ 时，有人提出了带方差缩减的随机梯度方法；在非凸情形下，例如可参见 Ghadimi 和 Lan（2013）；Reddi 等人（2016）；Allen-Zhu 和 Hazan（2016）；Nguyen 等人（2017）；Allen-Zhu（2018）；Fang 等人（2018）；Zhou 等人（2020）。这些算法通过令 B 等于单位矩阵 I_q，来求解问题 (1)-(2)，并使用无偏随机预言机 $\mathcal{G}(Z, s)$ 近似前向算子。

对于非凸复合优化（$g \neq 0$），我们可以引用 Ghadimi 等人（2016）和 Karimi 等人（2016）的研究，他们提出了使用无偏预言机 $\mathcal{G}(Z, s)$（见（4））的随机近端梯度算法。Li 和 Li（2018）、Wang 等人（2019）、Zhang 和 Xiao（2019）、Nhan 等人（2020）以及 Metel 和 Takeda（2021）提出了采用无偏随机预言机并包含方差缩减方案的随机近端梯度方法。在 Metel 和 Takeda（2021）的研究中， g 可以是非凸的，但其邻近算子可高效计算。Atchad´e 等人（2017）允许对前向算子 $n^{-1}\sum_{i=1}^{n} G_i(s)$进行确定性或随机近似；当扰动是随机的时，收敛性分析涵盖有偏和无偏预言机，并包括 Nesterov 加速方案，但仅适用于凸优化。再次，所有这些算法都通过选择1）-（2）来处理问题（ $B = \mathrm{I}_q$。

前向-后向算法存在收敛速度慢的问题。Chen 和 Rockafellar（1997）提出了变度量前向-后向算法（VMFB），以加快收敛（另见 Chouzenoux 等（2014）中的参考文献 11 至 16）。VMFB 在每次迭代中通过使用乘于前向算子的对称正定缩放矩阵来改变度量。它是惯性方法的一种替代方案，例如重球法或 Nesterov 加速法；后者会利用先前迭代点的信息。求解包含关系 (3) 时，VMFB 在第 #t 次迭代的前向步中使用迭代相关的预条件矩阵 B_t^{-1}对梯度进行预条件处理，并相应地缩放近端步。Chouzenoux 等（2014）和 Repetti 等（2014）给出了示例，表明 VMFB 比前向-后向算法以及带惯性方案的前向-后向算法更高效。变度量B_t的定义有多种策略；例如，它可以是依赖于算法过去历史的对角矩阵（关于变标量度量，参见 Park 等（2019）及其中的参考文献；在情形 $g = 0$ 下，另见 Chen 等（2019）），也可以继承自牛顿型方法（复合凸情形下，参见 Becker 和 Fadili（2012）；Lee 等（2014）；Becker 等（2019）；有限和结构的光滑凸情形下，参见 Kolte 等（2015）；Moritz 等（2016）；Gower 等（2016）；最后，有限和结构的光滑非凸情形下，参见 Zhang 等（2022）），或者通过主化-最小化策略定义，使后向算子显式化（例如参见 Chouzenoux 等（2014）和 Repetti 与 Wiaux（2021））。VMFB 在凸情形下（例如参见 Combettes 和 V˜u（2014）及 Bonettini 等（2021）；强凸情形下另见 Park 等（2019））和非凸情形下（例如参见 Chouzenoux 等（2014）和 Repetti 与 Wiaux（2021））均有收敛性结果。Yun 等（2021）研究了非凸情形下的随机 VMFB；精确梯度由随机预言机的线性组合近似，并采用指数遗忘，且假设这些预言机无偏且有界。

3P-SPIDER 通过使用近端梯度算法来处理具有有限和结构的非凸复合优化，并借助……加速

变度量技巧。这是一种随机的VMFB，其中包含方差缩减技术，以应对有限和情形；它还允许预条件前向算子的预言机，而这些预言机在随机情形下可以是有偏或无偏的（例如，参见Atchad´等人(2017）和 Fort 等人 (2018），其中的样本为梯度的有偏随机近似提供了动机）。据我们所知，这两种扰动源的组合是一种全新的设定，现有文献尚未对此进行研究。

3P-SPIDER 使用预条件矩阵，这些矩阵可能取决于迭代点的当前值，因此也可能是随机的。第 4 节推导的非渐近收敛性分析将基于对缩放矩阵谱的较弱小化假设（见 A3），不同于Yun 等人(2021年)；它不要求像 Combettes 和 V˜u (2014) 以及 Bonettini 等人(2021年) 那样对缩放矩阵序列作序假设，也不要求像 Chouzenoux 等人(2014) 那样对目标函数作 Kurdyka-Lojasiewicz 条件。因此，3P-SPIDER 的收敛性分析中 Lyapunov 不等式的构造与这些先前研究不同。

3P-SPIDER 要求后向算子是显式的，这可能是一个很强的假设，尤其对于变度量邻近算子而言（见第 3.5节）；将收敛性分析扩展到不精确邻近算子的情形不在本文讨论范围内。

2.2 曲指数族的期望最大化

考虑如下参数统计模型：观测值相互独立，且其密度为

$$y \mapsto \int_{\mathsf{Z}} p(y, z;\, \theta) \mu_{lv}(\mathrm{d}z)\,,$$

关于（w.r.t.）一个 σ-有限正测度μ_o 上的 $\mathbb{R}^{d_y}$。在该模型中， z 充当一个 潜在的 取值于可测集（Z $\mathcal{Z}$）中的变量，该集合赋有一个 σ-有限正测度 μ_{lv} （潜变量模型的样本示例见 Everitt(1984）。目标是根据 $\theta \in \Theta \subseteq \mathbb{R}^d$ 学习参数 n 观测值 Y_1、$\cdots$、Y_n，方法是

最小化负归一化对数似然

$$F(\theta) \stackrel{\text{def}}{=} -\frac{1}{n}\sum_{i=1}^{n} \log \int_{\mathsf{Z}} p(Y_i, z;\, \theta) \mu_{lv}(\mathrm{d}z) \qquad (5)$$

在 Θ。遗憾的是，这是一个非凸问题，通常针对（5）进行最小化的优化算法，至多只能收敛到目标函数的一个临界点（参见 Wu(1983）。

一种常见模型是完全数据似然 $(y, z) \mapsto p(y, z;\ \theta)$ 具有如下形式的情形

$$p(y, z;\, \theta) \stackrel{\text{def}}{=} H(y, z)\, \exp\left(\langle S(y, z), \phi(\theta)\rangle - \psi(\theta)\right)$$

其中 $H:\ \mathbb{R}^{d_y} \times \mathsf{Z} \to \mathbb{R}_+$, $S:\ \mathbb{R}^{d_y} \times \mathsf{Z} \to \mathbb{R}^q$, $\phi:\ \Theta \to \mathbb{R}^q$, $\psi:\ \Theta \to \mathbb{R}$；它对应于所谓的曲指数族假设。只要混合模型中的各个成分来自曲指数族，该假设便成立。有关曲指数分布族的介绍，参见Brown (1986)；有关此类潜变量模型的样本，参见McLachlan andKrishnan(2008)。

曲指数族的期望最大化迭代定义了一个取值于 Θ 的序列 $\{\theta_t, t \geq 0\}$，其机制如下：给定 θ_t，

•(E步) 计算 $\bar{s}(\theta_t)$，即关于后验分布的充分统计量的期望

$$\bar{s}(\theta_t) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \bar{s}_i(\theta_t)\,,$$
$$\bar{s}_i(\theta_t) \stackrel{\text{def}}{=} \int_{\mathsf{Z}} S(Y_i, z) \frac{p(Y_i, z;\, \theta_t)\, \mu_{lv}(\mathrm{d}z)}{\int_{\mathsf{Z}} p(Y_i, u;\, \theta_t) \mu_{lv}(\mathrm{d}u)}\,.$$

• (M步) 更新参数

$$\theta_{t+1} \stackrel{\text{def}}{=} \operatorname{argmin}_{\theta\in\Theta} \left(\psi(\theta) - \langle \bar{s}(\theta_t), \phi(\theta)\rangle\right)\,.$$

该算法交替执行参数空间中的一步 Θ （计算 $\theta_{t+1} \in \mathbb{R}^d$时），以及统计量空间中的一步，在计算时$\bar{s}$（$\theta_t$） $\in \mathbb{R}^q$。命题2.1 指出，在参数空间中运行的期望最大化算法的极限点 Θ是映射到 Θ上的算子的固定点；求出这样的固定点等价于求出映射到统计量空间 $\bar{s}$（Θ） $\subseteq \mathbb{R}^q$上的算子的固定点。

命题 2.1 *Assume that for any* $s \in \mathcal{S} \supseteq \bar{s}(\Theta)$,

$$\mathsf{T}(s) \stackrel{\text{def}}{=} \operatorname{argmin}_{\theta\in\Theta} \left(\psi(\theta) - \langle s, \phi(\theta)\rangle\right)$$

exists and is unique. Set $\mathcal{L}_\theta$ 定义= $\{\theta \in \Theta : \mathsf{T}(\bar{s}(\theta)) = \theta\}$*and* $\mathcal{L}_s$ 定义= $\{s \in \bar{s}(\Theta) : \bar{s}(\mathsf{T}(s)) = s\}$.

$\mathcal{L}_\theta$ *is the set of the limiting points of EM. If* $\theta_\star \in \mathcal{L}_\theta$, *then* $s_\star$ 定义= $\bar{s}(\theta_\star)$ *is in* $\mathcal{L}_s$. *Conversely, if* $s_\star \in \mathcal{L}_s$*then* $\theta_\star \stackrel{\text{def}}{=} \mathsf{T}(s_\star)$ *is in* $\mathcal{L}_\theta$.

例如参见Delyon 等人(1999) 关于证明。命题 2.1 的一个算法推论是，期望最大化等价于任何在统计量空间中运行并旨在求出函数根的算法

$$s \mapsto \frac{1}{n}\sum_{i=1}^{n} \bar{\mathsf{h}}_i(s)\,, \quad \text{where} \quad \bar{\mathsf{h}}_i(s) \stackrel{\text{def}}{=} \bar{s}_i(\mathsf{T}(s)) - s\,,$$

在子集 $\bar{s}(\Theta)$ 上，属于 $\mathbb{R}^q$。在统计模型满足正则性条件时，Delyon 等人 (1999)（另见（Fort 等人，2020，命题 1）中的陈述）证明了存在一个 $q\times q$正定矩阵$\mathsf{B}(s)$，使得

$$\nabla\left(F \circ \mathsf{T}\right)(s) = -\mathsf{B}(s)\left(\frac{1}{n}\sum_{i=1}^{n} \bar{\mathsf{h}}_i(s)\right)\,, \qquad (6)$$

其中 F 是负归一化对数似然（见（5））。因此，$n^{-1}\sum_{i=1}^{n} \bar{\mathsf{h}}_i(s)$ 在 $\bar{s}(\Theta)$ 上的根，就是 $n^{-1}\sum_{i=1}^{n} G_i(s)$ 在 $\bar{s}(\Theta)$ 上的根，其中 $G_i(s)$ 定义为 $= -\mathsf{B}(s)\,\bar{\mathsf{h}}_i(s)$。这也意味着，$n^{-1}\sum_{i=1}^{n} \bar{\mathsf{h}}_i(s)$ 的根就是 $F\circ\mathsf{T}$ 的梯度的根；目标函数通过映射 $\mathsf{T}: \Theta \to \mathbb{R}^q$ 转换到统计量空间上。

总之，统计量空间中的 EM 是问题 (1)-(2) 的一个例子，其中函数 g 汇集了关于 s 的约束，例如 $s \in \mathcal{S} \supseteq \bar{s}(\Theta)$：(i) 其旨在求解满足约束 $s \in \mathcal{S}$ 的方程 $n^{-1}\sum_{i=1}^{n} G_i(s) = \nabla(F \circ \mathsf{T})(s)$；(ii) 它使用量 $\mathsf{h}_i(s)$，该量是预条件前向算子，因为存在 $B(s)$ 使得 $\bar{\mathsf{h}}_i(s) = -\mathsf{B}(s)^{-1} G_i(s)$（见 (6)）；(iii) 该预条件前向算子至少有两个原因导致难以计算：首先，计算 $\bar{s}_i$ 时需要对集合 Z 进行内层积分，而 $p(y, z; \theta)$ 和 Z 通常过于复杂，无法显式求出积分；其次，计算 n 样本上的时需要进行外层积分，这在大规模学习中计算成本高得令人难以承受。不过，无论 i 和 s 取何值，始终可以对 $\mathsf{h}_i(s)$ 进行蒙特卡洛近似。理解随机 EM 版本的关键就在于此（例如参见 Celeux 和 Diebolt (1985)；Wei 和 Tanner (1990)；Delyon 等人 (1999)；Fort 和 Moulines (2003)，这些工作针对内层求和难以计算的问题；以及 Neal 和 Hinton (1998)；Ng 和 McLachlan (2003)；Capp´e 和 Moulines (2009)；Chen等人 (2018)；Karimi 等人 (2019)；Fort 等人 (2020, 2021a)，这些工作针对外层求和难以计算的问题）。其做法是运行一个平均场为 $n^{-1}\sum_{i=1}^{n} \bar{\mathsf{h}}_i(s)$ 的随机逼近（SA）算法（关于 SA 的介绍，例如参见 Benveniste 等人 (1990) 或 Borkar (2008)）；由此得到 EM 过程中的 SA。它们的区别在于用于近似平均场的随机场的构造方式不同（关于 EM 算法中的一些 SA 的说明，见 Fort 等人 (2021a, 2.2节)）。

3P-SPIDER 是期望最大化算法中的随机逼近算法之一。与以往的随机EM方法相比，它包含两种随机近似（分别针对 i 中的求和以及 Z 上的积分）和一个方差缩减步骤，并且允许采用比集合的$\{0, +\infty\}$值示性函数更一般的惩罚项 g 。

3 3P-SPIDER 算法

我们提出一种名为扰动近端预条件 *SPIDER*（3P-SPIDER）的新型算法，用于求解(1)，并满足(c1)、(c2)和 g 以下假设

一个**1.** g： $\mathbb{R}^q \to (-\infty, +\infty]$ 是正常、下半连续且凸的。记其定义域为 $\mathcal{S}$ $\mathcal{S}$ $\stackrel{\text{def}}{=}$ $\{s \in \mathbb{R}^q : g(s) < +\infty\}$。

在此假设下，我们定义一个变度量邻近算子。对于任意 $\gamma > 0$和$B \in \mathcal{P}_+^q$，正常下半连续凸函数 $\gamma g : \mathbb{R}^q \to (-\infty, +\infty]$ 相对于由B 所诱导的度量的邻近算子定义如下（参见例如（Hiriart-Urruty andLemar´echal，1996，第XV.4节））

$$\operatorname{prox}_{\gamma g}^{B}(s) \stackrel{\text{def}}{=} \operatorname{argmin}_{\mathbb{R}^q}\left(\gamma g(\cdot) + \frac{1}{2}\|\cdot - s\|_B^2\right). \quad (7)$$

当 $B = \mathrm{I}_q$ 时，我们简单地记作 $\mathrm{prox}_{\gamma g}(s)$，这就是莫罗(1965）最初定义的邻近算子。引理3.1表明，在A1条件下，$\mathrm{prox}^B_{\gamma g}(s)$ 对所有 $s \in \mathbb{R}^q$、$\gamma > 0$ 和 $B \in \mathcal{P}^q_+$ 均存在且唯一，并给出了该点的刻画。证明见第6.1节。

引理 3.1 *Assume*A*1*。

1. 对任意 $\gamma > 0$，$B \in \mathcal{P}^q_+$ 和 $s \in \mathbb{R}^q$，优化问题 (7)有唯一解，其特征为满足条件的唯一点 $\mathsf{p} \in \mathcal{S}$满足

$$-\gamma^{-1} B\,(\mathsf{p} - s) \in \partial g(\mathsf{p})\,.$$

2. 对任意 $\gamma > 0$、$B \in \mathcal{P}^q_+$、$s \in \mathcal{S}$ 和 $h \in \mathbb{R}^q$，

$$s = \mathrm{prox}^B_{\gamma g}(s + \gamma h) \quad \textit{iff} \quad Bh \in \partial g(s). \tag{8}$$

对于 $s \in \mathbb{R}^q$ 和 $B \in \mathcal{P}^q_+$，令

$$\mathsf{h}_i(s, B) \stackrel{\mathrm{def}}{=} -B^{-1} G_i(s)\,,$$
$$\mathsf{h}(s, B) \stackrel{\mathrm{def}}{=} n^{-1} \sum_{i=1}^n \mathsf{h}_i(s, B)\,. \tag{9}$$

根据引理 3.1-第 2 项，对于任意 $B \in \mathcal{P}^q_+$ 和 $\gamma > 0$ $s =$ $^B_{\gamma g}$ $s + \gamma$ s B $-B$ s B $\in$：近端算子 ∂g s 9 $B \in \mathcal{P}^q_+$ (h(,)) 当且仅当 h(,) ()。由 ()，可得对于任意、和$\gamma > 0$:

$$s_\star = \mathrm{prox}^B_{\gamma\, g}\,(s_\star + \gamma\ \mathsf{h}(s_\star, B)) \quad \text{iff} \quad s_\star \ \text{solves}\ (1). \tag{10}$$

3.1 变度量近端与预条件梯度

(10) 表明，在求解复合优化问题（1）时，只要使用算子 $s \mapsto n^{-1} \sum_{i=1}^n G_i(s)$ 的预条件版本（其预条件矩阵为 B^{-1}），就必须使用 g 相对于由矩阵 B 诱导的度量的近端算子。

基于刻画（10），在条件1和条件 (c1) 下求解（$\sigma_0 \in \mathcal{S}$，给定一个正的步长序列 $\{\gamma_{k+1}, k \geq 0\}$ 以及一个 $\mathcal{P}^q_+$ 值序列 $\{B_{k+1}, k \geq 0\}$，重复

$$\sigma_{k+1} = \mathrm{prox}^{B_{k+1}}_{\gamma_{k+1} g}\,(\sigma_k + \gamma_{k+1} \mathsf{h}(\sigma_k, B_{k+1}))\,. \tag{11}$$

这对应于变度量前向-后向算法（例如参见 Chen 和 Rockafel- lar(1997）； Combettes 和 V˜u (2014）。

在大规模学习的情形下，每次迭代（11）时都无法计算函数 h_i（见（9））的完整求和。此外，$\mathsf{h}_i(s)$ 可能无法显式表示（例如，增量 EM 算法中的情形，见 2.2节）。因此，一个自然的想法是提出由算法 1 定义的（11）的非精确版本：近端步骤保持不变（见第 8 行）；第 7 行中的随机逼近步使用精确平均场 $n^{-1} \sum_{i=1}^n \mathsf{h}_i(\widehat{S}_k)$ 的随机近似值 S_{k+1}；该近似值由第 6 行定义，结合了完整求和的小批量近似（见第 3 行）以及可能经过近似的项$\delta_{k+1,i}$（见第 5 行）。

算法 1 A 随机的 变量 度量前向-后向算法

要求： $k^{\mathrm{out}} \in \mathbb{N}^\star$，$\gamma_k > 0$ 对 $k \in [k^{\mathrm{out}}]^\star$，$\mathsf{b} \in \mathbb{N}^\star$，$\widehat{S}_{\mathrm{init}} \in \mathcal{S}$。

确保： 序列 $\{$ $\widehat{S}_k, k \in [k^{\mathrm{out}}]\}$。

1: $\widehat{S}_0 = \widehat{S}_{\mathrm{init}}$
2: **对** $k = 0, \ldots, k^{\mathrm{out}} - 1$ **执行**
3: 抽样一个大小为 b 的批次 $\mathcal{B}_{k+1}$，在 $[n]^\star$
4: 选择 $B_{k+1} \in \mathcal{P}^q_+$
5: 对于 $i \in \mathcal{B}_{k+1}$，计算一个近似值 $\delta_{k+1,\,i}$ $\mathsf{h}_i(\widehat{S}_k,\ B_{k+1})$。
6: $\mathsf{S}_{k+1} = \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{k+1}} \delta_{k+1,\,i}$
7: $\widehat{S}_{k+1/2} = \widehat{S}_k + \gamma_{k+1}\, \mathsf{S}_{k+1}$
8: $\widehat{S}_{k+1} =$ 近端算子$^{B_{k+1}}_{\gamma_{k+1} g}(\widehat{S}_{k+1/2})$。
9: **结束循环**

3.2 SPIDER 方差缩减技术

3P-SPIDER 利用算法 1 以及方差缩减技术 SPIDER来定义域S_{k+1}，其近似于 $n^{-1} \sum_{i=1}^n \mathsf{h}_i(\widehat{S}_k,\ B_{k+1})$。SPIDER 代表随机路径积分微分估计器，最初由Fang 等人 (2018) 在随机梯度下降文献中提出（另见 Nguyen 等人 (2017)；Wang 等人 (2019）。

我们介绍 SA 随机逼近设定下 SPIDER 的直觉，该设定涵盖随机梯度设定。

SA 方案求解一个求根问题 $\xi(s)=0$，其中 $\mathbb{R}^q$ 满足：给定初始值 $s_0 \in \mathbb{R}^q$ 和步长序列 $\{\gamma_{k+1}, k \geq 0\}$，重复执行 $s_{k+1}=s_k+\gamma_{k+1}\Xi_{k+1}$；在每次迭代中，#$(k+1)$，$\Xi_{k+1}$ 是 $\xi(s_k)$的一个随机近似值。通常要求，给定算法过去的信息，Ξ_{k+1} 的期望为 $\xi(s_k)$；在这种情况下，可以将 Ξ_{k+1} 替换为 $\mathsf{S}_{k+1} \overset{\text{定义}}{=} \Xi_{k+1}+V_{k+1}$，其中给定过去的信息时，$V_{k+1}$ 是中心化的。SPIDER 利用了这一观察以及控制变量技术：它提出了一种巧妙的构造，构造出一个与 Ξ_{k+1}相关、近似于零的随机变量 V_{k+1}。

具体做法如下：假设在第 #k 次迭代中，S_k 是 $\mathsf{h}(\widehat{S}_{k-1}, B_k)$ 的随机近似值。然后定义 S_{k+1}，使其满足 $\mathsf{S}_{k+1} \overset{\text{def}}{=} H_{k+1}+V_{k+1}$，其中

$$H_{k+1} \overset{\text{def}}{=} \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{k+1}} \mathsf{h}_i(\widehat{S}_k, B_{k+1}),$$

$$V_{k+1} \overset{\text{def}}{=} \mathsf{S}_k - \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{k+1}} \mathsf{h}_i(\widehat{S}_{k-1}, B_k),$$

并且 $\mathcal{B}_{k+1}$ 是从 $[n]^\star$ 中随机抽样得到的。随机变量 V_{k+1} 近似于零，因为 $\mathsf{b}^{-1}\sum_{i\in\mathcal{B}_{k+1}}\mathsf{h}_i(\widehat{S}_{k-1}, B_k)$ 和 S_k 都近似于 $n^{-1}\sum_{i=1}^n \mathsf{h}_i(\widehat{S}_{k-1}, B_k)$；$V_{k+1}$ 和 H_{k+1} 通过 $\mathcal{B}_{k+1}$ 相关联。

遗憾的是，随机变量 S_{k+1} 并不是 $n^{-1}\sum_{i=1}^n \mathsf{h}_i(\widehat{S}_k, B_{k+1})$ 的无偏近似值（参见命题 7.3），当 B_{k+1} 具有$\widehat{\mathsf{B}}(S_k$的形式时）。为了消除偏差，SPIDER会定期重启控制变量机制：每隔 k^{in} 次迭代，对 n 项进行完整求和，并令 $\mathsf{S}_{k^{\text{in}}+1} = n^{-1}\sum_{i=1}^n \mathsf{h}_i(\widehat{S}_{k^{\text{in}}}, B_{k^{\text{in}}+1})$。

3.3 3P-SPIDER

3P-SPIDER 由算法 2给出。迭代索引为 (t, k)，其中 t 是当前外循环的索引，范围为 1 到 k^{out}，而 k 是当前内循环的索引。在外循环 #t中，有 k_t^{in}次内循环迭代。内循环迭代采用算法 1 (参见算法 2 的第 8、9、12 和 13 行)，并结合 SPIDER 方差缩减技巧（参见11 行）算法 2），该算法适用于 quan-$\widehat{\ }$量 $\mathsf{h}_i(S_{t,k}, B_{t,k+1}) - \mathsf{h}_i(S_{t,k-1}, B_{t,k})$ 无法精确计算（参见第 10 行）。

当 G_i 是梯度且 $B(s)=\mathrm{I}_q$ 时，针对 SPIDER 中 b'_t 和 k_t^{in} 的选择，已有不同策略。在 Fang 等人（2018）、Nguyen 等人（2017）和 Wang 等人（2019）中，内循环的次数为常数（对任意 $t \geq 1$ 均为 $k_t^{\text{in}} = k^{\text{in}}$），且 $\mathsf{b}'_t = n$；Nguyen 等人（2017）还考虑了根据算法历史调整 k_t^{in}、同时对其施加上界的情形；在 Horv´ath 等人（2022）中，b'_t 是确定性的，且取决于 t，b 取决于 t，而 k_t^{in} 是几何随机变量，其期望取决于 t；在 Li 等人（2021）中，b'_t 不依赖于 t，且 k_t^{in} 是随机的。

对于 EM 问题（见 2.2节），Fort 等人（2020）提出了 SPIDER-EM，这是一种为大规模学习设计的方差缩减随机EM，¯适用于所有 s、i 下 $\mathsf{h}_i(s)$ 的计算都精确的情形。Fort 等人（2021b）讨论了该算法中采用递增批量大小 $t \mapsto \mathsf{b}'_t$ 以及内循环次数为几何分布、且期望随时间变化的情形 k_t^{in}。结论是，最佳策略是采用确定性的递增序列 b'_t，以便在刷新变量 S. 时提高准确率，并使用固定次数的内循环 $k_t^{\text{in}} = k^{\text{in}}$。本文允许 b'_t 和 k_t^{in} 随 t 变化：它们可以是 t 的确定性函数，也可以是随机的。

矩阵 $\{B_{t,k+1}, t \in [k^{\text{out}}]^\star, k \in [k^{\text{in}}-1]\}$可以是确定性的，也可以是随机的。它们可以在算法运行前选定；更高效的策略则是根据算法运行的历史，在运行过程中不断调整该矩阵。在 EM（见第2.2节）中，$B_{t,k+1}$ 的形式为$\widehat{\mathsf{B}}(S_{t,k})$，其中 B 由统计模型定义。

经过 k_t^{in} 次内循环迭代后，外循环 #$(t+1)$ 开始：随机平均场 $\mathsf{S}_{t+1,0}$会被刷新（见第3 行至第6行）。同样，这里允许对原始 SPIDER 算法进行两种近似：第一种是在计算 $\widehat{\ }\mathsf{h}_i(S_{t,0}, B_{t,1})$ 时采用近似；第二种则避免扫描完整数据集（可以选择 $\mathsf{b}'_t < n$）。

3P-SPIDER 的输入变量包括外循环次数 k^{out}、内循环次数 k_t^{in}、SA 步骤的步长序列 $\{\gamma_{t,k}, t \geq 1, k \geq 1\}$、小批量大小 b 和

算法 2扰动近端预条件 SPIDER 算法（3P-SPIDER）

要求： $k^{\mathrm{out}} \in \mathbb{N}^\star$, $k_t^{\mathrm{in}} \in \mathbb{N}^\star$ 对于 $t \in [k^{\mathrm{out}}]^\star$, $\gamma_{t,k+1} > 0$ 对于 $t \in [k^{\mathrm{out}}]^\star, k \in [k_t^{\mathrm{in}}]$, $\mathsf{b} \in \mathbb{N}^\star$, $\mathsf{b}'_t \in \mathbb{N}^\star$ 对于 $t \in [k^{\mathrm{out}}]^\star$, $\widehat{S}_{\mathrm{init}} \in \mathcal{S}$ 和 $B_{\mathrm{init}} \in \mathcal{P}_+^q$

确保： 序列 $\{\widehat{S}_{t,k}, t \in [k^{\mathrm{out}}]^\star, k \in [k_t^{\mathrm{in}}]^\star\}$。

$\widehat{S}_{0,k_0^{\mathrm{in}}} = \widehat{S}_{\text{初始化}}$, $B_{0,k_0^{\mathrm{in}}} \overset{\text{在}}{=} B_{\text{中}}$ init

对于 $t = 1, \cdots, k^{\mathrm{out}}$ **执行**

  $\widehat{S}_{t,0} = \widehat{S}_{t-1,k_{t-1}^{\mathrm{in}}}$, $\widehat{S}_{t,-1} = \widehat{S}_{t-1,k_{t-1}^{\mathrm{in}}}$, $B_{t,0} = B_{t-1,k_{t-1}^{\mathrm{in}}}$

  抽取一个大小为 b'_t 的批次 $\mathcal{B}_{t,0}$ 从 $[n]^\star$中抽样，采用放回抽样或不放回抽样。

  对于所有 $i \in \mathcal{B}_{t,0}$，计算 $\delta_{t,0,i}$，使其等于或近似于 $\mathsf{h}_i(\widehat{S}_{t,0}, B_{t,0})$。

  $\mathsf{S}_{t,0} = (\mathsf{b}'_t)^{-1} \sum_{i \in \mathcal{B}_{t,0}} \delta_{t,0,i}$

  对于 $k = 0, \cdots, k_t^{\mathrm{in}} - 1$ **执行**

    在 $[n]^\star$ 中抽取大小为 b 的小批量 $\mathcal{B}_{t,k+1}$，可采用放回抽样或不放回抽样。

    选择 $B_{t,k+1} \in \mathcal{P}_+^q$。

    对于所有 $i \in \mathcal{B}_{t,k+1}$，计算 $\delta_{t,k+1,i}$，使其等于或近似于 $\mathsf{h}_i(\widehat{S}_{t,k}, B_{t,k+1}) - \mathsf{h}_i(\widehat{S}_{t,k-1}, B_{t,k})$.

    $\mathsf{S}_{t,k+1} = \mathsf{S}_{t,k} + \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{t,k+1}} \delta_{t,k+1,i}$

    $\widehat{S}_{t,k+1/2} = \widehat{S}_{t,k} + \gamma_{t,k+1} \mathsf{S}_{t,k+1}$

    $\widehat{S}_{t,k+1} \overset{\text{近端算子}_{t,k}(S_{t,k+1/2}),\ \text{其中近端算子}_{t,k} \overset{\text{定义}}{=} \text{近端算子}_{B_{t,k+1}}}{=} \gamma_{t,k+1} g$。

  结束循环

结束循环

$\widehat{\mathsf{b}}'_t$，以及分别在 $\mathcal{S}$ 和 $\mathcal{P}_+^q$ 中的迭代点初始值 S_{init} 和度量 B_{init}。

3.4 关于 $\mathsf{h}_i(s, B)$ 的蒙特卡洛近似

令 $\vartheta \overset{\text{定义}}{=} (s, i, B) \in \mathcal{S} \times [n]^\star \times \mathcal{P}_+^q$。在某些应用中，存在一个可测函数 H_ϑ 和一个定义在可测集 $(\mathsf{Z}, \mathcal{Z})$ 上的概率测度 π_ϑ，使得

$$\mathsf{h}_i(s, B) = \int_{\mathsf{Z}} H_\vartheta(z) \pi_\vartheta(\mathrm{d}z) . \tag{12}$$

这是统计量空间中的期望最大化（参见第 2.2节），其中 $H_\vartheta(z) = S(Y_i, z) - s$ 和

$$\pi_\vartheta(\mathrm{d}z) \overset{\mathrm{def}}{=} \frac{p(Y_i, z; \mathsf{T}(s))}{\int_{\mathsf{Z}} p(Y_i, u; \mathsf{T}(s)) \mu_{lv}(\mathrm{d}u)} \mu_{lv}(\mathrm{d}z) .$$

当（12）中的积分难以求解时，可以采用蒙特卡罗积分来定义近似值 $\delta_{t,k+1,i}$ 和 $\delta_{t,0,i}$（例如，参见 Devroye（1986）关于精确抽样方法的介绍，以及 Robert 和 Casella （2004）关于马尔可夫链蒙特卡罗方法的介绍）。如果 $\{Z_m^\vartheta, m \geq 0\}$ 是服从分布 π_ϑ（$\mathrm{d}z$）的独立抽样，或者是具有唯一不变分布 $\pi_\vartheta(\mathrm{d}z)$ 的遍历马尔可夫链的路径，那么可以设定

$$\mathsf{h}_i(\widehat{S}_{t,k}, B_{t,k+1}) \approx \frac{1}{M} \sum_{m=1}^{M} H_{\vartheta_{t,k+1,i}}(Z_m^{\vartheta_{t,k+1,i}}) ,$$

其中 $\vartheta_{t,k+1,i} \overset{\mathrm{def}}{=} (\widehat{S}_{t,k}, i, B_{t,k+1})$。我们将在第 5 节通过数值结果展示，在近似 $\widehat{\ }$ $\widehat{\ }$ 时，若计算差值 $\mathsf{h}_i(S_{t,k}, B_{t,k+1}) - \mathsf{h}_i(S_{t,k-1}, B_{t,k})$，对两个序列$\{Z_m^{\vartheta_{t,k+1,i}}, m \geq 0\}$ 和 $\{Z_m^{\vartheta_{t,k,i}} m \geq 0\}$ 进行关联会带来收益；这会增强 SPIDER 控制变量的作用（见第 3.2 节）。

3.5 $\mathrm{prox}_{\gamma g}^B$ 的计算

当 $g = 0$ 时，$\mathrm{prox}_{\gamma g}^B(s) = s$。当 $g \neq 0$ 时，$\mathsf{p} \overset{\mathrm{def}}{=} \mathrm{prox}_{\gamma g}^B(s)$ 求解 $0 \in \mathsf{p} - s + \gamma B^{-1} \partial g(\mathsf{p})$，且 p 不一定总有显式表达式。

当 $B = \mathrm{I}_q$，(Combettes和Pesquet, 2011，表 10.1 和 10.2 ）给出了 $\mathrm{prox}_{\gamma g}$ 的性质，以及许多函数的邻近算子表达式 g。

当 B 是一个对角矩阵与一个秩一矩阵之和时，(Becker 和 Fadili，2012，第 3 节）介绍了用于计算p的迭代算法。对于一般的正定矩阵 B，我们根据（Combettes 和 V˜u，2014，

例 3.9)

$$\operatorname{prox}_{\gamma g}^{B}(s)=\sqrt{B}^{-1}\operatorname{prox}_{\gamma g(\sqrt{B}^{-1}\cdot)}(\sqrt{B}s)\,,$$

其中 $\sqrt{B}$ 是矩阵 B的平方根。(Becker 和 Fadili，2012，引理 5）（另见Combettes和V˜u（2014））建立了一个莫罗恒等式，即将近端算子 $\operatorname{prox}_{\gamma g}^{B}$ 表示为 g 的 Fenchel 共轭的邻近算子。

在特殊情形 g 是 $\{0,+\infty\}$-值闭凸集 $\mathcal{S}$的示性函数时，投影 *Landweber* 方法是一种用于计算 p 的迭代算法（参见Eicke（1992），另见（Combettes和Pesquet，2011，例 10.10））。

最后，对于包含度量选择步骤的应用，定义 B 的度量选择策略可见于（Park 等人，2019，第 3 节），适用于对角可变度量；以及 Repetti 和 Wiaux（2021年），针对特定函数 g，这些函数避免了通常颇具挑战性的近端算子 $\operatorname{prox}_{\gamma g}^{B}$的计算。

4 非渐近收敛性分析

本节旨在给出 3P-SPIDER 期望意义下收敛的显式非渐近界。我们将限定在存在 B: $\mathcal{S}\to\mathcal{P}_{+}^{q}$ 且

$$B_{t,k+1}\overset{\text{def}}{=}\mathsf{B}(\widehat{S}_{t,k})\,.$$

该框架涵盖 EM 问题（见第 2.2 节）以及任何预条件梯度算法（见第 2.1 节），其中预条件矩阵通过迭代点的当前值依赖于算法的历史信息。我们还将使用以下记号

$$\bar{\mathsf{h}}_i(s)\overset{\text{def}}{=}\mathsf{h}_i(s,\mathsf{B}(s))\,,\quad \bar{\mathsf{h}}(s)\overset{\text{def}}{=}\mathsf{h}(s,\mathsf{B}(s))\,. \tag{13}$$

3P-SPIDER 旨在求解满足约束 c1 至 c3 的 (1)。因此，根据 (10)，我们关注如何控制以下量：$\operatorname{prox}_{t,k}\left(\widehat{S}_{t,k}+\gamma_{t,k+1}\bar{\mathsf{h}}(\widehat{S}_{t,k})\right)-\widehat{S}_{t,k}$ 其中

$$\operatorname{prox}_{t,k}(s)\overset{\text{def}}{=}\operatorname{prox}_{\gamma_{t,k+1}g}^{\mathsf{B}(\widehat{S}_{t,k})}(s)\,.$$

粗略地说，这些量用于评估算法在第 #(t,k) 次迭代时偏离极限集的程度。更准确地说，我们将控制累积的“到驻点的距离”$\sum_{t=1}^{k^{\text{out}}}\sum_{k=0}^{k_t^{\text{in}}-1}\Delta_{t,k+1}^{\star}$ 其中$\Delta_{t,k+1}^{\star}$ 等于

$$\frac{\|\operatorname{prox}_{t,k}(\widehat{S}_{t,k}+\gamma_{t,k+1}\bar{\mathsf{h}}(\widehat{S}_{t,k}))-\widehat{S}_{t,k}\|^2_{\mathsf{B}(\widehat{S}_{t,k})}}{\gamma_{t,k+1}^2}\,; \tag{14}$$

h 由（9）定义。

在假设 A2至 A 4下，可以得到累积距离的期望控制。A2是关于函数 h_i 的光滑性假设，A3 假设 $n^{-1}\sum_{i=1}^{n}G_i(s)$ 是某个所谓的 Lyapunov 函数的梯度算子，并且矩阵 $\mathsf{B}(s)$ 的谱在 s 中一致有界。A4是关于近似值 $\delta_{t、k+1、i}$ 的假设。

¯假设**2.** 对于所有 $i\in[n]^{\star}$，函数 h_i 在 $\mathcal{S}$ 上全局满足利普希茨条件，其常数为 L_i：存在¯正常数 L_i 使得 $\forall s,s'\in\mathcal{S}$, $\|\mathsf{h}_i(s)-\bar{\mathsf{h}}_i(s')\|\le L_i\|s-s'\|$。令 $L^2\overset{\text{定义为}}{=}n^{-1}\sum_{i=1}^{n}L_i^2$。

假设2 仅要求在 $\mathcal{S}$ 上满足利普希茨条件；这比假设在整个空间 $\mathbb{R}^q$ 上满足利普希茨条件更弱，而文献中有时会采用后一假设（例如参见 Combettes 和 Wajs (2005)). 假设2 例如在以下情形下成立： $\mathcal{S}$ 是紧致的，且对于所有$i\in[n]^{\star}$，梯度 $\nabla\mathsf{h}_i$ 存在并且在 $\mathcal{S}$上连续。

假设**3.** *1.* 存在一个函数 $\mathrm{W}:\mathbb{R}^q\to\mathbb{R}$，它在 $\mathcal{S}$ 上连续可微，并且满足

$$\forall s\in\mathcal{S},\qquad \nabla\mathrm{W}(s)=\frac{1}{n}\sum_{i=1}^{n}G_i(s)\,;$$

此外，$\bar{\mathsf{h}}_i(s)=-\mathsf{B}(s)^{-1}G_i(s)$，其中$\mathsf{B}(s)\in\mathcal{P}_{+}^{q}$.

2. $\nabla\mathrm{W}$ 在全局范围内是 L -W-利普希茨的。*3.* 存在 $0<v_{\min}\le v_{\max}<+\infty$，使得对于任意 $s\in\mathcal{S}$, $v_{\min}\|\cdot\|^2\le\|\cdot\|^2_{\mathsf{B}(s)}\le v_{\max}\|\cdot\|^2$。

同样，这里要求矩阵 $\mathsf{B}(s)$ 的利普希茨性质以及其谱的有界性条件作用于 $\mathcal{S}$，而不是整个空间 $\mathbb{R}^q$。当 $\mathsf{B}(s)$ 不依赖于

s （对于任意 $s \in \mathcal{S}$，$\mathsf{B}(s) = B$ ）时，我们有 $L_{\dot{\mathsf{W}}} \leq v_{\max} n^{-1} \sum_{i=1}^{n} L_i$。

最后一个假设涉及近似 $\mathsf{h}_i(S_{t,k}) - \bar{\mathsf{h}}_i(\widehat{S}_{t,k-1})$ 时误差的波动：令 $\xi_{t,k+1,i} \stackrel{\text{def}}{=} \delta_{t,k+1,i} - \bar{\mathsf{h}}_i(\widehat{S}_{t,k}) + \bar{\mathsf{h}}_i(S_{t,k-1})$，并定义其条件偏差和方差（以由 $\mathcal{B}_{t,k+1}$，$\widehat{S}_{t,k}$ 和 $\widehat{S}_{t,k-1}$ 生成的 σ-域为条件）。令 $\mathcal{P}_{t,k+1/2} \stackrel{\text{def}}{=} \sigma\left(\mathcal{B}_{t,k+1}, \widehat{S}_{t,k}, \widehat{S}_{t,k-1}\right)$。

$$\mu_{t,k+1,i} \stackrel{\text{def}}{=} \mathbb{E}\left[\xi_{t,k+1,i} | \mathcal{P}_{t,k+1/2}\right]$$
$$\sigma^2_{t,k+1,i} \stackrel{\text{def}}{=} \mathbb{E}\left[\|\xi_{t,k+1,i} - \mu_{t,k+1,i}\|^2 | \mathcal{P}_{t,k+1/2}\right] .$$

我们假设

A4. *1.* 在给定 $\mathcal{B}_{t,k+1}$，$S_{t,k}$ 和 $S_{t,k-1}$ 的条件下，近似值 $\{\delta_{t,k+1,i}, i \in \mathcal{B}_{t,k+1}\}$ 相互独立。

2. 存在一个非负常数 C_b，并且对于任意 $t \in [k^{\text{out}}]^\star$，存在一个非递减的确定性序列 $\{m_{t,k}, k \geq 1\}$，使得对于任意 $k \in [k_t^{\text{in}} - 1]$，几乎必然有：

$$\|\frac{1}{n}\sum_{i=1}^{n} \mu_{t,k+1,i}\| \leq \frac{C_b}{m_{t,k+1}} .$$

3. 存在非负常数 C_v 和 C_{vb}，并且对于任意 $t \in [k^{\text{out}}]^\star$，存在非递减的确定性序列 $\{M_{t,k}, k \geq 1\}$ 和 $\{\bar{M}_{t,k}, k \geq 1\}$，使得对于任意 $k \in [k_t^{\text{in}} - 1]$，几乎必然有：

$$\frac{1}{n}\sum_{i=1}^{n} \sigma^2_{t,k+1,i} \leq \frac{C_v}{M_{t,k+1}} ,$$
$$\frac{1}{n}\sum_{i=1}^{n} \|\mu_{t,k+1,i} - \frac{1}{n}\sum_{j=1}^{n} \mu_{t,k+1,j}\|^2 \leq \frac{C_{vb}^2}{\bar{M}^2_{t,k+1}} .$$

我们允许误差 $\xi_{t,k+1,i}$ 是确定性的或随机的。当不存在误差（$\xi_{t,k+1,i} = 0$）时，则有 $C_b = C_v = C_{vb} = 0$。当误差是确定性的时，我们有 $\xi_{t,k+1,i} = \mu_{t,k+1,i}$，以及 $\sigma^2_{t,k+1,i} = 0$。当误差是随机且无偏的时，则有 $\mu_{t,k+1,i} = 0$。因此，常数 C_b、C_v 或 C_{vb} 中有些可以为零，如表 1 所示。

在第 A 节中，我们讨论在 A4 经验证的情形：$\mathsf{h}_i(s') - \mathsf{h}_i(s)$ 是一个期望

	C_b	C_v	C_{vb}
精确的	0	0	0
确定性的	≥ 0	0	≥ 0
随机的，无偏	0	≥ 0	0
随机的，有偏的	> 0	≥ 0	≥ 0

表 1 在 $\mathsf{h}_i(s)'$ 上没有近似值时（精确情形）以及存在近似值时，常数 C_b, C_v, C_{vb} 的符号。

在某个可能依赖于 (s, s', i) 的分布下（见第 3.4 节），并且 $\delta_{t,k+1,i}$ 是一个蒙特卡罗近似值。

定理 4.1 给出了算法进行 $\sum_{t=1}^{k^{\text{out}}} k_t^{\text{in}}$ 次迭代时到稳态的累积距离 $\Delta^\star_{t,k+1}$ 的显式上界（见（14））。它还给出了累积误差 $\mathcal{D}^\star_{t,k+1}$ 的上界，定义如下：

$$\frac{\|\widehat{S}_{t,k+1} - \mathrm{prox}_{t,k}(\widehat{S}_{t,k} + \gamma_{t,k+1}\bar{\mathsf{h}}(\widehat{S}_{t,k}))\|^2_{\mathsf{B}(\widehat{S}_{t,k})}}{\gamma^2_{t,k+1}} ,$$

其中 $\bar{\mathsf{h}}$ 由（13）定义。给定当前迭代 $\widehat{S}_{t,k}$，$\mathcal{D}^\star_{t,k+1}$ 比较两次迭代：理想的近端算子 $\mathrm{prox}_{t,k}(S_{t,k} + \gamma_{t,k+1}\bar{\mathsf{h}}(S_{t,k}))$ 与可用的近端算子 $\mathrm{prox}_{t,k}(S_{t,k} + \gamma_{t,k+1}\mathsf{S}_{t,k+1})$。

定理 4.1 *Assume 1, 2, 3 and 4. Let* $\mathsf{A}\mathsf{A}\mathsf{A}\mathsf{A}$, $\{k_t^{\text{in}}, t \in [k^{\text{out}}]^\star\}$ *be a deterministic positive sequence. For any* $t \in [k^{\text{out}}]^\star$ *and* $k \in [k_t^{\text{in}} - 1]$ *define* $\Lambda_{t\,k+1}$ *by*

$$\frac{\gamma_{t,k} L_{\dot{\mathsf{W}}}}{v_{\min}} + \gamma^2_{t,k} L^2 \frac{2 v_{\max} k_t^{\text{in}}}{v_{\min} \mathsf{b}} \left(1 + \frac{2\, C_{vb}}{\sqrt{\mathsf{b}}\, \bar{M}_{t,k+1}}\right) . \quad (15)$$

Let $\{\widehat{S}_{t,k}, t \in [k^{\text{out}}]^\star, k \in [k_t^{\text{in}}]^\star\}$ *be the sequence given by Algorithm 2 when the stepsize sequence* $\{\gamma_{t,k+1}, t \in [k^{\text{out}}]^\star, k \in [k^{\text{in}} - 1]\}$ *satisfies*

$$\gamma_{t,k+1}\left(1 + \frac{2C_b}{m_{t,k+1}}\right) \leq \gamma_{t,k} , \quad \Lambda_{t,k+1} \in (0, 1/2) . \quad (16)$$

Then,

$$\sum_{t=1}^{k^{\text{out}}} \sum_{k=1}^{k_t^{\text{in}}} \gamma_{t,k} \left(\frac{1}{2} - \Lambda_{t,k+1}\right) \left\{\mathbb{E}\left[\Delta^\star_{t,k}\right] + \mathbb{E}\left[\mathcal{D}^\star_{t,k}\right]\right\}$$
$$\leq \mathbb{E}\left[\mathrm{W}(\widehat{S}_{1,0}) + g(\widehat{S}_{1,0})\right] - \min_{\mathcal{S}} (\mathrm{W} + g)$$
$$+ v_{\max} \sum_{t=1}^{k^{\text{out}}} \gamma_{t,0}\, k_t^{\text{in}}\, \mathbb{E}\left[\|\mathcal{E}_t\|^2\right]$$
$$+ v_{\max} \sum_{t=1}^{k^{\text{out}}} \sum_{k=1}^{k_t^{\text{in}}} \left(k_t^{\text{in}} - k + 1\right) \gamma_{t,k}\, \mathcal{U}_{t,k} ,$$

where $\mathcal{E}_t \overset{\text{def}}{=} \mathsf{S}_{t,0} - \mathsf{h}(\widehat{S}_{t,0})$ *and*

$$\mathcal{U}_{t,k} \overset{\text{def}}{=} \frac{2\,C_b}{m_{t,k}} + \frac{C_b^2}{m_{t,k}^2} + \frac{C_v}{\mathsf{b}\,M_{t,k}} + \frac{2\,C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k}} + \frac{C_{vb}^2}{\mathsf{b}\,\bar{M}_{t,k}^2}\,.$$

定理的证明4.1见第7.5节。注意，$\mathcal{U}_{t,\,k+1} = 0$ 当算法在每次 $\bar{\widehat{\ }}$ 迭代中使用精确预条件梯度时：$\delta_{t,\,0,\,i} = \mathsf{h}_i(S_{t,\,0})$以及$\delta_{t,\,k+1,\,i} = \mathsf{h}_i(S_{t,\,k}) - \bar{\widehat{\ }}\mathsf{h}_i(S_{t,\,k-1})$对所有 i，t，k均成立。

随机内循环次数 k_t^{in}。当内循环次数 k_t^{in}在外循环 #t处为随机数时，我们认为它是在算法运行前抽取的。因此，定理 4.1 中的期望是相对于随机序列 $\{k_t^{\text{in}}, t \in [k^{\text{out}}]^\star\}$的条件期望。关于 k_t^{in} 随机性的期望可由定理 4.1直接得到；此处省略细节。

步长 $\gamma_{t,\,k}$。序列 $\{\gamma_{t,k+1}, t \in [k^{\text{out}}]^\star, k \in [k_t^{\text{in}} - 1]\}$ 上的条件满足，且

$$\gamma_{t,k+1} \overset{\text{def}}{=} \prod_{j=0}^{k} \left(1 + \frac{2\,C_b}{m_{t,j+1}}\right)^{-1} \gamma_{t,0}$$

其中 $\gamma_{t,\,0}$ 为正数且严格小于

$$\frac{1}{4Lv_{\max}\upsilon}\frac{\mathsf{b}}{k_t^{\text{in}}}\left(\sqrt{\frac{L_{\dot{\mathrm{W}}}^2}{L^2} + 4v_{\min}v_{\max}\frac{k_t^{\text{in}}}{\mathsf{b}}\upsilon} - \frac{L_{\dot{\mathrm{W}}}}{L}\right); \tag{17}$$

υ 定义为 $= 1 + 2C_{vb}/(\sqrt{\mathsf{b}}$ 下确界$_{t,\,k}\,\bar{M}_{t,\,k+1})$（见 C.1 节中的证明）。首先注意，当 $C_b = 0$时，步长可以是内循环索引 k loop 的常数函数（$\gamma_{t,\,k+1} = \gamma_{t,\,0}$ 对任意 k均成立）。相反，当 $C_b > 0$ 时，即对于确定性近似或有偏随机近似（见表 1），步长序列是内循环索引 k的严格递减函数。

其次，当 $C_{vb} = 0$ 时，$\gamma_{t,0}$ 的最大值大于当 $C_{vb} > 0$时的最大值。这里同样，确定性的近似或无偏随机近似需要采用更激进的步长。

外循环初始化。设定$\mathcal{N}_t \overset{\text{def}}{=} \|\mathcal{E}_t\|^2$。当 $\mathcal{B}_{t,\,0} = \{1, \cdots, n\}$ 和 $\delta_{t,0,i} = \bar{\widehat{\ }}\mathsf{h}_i(S_{t,0})$ 对所有 i都成立时，则 $\mathcal{N}_t = 0$；否则，$\mathcal{N}_t$ 为正数。

我们来讨论当 $\delta_{t,0,i}$ $\bar{\widehat{\ }}$ 是 $\mathsf{h}_i(S_{t,0})$ 的无偏随机近似，且方差记为 $\sigma_{t,0,i}^2$时，$\mathcal{N}_t$ 的行为。当 $\mathcal{B}_{t,0} = \{1, \cdots, n\}$ 时，

$$\mathbb{E}\left[\mathcal{N}_t\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_{t,0,i}^2\,. \tag{18}$$

然而，策略 $\mathcal{B}_{t,0} = \{1, \cdots, n\}$ 的计算成本较高；抽取大小为 b'_t 的子集可降低计算成本，但会增大误差的平方范数：我们有

$$\mathbb{E}\left[\mathcal{N}_t\right] \le \frac{1}{\mathsf{b}'_t n}\sum_{i=1}^{n}\left(\sigma_{t,0,i}^2 + \|\bar{\mathsf{h}}_i(s) - \bar{\mathsf{h}}(s)\|^2\right)\,, \tag{19}$$

若 $\mathcal{B}_{t,0}$ 是在 $\{1, \cdots, n\}$中有放回抽样，则等号成立。详细计算见第 C.2 节。从数值计算的角度来看，一种高效策略是随着外循环索引b'_t 的增大而增大 t （关于 3.3 节中应用于 EM 的3P-SPIDER 算法，请参见其中的参考文献）。

算法的随机停止时间。在非凸优化中，最后一个迭代点 $\bar{\widehat{\ }}S_{k^{\text{out}},k_k^{\text{in}}\ \text{out}}$ 未必是序列 $\{\widehat{S}_{t,k}, t \in [k^{\text{out}}]^\star, k \in [k_t^{\text{in}}]^\star\}$中与（1）的解集距离最小的点。受（$\Delta^\star$）启发的量10无法在我们的框架中精确计算，因此无法根据这一准则识别“最佳”迭代点。因此，分析算法在随机时刻停止时的性质是一种常见做法（例如参见（Lan，2020，第 6 章））。为简单起见，我们考虑对任意 $k_t^{\text{in}} = k^{\text{in}}$，$t$ 及 $C_b = 0$ 的情形。我们有以下推论：

推论 **4 2** 4 1 *Assume that* $k_t^{\text{in}} =$ 。(定理。) $k^{\text{in}}C_b = 0$ *and the stepsize sequence is constant* , $\gamma_{t\,k} = \gamma_\star$. *Let* τ K *be a uniform random variable on* , (,) $[k^{\text{out}}]^\star \times [k^{\text{in}}]^\star$, *independent of the algorithm. Then*

$$\inf_{(t,k)\in[k^{\text{out}}]^\star\times[k^{\text{in}}]^\star}\left(\frac{1}{2} - \Lambda_{t,k}\right)\mathbb{E}\left[\Delta_{\tau,K}^\star + \mathcal{D}_{\tau,K}^\star\right]$$
$$\le \frac{\mathbb{E}\left[\mathrm{W}(\widehat{S}_{1,0}) + g(\widehat{S}_{1,0})\right] - \min_{\mathcal{S}}\left(\mathrm{W} + g\right)}{k^{\text{out}}k^{\text{in}}\gamma_\star}$$
$$+ v_{\max}\mathbb{E}\left[\|\mathcal{E}_\tau\|^2\right] + v_{\max}\,\mathbb{E}\left[\left(k^{\text{in}} - K + 1\right)\mathcal{U}_{\tau,K}\right]\,.$$

根据式（15），可以轻松得到 $\Lambda_{t,\,k}$ 的上界，它是 $\mathrm{W}\,L$, $v_{\min}$, $v_{\max}k^{\mathrm{in}}$ $\gamma_\star$ C_{vb} $t\,k\,\bar{M}_t\,k{+}1$ ，b 和 inf, , 的函数。

推论 4.2表明，即使采用简单规则停止 3P-SPIDER ，右端的第一项也与最大迭代次数成反比 $k^{\mathrm{out}}k^{\mathrm{in}}$。

当 $\mathcal{E}_t = 0$, $\mathcal{U}_{t,\,k}$ =0 且 $k_t^{\mathrm{in}} = k^{\mathrm{in}}$ 时。对于光滑的一阶优化，通常通过比较算法的复杂度来衡量其满足 ϵ阶平稳条件的复杂度。在随机复合优化中，这一准则自然扩展为由以下内容定义的近似 ϵ-平稳条件：

$$\mathbb{E}\left[\Delta^\star_{\tau,K}\right] \le \epsilon\,,$$

其中 $(\tau,\ K)$ 是一个取值于 $[k^{\mathrm{out}}]^\star \times [k^{\mathrm{in}}]^\star$ 的随机变量；例如参见 (Ghadimi 等人，2016，第 4 节)、(Wang 等人，2019，第 3 节) 以及 Fort 和 Moulines (2021年)。

推论 4.3 研究了近端复杂度$\mathcal{K}_{\mathrm{prox}}$ 其定义为调用近端算子的次数，以满足近似的 ϵ-¯平稳条件；随机h-复杂度$\mathcal{K}_{\bar{\mathsf{h}}}$ 定义为调用一个h_i的次数；以及迭代总次数 $k^{\mathrm{in}}k^{\mathrm{out}}$。同样，为了简化并将我们的结果与现有文献进行比较，我们考虑一个简化的设置。

推论 4 3 *4 2 Assume in 。addition that $\mathcal{E}_t = 0$ and $\mathcal{U}_{t\,k} = 0$. The*（推论 。），*approximate ϵ-stationary condition is satisfied with* $\gamma_\star = v_{\min}/(4\,\underline{L}\cdot\mathrm{W})$ $k^{\mathrm{in}}/\mathsf{b} = L_{\mathrm{W}}^2/(v_{\min}v_{\max}L^2)$, $\mathsf{b} = O\left(\sqrt{n}\sqrt{v_{\min}v_{\max}}L/L\cdot\right.$ *and* $k^{\mathrm{out}}k^{\mathrm{in}} = (\mathrm{W})$ $\left(O\ L\cdot/\ \epsilon v_{\min}\right.$. *Moreover* $\mathcal{K}_{\mathrm{prox}} = O\ L\cdot/\ v_{\min}\,\epsilon$ $\mathrm{W}(\,)\,)$, $(\mathrm{W}(\,)\,)$ *and* $\mathcal{K}_{\bar{\mathsf{h}}} = O\left(\sqrt{v_{\max}}L\sqrt{n}/(\epsilon\sqrt{v_{\min}})\right)$。

证明见第7.6节。该结果表明，步长 $\gamma_\star$ 和内循环次数k^{in} 均独立于准确率 ϵ。

应用于随机梯度下降时，3P-SPIDER 在推论 4.3 的设定下就是Prox-SpiderBoost 算法，该算法由Wang 等人(2019) 研究：推论 4.3 与（Wang 等人，2019，定理 2）给出了相同的复杂度结果。（Wang 等人，2019，表 1）将Prox-SpiderBoost与其他随机梯度用于复合非凸有限和优化的算法。研究表明，基于 SPIDER阶的方差缩减优于其他方差缩减策略，例如SVRG 策略和 SAGA 策略，分别由Johnson 和 Zhang (2013) 以及 Defazio 等人(2014) 提出。因此，3P-SPIDER在为求解有限和非凸复合优化而设计的近端随机梯度算法中达到了最先进水平。

应用于 EM 时，推论 4.3 所设定的 3P-SPIDER 是对 Fort 等人（2020）研究的 SPIDER-EM 算法的扩展，适用于存在近端步来处理约束 g的情况。同样，比较推论 4.3 与（Fort 等人，2020，定理 2）可知，在采用方差缩减的增量 EM 算法中，3P-SPIDER 达到了当前最先进水平，其中包括 Chen 等人（2018）和 Karimi 等人（2019）分别提出的 sEM-VR 和 FIEM（另见 Fort 等人（2021a））。关于与现有文献的比较，参见 Fort 等人（2020）。

除这两种应用外，据我们所知，推论 4.3是首个针对非凸有限和复合优化、且为求解约束（1）和（2）下问题而设计的算法的复杂度结果。

近似的 ϵ-平稳条件：非精确预条件前向算子的代价。我们来讨论非精确¯$\mathsf{h}_i(s)$ 在近似值无偏且为随机的情况下的代价（因此 $C_b = C_{vb} = 0$, 见表1）：它是否会增加近端复杂度 $\mathcal{K}_{\mathrm{prox}}$以及调用预条件前向算子预言机的次数¯$\mathsf{h}_i$ (下文仍记为 $\mathcal{K}_{\bar{\mathsf{h}}}$)? 下文各项结论的详细计算见第 7.7节。

若对于某个 $\mathsf{a}' \in [0,1)$, $\mathbb{E}\left[\|\mathcal{E}_t\|^2\right] = O(\epsilon^{1-\mathsf{a}'}/(\sqrt{nt})^{\mathsf{a}'})$, 且

$$M_{t,k+1} = O\left(\frac{n^{(\mathsf{a}-\bar{\mathsf{a}})/2}}{\epsilon^{1-\mathsf{a}}}\,t^{\mathsf{a}}\,(k+1)^{\bar{\mathsf{a}}}\right)$$

对于某个 a ¯a $\in [0,1)$，则近似 ϵ- 驻点条件以¯$k_t^{\mathrm{in}} = O(\sqrt{n})$ 的形式成立，$\mathsf{b} = O(\sqrt{n})$ 和 $k^{\mathrm{out}} = O(1/(\sqrt{n}\epsilon))$。此外，$\mathcal{K}_{\mathrm{prox}} = O(1/\epsilon)$ 和 $\mathcal{K}_{\bar{\mathsf{h}}} = O(\sqrt{n}/\epsilon)$。因此，推论 4.3 的结论仍然成立，并且 h_i 的近似值不会变差

当近似值足够小时，算法的复杂度表现。

现在，我们来评估无偏随机近似由独立同分布（i.i.d.）样本计算得到的蒙特卡洛近似时的计算成本。在这种情况下，$M_{t,\cdot}$ 是蒙特卡洛求和中的项数（见第A节）。蒙特卡洛复杂度 $\mathcal{K}_{\mathsf{MC}}$ 定义为满足近似 ϵ-平稳条件所需的蒙特卡洛抽样总数：$\mathcal{K}_{\mathsf{MC}} = O(\sqrt{n}/\epsilon^2)$，对任意$\mathsf{a}$，$\mathsf{a}'$，$\bar{\mathsf{a}} \in [0,1)$。

–

据我们所知，这是首次针对预条件前向算子$\bar{\mathsf{h}}_i$'的此类蒙特卡洛近似进行复杂度分析。

5 应用：带随机效应的惩罚逻辑回归

5.1 模型

受分类应用的启发，我们考虑带有随机效应的逻辑回归模型。

设有 n 对样本 $\{(X_i, Y_i)\ i \in [n]^\star\}$，其中 $X_i \in \mathbb{R}^d$ 包含 d 解释变量，而 Y_i 是取值于 $\{-1,1\}$ 的二元响应变量。我们假设，给定 $\{X_i, i \in [n]^\star\}$ 后，二元观测值 $\{Y_i, i \in [n]^\star\}$ 相互独立，且服从如下分布

$$\{-1,1\} \ni y_i \mapsto \int_{\mathbb{R}^d} (1+\exp(-y_i \langle X_i, z_i\rangle))^{-1} \times \frac{1}{\sqrt{2\pi}^d \sigma^d} \exp\left(-(2\sigma^2)^{-1}\|z_i-\theta\|^2\right) \mathrm{d}z_i \,.$$

换言之，每个样本 $\#i$ 都有一个个体回归向量 Z_i，其取值于 $\mathbb{R}^d$；给定 Z_i 后，成功概率 $\mathbb{P}(Y_i = 1 \mid Z_i)$ 为 $(1+\exp(-\langle X_i, Z_i\rangle))^{-1}$。回归向量$Z_1, \cdots, Z_n$ 相互独立，且服从高斯分布 $\mathcal{N}(\theta, \sigma^2\mathrm{I}_d)$。$\theta$ 假定为未知，σ^2 已知。

目标是通过最大化带岭惩罚项的对数似然准则来估计 θ，其中岭惩罚项为 $\mathrm{pen}(\theta) \overset{\text{定义}}{=} n\tau\|\theta\|^2$，且 $\tau > 0$。通过变量替换，我们得到该准则待最小化的是（见引理B.1）

$$F : \theta \mapsto -\frac{1}{n}\sum_{i=1}^n \log \int_{\mathbb{R}} \frac{\exp\left(x \langle X_i, \theta\rangle / (\sigma^2\|X_i\|)\right)}{1+\exp(-y_i\|X_i\|x)} \times \exp\left(-x^2/(2\sigma^2)\right) \mathrm{d}x + \|\theta\|_U^2 \,,$$

其中

$$U \overset{\text{def}}{=} \tau \mathrm{I}_d + \frac{1}{2\sigma^2}\frac{1}{n}\sum_{i=1}^n \frac{X_i X_i^\top}{\|X_i\|^2} \,.$$

以下引理表明，F 的极小化子属于 $\mathcal{K}$ 的 $\mathbb{R}^d$ 一个紧集，因此优化问题可以约束在 $\mathcal{K}$上。证明见第B.2节。

引理 5：1 *The minimizers of F are in the set* $\mathcal{K}$ 定义= $\{\theta \in \mathbb{R}^d \quad \|\theta\|^2 \leq \quad 4\ /\tau\}$. : （ln）

为求解该优化问题，我们提出两种方法：一种是梯度方法，在原始空间中求解 $\theta \in \mathbb{R}^d$（见第2.1节）；另一种是期望最大化方法，在统计量空间中求解（见第2.2节）。第5.2 节和第5.3 节的讨论表明，期望最大化是寻找 $s \mapsto F$的临界点的梯度方法（$U^{-1}s/2$）。

5.2 梯度方法

我们旨在寻找 F在 $\mathcal{K}$中的一个临界点。等价地，我们希望求解

$$0 \in \frac{1}{n}\sum_{i=1}^n G_i(\theta) + \partial g(\theta)$$

其中，g 是集合 $\mathcal{K}$ 的 $\{0,+\infty\}$ 值示性函数，且

$$G_i(\theta) \overset{\text{def}}{=} 2U\theta - \frac{X_i}{\sigma^2\|X_i\|}\int_{\mathbb{R}} z\ \pi_{\theta,i}(z)\mathrm{d}z \,;$$

$\pi_{\theta,i}(z)$ 是正比于.....的概率密度

$$\frac{\exp\left(z \langle X_i, \theta\rangle / (\sigma^2\|X_i\|) - z^2/(2\sigma^2)\right)}{1+\exp(-y_i\|X_i\|z)} \,. \qquad (20)$$

我们使用 3P-SPIDER，其中 $B \overset{\text{定义}}{=} \mathrm{I}_q$，且 $\mathsf{h}_i(\theta, \mathrm{I}_q) \overset{\text{定义}}{=} -G_i(\theta)$；注意，$\mathrm{prox}_{\gamma g}(\theta) = \mathrm{argmin}_{x\in\mathcal{K}}\|x -$

$\theta\|^2$。 h_i 是显式项与一个没有闭式的积分之和：我们将使用基于马尔可夫链蒙特卡罗（MCMC）抽样器的蒙特卡罗方法对其进行近似（见下文第 5.4 节）。因此，$\delta_{t,\,k,\,i}$ 将是有偏随机近似。

5.3 一种期望最大化方法

要最小化的准则 F 具有如下形式（5），其中 $\mathsf{Z} = \mathbb{R}$, $\mu_{lv}(\mathrm{d}z) = \mathrm{d}z$ 以及 $p(Y_i,\ z;\ \theta)$ 等于

$$\frac{\exp\left(z\,\langle X_i,\theta\rangle\,/(\sigma^2\|X_i\|) - z^2/(2\sigma^2) - \|\theta\|_U^2\right)}{1+\exp(-Y_i\|X_i\|z)}\,.$$

满足完整数据模型的曲指数族假设：$p(Y_i,\ z;\ \theta) = H(Y_i,\ z)\,\exp\left(\langle S(Y_i,\ z)\ \phi(\theta)\rangle - \psi(\theta)\right)$，其中 $\phi(\theta)$ 定义为 $=\theta$，$\psi(\theta) \stackrel{\mathrm{def}}{=} \|\theta\|_U^2$ 以及

$$S(Y_i,z) \stackrel{\mathrm{def}}{=} z\,\frac{X_i}{\sigma^2\,\|X_i\|}\,.$$

根据第2.2节，统计量空间中的期望最大化算法具有形式（1）- （2）：它求解 $0 \in n^{-1}\sum_{i=1}^n \bar{G}_i(s) + \partial\bar{g}(s)$，其中 $\bar{G}_i(s) = B\,G_i(Bs)$，$B$ 定义为 $= U^{-1}/2$ 并且 $\bar{g}(s)$ 是集合 $\{s\in\mathbb{R}^d$ 的 $\{0,+\infty\}$ 值示性函数：$\mathsf{T}(s) \in \mathcal{K}\}$，其中 $\mathsf{T}(s)$ 定义为 $= Bs$；它使用

$$\bar{\mathsf{h}}_i(s) \stackrel{\mathrm{def}}{=} \frac{X_i}{\sigma^2\,\|X_i\|}\int_{\mathbb{R}} z\,\pi_{Bs,i}(z)\mathrm{d}z - s\,, \qquad (21)$$

以及由 $\mathsf{B}(s)$ def$=$ B 诱导的度量。详细计算请参见附录 B.3。与梯度方法一样，h_i 需要对分布 $\pi_{\cdot,i}$ 取期望（见式 (20)），而该期望没有闭式表达。我们将使用 $B(s) \leftarrow B$ 运行 3P-SPIDER，并对 $\mathsf{h}_i(s)$ 的 $\bar{}$ 采用有偏随机近似（见第 5.4 节）；注意，$\mathrm{prox}^B_{\gamma\bar{g}}(s) = B^{-1}\mathrm{argmin}_{x\in\mathcal{K}}\left((x-Bs)^\top B^{-1}(x-Bs)\right)$。

5.4 $\bar{\mathsf{h}}_i$ 的 MCMC近似

我们讨论如何为……的近似值设计高效的 MCMC采样器

$$\mathcal{I}_i(\theta) \stackrel{\mathrm{def}}{=} \int_{\mathbb{R}} z\,\pi_{\theta,i}(z)\,\mathrm{d}z\,,\quad \theta\in\mathbb{R}^d\,,$$

其中，$\pi_{\theta,\,i}$ 定义为（相差一个归一化常数），如（20）所示。通过分部积分并应用（Polson 等人, 2013，定理 1），我们证明可以采用数据增广方案来近似关于 $\pi_{\theta,\,i}(z)$ 的积分。

引理 5.2 *For any $i\in[n]^\star$ and $\theta\in\mathbb{R}^d$, it holds*

$$\mathcal{I}_i(\theta) = \left\langle \frac{X_i}{\|X_i\|}, \theta\right\rangle + y_i\|X_i\|\sigma^2\int_{\mathbb{R}}\int_0^{+\infty}\frac{\bar{\pi}_{\theta,i}(z,\omega)}{1+\exp\left(y_i\|X_i\|z\right)}\,\mathrm{d}z\mathrm{d}\omega\,,$$

where $\bar{\pi}_{\theta,\,i}(z,\ \omega)$ is a probability density on $\mathbb{R}\times(0,\ +\infty)$。 The conditional distribution of z given ω is a Gaussian distribution with parameters

$$\frac{\langle X_i,\theta\rangle\,/\|X_i\| + y_i\|X_i\|\sigma^2/2}{1+\omega\sigma^2\|X_i\|^2}\,,\qquad \frac{\sigma^2}{1+\omega\sigma^2\|X_i\|^2}\,;$$

the conditional distribution of ω given z is a Polya-Gamma distribution with parameters 1 $\|X_i\|z$. (,)

证明见 B.4 节。因此，关于$\pi_{\theta,\,i}$ 的积分的蒙特卡洛近似值可通过以分布$\bar{\pi}_{\theta,\,i}(z,\ \omega)$为目标分布的吉布斯采样器获得：它会生成一系列成对样本 $\{(Z_r,\ \Omega_r)\ \ r\geq 0\}$，而蒙特卡洛近似中仅保留 Z_r。例如，$\mathsf{h}_i(s)$，由（21）给出，可近似为

$$\bar{\mathsf{h}}_i(s) \approx -s + \frac{X_i}{\sigma^2\|X_i\|^2}\,\langle X_i, Bs\rangle + y_i\|X_i\|\frac{1}{m}\sum_{r=1}^m\left(1+\exp(y_i\|X_i\|Z_r^{s,i})\right)^{-1}\,. \qquad (22)$$

该吉布斯采样器是一致遍历的（见（Choi and Hobert, 2013, 命题 3.1））；因此，注意到 $z\mapsto H_i(z)$ def $= (1+\exp(y_i\|X_i\|z))^{-1}$ 关于 i 和 z 一致地以 1 为界，便可验证 A 节中的条件 A5，其中 U 等于常函数 1，且几何收敛速率为 $\rho(r) \stackrel{\mathrm{def}}{=} \upsilon^r$，其中某个$\upsilon\in(0,1)$（请记住，在我们的应用中，$\mathcal{S}$ 是一个紧集）；详情见 B.5 节。

因此，A4 已得到验证，且速率 $m_{t,\ k+1}$，$M_{t,\ k+1}$ 和$M_{t,\ k+1}$ 均相等，且等于蒙特卡洛求和中的点数（见命题 A.1）。

5.5 数值示例

我们运行 3P-SPIDER 来最小化准则 F；根据此前对比方差缩减期望最大化算法与方差缩减梯度算法的结果（例如参见（Chen等人，2018，第4节）），我们将注意力集中在期望最大化方法上。在这一数值应用中，$n = 24\,989$ 和 $d = 21$；我们选择$\tau = 1$ 和 $\sigma^2 = 0.05$。

数据集。 n 对（y_i，X_i）由 MNIST 数据集构成。标记为 $y_i = -1$ 的 13 007 个样本，是 MNIST 训练数据集中标记为 1 或 7 的样本；标记为 $y_i = 1$ 的 11 982 个样本，是 MNIST 训练数据集中标记为 3 或 8 的样本。协变量X_i按如下方式获得。令 X^{im} 为一个矩阵，其中汇集了每幅图像的 784 个像素。像素取值范围为 $[0, 1]$。然后将 X^{im} 的各行中心化；通过主成分分析，将每幅图像降至 $\mathbb{R}^{20}$中的一个向量。由此得到 $\mathsf{X}^{\mathrm{red}} \in \mathbb{R}^{20\times n}$。最后，在 $\mathsf{X}^{\mathrm{red}}$ 中添加一行 1，从而得到 $\mathsf{X} \in \mathbb{R}^{21\times n}$。$\mathsf{X}$ 的列为X_i。

算法。 我们比较四种算法。 EM 表示 SAEM 算法（Delyon 等人 (1999)）结合近端步：每次迭代处理完整数据集，因此每个训练周期包含一次 EM 迭代：

$$\widehat{S}^{\mathtt{EM}}_{r+1} \overset{\mathrm{def}}{=} \mathrm{prox}^{B}_{\gamma g}\Big(\widehat{S}^{\mathtt{EM}}_{r} + \frac{\gamma}{n}\sum_{i=1}^{n} \widehat{\bar{\mathsf{h}}_i(\widehat{S}^{\mathtt{EM}}_{r})}\Big)\,.$$

在线 EM是 Cappé 和 Moulines(2009) 提出的算法，结合了近端步；每次迭代处理b 个样本，下面我们将运行 $k^{\mathrm{in}} \overset{\mathrm{def}}{=} \lceil n/\mathsf{b}\rceil$ 次迭代：

$$\widehat{S}^{\mathtt{OEM}}_{r+1} \overset{\mathrm{def}}{=} \mathrm{prox}^{B}_{\gamma g}\Big(\widehat{S}^{\mathtt{OEM}}_{r} + \frac{\gamma}{\mathsf{b}}\sum_{i\in\mathcal{B}_{r+1}} \widehat{\bar{\mathsf{h}}_i(\widehat{S}^{\mathtt{OEM}}_{r})}\Big)\,.$$

$\widehat{\bar{\mathsf{h}}_i(S^{\bullet}_t)}$ 对于 EM 和在线 EM，$\mathsf{h}_i(S^{\bullet}_t)$ 是使用 m^t 个点计算得到的 $\bar{\mathsf{h}}_i(\widehat{S}^{\bullet}_t)$ 的蒙特卡洛近似。3P-SPIDER 是算法 2；我们选择 $k^{\mathrm{in}}_t = k^{\mathrm{in}}$ 和 $k^{\mathrm{in}} = \lceil n/\mathsf{b}\rceil$，使得一个训练周期对应于 k^{in} 个内循环；我们选择 $\mathsf{b}'_t = n$，使得每个外循环的初始化就是一个训练周期；δ_t、k、i 通过蒙特卡罗求和（见（22））计算，其中 m^0 个点用于 δ_t、0、i，m^t 个点用于 $\widehat{\delta_t}$、$k+1$、i；由于 $S_{t0} = S_{t-1}$，我们为 δ_t、1、$i = 0$ 设置所有 i，从而 $\mathsf{S}_{t,1} = \mathsf{S}_{t,0} = n^{-1}\sum_{i=1}^{n}\delta_{t,0,i}$。最后，3P-SPIDER 与 3P-SPIDER-corr 的区别如下：蒙特卡洛近似 $\delta_{t,k+1,i}$ 需要对 $\mathsf{h}_i(S_{t,k})$ $\widehat{\bar{}}$ 进行一次蒙特卡洛近似，并对 $\mathsf{h}_i(S_{t,k-1})$ 进行另一次蒙特卡洛近似。在 3P-SPIDER 中，蒙特卡洛近似基于两条相互独立的链（见（22）），而在 3P-SPIDER-corr 中，这些链彼此相关。

所有算法均初始化为零向量 $\widehat{S}_{\mathrm{init}} = 0 \in \mathbb{R}^d$。前六个训练周期的步长等于$\gamma = 0.4$，之后则等于 $\gamma = 0.1$。所有路径的长度均为 20 个训练周期。在除图 3之外的所有图中，我们报告每种算法 25 次独立运行所得的均值；阴影区域的边界由这些运行中所显示准则的最小值和最大值确定。

分析。 大多数比较都基于以下指标随训练周期数的变化情况：

$$\Delta_{t,k+1} \overset{\mathrm{def}}{=} \frac{\|\mathrm{prox}^{B}_{\gamma g}(\widehat{S}_{t,k} + \gamma\,\mathsf{S}_{t,k+1}) - \widehat{S}_{t,k}\|^2_B}{\gamma^2}$$

该指标是 $\Delta^{\star}_{t,k+1}$ 的近似值（见（14）），由于在此应用中 h 没有闭式表达，因此无法计算。对于 3P-SPIDER 和 3P-SPIDER-corr，该指标$\Delta_{t,k+1}$ 与$\Delta^{\mathtt{EM}}_{r}$ 进行比较，后者定义为

$$\frac{\|\mathrm{prox}^{B}_{\gamma g}(\widehat{S}^{\mathtt{EM}}_{r} + \gamma n^{-1}\sum_{i=1}^{n} \widehat{\bar{\mathsf{h}}_i(\widehat{S}^{\mathtt{EM}}_{r})}) - \widehat{S}^{\mathtt{EM}}_{r}\|^2_B}{\gamma^2}\,;$$

以及与 $\Delta^{\mathtt{OEM}}_{r}$进行比较，后者定义为

$$\frac{\|\mathrm{prox}^{B}_{\gamma g}(\widehat{S}^{\mathtt{OEM}}_{r} + \gamma\mathsf{b}^{-1}\sum_{i\in\mathcal{B}_{r+1}} \widehat{\bar{\mathsf{h}}_i(\widehat{S}^{\mathtt{OEM}}_{r})}) - \widehat{S}^{\mathtt{OEM}}_{r}\|^2_B}{\gamma^2}\,.$$

最佳算法对应的$\Delta_{t,k+1}$值最小。

我们首先研究 3P-SPIDER 的一些设计参数的作用，例如计算 $\delta_{t,0,i}$ 时使用的蒙特卡罗点数量（记作m^0）和$\delta_{t,k+1,i}$（记作 m^t），以及 k^{in} 与满足 $k^{\mathrm{in}}\mathsf{b} \approx n$ 的 b 之间的平衡。

在图 1中，选择了两种策略：首先，$m^0 = m^t = 2\lceil\sqrt{n}\rceil$；然后 $m^0 = \overline{m}^t = 5\lceil\sqrt{n}\rceil$；在所有情况下，$\overline{k}^{\mathrm{in}} = \lceil\sqrt{n/10}\rceil$ 以及$\mathsf{b} = \lceil n/\overline{k}^{\mathrm{in}}\rceil$。作为对比，期望最大化 和 在线 EM 也进行了运行，每次迭代使用的蒙特卡罗样本点数量为 m^t。

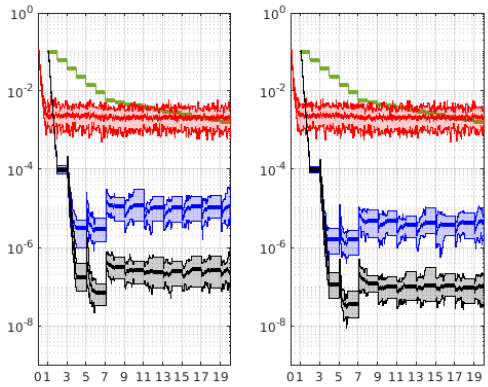


图 1 用于近似时蒙特卡罗点数量的不同策略 $\bar{\mathsf{h}}_i(s)$ - 参见 (22)。Δ_r^{EM} 的演变（绿色），Δ_r^{OEM}（红色），$\Delta_{t,\,k+1}$ 对应 3P-SPIDER（蓝色）以及 $\Delta_{t,\,k+1}$ 对应 3P-SPIDER corr（黑色），随训练周期数的变化。[ ] $m^0 = m^t = 2\lceil\sqrt{n}\rceil$ [ ] 左，右。

在图 2中，将 $k^{\mathrm{in}} = \lceil\sqrt{n/10}\rceil$ 的情况与 $k^{\mathrm{in}} = \lceil\sqrt{n/2}\rceil$的情况进行了比较；在这两种情况下，$\mathsf{b} = \lceil n/k^{\mathrm{in}}\rceil$ 以及 $m^0 = m^t = 2\lceil\sqrt{n}\rceil$。

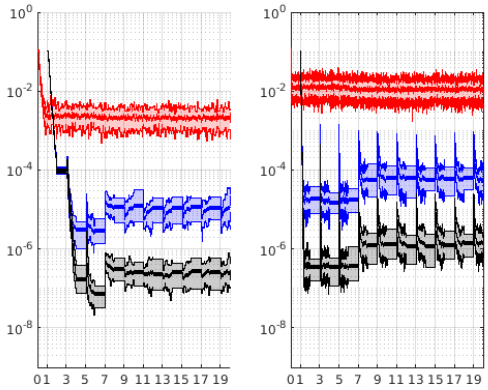


图 2 每个训练周期的内循环次数。Δ_r^{OEM}的演变（红色），$\Delta_{t,\,k+1}$ 对应 3P-SPIDER（蓝色）以及 $\Delta_{t,\,k+1}$对应 3P-SPIDER corr（黑色），随训练周期数的变化。[ ] $k^{\mathrm{in}} = \lceil\sqrt{n/10}\rceil$ $= \lceil n/k^{\mathrm{in}}\rceil$ [ ] 左和 b。右和 b。$k^{\mathrm{in}} = \lceil\sqrt{n/2}\rceil$ $= \lceil n/k^{\mathrm{in}}\rceil$

每个算法都会返回一列位于 s-空间中的点，由此可通过以下公式推导出一列位于 θ-空间中的点：$\theta = \mathsf{T}(s) = Bs \in \mathbb{R}^d$。图 3展示了该 θ-序列中的三个分量随训练周期数的变化。

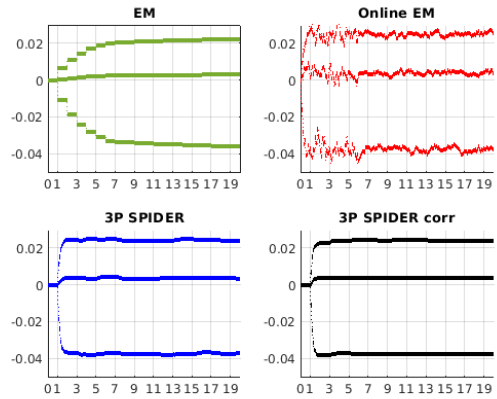


图 3 三个参数的估计。三个分量随训练周期数的变化：θ 由 EM 表示，绿色（左上），OEM红色（右上），3P-SPIDER 蓝色（左下），以及3P-SPIDER corr 黑色（右下）。

最后，我们还在图 4 中展示了 3P-SPIDER 和 3P-SPIDER-corr 得到的迭代点$\|\widehat{S}_{t,\,k}\|^2$，以及分别由 EM 和在线 EM 得到的$\|\widehat{S}_r^{\mathrm{OM}}\|^2$ 和 $\|\widehat{S}_r^{\mathrm{OEM}}\|^2$ 的平方范数随训练周期数的变化。

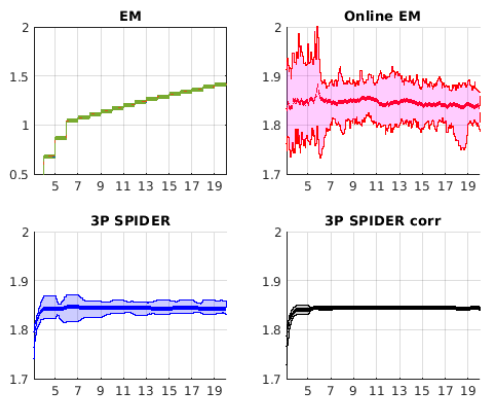


图 4 迭代点的平方范数。$\|\widehat{S}_r^{\mathrm{EM}}\|^2$绿色（左上）、$\|\widehat{S}_r^{\mathrm{OEM}}\|^2$ 红色（右上）、$\|\widehat{S}_{t,\,k}\|^2$ 对于3P-SPIDER 蓝色（左下），以及$\|\widehat{S}_{t,\,k}\|^2$ 对于 3P-SPIDERcorr黑色（右下），随训练周期数的变化。

结论。 期望最大化 的收敛速率较慢，甚至在 20 个训练周期内仍未收敛，这与其他算法相反（例如，参见图 4）：每个训练周期只更新一次迭代点是不够的，尤其是在前几次迭代中，此时需要更多

即使每次只基于部分数据集进行更新，也是更好的策略（例如，参见在线 EM 的表现，它每个训练周期包含 k^{in} 次更新）。
在线 EM、3P-SPIDER 和 3P-SPIDER-corr 在每次迭代时处理部分数据集；与在线 EM 相比，3P-SPIDER 方法包含方差缩减。所有图表都展示了这种方差缩减的益处，它降低了收敛时的变异性。

.....的选择 γ 会影响这种变异性：例如，参见图1、2 和图 4，其中在训练周期 #7 处发生了变化（请记住，从训练周期 2ℓ 到 $2\ell+1$，在线 EM 会对迭代点进行 k^{in} 次更新，而 3P-SPIDER 方法不会更新迭代点，因为它们会计算 $\mathsf{S}_{t,\,0}$）。
3P-SPIDER-corr 优于 3P-SPIDER。相关性越强，控制变量的影响越大，所有图表都说明了这一点。它降低了小批次（$\mathsf{b} < n$）引入的变异性，以及蒙特卡洛近似 $\delta_{t,\,k+1,\,i}$ 引入的变异性。

在给定每个外循环处理的 n 样本预算下，图2 显示，在收敛时，较大的小批量大小能够提高准确率，从而减少内循环的次数。较多的蒙特卡罗点会降低收敛时的变异性，这并不令人意外（见图1）。

第3节证明

引理3.1的证明

引理6.1 汇总了引理3.1 的两个陈述及第三个性质。

引理 6.1 *Assume*A*1*。
*1.*对于任意 $\gamma > 0$、$B \in \mathcal{P}_+^q$ 和 $s \in \mathbb{R}^q$，优化问题(7) 具有唯一解，其特征是唯一的点$\mathsf{p} \in \mathcal{S}$ 满足 $-\gamma^{-1} B(\mathsf{p} - s) \in \partial g(\mathsf{p})$。

2. 对于任意的 $\gamma > 0$、$B \in \mathcal{P}_+^q$、$s \in \mathcal{S}$ 和 $h \in \mathbb{R}^q$，

$$s = \operatorname{prox}_{\gamma g}^B(s + \gamma h) \quad \textit{iff} \quad Bh \in \partial g(s). \quad (23)$$

3. 令 $\gamma > 0$ 和 $B \in \mathcal{P}_+^q$。算子 近端算子$\operatorname{prox}_{\gamma g}^B$ 是坚非扩张的；这意味着对于任意s、$s' \in \mathbb{R}^q$，

$$\|\operatorname{prox}_{\gamma g}^B(s') - \operatorname{prox}_{\gamma g}^B(s)\|_B^2 \le \left\langle \operatorname{prox}_{\gamma g}^B(s') - \operatorname{prox}_{\gamma g}^B(s), s' - s \right\rangle_B .$$

Proof 存在性、唯一性及刻画见 Hiriart-Urruty 和 Lemar´echal (1996，第 XV 章，引理 4.1.1）。陈述（23）由刻画可得；注意，对于任意，$\operatorname{prox}_{\gamma g}^B($ $s \in \mathcal{S}$ $s \in \mathbb{R}^q$)。
坚非扩张性质是 Hiriart-Urruty 和 Lemar´echal (1996，第 XV 章，定理 4.1.4）的推论。□

第 7 节的证明4

7.1 符号

对任意 $s \in \mathcal{S}$，定义

$$\bar{\mathsf{h}}(s) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \bar{\mathsf{h}}_i(s)\,, \qquad \bar{\mathsf{h}}_{\mathcal{B}} \stackrel{\text{def}}{=} \frac{1}{\mathsf{b}}\sum_{i\in\mathcal{B}} \bar{\mathsf{h}}_i\,,$$

其中，$\mathcal{B}$ 是由 $[n]^\star$ 中基数为 b 的元素（可重复或不可重复）构成的 n 元组。
所有随机变量均定义在概率空间 $(\Omega\ \mathcal{A},\ \mathbb{P})$ 上。该空间赋有以下滤子，对于 $t \ge 0$ 和 $k \ge 0$，

$$\begin{aligned}
\mathcal{F}_{0,k_0^{\text{in}}} &\stackrel{\text{def}}{=} \sigma(\widehat{S}_{\text{init}}),\\
\mathcal{F}_{t,0} &\stackrel{\text{def}}{=} \mathcal{F}_{t-1,k_{t-1}^{\text{in}}} \bigvee \sigma\left(\mathcal{B}_{t,0}, \delta_{t,0,i} \text{ for all } i\right),\\
\mathcal{F}_{t,k+\frac{1}{2}} &\stackrel{\text{def}}{=} \mathcal{F}_{t,k} \bigvee \sigma(\mathcal{B}_{t,k+1}),\\
\mathcal{F}_{t,k+1} &\stackrel{\text{def}}{=} \mathcal{F}_{t,k+\frac{1}{2}} \bigvee \sigma\left(\delta_{t,k+1,i} \text{ for all } i \in \mathcal{B}_{t,k+1}\right) .
\end{aligned}$$

对于任意 $t \in [k^{\text{out}}]^\star$，令

$$\mathcal{E}_t \stackrel{\text{def}}{=} \mathsf{S}_{t,0} - \bar{\mathsf{h}}(\widehat{S}_{t,0}) = \frac{1}{\mathsf{b}'_t}\sum_{i\in\mathcal{B}_{t,0}} \delta_{t,0,i} - \bar{\mathsf{h}}(\widehat{S}_{t,0})\,.$$

$\mathcal{E}_t$ 是使用 $\bar{\ }$ $\widehat{\ }$个精确项 $\mathsf{h}_i(S_{t,\,0})$ 替代完整总和时产生的误差，其中可能使用大小为 $\mathsf{b}'_t < n$ 的 $\bar{\ }$ $\widehat{\ }$子和，并采用 $\mathsf{h}_i(S_{t,\,0})$ 的近似值。请记住，

$$\xi_{t,k+1,i} \stackrel{\text{def}}{=} \delta_{t,k+1,i} - \bar{\mathsf{h}}_i(\widehat{S}_{t,k}) + \bar{\mathsf{h}}_i(\widehat{S}_{t,k-1})\,,$$

and

$$\begin{aligned}
\mu_{t,k+1,i} &\stackrel{\text{def}}{=} \mathbb{E}\left[\xi_{t,k+1,i} | \mathcal{F}_{t,k+1/2}\right]\,,\\
\sigma^2_{t,k+1,i} &\stackrel{\text{def}}{=} \mathbb{E}\left[\|\xi_{t,k+1,i} - \mu_{t,k+1,i}\|^2 | \mathcal{F}_{t,k+1/2}\right]\,.
\end{aligned}$$

最后，设

$$\eta_{t,k+1} \stackrel{\text{def}}{=} \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t,k+1}} \xi_{t,k+1,i}.$$

在整个证明过程中，我们将使用简写符号

$$B_{t,k} \stackrel{\text{def}}{=} \mathsf{B}(\widehat{S}_{t,k}) \,.$$

7.2 预备引理

引理 7 1 *Let $\mathcal{B}$ be a batch of $[n]^\star$ of size b sampled at random (with or without replacement).*

1. *For any family $\{f_1, \cdots, f_n\}$, $\mathbb{E}\left[\mathsf{b}^{-1} \sum_{i \in \mathcal{B}} f_i\right] = n^{-1} \sum_{i=1}^n f_i$.*
2. *For any family $\{f_1, \cdots, f_n\}$,*

$$\mathbb{E}\left[\left\| \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}} f_i - \frac{1}{n} \sum_{i=1}^n f_i \right\|^2\right] \leq \frac{1}{\mathsf{b}\, n} \sum_{i=1}^n \| f_i - \frac{1}{n} \sum_{j=1}^n f_j \|^2 \,.$$

3. 假设 A2。对于 s， $s' \in \mathcal{S}$，有

$$\mathbb{E}\left[\left\| \left\{\bar{\mathsf{h}}_{\mathcal{B}}(s) - \bar{\mathsf{h}}_{\mathcal{B}}(s')\right\} - \left\{\bar{\mathsf{h}}(s) - \bar{\mathsf{h}}(s')\right\} \right\|^2\right] \leq \frac{1}{\mathsf{b}} \left(L^2 \|s - s'\|^2 - \|\bar{\mathsf{h}}(s) - \bar{\mathsf{h}}(s')\|^2\right) \,.$$

Proof 证明思路与（Fort 等人，2020，引理 4）的证明相同。详细证明见第C.3节。□

引理 7.2 *Assume A4-item 1 and A4-item 3. For any $t \in [k^{\text{out}}]^\star$ and $k \in [k_t^{\text{in}} - 1]$, it holds*

$$\mathbb{E}\left[\eta_{t,k+1} | \mathcal{F}_{t,k+1/2}\right] = \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t,k+1}} \mu_{t,k+1,i} \,,$$

$$\mathbb{E}\left[\eta_{t,k+1} | \mathcal{F}_{t,k}\right] = \frac{1}{n} \sum_{i=1}^n \mu_{t,k+1,i} \,,$$

$$\mathbb{E}\left[\| \eta_{t,k+1} - \mathbb{E}\left[\eta_{t,k+1} | \mathcal{F}_{t,k}\right] \|^2 | \mathcal{F}_{t,k}\right] \leq \frac{1}{\mathsf{b}} \left(\frac{C_v}{M_{t,k+1}} + \frac{C_{vb}^2}{\bar{M}_{t,k+1}^2}\right) \,.$$

Proof 设 $t \in [k^{\text{out}}]^\star$ 和 $k \in [k_t^{\text{in}} - 1]$。我们有

$$\mathbb{E}\left[\eta_{t,k+1} | \mathcal{F}_{t,k+1/2}\right] = \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t,k+1}} \mu_{t,k+1,i} \,,$$

由于$\mathcal{B}_{t,\ k+1} \in \mathcal{F}_{t,\ k+1/2}$；并由引理 7.1，

$$\mathbb{E}\left[\eta_{t,k+1} | \mathcal{F}_{t,k}\right] = \frac{1}{n} \sum_{i=1}^n \mu_{t,k+1,i} \,.$$

记

$$\begin{aligned} &\eta_{t,k+1} - \mathbb{E}\left[\eta_{t,k+1} | \mathcal{F}_{t,k}\right] \\ &\quad = \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t,k+1}} \xi_{t,k+1,i} - \frac{1}{n} \sum_{i=1}^n \mu_{t,k+1,i} \\ &\quad = \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t,k+1}} \left\{\xi_{t,k+1,i} - \mu_{t,k+1,i}\right\} \\ &\quad + \frac{1}{\mathsf{b}} \sum_{i \in \mathcal{B}_{t,k+1}} \mu_{t,k+1,i} - \frac{1}{n} \sum_{i=1}^n \mu_{t,k+1,i} \,. \end{aligned}$$

右端具有 $U + V$ 的形式，我们记 $\|U+V\|^2 = \|U\|^2 + \|V\|^2 + 2\langle U, V\rangle$，其中 $U \leftarrow \mathsf{b}^{-1} \sum_{i \in \mathcal{B}_{t,\ k+1}} \left\{\xi_{t,k+1,i} - \mu_{t,k+1,i}\right\}$。根据条件化以及$\sigma_t^2$的定义，$k+1$，$i$，可得

$$\mathbb{E}\left[\|U\|^2 | \mathcal{F}_{t,k}\right] = \frac{1}{\mathsf{b}^2} \mathbb{E}\left[\sum_{i \in \mathcal{B}_{t,k+1}} \sigma_{t,k+1,i}^2 | \mathcal{F}_{t,k}\right] \,.$$

在条件 A4-item1下，由引理 7.1

$$\mathbb{E}\left[\|U\|^2 | \mathcal{F}_{t,k}\right] = \frac{1}{\mathsf{b}\, n} \sum_{i=1}^n \sigma_{t,k+1,i}^2 \leq \frac{C_v}{\mathsf{b}\, M_{t,k+1}} \,.$$

根据引理7.1 再次可得

$$\mathbb{E}\left[\|V\|^2 | \mathcal{F}_{t,k}\right] \leq \frac{1}{\mathsf{b}\, n} \sum_{i=1}^n \|\mu_{t,k+1,i} - \frac{1}{n} \sum_{j=1}^n \mu_{t,k+1,j}\|^2 \,,$$

由此可得

$$\mathbb{E}\left[\|V\|^2 | \mathcal{F}_{t,k}\right] \leq \frac{C_{vb}^2}{\mathsf{b}\, \bar{M}_{t,k+1}^2} \,.$$

最后，注意到 $\mathbb{E}\left[U | \mathcal{F}_{t,k+1/2}\right] = 0$ 和 $V \in \mathcal{F}_{t,\ k+1/2}$，可得

$$\mathbb{E}\left[\langle U, V\rangle | \mathcal{F}_{t,k}\right] = \mathbb{E}\left[\left\langle \mathbb{E}\left[U | \mathcal{F}_{t,k+1/2}\right], V \right\rangle | \mathcal{F}_{t,k}\right] = 0 \,.$$

至此，证明完毕。□

7.3 关于变量 $\mathsf{S}_{t,\ k}$的结果

命题7.3 研究了变量 $\mathsf{S}_{t,k+1}$ 的偏差。它表明，$\mathsf{S}_{t,\ k+1}$ 是 $\mathsf{h}(S_{t,k})$ 的有偏近似：

$$\mathbb{E}\left[\mathsf{S}_{t,k+1} | \mathcal{F}_{t,k}\right] \neq \bar{\mathsf{h}}(\widehat{S}_{t,k}).$$

¯ ^当 $k = 0$ 时，我们可能有 $\mathbb{E}[\mathsf{S}_{t,1}|\mathcal{F}_{t,0}] = \mathsf{h}(S_t,0)$^，若 $\delta_{t,0,i} = \mathsf{h}_i(S_t,0)$ 且 $\mathcal{B}_{t,0} = [n]^\star$。选择 $\mathcal{B}_{t,0} = [n]^\star$ 是 Wang 等人（2019）为 SpiderBoost 提出的策略；它的计算成本很高，但优点是在每个外循环开始时消除变量 $\mathsf{S}_\cdot$ 的偏差。在内循环中，会出现（带符号的）偏差。

命题 7.3 *For any $t \in [k^{\mathrm{out}}]^\star$ and $k \in [k_t^{\mathrm{in}} - 1]$, it holds*

$$\mathbb{E}\left[\mathsf{S}_{t,k+1}|\mathcal{F}_{t,k}\right] - \bar{\mathsf{h}}(\widehat{S}_{t,k}) = \mathsf{S}_{t,k} - \bar{\mathsf{h}}(\widehat{S}_{t,k-1}) + \mathbb{E}\left[\eta_{t,k+1}|\mathcal{F}_{t,k}\right] ,$$

and

$$\mathbb{E}\left[\mathsf{S}_{t,k+1} - \bar{\mathsf{h}}(\widehat{S}_{t,k})|\mathcal{F}_{t,0}\right] = \mathcal{E}_t + \sum_{j=1}^{k+1} \mathbb{E}\left[\eta_{t,j}|\mathcal{F}_{t,0}\right] .$$

Proof 令 $t \in [k^{\mathrm{out}}]^\star$ 且 $k \in [k_t^{\mathrm{in}} - 1]$。记^ ^ $\mathsf{S}_{t,k+1} = \mathsf{S}_{t,k} + h_{\mathcal{B}_{t,k+1}}(S_{t,k}) - h_{\mathcal{B}_{t,k+1}}(S_{t,k-1}) + \eta_{t,k+1}$，$\mathsf{S}_\cdot(\cdot)$ 由引理 7.1。

$$\mathbb{E}\left[\mathsf{S}_{t,k+1}|\mathcal{F}_{t,k}\right] = \mathsf{S}_{t,k} + \bar{\mathsf{h}}(\widehat{S}_{t,k}) - \bar{\mathsf{h}}(\widehat{S}_{t,k-1}) + \mathbb{E}\left[\eta_{t,k+1}|\mathcal{F}_{t,k}\right] .$$

由于 $\mathcal{F}_{t,0} \subseteq \mathcal{F}_{t,k}$，我们有

$$\mathbb{E}\left[\mathsf{S}_{t,k+1} - \bar{\mathsf{h}}(\widehat{S}_{t,k})|\mathcal{F}_{t,0}\right] = \mathbb{E}\left[\mathsf{S}_{t,k} - \bar{\mathsf{h}}(\widehat{S}_{t,k-1})|\mathcal{F}_{t,0}\right] + \mathbb{E}\left[\eta_{t,k+1}|\mathcal{F}_{t,0}\right] .$$

从 $j = 0$ 到 $j = k$ 求和可得

$$\mathbb{E}\left[\mathsf{S}_{t,k+1} - \bar{\mathsf{h}}(\widehat{S}_{t,k})|\mathcal{F}_{t,0}\right] = \mathsf{S}_{t,0} - \bar{\mathsf{h}}(\widehat{S}_{t,-1}) + \sum_{j=0}^{k} \mathbb{E}\left[\eta_{t,j+1}|\mathcal{F}_{t,0}\right] .$$

^ ^通过使用 $S_{t,0} = S_{t,-1}$ 以及 $\mathcal{E}_t$ 的定义，证明得证；注意 $\mathcal{E}_t \in \mathcal{F}_{t,0}$. □

命题 7.4 给出了 $\mathsf{S}_{t,k}$ 的条件方差的一个控制上界。

命题 7.4 *Assume 2, 4-item 1 and 4-item 3. For any $t \in [k^{\mathrm{out}}]^\star$ and $k \in [k_t^{\mathrm{in}} - 1]$ it holds*

$$\mathbb{E}\left[\left\|\mathsf{S}_{t,k+1} - \mathbb{E}\left[\mathsf{S}_{t,k+1}|\mathcal{F}_{t,k}\right]\right\|^2|\mathcal{F}_{t,k}\right] \le \frac{L^2}{\mathsf{b}}\left(1 + \frac{2C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k+1}}\right)\|\widehat{S}_{t,k} - \widehat{S}_{t,k-1}\|^2 + \frac{C_v}{\mathsf{b}\,M_{t,k+1}} + \frac{C_{vb}^2}{\mathsf{b}\,\bar{M}_{t,k+1}^2} + \frac{2C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k+1}} .$$

Proof 令 $t \in [k^{\mathrm{out}}]^\star$, $k \in [k_t^{\mathrm{in}} - 1]$。根据引理 7.1、命题 7.3，以及 $\mathsf{S}_{t,k+1}$ 和滤过 $\mathcal{F}_{t,k}$ 的定义，

$$\begin{aligned}
&\mathsf{S}_{t,k+1} - \mathbb{E}\left[\mathsf{S}_{t,k+1}|\mathcal{F}_{t,k}\right] \\
&= \mathsf{S}_{t,k+1} - \bar{\mathsf{h}}(\widehat{S}_{t,k}) - \mathsf{S}_{t,k} + \bar{\mathsf{h}}(\widehat{S}_{t,k-1}) - \mathbb{E}\left[\eta_{t,k+1}|\mathcal{F}_{t,k}\right] \\
&= \eta_{t,k+1} - \mathbb{E}\left[\eta_{t,k+1}|\mathcal{F}_{t,k}\right] \\
&+ \bar{\mathsf{h}}_{\mathcal{B}_{t,k+1}}(\widehat{S}_{t,k}) - \bar{\mathsf{h}}_{\mathcal{B}_{t,k+1}}(\widehat{S}_{t,k-1}) - \bar{\mathsf{h}}(\widehat{S}_{t,k}) + \bar{\mathsf{h}}(\widehat{S}_{t,k-1}) .
\end{aligned}$$

The RHS is of the form $U + V$ with $U \leftarrow \eta_{t,k+1} - \mathbb{E}\left[\eta_{t,k+1}|\mathcal{F}_{t,k}\right]$ and $V \in \mathcal{F}_{t,k+1/2}$. Then, we write $\mathbb{E}\left[\|U+V\|^2|\mathcal{F}_{t,k}\right] = \mathbb{E}\left[\|U\|^2|\mathcal{F}_{t,k}\right] + \mathbb{E}\left[\|V\|^2|\mathcal{F}_{t,k}\right] + 2\mathbb{E}\left[\left\langle \mathbb{E}\left[U|\mathcal{F}_{t,k+1/2}\right], V\right\rangle|\mathcal{F}_{t,k}\right]$.

项 $\mathbb{E}\left[\|V\|^2|\mathcal{F}_{t,k}\right]$ 可由引理 7.1 控制：其上界为 $L^2\mathsf{b}^{-1}\|\widehat{S}_{t,k} - \widehat{S}_{t,k-1}\|^2$。项 $\mathbb{E}\left[\|U\|^2|\mathcal{F}_{t,k}\right]$ 可由引理 7.2 控制：其上界为 $C_v/(\mathsf{b}M_{t,k+1}) + C_{vb}^2/(\mathsf{b}\,\bar{M}_{t,k+1}^2)$。注意到 $V \in \mathcal{F}_{t,k+1/2}$，并使用引理 7.2 和引理 7.1，可将标量积上界估计为

$$\begin{aligned}
&2\,\mathbb{E}\left[\|V\|\left\|\mathbb{E}\left[U|\mathcal{F}_{t,k+1/2}\right]\right\|\middle|\mathcal{F}_{t,k}\right] \\
&\le 2\frac{C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k+1}}\left\{\mathbb{E}\left[\|V\|^2\middle|\mathcal{F}_{t,k}\right]\right\}^{1/2} \\
&\le 2\frac{C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k+1}}\left(1 + \mathbb{E}\left[\|V\|^2|\mathcal{F}_{t,k}\right]\right) ,
\end{aligned}$$

其中我们用到了 $\mathsf{a} \le 1 + \mathsf{a}^2$。 □

命题 7.5 给出了二次误差 $\|\mathsf{S}_{t,k+1} - \bar{\mathsf{h}}(\widehat{S}_{t,k})\|^2$ 条件期望的上界。

命题 7.5 *Assume A 2 and A 4. For any $t \in [k^{\mathrm{out}}]^\star$ and $k \in [k_t^{\mathrm{in}} - 1]$, it holds*

$$\begin{aligned}
&\mathbb{E}\left[\|\mathsf{S}_{t,k+1} - \bar{\mathsf{h}}(\widehat{S}_{t,k})\|^2|\mathcal{F}_{t,k}\right] \\
&\le \left(1 + \frac{2C_b}{m_{t,k+1}}\right)\|\mathsf{S}_{t,k} - \bar{\mathsf{h}}(\widehat{S}_{t,k-1})\|^2 \\
&+ \frac{L^2}{\mathsf{b}}\left(1 + \frac{2C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k+1}}\right)\|\widehat{S}_{t,k} - \widehat{S}_{t,k-1}\|^2 \\
&+ \mathcal{U}_{t,k+1} ,
\end{aligned}$$

where

$$\mathcal{U}_{t,k} \stackrel{\mathrm{def}}{=} \frac{2C_b}{m_{t,k}} + \frac{C_b^2}{m_{t,k}^2} + \frac{C_v}{\mathsf{b}\,M_{t,k}} + \frac{2C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k}} + \frac{C_{vb}^2}{\mathsf{b}\,\bar{M}_{t,k}^2} .$$

Proof 令 $t \in [k^{\text{out}}]^\star$ 和 $k \in [k_t^{\text{in}} - 1]$。根据条件期望的定义，对于任意随机变量$U$，$V$以及满足 $V \in \mathcal{F}$ 的任意 σ-域 $\mathcal{F}$：

$$\mathbb{E}\left[\|U - V\|^2 | \mathcal{F}\right] = \mathbb{E}\left[\|U - \mathbb{E}[U|\mathcal{F}]\|^2 | \mathcal{F}\right] + \|\mathbb{E}[U|\mathcal{F}] - V\|^2 .$$

我们将此等式应用于 $U \leftarrow S_{t,k+1}$, $V \leftarrow \bar{\mathsf{h}}(S_{t,k})$ 和 $\mathcal{F} \leftarrow \mathcal{F}_{t,k}$。命题 7.4 控制第一项。对于第二项，根据命题 7.3、引理 7.2和 A4-第 2 项 可得

$$\begin{aligned}
&\|\mathbb{E}\left[\mathsf{S}_{t,k+1}|\mathcal{F}_{t,k}\right] - \bar{\mathsf{h}}(\widehat{S}_{t,k})\|^2 \\
&= \|\mathsf{S}_{t,k} - \bar{\mathsf{h}}(\widehat{S}_{t,k-1}) + \mathbb{E}\left[\eta_{t,k+1}|\mathcal{F}_{t,k}\right]\|^2 \\
&\le \|\mathsf{S}_{t,k} - \bar{\mathsf{h}}(\widehat{S}_{t,k-1})\|^2 + \frac{C_b^2}{m_{t,k+1}^2} \\
&+ 2\frac{C_b}{m_{t,k+1}}\|\mathsf{S}_{t,k} - \bar{\mathsf{h}}(\widehat{S}_{t,k-1})\| .
\end{aligned}$$

我们通过使用 $\|a\| \le 1 + \|a\|^2$ with $a \leftarrow \|\mathsf{S}_{t,k} - \bar{\mathsf{h}}(\widehat{S}_{t,k-1})\|$. □

推论 7.6 (命题 7.5 的推论) *Assume also item 3. For* $t \in [k^{\text{out}}]^\star$ *and* $k \in [k_t^{\text{in}} - 1]$ *define* $\mathcal{D}_{t,k+1} = \|\widehat{S}_{t,k+1} - \text{prox}^{B_{t,k}}_{\gamma_{t,k+1}g}(\widehat{S}_{t,k} + \gamma_{t,k+1}\bar{\mathsf{h}}(\widehat{S}_{t,k}))\|^2_{B_{t,k}}$, *and* $\mathcal{D}_{t,0} \overset{\text{def}}{=} 0$. *For* $t \in [k^{\text{out}}]^\star$ *and* $k \in [k_t^{\text{in}} -]$, *it holds*

$$\begin{aligned}
&\mathbb{E}\left[\|\mathsf{S}_{t,k+1} - \bar{\mathsf{h}}(\widehat{S}_{t,k})\|^2 | \mathcal{F}_{t,k}\right] \\
&\le \left(1 + \frac{2C_b}{m_{t,k+1}}\right)\|\mathsf{S}_{t,k} - \bar{\mathsf{h}}(\widehat{S}_{t,k-1})\|^2 \\
&+ \gamma_{t,k}^2 \frac{2}{v_{\min}}\frac{L^2}{\mathsf{b}}\left(1 + \frac{2C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k+1}}\right)\Delta^\star_{t,k} \\
&+ \frac{2}{v_{\min}}\frac{L^2}{\mathsf{b}}\left(1 + \frac{2C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k+1}}\right)\mathcal{D}_{t,k} \\
&+ \mathcal{U}_{t,k+1} .
\end{aligned}$$

By convention, $\Delta^\star_{t,0} \overset{\text{def}}{=} 0$.

Proof 证明在于对 $\|S_{t,k} - \widehat{S}_{t,k-1}\|^2$给出一个上界。设 $s \in \mathcal{S}$, H, $h \in \mathbb{R}^q$, $\gamma > 0$ 和 B 为一个 $q \times q$ 正定矩阵。对于任意 $\beta > 0$，有

$$\begin{aligned}
\|\text{prox}^B_{\gamma g}(s + \gamma H) - s\|^2_B &\le (1 + \frac{1}{\beta})\|\text{prox}^B_{\gamma g}(s + \gamma h) - s\|^2_B \\
&+ (1 + \beta)\|\text{prox}^B_{\gamma g}(s + \gamma H) - \text{prox}^B_{\gamma g}(s + \gamma h)\|^2_B .
\end{aligned}$$

我们将这些不等式应用于 $\gamma \leftarrow \gamma_{t,k}$, $B \leftarrow B_{t,k-1}$, $s \leftarrow \widehat{S}_{t,k-1}$, $H \leftarrow \mathsf{S}_{t,k}$ 以及 $h \leftarrow \bar{\mathsf{h}}(\widehat{S}_{t,k-1})$。然后，对于任意 $k > 0$,

$$\begin{aligned}
&\|\widehat{S}_{t,k} - \widehat{S}_{t,k-1}\|^2_{B_{t,k-1}} \\
&\le (1 + \beta^{-1})\gamma_{t,k}^2\Delta^\star_{t,k} + (1 + \beta)\mathcal{D}_{t,k} . \qquad (24)
\end{aligned}$$

我们选择 $\beta = 1$ 并通过 A-项：最小值 $\|\cdot\|^2_{B_{t,k-1}}$ 得出结论。

当 $k = 0$, $\|\widehat{S}_{t,k} - \widehat{S}_{t,k-1}\|^2_{B_{t,k-1}} = 0$ 时，由定义可知 $\widehat{S}_{t,0} = \widehat{S}_{t,-1}$。因此，(24)仍然成立，因为按约定 $\mathcal{D}_{t,0} = 0$ 且 $\Delta^\star_{t,0} = 0$。证明到此结束。□

推论 7.7 (推论 7.6的推论) *Let* $\{\rho_{t,k}, t \ge 1, k \ge 0\}$ *be a positive sequence satisfying*

$$\rho_{t,k+1}\left(1 + \frac{2C_b}{m_{t,k+1}}\right) \le \rho_{t,k} . \qquad (25)$$

For any $t \in [k^{\text{out}}]^\star$, $k \in [k_t^{\text{in}} - 1]$, *it holds*

$$\begin{aligned}
&\rho_{t,k+1}\mathbb{E}\left[\|\mathsf{S}_{t,k+1} - \bar{\mathsf{h}}(\widehat{S}_{t,k})\|^2 | \mathcal{F}_{t,0}\right] \\
&\le \rho_{t,0}\|\mathcal{E}_t\|^2 + \sum_{\ell=1}^{k+1}\rho_{t,\ell}\,\mathcal{U}_{t,\ell} + \frac{2}{v_{\min}}\frac{L^2}{\mathsf{b}}\cdots \\
&\times \left\{\sum_{\ell=1}^{k}\gamma_{t,\ell}^2\rho_{t,\ell+1}\left(1 + \frac{2C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,\ell+1}}\right)\mathbb{E}\left[\Delta^\star_{t,\ell}|\mathcal{F}_{t,0}\right]\right. \\
&\left. + \sum_{\ell=1}^{k}\rho_{t,\ell+1}\left(1 + \frac{2C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,\ell+1}}\right)\mathbb{E}\left[\mathcal{D}_{t,\ell}|\mathcal{F}_{t,0}\right]\right\} .
\end{aligned}$$

Proof 在推论7.6中，结论的形式为

$$\begin{aligned}
&\mathbb{E}\left[\|\mathsf{S}_{t,k+1} - \bar{\mathsf{h}}(\widehat{S}_{t,k})\|^2 | \mathcal{F}_{t,k}\right] \\
&\le \left(1 + \frac{2C_b}{m_{t,k+1}}\right)\|\mathsf{S}_{t,k} - \bar{\mathsf{h}}(\widehat{S}_{t,k-1})\|^2 + A_k .
\end{aligned}$$

利用条件（25），可得

$$\begin{aligned}
&\rho_{t,k+1}\mathbb{E}\left[\|\mathsf{S}_{t,k+1} - \bar{\mathsf{h}}(\widehat{S}_{t,k})\|^2 | \mathcal{F}_{t,k}\right] \\
&\le \rho_{t,k}\|\mathsf{S}_{t,k} - \bar{\mathsf{h}}(\widehat{S}_{t,k-1})\|^2 + \rho_{t,k+1}A_k .
\end{aligned}$$

利用 $\mathbb{E}[U|\mathcal{F}_{t,0}] = \mathbb{E}\left[\mathbb{E}[U|\mathcal{F}_{t,k}]|\mathcal{F}_{t,0}\right]$ 并从 $\ell = 0$ 求和至 $\ell = k$，可得

$$\begin{aligned}
&\rho_{t,k+1}\mathbb{E}\left[\|\mathsf{S}_{t,k+1} - \bar{\mathsf{h}}(\widehat{S}_{t,k})\|^2 | \mathcal{F}_{t,0}\right] \\
&\le \sum_{\ell=0}^{k}\rho_{t,\ell+1}\mathbb{E}\left[A_\ell|\mathcal{F}_{t,0}\right] + \rho_{t,0}\|\mathsf{S}_{t,0} - \bar{\mathsf{h}}(\widehat{S}_{t,-1})\|^2 ;
\end{aligned}$$

然后利用等式 $S_{t,-1} = S_{t,0}$ 以及 $\mathcal{E}_t$ 的定义即可得出结论。另请注意，$\Delta^\star_{t,0} = 0$ 以及 $\mathcal{D}_{t,0} = 0$。□

7.4 W、 g和 W $+g$的李雅普诺夫不等式

引理7.8，尽管这是光滑优化中的经典结果，此处仍予以陈述，以使内容自成一体。

引理 7.8 *Assume* A3. *For any* s， $s' \in \mathcal{S}$ *and* $\gamma > 0$,

$$\mathrm{W}(s') \le \mathrm{W}(s) - \left\langle \bar{\mathsf{h}}(s), s' - s \right\rangle_{B(s)} + \frac{L_{\dot{\mathrm{W}}}}{2}\|s' - s\|^2 .$$

Proof $\mathcal{S}$ 是凸的，因为它是某个凸函数的定义域。根据 A3，W 在 $\mathcal{S}$ 上连续可微，且其梯度为 L -W-利普希茨连续。因此，对于任意 s， $s' \in \mathcal{S}$,

$$\mathrm{W}(s') - \mathrm{W}(s) \le \left\langle \nabla \mathrm{W}(s), s' - s \right\rangle + \frac{L_{\dot{\mathrm{W}}}}{2}\|s' - s\|^2 .$$

¯我们利用 $\nabla \mathrm{W}(s) = -B(s)\bar{\mathsf{h}}(s)$，因此

$$\left\langle \nabla \mathrm{W}(s), s' - s \right\rangle = -\left\langle \bar{\mathsf{h}}(s), s' - s \right\rangle_{B(s)} .$$

□

引理 7 9 *Assume* 1. *Let* B *be a* $q \times q$ *positive*. *definite matrix. For any* $s \in \mathcal{S}$ $\gamma > 0$ H $h \in \mathbb{R}^q$ *and* A,,,$\beta > 0$,

$$\begin{aligned} g\left(\mathrm{prox}^B_{\gamma g}(s + \gamma H)\right) &\le g(s) \\ &- \frac{1}{\gamma}\left(1 - \frac{\beta}{4}\right)\|\mathrm{prox}^B_{\gamma g}(s + \gamma h) - s\|^2_B \\ &- \frac{1}{\gamma}(1 - \frac{1}{\beta})\|\mathrm{prox}^B_{\gamma g}(s + \gamma h) - \mathrm{prox}^B_{\gamma g}(s + \gamma H)\|^2_B \\ &- \left\langle h, s - \mathrm{prox}^B_{\gamma g}(s + \gamma h)\right\rangle_B \\ &+ \left\langle H, \mathrm{prox}^B_{\gamma g}(s + \gamma H) - \mathrm{prox}^B_{\gamma g}(s + \gamma h)\right\rangle_B . \end{aligned}$$

Proof 在此证明中，我们使用简写符号$\mathsf{p}_H \overset{\text{def}}{=}$ 近端算子$^B_{\gamma g}(s + \gamma H)$ 和 $\mathsf{p}_h \overset{\text{def}}{=}$ 近端算子$^B_{\gamma g}(s + \gamma h)$。根据引理 3.1 以及点处次微分的定义，可得

$$\begin{aligned} g(\mathsf{p}_h) &\ge g(\mathsf{p}_H) - \gamma^{-1}\left\langle \mathsf{p}_H - s - \gamma H, \mathsf{p}_h - \mathsf{p}_H\right\rangle_B \\ g(s) &\ge g(\mathsf{p}_h) - \gamma^{-1}\left\langle \mathsf{p}_h - s - \gamma h, s - \mathsf{p}_h\right\rangle_B . \end{aligned}$$

由此可得

$$\begin{aligned} g(\mathsf{p}_H) &\le g(s) - \gamma^{-1}\|\mathsf{p}_h - s\|^2_B - \left\langle h, s - \mathsf{p}_h\right\rangle_B \\ &\quad - \left\langle H, \mathsf{p}_h - \mathsf{p}_H\right\rangle_B + \gamma^{-1}\left\langle \mathsf{p}_H - s, \mathsf{p}_h - \mathsf{p}_H\right\rangle_B . \end{aligned}$$

对于最后一项，对任意 $\beta > 0$，我们有

$$\begin{aligned} &\gamma^{-1}\left\langle \mathsf{p}_H - s, \mathsf{p}_h - \mathsf{p}_H\right\rangle_B + \gamma^{-1}\|\mathsf{p}_h - \mathsf{p}_H\|^2_B \\ &= \gamma^{-1}\left\langle \mathsf{p}_h - s, \mathsf{p}_h - \mathsf{p}_H\right\rangle_B \\ &\le 2\left\langle (\mathsf{p}_h - s)\frac{\sqrt{\beta}}{2\sqrt{\gamma}}, (\mathsf{p}_h - \mathsf{p}_H)\frac{1}{\sqrt{\beta\gamma}}\right\rangle_B \\ &\le \frac{\beta}{4\gamma}\|\mathsf{p}_h - s\|^2_B + \frac{1}{\beta\gamma}\|\mathsf{p}_h - \mathsf{p}_H\|^2_B . \end{aligned}$$

证明到此结束。 □

命题 7.10 *Assume* A1, A2 *and*A3. *For anyt* $\in [k^{\text{out}}]^\star$, $k \in [k_t^{\text{in}} - 1]$ *and* $\beta > 0$,

$$\begin{aligned} &\mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k+1}) + g(\widehat{S}_{t,k+1})|\mathcal{F}_{t,0}\right] \\ &\le \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k}) + g(\widehat{S}_{t,k})|\mathcal{F}_{t,0}\right] \\ &- \gamma_{t,k+1}\left(1 - \frac{\beta}{4} - \frac{L_{\dot{\mathrm{W}}}\gamma_{t,k+1}}{v_{\min}}\right)\mathbb{E}\left[\Delta^\star_{t,k+1}|\mathcal{F}_{t,0}\right] \\ &- \frac{1}{\gamma_{t,k+1}}\left(1 - \frac{1}{\beta} - \frac{L_{\dot{\mathrm{W}}}}{v_{\min}}\gamma_{t,k+1}\right)\mathbb{E}\left[\mathcal{D}_{t,k+1}|\mathcal{F}_{t,0}\right] \\ &+ \gamma_{t,k+1}\mathbb{E}\left[\|\mathsf{S}_{t,k+1} - \bar{\mathsf{h}}(\widehat{S}_{t,k})\|^2_{B_{t,k}}|\mathcal{F}_{t,0}\right] , \end{aligned}$$

where $\mathcal{D}_{t,k+1}$ *is defined in Corollary 7.6.*

Proof 设 $\gamma > 0, s \in \mathcal{S}$ 和 $H \in \mathbb{R}^q$。应用引理 7.8，取 $s' \leftarrow$ 近端算子$^{B(s)}_{\gamma g}(s+\gamma H) \in \mathcal{S}$；并应用引理 7.9，取 $h \leftarrow$ ¯ $\bar{\mathsf{h}}(s)$。这对任意 $\beta > 0$,

$$\begin{aligned} &\mathrm{W}(\mathrm{prox}^B_{\gamma g}(s + \gamma H)) + g(\mathrm{prox}^B_{\gamma g}(s + \gamma H)) \\ &\le \mathrm{W}(s) + g(s) - \frac{1}{\gamma}(1 - \frac{\beta}{4})\|\mathrm{prox}^B_{\gamma g}(s + \gamma\bar{\mathsf{h}}(s)) - s\|^2_B \\ &- \frac{1}{\gamma}\left(1 - \frac{1}{\beta}\right)\|\mathrm{prox}^B_{\gamma g}(s+\gamma H) - \mathrm{prox}^B_{\gamma g}(s+\gamma\bar{\mathsf{h}}(s))\|^2_B \\ &- \left\langle \bar{\mathsf{h}}(s) - H, \mathrm{prox}^B_{\gamma g}(s + \gamma H) - \mathrm{prox}^B_{\gamma g}(s + \gamma\bar{\mathsf{h}}(s))\right\rangle_B \\ &\qquad + \frac{L_{\dot{\mathrm{W}}}}{2}\|\mathrm{prox}^B_{\gamma g}(s + \gamma H) - s\|^2 . \end{aligned}$$

由于近端算子$^B_{\gamma g}$ 是坚非扩张的（参见引理 6.1），标量积的上界为 $\gamma\|H - \bar{\mathsf{h}}(s)\|^2_B$。根据 A3-第 3 项，我们写成

$$\|\mathrm{prox}^B_{\gamma g}(s+\gamma H) - s\|^2 \le \frac{1}{v_{\min}}\|\mathrm{prox}^B_{\gamma g}(s+\gamma H) - s\|^2_B ;$$

然后我们利用 $\|a + b\|^2_B \le 2\|a\|^2_B + 2\|b\|^2_B$，其中 $a \leftarrow$ 近端算子$^B_{\gamma g}(s + \gamma H)$ − 近端算子$^B_{\gamma g}(s + \gamma$ ¯$\bar{\mathsf{h}}(s))$. 由此得到

$$\begin{aligned} &\frac{L_{\dot{\mathrm{W}}}}{2}\|\mathrm{prox}^B_{\gamma g}(s + \gamma H) - s\|^2 \\ &\le \frac{L_{\dot{\mathrm{W}}}}{v_{\min}}\|\mathrm{prox}^B_{\gamma g}(s + \gamma H) - \mathrm{prox}^B_{\gamma g}(s + \gamma\bar{\mathsf{h}}(s))\|^2_B \\ &+ \frac{L_{\dot{\mathrm{W}}}}{v_{\min}}\|\mathrm{prox}^B_{\gamma g}(s + \gamma\bar{\mathsf{h}}(s)) - s\|^2_B . \end{aligned}$$

^我们将这些不等式用于 $s \leftarrow S_{t,k}$, $\gamma \leftarrow \gamma_{t,k+1}$, $H \leftarrow \mathsf{S}_{t,k+1}$, $s' \leftarrow \widehat{S}_{t,k+1}$ 以及 $B \leftarrow B_{t,k}$。注意，近端算子 $\mathrm{prox}^B_{\gamma g}(s+\gamma H) = \widehat{S}_{t,k+1}$。证明至此完成。

□

定理 4.1 的 7.5 证明

设 $t \in [k^{\text{out}}]^{\star}$。设 $\mu \in (0,1)$。在整个证明过程中，令

$$A_{t,k+1} \stackrel{\text{def}}{=} \left(1 + \frac{2C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k+1}}\right) .$$

由推论 7.7 应用 $C_{t,\,k+1} \leftarrow \gamma_{t,\,k+1}$以及命题 7.10 应用 $\beta \leftarrow 4\mu$，对任意 $k \in [k_t^{\text{in}} - 1]$,

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k+1}) + g(\widehat{S}_{t,k+1})|\mathcal{F}_{t,0}\right] \\
&\le \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k}) + g(\widehat{S}_{t,k})|\mathcal{F}_{t,0}\right] \\
&- \gamma_{t,k+1}\left(1 - \mu - \frac{L_{\dot{\mathrm{W}}}}{v_{\min}}\gamma_{t,k+1}\right)\mathbb{E}\left[\Delta^{\star}_{t,k+1}|\mathcal{F}_{t,0}\right] \\
&- \frac{1}{\gamma_{t,k+1}}\left(1 - \frac{1}{4\mu} - \frac{L_{\dot{\mathrm{W}}}}{v_{\min}}\gamma_{t,k+1}\right)\mathbb{E}\left[\mathcal{D}_{t,k+1}|\mathcal{F}_{t,0}\right] \\
&+ \gamma_{t,0} v_{\max}\|\mathcal{E}_t\|^2 + v_{\max}\sum_{\ell=1}^{k+1}\gamma_{t,\ell}\mathcal{U}_{t,\ell} \\
&+ \frac{2v_{\max}}{v_{\min}}\frac{L^2}{\mathsf{b}}\sum_{\ell=1}^{k}\gamma_{t,\ell}^3 A_{t,\ell+1}\mathbb{E}\left[\Delta^{\star}_{t,\ell}|\mathcal{F}_{t,0}\right] \\
&+ \frac{2v_{\max}}{v_{\min}}\frac{L^2}{\mathsf{b}}\sum_{\ell=1}^{k}\gamma_{t,\ell+1} A_{t,\ell+1}\mathbb{E}\left[\mathcal{D}_{t,\ell}|\mathcal{F}_{t,0}\right] .
\end{aligned}$$

上文中，我们用到了 $\gamma_{t,\,k+1} \le \gamma_{t,\,\ell}$ 对任意 $\ell \in [k+1]$均成立。现在，我们从 $k = 0$ 求和至 $k = k_t^{\text{in}} - 1$。由此可得,

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k_t^{\text{in}}}) + g(\widehat{S}_{t,k_t^{\text{in}}})|\mathcal{F}_{t,0}\right] \\
&\quad\le \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,0}) + g(\widehat{S}_{t,0})|\mathcal{F}_{t,0}\right] \\
&- \sum_{k=1}^{k_t^{\text{in}}}\gamma_{t,k}\left(1 - \mu - \frac{L_{\dot{\mathrm{W}}}}{v_{\min}}\gamma_{t,k}\right)\mathbb{E}\left[\Delta^{\star}_{t,k}|\mathcal{F}_{t,0}\right] \\
&- \sum_{k=1}^{k_t^{\text{in}}}\frac{1}{\gamma_{t,k}}\left(1 - \frac{1}{4\mu} - \frac{L_{\dot{\mathrm{W}}}}{v_{\min}}\gamma_{t,k}\right)\mathbb{E}\left[\mathcal{D}_{t,k}|\mathcal{F}_{t,0}\right] \\
&+ \gamma_{t,0} v_{\max} k_t^{\text{in}}\|\mathcal{E}_t\|^2 \\
&+ v_{\max}\sum_{\ell=1}^{k_t^{\text{in}}}(k_t^{\text{in}} - \ell + 1)\gamma_{t,\ell}\mathcal{U}_{t,\ell} \\
&+ \frac{2v_{\max}}{v_{\min}}k_t^{\text{in}}\sum_{k=1}^{k_t^{\text{in}}-1}\gamma_{t,k}^3 A_{t,k+1}\mathbb{E}\left[\Delta^{\star}_{t,k}|\mathcal{F}_{t,0}\right] \\
&+ \frac{2v_{\max}}{v_{\min}}k_t^{\text{in}}\sum_{k=1}^{k_t^{\text{in}}-1}\gamma_{t,k+1} A_{t,k+1}\mathbb{E}\left[\mathcal{D}_{t,k}|\mathcal{F}_{t,0}\right] .
\end{aligned}$$

注意到，$\mathbb{E}\left[\Delta^{\star}_{t,k}|\mathcal{F}_{t,0}\right]$ 前面的系数为 $\gamma_{t,\,k}(1 - \mu - \Lambda_{t,\,k+1})$；而 $\mathbb{E}\left[\mathcal{D}_{t,k}|\mathcal{F}_{t,0}\right]$ 前面的项为 $\gamma_{t\,k}^{-1}$, $(1 - 1/(4\mu) - \Lambda_{t,\,k+1})$。根据对称性，我们选择 $\mu = 1/2$，使得$\mu = 1/(4\mu)$。由此得到

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,k_t^{\text{in}}}) + g(\widehat{S}_{t,k_t^{\text{in}}})|\mathcal{F}_{t,0}\right] \\
&\quad\le \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,0}) + g(\widehat{S}_{t,0})|\mathcal{F}_{t,0}\right] \\
&\quad - \sum_{k=1}^{k_t^{\text{in}}}\gamma_{t,k}\left(\frac{1}{2} - \Lambda_{t,k+1}\right)\mathbb{E}\left[\Delta^{\star}_{t,k}|\mathcal{F}_{t,0}\right] \\
&\quad - \sum_{k=1}^{k_t^{\text{in}}}\frac{1}{\gamma_{t,k}}\left(\frac{1}{2} - \Lambda_{t,k+1}\right)\mathbb{E}\left[\mathcal{D}_{t,k}|\mathcal{F}_{t,0}\right] \\
&\quad + \gamma_{t,0} v_{\max} k_t^{\text{in}}\|\mathcal{E}_t\|^2 \\
&\quad + v_{\max}\sum_{\ell=1}^{k_t^{\text{in}}}(k_t^{\text{in}} - \ell + 1)\gamma_{t,\ell}\mathcal{U}_{t,\ell} .
\end{aligned}$$

现在，我们对 $t = 1$ 求和至 $t = k^{\text{out}}$，并计算^期望。利用$S_{t+1,\,0} =\widehat{} S_{t,\,k^{\text{in}}}$，可得,

$$\begin{aligned}
&\sum_{t=1}^{k^{\text{out}}}\sum_{k=1}^{k_t^{\text{in}}}\gamma_{t,k}\left(\frac{1}{2} - \Lambda_{t,k+1}\right)\mathbb{E}\left[\Delta^{\star}_{t,k}\right] \\
&+ \sum_{t=1}^{k^{\text{out}}}\sum_{k=1}^{k_t^{\text{in}}}\frac{1}{\gamma_{t,k}}\left(\frac{1}{2} - \Lambda_{t,k+1}\right)\mathbb{E}\left[\mathcal{D}_{t,k}\right] \\
&\le \quad \mathbb{E}\left[\mathrm{W}(\widehat{S}_{t,0}) + g(\widehat{S}_{t,0})\right] \\
&\quad - \mathbb{E}\left[\mathrm{W}(\widehat{S}_{k^{\text{out}},k_{k^{\text{out}}}^{\text{in}}}) + g(\widehat{S}_{k^{\text{out}},k_{k^{\text{out}}}^{\text{in}}})\right] \\
&\quad + v_{\max}\sum_{t=1}^{k^{\text{out}}}\gamma_{t,0} k_t^{\text{in}}\mathbb{E}\left[\|\mathcal{E}_t\|^2\right] \\
&\quad + v_{\max}\sum_{t=1}^{k^{\text{out}}}\sum_{\ell=1}^{k_t^{\text{in}}}\left(k_t^{\text{in}} - \ell + 1\right)\gamma_{t,\ell}\mathcal{U}_{t,\ell} .
\end{aligned}$$

注意到$\mathbb{E}\left[\mathrm{W}(\widehat{S}_{k^{\text{out}},k_{k^{\text{out}}}^{\text{in}}}) + g(\widehat{S}_{k^{\text{out}},k_{k^{\text{out}}}^{\text{in}}})\right] \ge \min_{\mathcal{S}}(W + g)$，证明得以完成。

7.6 推论 4.3的证明

由于$\mathcal{U}_{t,\,k}=0$，我们有 $C_b=C_{vb}=0$。此外，$k_t^{\mathrm{in}}=k^{\mathrm{in}}$ 对于任意 t。因此，我们可以考虑一个常数步长序列$\gamma_{t,\,k}=\gamma_\star$ 其中 $\gamma_\star$满足（见（15）和（16））

$$\gamma_\star\frac{L_{\dot{\mathrm{W}}}}{v_{\min}}+\gamma_\star^2\frac{2v_{\max}}{v_{\min}}L^2\frac{k^{\mathrm{in}}}{\mathsf{b}}\in(0,1/2)\ .$$

通过选择可满足该条件

$$\frac{k^{\mathrm{in}}}{\mathsf{b}}\overset{\mathrm{def}}{=}\frac{1}{v_{\min}v_{\max}}\frac{L_{\dot{\mathrm{W}}}^2}{L^2}\ ,$$
$$\gamma_\star\overset{\mathrm{def}}{=}\frac{1}{4v_{\max}}\frac{L_{\dot{\mathrm{W}}}}{L^2}\frac{\mathsf{b}}{k^{\mathrm{in}}}=\frac{v_{\min}}{4L_{\dot{\mathrm{W}}}}\ .$$

这样的选择意味着 $\inf_{t,\,k}(1/2-\Lambda_{t,\,k})=1/2-3/8=1/8$。由于 $\mathcal{E}_t=\mathcal{U}_{t,\,k}=0$，根据推论 4.2 可得

$$\mathbb{E}\left[\Delta_{\tau,K}^\star+\mathcal{D}_{\tau,K}^\star\right]\le\frac{32\,L_{\dot{\mathrm{W}}}}{v_{\min}}\frac{\left(\mathbb{E}\left[\mathrm{W}(\widehat{S}_{1,0})+g(\widehat{S}_{1,0})\right]-\min_{\mathcal{S}}\left(\mathrm{W}+g\right)\right)}{k^{\mathrm{out}}k^{\mathrm{in}}}\ .$$

通过选择 $k^{\mathrm{out}}k^{\mathrm{in}}=O(L\cdot\mathrm{W}/(v_{\min}\,\epsilon))$，可满足近似的 ϵ-平稳条件。近端算子的调用次数为$k^{\mathrm{out}}k^{\mathrm{in}}$，因此 $\mathcal{K}_{\mathrm{prox}}=O(L\cdot\mathrm{W}/(v_{\min}\,\epsilon))$。最后，我们有 $\mathsf{b}'_t=n$，因此调用某个 $\bar{\mathsf{h}}_i$'的次数为 $k^{\mathrm{out}}n+2k^{\mathrm{out}}k^{\mathrm{in}}\mathsf{b}$。我们可以选择 $\mathsf{b}=O\left(\sqrt{n}\sqrt{v_{\min}v_{\max}}L/L_{\dot{\mathrm{W}}}\right)$。这给出 $k^{\mathrm{out}}=O(L\sqrt{v_{\max}}/(\sqrt{v_{\min}}\epsilon\sqrt{n}))$，以及 $\mathcal{K}_{\bar{\mathsf{h}}}=O(\sqrt{v_{\max}}L\sqrt{n}/(\epsilon\sqrt{v_{\min}}))$。

7.7 关于 h_i 的近似值成本

根据推论 4.3 得到的速率，令 $k_t^{\mathrm{in}}=O(\sqrt{n})$、$\mathsf{b}=O(\sqrt{n})$ 和 $k^{\mathrm{out}}=O(1/(\sqrt{n}\epsilon))$，并证明我们可以定义随机近似 $\delta_{t,0,i}$ 和 $\delta_{t,k+1,i}$，使得近似的 ϵ-驻点条件得以满足。

关于术语 $\mathbb{E}\left[\|\mathcal{E}_\tau\|^2\right]$。我们记作$\mathbb{E}\left[\|\mathcal{E}_\tau\|^2\right]=(k^{\mathrm{out}})^{-1}\sum_{t=1}^{k^{\mathrm{out}}}\mathbb{E}\left[\|\mathcal{E}_t\|^2\right]$ 和

$$\frac{1}{k^{\mathrm{out}}}\sum_{t=1}^{k^{\mathrm{out}}}\frac{\epsilon^{1-\mathsf{a}'}}{\sqrt{n}^{\mathsf{a}'}t^{\mathsf{a}'}}=\epsilon\,O(1)\ .$$

我们来计算在情形 $\mathcal{E}_t=n^{-1}\sum_{i=1}^n\{\delta_{t,0,i}-\bar{\mathsf{h}}_i(S_{t,0})\}$ 且 $\delta_{t,0,i}$ 等于由 $m_{t,0}$ 个独立同分布样本构成的蒙特卡罗和时相应的蒙特卡洛复杂度。则 $\mathbb{E}\left[\|\mathcal{E}_t\|^2\right]=n^{-1}m_t^{-1},_0O(1)$。当 $m_{t,0}=O(n^{\mathsf{a}'/2-1}t^{\mathsf{a}'}/\epsilon^{1-\mathsf{a}'})$ 时，它等于 $O(\epsilon^{1-\mathsf{a}'}/(\sqrt{n}t)^{\mathsf{a}'})$。因此，蒙特卡洛成本为

$$\sum_{t=1}^{k^{\mathrm{out}}}nm_{t,0}=O\left(\frac{1}{\sqrt{n}\epsilon^2}\right)\ .$$

关于项 $\mathbb{E}\left[\left(k^{\mathrm{in}}-K+1\right)\mathcal{U}_{\tau,K}\right]$。该项的上界为 $k^{\mathrm{in}}\,\mathbb{E}\left[\mathcal{U}_{\tau,K}\right]$，记为

$$k^{\mathrm{in}}\,\mathbb{E}\left[\mathcal{U}_{\tau,K}\right]\le\frac{1}{k^{\mathrm{out}}}\sum_{t=1}^{k^{\mathrm{out}}}\sum_{k=1}^{k^{\mathrm{in}}}O\left(\frac{1}{\mathsf{b}\,M_{t,k+1}}\right)\ .$$

右端为 $O(\epsilon)$。相应的蒙特卡洛复杂度为

$$2\mathsf{b}\sum_{t=1}^{k^{\mathrm{out}}}\sum_{k=1}^{k^{\mathrm{in}}}M_{t,k}=O\left(\frac{\sqrt{n}}{\epsilon^2}\right)\ ,$$

无论 $\mathsf{a}\in[0,1)$。

致谢。本研究部分由西蒙娜与奇诺·德尔·杜卡基金会通过 OpSiMorE 项目资助；并由法国国家研究署（ANR）通过 ANR-19-CE23 MASDOL 项目资助。

附录A 条件A4 蒙特卡洛情形

按照第3.4节中详细介绍的框架，假设*(i)* 难以求解的量$\widehat{\mathsf{h}}_i(S_{t,k},B_{t,k+1})$ 和 $\mathsf{h}_i(S_{t,k-1},B_{t,k})$ 的形式为

$$\mathsf{h}_i(s,B)=\int_{\mathsf{Z}}H_\vartheta(z)\pi_\vartheta(\mathrm{d}z)\ ,\qquad(\mathrm{A1})$$

其中 $\vartheta\overset{\text{定义为}}{=}(s,\ i,\ B)$; 且 *(ii)* 这些积分通过蒙特卡洛求和近似：设$\vartheta_{t,\,\ell+1,\,i}\overset{\text{定义为}}{=}(\widehat{S}_{t,\,\ell},\ i,\ B_{t,\,\ell+1})$ 并

$$\delta_{t,k+1,i} \stackrel{\text{def}}{=} \frac{1}{m_{t,k+1}} \sum_{r=1}^{m_{t,k+1}} \Big\{ H_{\vartheta_{t,k+1,i}}(Z_r^{\vartheta_{t,k+1,i}}) - H_{\vartheta_{t,k,i}}(Z_r^{\vartheta_{t,k,i}}) \Big\} , \quad \text{(A2)}$$

在给定$S_{t,\,k-1}$, $S_{t,\,k}$, $B_{t,\,k}$ 和 $B_{t,\,k+1}$的条件下，样本 $\{Z_r^{\vartheta_{t,\ell,i}}, r \geq 1\}$ 构成一个具有唯一平稳分布的马尔可夫链$\pi_{\vartheta_{t,\,\ell,\,i}}(\mathrm{d}z)$; $\ell \in \{k, k+1\}$。下文将说明 A4 在马尔可夫链足够遍历时成立。首先介绍马尔可夫链理论中的一些记号（例如参见Meynand Tweedie (1993)）。

设 P 为作用于可测空间 $(\mathsf{Z}\ \mathcal{Z})$ 上的转移核，且 λ, π 为 $(\mathsf{Z}\ \mathcal{Z})$ 上的概率测度。对于可测函数 $\xi\colon \mathsf{Z} \to [0, +\infty)$，定义

$$\pi(\xi) \stackrel{\text{def}}{=} \int_{\mathsf{Z}} \xi(z)\, \pi(\mathrm{d}z) .$$

对于任意 $r \in \mathbb{N}$, r-迭代转移核 P^r通过归纳法定义为:

$$P^{r+1}(z, A) \stackrel{\text{def}}{=} \int_{\mathsf{Z}} P^r(z, \mathrm{d}y)\, P(y, A) = \int_{\mathsf{Z}} P(z, \mathrm{d}y)\, P^r(y, A) ,$$

对所有 $z \in \mathsf{Z}$ $A \in \mathcal{Z}$成立；约定 $P^0(z,\ A)$ 定义为 $=\chi_A(z)$, 即取值于 $\{0, 1\}$ 的示性函数，且$P^0(z,\ A) = \delta_z(A)$，即零点处的狄拉克测度。给定 $(\mathsf{Z}\ \mathcal{Z})$ 上的概率测度 λ， λP 表示定义在 $(\mathsf{Z}\ \mathcal{Z})$ 上、由下式给出的概率测度

$$\lambda P(A) \stackrel{\text{def}}{=} \int_{\mathsf{Z}} \lambda(\mathrm{d}y) P(y, A) , \qquad \forall A \in \mathcal{Z} .$$

对于满足$\lambda P^r(U) + \pi(U) < +\infty$的函数 $U : \mathsf{Z} \to [1, +\infty)$，定义可测函数 $\xi : \mathsf{Z} \to \mathbb{R}^q$ 的 U-范数

$$\|\xi\|_U \stackrel{\text{def}}{=} \sup_{\mathsf{Z}} \frac{\|\xi\|}{U} ;$$

以及符号测度 $\lambda P^r - \pi$ 的 U-范数

$$\|\lambda P^r - \pi\|_U \stackrel{\text{def}}{=} \sup_{\xi : \|\xi\|_U \leq 1} \|\lambda P^r(\xi) - \pi(\xi)\| .$$

我们回到验证A4的充分条件。记 P_ϑ 为一个马尔可夫转移核，其不变分布 $\pi_\vartheta(\mathrm{d}z)$：在第 $(t, k+1)$ 次迭代时，在给定 $(S_{t,k-1}, S_{t,k}, B_{t,k}, B_{t,k+1})$ 的条件下，链 $\{Z_r^{\vartheta_{t,k}}, r \geq 0\}$ 和 $\{Z_r^{\vartheta_{t,k+1}}, r \geq 0\}$ 分别是转移核为 $P_{\vartheta_{t,k}}$ 和 $P_{\vartheta_{t,k+1}}$ 的马尔可夫链。它们具有相同的初始值 λ。假设

A5. *1.* 存在一个可测函数 $U : \mathsf{Z} \to [1, +\infty)$，使得

$$H_\star \stackrel{\text{def}}{=} \sup_{(s,i,B) \in \mathcal{S} \times [n]^\star \times \mathcal{P}_+^q} \|H_\vartheta\|_U < +\infty ,$$

其中 H_ϑ 定义为 （A1）。

2. 存在一个函数 $\rho : \mathbb{N} \to [0, 1]$ 和一个正常数 C_{MC}，使得对于任意 $r \in \mathbb{N}$,

$$\sup_{\vartheta \in \mathcal{S} \times [n]^\star \times \mathcal{P}_+^q} \|\lambda P_\vartheta^r - \pi_\vartheta\|_U \leq C_{\mathsf{MC}}\, \rho(r) .$$

此外， $\sum_{r \geq 1} \rho\ (r) < +\infty$。

3. 设 $\vartheta \in \mathcal{S} \times [n]^\star \times \mathcal{P}_+^q$。设 $\{Z_r^\vartheta, r \geq 1\}$ 为一个转移核为 P_ϑ 、初始分布为 λ 的马尔可夫链。存在一个正常数 C'_{MC}，使得对于任意 $\vartheta \in \mathcal{S} \times [n]^\star \times \mathcal{P}_+^q$ 和 $m' \in \mathbb{N}_\star$,

$$\mathbb{E}\left[\Big\| \sum_{r=1}^{m'} \{H_\vartheta(Z_r^\vartheta) - \pi_\vartheta(H_\vartheta)\} \Big\|^2 \right] \leq H_\star^2\, C'_{\mathsf{MC}}\, m' .$$

假设 5-条件 2 是关于s 遍历性的条件。关于该条件的充分条件，在几何速率 Fort 等人, 2011，引理 2.3）中给出，适用于 $\rho(r) = \kappa^r$ 的情形，其中 $\kappa \in (0,\ 1)$。通过调整（Andrieu 等人, 2015，定理 1），在次几何速率 $\rho(r)$ 的情形下也可得到类似条件。关于假设 A5-条件 3 的充分条件，可通过对（Fort 和 Moulines, 2003, 命题 12）作简单调整得到。

我们证明如下结果。

命题 A.1 *Assume* **A** *5. Let* $\delta_{t,k+1,i}$ *be given by* **(A2)** *where conditionally to* $(\widehat{S}_{t,k-1}, \widehat{S}_{t,k}, B_{t,k}, B_{t,k+1})$ $\{Z_r^{\vartheta_{t,\ell,i}}, r \geq 0\}$ *is a Markov chain with transition kernel* $P_{\vartheta_{t\,\ell\,i}}$ *and initial distribution* λ *for* $\ell \in \{k, k+1\}$. *Then 4 is* **A** *verified with* $m_{t,k+1} = M_{t,k+1} = M_{t,k+1} \leftarrow m_{t,k+1}$,

$C_b = C_{vb} \stackrel{\text{def}}{=} 2\,H_\star\,C_{\mathsf{MC}}\sum_{r\geq 1}\rho(r)$ *and* $C_v \stackrel{\text{def}}{=} 2H_\star^2\,C'_{\mathsf{MC}}$.

Proof 我们将使用以下记号

$P_{\ell,i} \stackrel{\text{def}}{=} P_{\vartheta_{t,\ell+1,i}}$, $\pi_{\ell,i} \stackrel{\text{def}}{=} \pi_{\vartheta_{t,\ell+1,i}}$, $H_{\ell,i} \stackrel{\text{def}}{=} H_{\vartheta_{t,\ell+1,i}}$.

- **μ_t的表达式**, $k{+}1$。我们有

$$\mu_{t,k+1,i} \stackrel{\text{def}}{=} \frac{1}{m_{t,k+1}}\sum_{r=1}^{m_{t,k+1}}\left(\lambda P_{k,i}^r H_{k,i} - \pi_{k,i}(H_{k,i})\right) - \frac{1}{m_{t,k+1}}\sum_{r=1}^{m_{t,k+1}}\left(\lambda P_{k-1,i}^r H_{k-1,i} - \pi_{k-1,i}(H_{k-1,i})\right).$$

- **条件 A4-项 2**。根据 A5，且由于 $\curvearrowright S_\bullet \in \mathcal{S}$，我们写作

$$\sup_{k,i}\|\lambda P_{k,i}^r H_{k,i} - \pi_{k,i}(H_{k,i})\| \leq H_\star\,C_{\mathsf{MC}}\,\rho(r).$$

这意味着

$$\|\mu_{t,k+1,i}\| \leq 2\,H_\star\,C_{\mathsf{MC}}\,\frac{1}{m_{t,k+1}}\sum_{r=1}^{m_{t,k+1}}\rho(r).$$

由于 $\sum_r \rho(r) < \infty$，右端具有 $C_b/m_{t,\ k+1}$ 的形式，其中 $C_b \stackrel{\text{定义}}{=} 2\,H_\star\,C_{\mathsf{MC}}\sum_r \rho(r)$。

- **条件 A4 项 3**。记作

$$\sigma^2_{t,k+1,i} \leq \mathbb{E}\left[\|\xi_{t,k+1,i}\|^2|\mathcal{P}_{t,k+1/2}\right].$$

于是，我们有

$$\mathbb{E}\left[\|\xi_{t,k+1,i}\|^2|\mathcal{P}_{t,k+1/2}\right] \leq 2\sup_{s,i}\mathbb{E}\left[\|\frac{1}{m_{t,k+1}}\sum_{r=1}^{m_{t,k+1}}H_{s,i}(Z_r^{s,i}) - \pi_{s,i}(H_{s,i})\|^2\right]$$

且右端的上界为 $2H_\star^2\,C'_{\mathsf{MC}}/m_{t,\ k+1}$ 由 A5 项 3。

我们还有

$$\frac{1}{n}\sum_{i=1}^n\|\mu_{t,k+1,i} - \frac{1}{n}\sum_{j=1}^n\mu_{t,k+1,j}\|^2 \leq \frac{1}{n}\sum_{i=1}^n\|\mu_{t,k+1,i}\|^2.$$

根据上文对 $\|\mu_{t,\ k+1,\ i}\|$ 的上界，我们有

$$\|\mu_{t,k+1,i}\|^2 \leq \frac{C_b^2}{m_{t,k+1}^2}.$$

证明到此结束。 □

附录 B 第 5 节的补充材料

B.1 惩罚对数似然准则

假设这些观测值相互独立，因此对数似然为

$$\theta \mapsto \sum_{i=1}^n \log\int_{\mathbb{R}^d}(1+\exp(-y_i\,\langle X_i, z_i\rangle))^{-1} \times \frac{1}{\sqrt{2\pi}^d\sigma^d}\exp\left(-(2\sigma^2)^{-1}\|z_i-\theta\|^2\right)\mathrm{d}z_i.$$

惩罚项为 $-n\tau\|\theta\|^2$。

引理 B.1 *The sum of the log-likelihood and the penalty term is equal to*

$$-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\theta^\top\sum_{i=1}^n\frac{X_iX_i^\top}{\|X_i\|^2}\theta - n\tau\|\theta\|^2 + \sum_{i=1}^n\log\int_{\mathbb{R}}\frac{\exp\left(x\,\langle X_i,\theta\rangle/(\sigma^2\|X_i\|)\right)}{1+\exp(-y_i\|X_i\|x)}\exp(-\frac{x^2}{2\sigma^2})\mathrm{d}x.$$

Proof 令 $i \in [n]^\star$。定义一个正交 $d\times d$ 矩阵 Q，其列记为 $(Q_1, \cdots, Q_d)$，使得 Q_1 定义为$= X_i/\|X_i\|$。我们有 $\langle z_i, X_i\rangle = \|X_i\|\,\langle Q_1, z_i\rangle$, $\langle z_i,\theta\rangle = \left\langle Q^\top z_i, Q^\top\theta\right\rangle$ 且 $\|z_i\|^2 = \|Q^\top z_i\|^2$。这意味着

$$-y_i\,\langle z_i, X_i\rangle = -y_i\|X_i\|\,\langle Q_1, z_i\rangle$$
$$\|z_i-\theta\|^2 = \|\theta\|^2 + \|Q^Tz_i\|^2 - 2\left\langle Q^\top z_i, Q^\top\theta\right\rangle$$

因此，观测值 Y_i 的对数似然为（除加性常数 C_1 定义为$= -d\ln\sigma - (d/2)\ln(2\pi)$ 外）

$$y_i \mapsto -\frac{\|\theta\|^2}{2\sigma^2} + \log\int_{\mathbb{R}^d}(1+\exp(-y_i\|X_i\|\,\langle Q_1, z_i\rangle))^{-1} \times\exp\left(-(2\sigma^2)^{-1}\left(\|Q^Tz_i\|^2 - 2\left\langle Q^\top z_i, Q^\top\theta\right\rangle\right)\right)\mathrm{d}z_i.$$

通过变量替换 $v = (v_1, \cdots, v_q) \leftarrow Q^\top z_i$，积分的对数等于

$$\log\int_{\mathbb{R}^d}\frac{\exp\left(-(2\sigma^2)^{-1}\left(\|v\|^2 - 2\left\langle v, Q^\top\theta\right\rangle\right)\right)}{1+\exp(-y_i\|X_i\|v_1)}\mathrm{d}v = \log\int_{\mathbb{R}}\frac{\exp\left(-(2\sigma^2)^{-1}\left(v_1^2 - 2v_1\,\langle Q_1,\theta\rangle\right)\right)}{1+\exp(-y_i\|X_i\|v_1)}\mathrm{d}v_1$$

$$+\sum_{u=2}^{d}\log\int_{\mathbb{R}}\exp\left(-\frac{1}{2\sigma^2}\{v_u^2-2v_u\langle Q_u,\theta\rangle\}\right)\mathrm{d}v_u\,.$$

最后的（$d-1$）个积分有闭式。注意到 $v_u^2-2v_u\langle Q_u,\theta\rangle=(v_u-\langle Q_u,\theta\rangle)^2-(\langle Q_u,\theta\rangle)^2$，因此除加性常数 C_2 定义=$(d-1)\{$对数$(2\pi)/2+$对数$\sigma\}$ 外

$$\begin{aligned}&\sum_{u=2}^{d}\log\int_{\mathbb{R}}\exp\left(-\frac{1}{2\sigma^2}\{v_u^2-2v_u\langle Q_u,\theta\rangle\}\right)\mathrm{d}v_u\\&=\sum_{u=2}^{d}\frac{(\langle Q_u,\theta\rangle)^2}{2\sigma^2}=\frac{\|\theta\|^2-(\langle Q_1,\theta\rangle)^2}{2\sigma^2}\\&=\frac{\|\theta\|^2-(\langle X_i,\theta\rangle)^2/\|X_i\|^2}{2\sigma^2}\,.\end{aligned}$$

证明至此结束；关于 θ 的常数等于 C_1+C_2。□

B.2 引理5.1 的证明

准则 F 等于 $-\mathcal{L}(\theta)-\log(2\pi\sigma^2)/2$，其中 $\mathcal{L}(\theta)$ 是归一化的惩罚对数似然。

似然是概率的乘积，取值在 (0，1) 内；因此，其对数为负。惩罚对数似然的上界为 $-\mathrm{pen}(\theta)=-n\tau\|\theta\|^2$。归一化惩罚对数似然的上界为 $-\tau\|\theta\|^2$。因此，准则的下界为 $\tau\|\theta\|^2-\ln(2\pi\sigma^2)/2$。

另一方面，准则的最小值小于其在$\theta=0$处的值。我们来证明该值为 $(\ln 4)/n-\ln(2\pi\sigma^2)/2$。这将说明准则的极小化子属于集合 $\{\theta\in\mathbb{R}^d:\tau\|\theta\|^2\le\ln 4\}$，从而完成证明。

我们有 $\mathrm{pen}(0)=0$。下面给出在 $\theta=0$处观测值 $Y_i=+1$的似然下界。该似然等于

$$\frac{1}{\sqrt{2\pi}^d\sigma^d}\int_{\mathbb{R}^d}\frac{\exp\left(-(2\sigma^2)^{-1}\|z_i\|^2\right)}{1+\exp(-\langle X_i,z_i\rangle)}\mathrm{d}z_i\,.$$

通过使用与引理证明中相同的变量替换B.1，可得

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}}\frac{\exp\left(-x^2/(2\sigma^2)\right)}{1+\exp(-\|X_i\|x)}\mathrm{d}x\left(\frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}}\exp\left(-x^2/(2\sigma^2)\right)\mathrm{d}x\right)^{d-1},$$

并且其下界为（注意，$(d-1)$ 个相同的积分等于 1）

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}^+}\frac{\exp\left(-x^2/(2\sigma^2)\right)}{1+\exp(-\|X_i\|x)}\mathrm{d}x\,,$$

由于对所有 $x\ge 0$都有 $1+\exp(-\|X_i\|x)\le 2$，因此该值进一步下界为 $1/4$。

情形 $Y_i=-1$ 的证明与此类似，故略去。

这意味着 n 变量的似然下界为 $1/4^n$；归一化对数似然的下界为 $-\ln 4$；准则的上界为 $\ln 4-\ln(2\pi\sigma^2)/2$。

B.3 将优化问题视为期望最大化

我们已证明，对于任意 $\theta\in\mathbb{R}^d$，$\nabla F(\theta)=n^{-1}\sum_{i=1}^n G_i(\theta)$，其中

$$G_i(\theta)\overset{\mathrm{def}}{=}2U\theta-\frac{X_i}{\sigma^2\|X_i\|}\int_{\mathbb{R}}z\,\pi_{\theta,i}(z)\mathrm{d}z\,,$$

且 $\pi_{\theta,\,i}(z)$ 是与 (20) 成正比的概率密度。

由 ϕ，ψ和 $S(Y_i,\,z)$ 的表达式可得，根据命题 2.1 定义的 T 为 $\mathsf{T}(s)$ 定义为 $=U^{-1}s/2$，对任意 $s\in\mathbb{R}^d$ 均成立。这意味着，对于任意 $s\in\mathbb{R}^d$，

$$\begin{aligned}\mathsf{h}_i(s,B)&\overset{\mathrm{def}}{=}\int_{\mathbb{R}}S(Y_i,z)\,\pi_{\mathsf{T}(s),i}(z)\,\mathrm{d}z-s\\&=\frac{X_i}{\sigma^2\|X_i\|}\int_{\mathbb{R}}z\,\pi_{Bs,i}(z)\,\mathrm{d}z-s\,.\end{aligned}$$

对于任意 $s\in\mathbb{R}^d$，我们来求满足 $\nabla(F\circ\mathsf{T})(s)=-n^{-1}\sum_{i=1}^n\mathsf{B}(s)\mathsf{h}_i(s)$ 的矩阵 $\mathsf{B}(s)$（见（6））。我们有 $\nabla(F\circ\mathsf{T})=\nabla F(B\cdot)=B^\top(\nabla F)(B\cdot)=B(\nabla F)(B\cdot)$。由此得到

$$\begin{aligned}\nabla(F\circ\mathsf{T})(s)&=B\frac{1}{n}\sum_{i=1}^n G_i(Bs)\\&=B\frac{1}{n}\sum_{i=1}^n\left(2UB\theta-\frac{X_i}{\sigma^2\|X_i\|}\int_{\mathbb{R}}z\,\pi_{Bs,i}(z)\mathrm{d}z\right)\\&=Bs-B\frac{1}{n}\sum_{i=1}^n\frac{X_i}{\sigma^2\|X_i\|}\int_{\mathbb{R}}z\,\pi_{Bs,i}(z)\mathrm{d}z\end{aligned}$$

$$= -B\frac{1}{n}\sum_{i=1}^{n}\mathsf{h}_i(s)\ .$$

这就得到 $\mathsf{B}(s) \stackrel{\text{def}}{=} B$，对任意 $s \in \mathbb{R}^d$均成立。

B.4 引理5.2 的证明

设 $i \in [n]^\star$ 和 $\theta \in \mathbb{R}^d$。**步骤 1.** 利用

$$\begin{aligned}-z^2/(2\sigma^2) + z\left\langle X_i, \theta\right\rangle/(\sigma^2\|X_i\|) \\ = -(z - \left\langle X_i, \theta\right\rangle/\|X_i\|)^2/(2\sigma^2) \\ + (\left\langle X_i, \theta\right\rangle)^2/(2\sigma^2\|X_i\|^2)\ ,\end{aligned}$$

我们写作

$$\pi_{\theta,i}(z) = \frac{\exp\left(-\left(z - \left\langle X_i, \theta\right\rangle/\|X_i\|\right)^2/(2\sigma^2)\right)}{Z_{\theta,i}\ \ 1 + \exp(-y_i\|X_i\|z)}$$

其中 $Z_{\theta,\ i}$ 是归一化常数。其次，我们使用 $z\,\pi_{\theta,\ i}(z) = (z - a_i)\pi_{\theta,\ i}(z) + a_i\,\pi_{\theta,\ i}(z)$，结合$a_i \leftarrow \left\langle X_i, \theta\right\rangle/\|X_i\|$，并且由于 $\int_{\mathbb{R}}\pi_{\theta,\ i}(z)\mathrm{d}z = 1$，我们得到

$$\mathcal{I}_i(\theta) = \frac{\left\langle X_i, \theta\right\rangle}{\|X_i\|} + \int_{\mathbb{R}}\left(z - \frac{\left\langle X_i, \theta\right\rangle}{\|X_i\|}\right)\pi_{\theta,i}(z)\,\mathrm{d}z\ .$$

最后，右端的积分具有如下形式

$$\sigma^2\int_{\mathbb{R}}\frac{f'(z)}{Z_{\theta,i}\ \ 1 + \exp(-y_i\|X_i\|z)}\mathrm{d}z$$

with

$$f(z) \stackrel{\text{def}}{=} -\exp\left(-\left(z - \left\langle X_i, \theta\right\rangle/\|X_i\|\right)^2/(2\sigma^2)\right)\ ,$$

我们使用分部积分。注意到 $z \mapsto 1/(1 + \exp(-y_i\|X_i\|z))$ 的导数为

$$y_i\|X_i\|\frac{\exp(-y_i\|X_i\|z)}{(1 + \exp(-y_i\|X_i\|z))^2}\ ,$$

我们写作

$$\begin{aligned}&\int_{\mathbb{R}}\frac{f'(z)}{1 + \exp(-y_i\|X_i\|z)}\mathrm{d}z \\ &= -y_i\|X_i\|\int f(z)\frac{\exp(-y_i\|X_i\|z)}{(1 + \exp(-y_i\|X_i\|z))^2}\mathrm{d}z\ .\end{aligned}$$

因此，第一步的结论是

$$\begin{aligned}\mathcal{I}_i(\theta) = \left\langle\frac{X_i}{\|X_i\|}, \theta\right\rangle \\ + y_i\|X_i\|\sigma^2\int_{\mathbb{R}}\frac{\pi_{\theta,i}(z)}{1 + \exp\left(y_i\|X_i\|z\right)}\mathrm{d}z\ .\end{aligned}$$

步骤 2。 这一步在马尔可夫链蒙特卡罗文献中是经典内容（例如，参见 ChoiandHobert(2013) 及其中的参考文献）。我们证明，对于任意 $z \in \mathbb{R}$,

$$\pi_{\theta,i}(z) = \int_0^{+\infty}\bar{\pi}_{\theta,i}(z,\omega)\mathrm{d}\omega\ .$$

根据（Polson 等人，2013，定理 1），有

$$\begin{aligned}\frac{1}{1 + \exp\left(-y_i\|X_i\|z\right)} = \frac{1}{2}\exp\left(y_i\|X_i\|z/2\right) \\ \times\int_0^{+\infty}\exp(-\omega\|X_i\|^2z^2/2)\mathsf{p}(\omega;\ 1)\mathrm{d}\omega\ ,\end{aligned}$$

其中 $\mathsf{p}(\omega;\ b)\mathrm{d}\omega$ 是参数为 b 的 Polya-Gamma 分布。这意味着 $\pi_{\theta,\ i}(z)$ 等于

$$\begin{aligned}&\exp\left(\frac{y_i\|X_i\|z}{2} - \frac{(z - \left\langle X_i, \theta\right\rangle/\|X_i\|)^2}{2\sigma^2}\right) \\ &\times\frac{1}{2Z_{\theta,i}}\int_0^{+\infty}\exp(-\omega\|X_i\|^2z^2/2)\mathsf{p}(\omega;\ 1)\mathrm{d}\omega\ .\end{aligned}$$

证明至此完成。

B.5 假设 A4 得到验证。

定义具有如下密度的马尔可夫核

$$P_{s,i}(z;\ z') \stackrel{\text{def}}{=} \left(\int_0^{\infty}\pi_2(\omega|z;\ i)\,\pi_1(z'|\omega;\ s, i)\ \mathrm{d}\omega\right)$$

相对于 $\mathbb{R}$ 上的勒贝格测度；其中，$\pi_1(z'|\omega;\ \ s,\ \ i)$ 是高斯分布的密度，其期望为 $\mathsf{m}_{s,\ i}(\omega)$，方差为 $v_i(\omega)$，定义为

$$\begin{aligned}v_i(\omega) &\stackrel{\text{def}}{=} \frac{\sigma^2}{1 + \omega\sigma^2\|X_i\|^2}\ , \\ \mathsf{m}_{s,i}(\omega) &\stackrel{\text{def}}{=} v_i(\omega)\left(\frac{1}{\sigma^2}\left\langle\frac{X_i}{\|X_i\|}, Bs\right\rangle + \frac{1}{2}y_i\|X_i\|\right)\ ;\end{aligned}$$

和 $\pi_2(\omega|z;\ \ i)$ 是参数为 $(1,\ \ \|X_i\|z)$ 的 Polya-Gamma 分布。由引理 5.2 所描述、以密度为目标的吉布斯核

分布 $\bar{\pi}_{Bs,i}(z,\omega)\mathrm{d}z\mathrm{d}\omega$，生成马尔可夫链 $\{(Z^{s,i}_r, \Omega^{s,i}_r), r \geq 0\}$，使得边缘$\{Z^{s,i}_r, r \geq 0\}$ 是一个转移核为 $P_{s,i}(z;\ z')\,\mathrm{d}z'$ 的马尔可夫链。我们应用 (Choi and Hobert, 2013, 命题 3.1) 的结果，

$y \in \{0,1\}$	$y_i \in \{-1,1\}$
n	1
$\Omega(\omega)$	ω
X	$\|X_i\|$
B	σ^2
b	$\langle X_i, Bs\rangle / \|X_i\|$

表 B1 [左] Choi and Hobert (2013)。 [右] 本文中的记号。

由此可得

$$\int_A P_{s,i}(z;\ z')\mathrm{d}z' \geq \varepsilon \int_A \exp(-0.5(x-\mathsf{m}_\star)^2/v_\star^2)\,\mathrm{d}x \qquad \text{(B3)}$$

其中

$$\varepsilon \stackrel{\text{def}}{=} \inf_{s\in\mathcal{S}, i\in[n]^\star} \frac{\exp\left(-\frac{1}{4} - \frac{\{\mathsf{m}_{s,i}(1/2)\}^2\,\sigma^2\|X_i\|^2}{4\,v_i(1/2)}\right)}{2\sqrt{1+\sigma^2\|X_i\|^2/2}}$$

且 $(\mathsf{m}_\star,\ v_\star)$ 对任意 $x \in \mathbb{R}$ 均满足，

$$\inf_{s\in\mathcal{S}, i\in[n]^\star} \exp(-0.5\,(x-\mathsf{m}_{s,i}(1/2))^2/v_i(1/2)) \geq \exp(-0.5(x-\mathsf{m}_\star)^2/v_\star^2)\,.$$

引理 B.2 *Since $\mathcal{S}$ is bounded, then $\varepsilon > 0$ and $\mathsf{m}_\star$, $v_\star$ exist in $\mathbb{R}\times(0,+\infty)$.*

小集化条件（B3）意味着核 $P_{s,\ i}(z,\ z')\mathrm{d}z'$ 关于 s、i 和 z一致遍历。根据（Meyn and Tweedie，1993，定理 16.0.2。）和（Fort 和 Moulines，2003，命题 1），A5-项 2 和 A5-项 3 均成立。

附录 C 详细证明

C.1 证明(17)

令 $t \in [k^{\text{out}}]^\star$。由 $\gamma_{t,\ k+1}$ 定义$=\prod_{j=0}^{k}\left(1+\frac{2C_b}{m_{t,j+1}}\right)^{-1}\gamma_{t,\ 0}$对任意 $k \geq 0$, 所给出的序列满足

$$\gamma_{t,k+1}\left(1+\frac{2C_b}{m_{t,k+1}}\right) \leq \gamma_{t,k}\,.$$

性质 $\Lambda_{t,k+1} \in (0,1/2)$ 成立的一个充分条件是 $\mathsf{a}\gamma_{t,\ 0}^2 + \bar{\mathsf{a}}\gamma_{t,0} - 1/2 < 0$，其中

$$\bar{\mathsf{a}} \stackrel{\text{def}}{=} \frac{L_{\dot{\mathrm{W}}}}{v_{\min}}\,,$$
$$\mathsf{a} \stackrel{\text{def}}{=} L^2\frac{2v_{\max}k_t^{\text{in}}}{v_{\min}\mathsf{b}}\left(1+\frac{2\,C_{vb}}{\sqrt{\mathsf{b}}\,\bar{M}_{t,k+1}}\right)\,.$$

函数 $x \mapsto \mathsf{a}x^2 + \bar{\mathsf{a}}x - 1/2$ 有两个根：一个为正，另一个为负。正根为 $(-\bar{\mathsf{a}} + \sqrt{\bar{\mathsf{a}}^2 + 2\mathsf{a}})/(2\mathsf{a})$; 它等于 (17)。

C.2 证明(18)和 (19)

我们记 $\mathsf{S}_{t,0} - \mathsf{h}(S_{t,0}) = U + V$其中

$$U \stackrel{\text{def}}{=} \frac{1}{\mathsf{b}'_t}\sum_{i\in\mathcal{B}_{t,0}}\left(\delta_{t,0,i} - \mathbb{E}\left[\delta_{t,0,i}|\widehat{S}_{t,0},\mathcal{B}_{t,0}\right]\right)\,,$$
$$V \stackrel{\text{def}}{=} \frac{1}{\mathsf{b}'_t}\sum_{i\in\mathcal{B}_{t,0}}\mathbb{E}\left[\delta_{t,0,i}|\widehat{S}_{t,0},\mathcal{B}_{t,0}\right] - \mathsf{h}(\widehat{S}_{t,0})\,.$$

根据条件期望的定义，我们有 $\mathbb{E}\left[\|U+V\|^2\right] = \mathbb{E}\left[\|U\|^2\right] + \mathbb{E}\left[\|V\|^2\right]$。由于$\delta_{t,0,i}$ 是 $\mathsf{h}_i(\widehat{S}_{t,0})$ 的无偏随机近似，因此我们有 $\mathbb{E}\left[\delta_{t,0,i}|\widehat{S}_{t,0},\mathcal{B}_{t,0}\right] = \mathsf{h}_i(\widehat{S}_{t,0})$。

在 $\mathcal{B}_{t,\ 0} = \{1,\cdots,n\}$ 的情况下，则 $V = 0$。因此，

$$\begin{aligned}&\mathbb{E}\left[\|\mathsf{S}_{t,0} - \mathsf{h}(\widehat{S}_{t,0})\|^2\right]\\ &= \frac{1}{n^2}\mathbb{E}\left[\|\sum_{i=1}^{n}\{\delta_{t,0,i} - \mathbb{E}\left[\delta_{t,0,i}|\widehat{S}_{t,0}\right]\}\|^2\right]\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2_{t,0,i}\,,\end{aligned}$$

其中我们利用了变量 $\{\delta_{t,0,i}, i \in [n]^\star\}^{\widehat{\ }}$ 在给定 $S_{t,0}$ 的条件下相互独立，且方差为 σ_t^2

在这种情况下$\mathcal{B}_{t,0}^{,0,i\cdot}$ 是 $[n]^\star$ 的一个子集，基数为 b'_t，则记作

$$\mathbb{E}\left[\|U\|^2\right] = \frac{1}{(\mathsf{b}'_t)^2}\mathbb{E}\left[\mathbb{E}\left[\|U\|^2|\mathcal{B}_{t,0}, \widehat{S}_{t,0}\right]\right]$$
$$= \frac{1}{\mathsf{b}'_t}\mathbb{E}\left[\frac{1}{\mathsf{b}'_t}\sum_{i\in\mathcal{B}_{t,0}}\sigma^2_{t,0,i}\right]$$

由引理 7.1得出结论。再次根据引理7.1，我们有

$$\mathbb{E}\left[\|V\|^2\right] \le \frac{1}{\mathsf{b}'_t\, n}\sum_{i=1}^{n}\|\mathsf{h}_i(\widehat{S}_{t,0}) - \mathsf{h}(\widehat{S}_{t,0})\|^2\,,$$

当 $\mathcal{B}_{t,0}$ 是有放回抽样时取等号。由此完成证明。

C.3 引理7.1的证明

为简化记号，

$$\mathcal{B} \overset{\text{def}}{=} \mathcal{B}_{t,k}, \qquad \mathsf{h}_{\mathcal{B}} \overset{\text{def}}{=} \frac{1}{\mathsf{b}}\sum_{i\in\mathcal{B}}\mathsf{h}_i\,.$$

C.3.1 有放回情形

我们记作 $\mathcal{B} = \{I_1, \cdots, I_{\mathsf{b}}\}$，其中随机变量 I_i 相互独立，且在 $[n]^\star$ 上服从均匀分布。

- 于是

$$\mathbb{E}\left[f_{\mathcal{B}}\right] = \frac{1}{\mathsf{b}}\sum_{\ell=1}^{\mathsf{b}}\mathbb{E}\left[f_{I_\ell}\right] = \mathbb{E}\left[f_{I_1}\right] = n^{-1}\sum_{i=1}^{n} f_i(u)\,.$$

- 设 $\bar{f}_{\text{定义}=} n^{-1}\sum_{i=1}^{n} f_i$。利用随机变量 $\{I_1, \cdots, I_{\mathsf{b}}\}$ 相互独立这一点，我们有，

$$\mathbb{E}\left[\|f_{\mathcal{B}} - \bar{f}\|^2\right] = \mathbb{E}\left[\|\frac{1}{\mathsf{b}}\sum_{\ell=1}^{\mathsf{b}}\left(f_{I_\ell} - \bar{f}\right)\|^2\right]$$
$$= \frac{1}{\mathsf{b}^2}\sum_{\ell=1}^{\mathsf{b}}\mathbb{E}\left[\|f_{I_\ell} - \bar{f}\|^2\right]$$
$$= \frac{1}{\mathsf{b}}\mathbb{E}\left[\|f_{I_1} - \bar{f}\|^2\right] = \frac{1}{\mathsf{b}\,n}\sum_{i=1}^{n}\|f_i - \bar{f}\|^2\,.$$

- 由于和的方差等于相互独立的随机变量的方差之和

$$\mathbb{E}\left[\|\mathsf{h}_{\mathcal{B}}(u) - \mathsf{h}_{\mathcal{B}}(u') - \{\mathsf{h}(u) - \mathsf{h}(u')\}\|^2\right]$$
$$= \frac{1}{\mathsf{b}^2}\sum_{\ell=1}^{\mathsf{b}}\mathbb{E}\left[\|\mathsf{h}_{I_\ell}(u) - \mathsf{h}_{I_\ell}(u') - \mathsf{h}(u) + \mathsf{h}(u')\|^2\right]\,.$$

于是，由于 I_ℓ 在 $[n]^\star$ 上服从均匀分布，

$$\mathbb{E}\left[\|\mathsf{h}_{I_\ell}(u) - \mathsf{h}_{I_\ell}(u') - \mathsf{h}(u) + \mathsf{h}(u')\|^2\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\mathsf{h}_i(u) - \mathsf{h}_i(u')\|^2\right] - \|\mathsf{h}(u) - \mathsf{h}(u')\|^2$$
$$\le \|u - u'\|^2\frac{1}{n}\sum_{i=1}^{n}L_i^2 - \|\mathsf{h}(u) - \mathsf{h}(u')\|^2\,. \tag{C4}$$

C.3.2 不放回情形

令 $\mathcal{B} = \{I_1, \cdots, I_{\mathsf{b}}\}$ 和 $\bar{f}_{\text{def}=} n^{-1}\sum_{i=1}^{n} f_i$。$I_1$ 是 $[n]^\star$ 上的均匀随机变量，因此 $\mathbb{E}\left[f_{I_1}\right] = \bar{f}$。

- 在给定 I_1的条件下，I_2 是 $[n]^\star \setminus \{I_1\}$上的均匀随机变量。因此

$$\mathbb{E}\left[f_{I_2}\right] = \frac{1}{n-1}\left(\sum_{j=1}^{n} f_j - \mathbb{E}\left[f_{I_1}\right]\right)$$
$$= \frac{n}{n-1}\bar{f} - \frac{1}{n-1}\bar{f} = \bar{f}\,.$$

根据归纳法，对于任意 $\ell \ge 2$,

$$\mathbb{E}\left[f_{I_\ell}\right]$$
$$= \frac{1}{n-\ell+1}\left(\sum_{j=1}^{n} f_j - \sum_{r=1}^{\ell-1}\mathbb{E}\left[f_{I_r}\right]\right)$$
$$= \frac{n}{n-\ell+1}\bar{f} - \frac{\ell-1}{n-\ell+1}\bar{f} = \bar{f}\,.$$

综上，$\mathsf{b}^{-1}\sum_{\ell=1}^{\mathsf{b}}\mathbb{E}\left[f_{I_\ell}\right] = \bar{f}$。

- 令 $u, u' \in \mathcal{S}$；设 $\phi(I_\ell)$ 定义为 $= \mathsf{h}_{I_\ell}(u) - \mathsf{h}(u) - \mathsf{h}_{I_\ell}(u') + \mathsf{h}(u')$。则 $\mathbb{E}\left[\phi(I_\ell)\right] = 0$。首先，我们用归纳法证明 $\mathbb{E}\left[\|\phi(I_\ell)\|^2\right] = \mathbb{E}\left[\|\phi(I_1)\|^2\right]$。注意到 I_1 是定义在 $[n]^\star$ 上的均匀随机变量，并利用归纳假设，

$$(n-\ell+1)\mathbb{E}\left[\|\phi(I_\ell)\|^2\right]$$

$$= \left(\sum_{i=1}^{n} \|\phi(i)\|^2 - \mathbb{E}\left[\sum_{r=1}^{\ell-1} \|\phi(I_r)\|^2 \right] \right)$$
$$= n\mathbb{E}\left[\|\phi(I_1)\|^2\right] - (\ell-1)\mathbb{E}\left[\|\phi(I_1)\|^2\right],$$

由此完成归纳证明。其次，我们证明对于任意 $\ell \geq 0$,

$$\mathbb{E}\left[\| \sum_{r=1}^{\ell+1} \phi(I_r)\|^2 \right] \leq (\ell+1)\mathbb{E}\left[\|\phi(I_1)\|^2\right] . \quad \text{(C5)}$$

由于 $\sum_{i=1}^{n} \phi(i) = n\mathbb{E}\left[\phi(I_1)\right] = 0$,

$$\mathbb{E}\left[\left\langle \sum_{p=1}^{\ell} \phi(I_p), \phi(I_{\ell+1}) \right\rangle \right]$$
$$= \mathbb{E}\left[\left\langle \sum_{p=1}^{\ell} \phi(I_p), \mathbb{E}\left[\phi(I_{\ell+1}) \middle| I_1, \cdots, I_\ell \right] \right\rangle \right]$$
$$= \frac{1}{n-\ell}\mathbb{E}\left[\left\langle \sum_{p=1}^{\ell} \phi(I_p), \sum_{i=1}^{n} \phi(i) - \sum_{p=1}^{\ell} \phi(I_p) \right\rangle \right]$$
$$= -\frac{1}{n-\ell}\mathbb{E}\left[\| \sum_{p=1}^{\ell} \phi(I_p)\|^2 \right] ,$$

因此

$$\mathbb{E}\left[\| \sum_{p=1}^{\ell+1} \phi(I_p)\|^2 \right]$$
$$= \left(1 - \frac{2}{n-\ell}\right) \mathbb{E}\left[\| \sum_{p=1}^{\ell} \phi(I_p)\|^2 \right] + \mathbb{E}\left[\|\phi(I_{\ell+1})\|^2\right]$$
$$\leq (\ell+1)\mathbb{E}\left[\|\phi(I_1)\|^2\right] .$$

第一个界的证明由(C5)和(C4)得出，因为这里同样有， I_1 在 $[n]^\star$ 上服从均匀分布。

- 第二个界的证明类似（更改 $\phi(I_\ell)$ 的定义）；此处省略。

参考文献

Allen-Zhu Z（2018）Natasha2：比随机梯度下降更快的非凸优化。载于：Bengio S、Wallach H、Larochelle H 等（编），《神经信息处理系统进展》，第31卷。Curran Associates, Inc.

Allen-Zhu Z、Hazan E（2016）用于加速非凸优化的方差缩减。载于：Balcan M、Weinberger K（编），第33届国际机器学习会议，ICML 2016。国际机器学习协会（IMLS），第33届国际机器学习会议，ICML 2016，第1093–1101页

Andrieu C, Fort G, Vihola M（2015）次几何遍历马尔可夫链的定量收敛速率。《应用概率学杂志》52(2):391–404。https://doi.org/10.1239/jap/1437658605

Atchad´e Y, Fort G, Moulines E（2017）关于受扰近端梯度算法。机器学习研究期刊 18(10):1–33 BauschkeHH, Combettes PL（2011）希尔伯特空间中的凸分析与单调算子理论，第1版。Springer Publishing Company, Incorporated, https://doi.org/10.1007/978-1-4419-9467-7Beck A（2017）优化中的一阶方法。工业与应用数学学会，宾夕法尼亚州费城, https://doi.org/10.1137/1.9781611974997Becker S, Fadili J（2012）一种拟牛顿近端分裂方法。载于：Pereira F, Burges C, Bottou L 等（编），神经信息处理系统进展，第25卷。Curran Associates, Inc. Becker S, Fadili J, Ochs P（2019）关于拟牛顿前向-后向分裂：近端微积分与收敛。SIAM 优化期刊 29(4):2445–2481。https://doi.org/10.1137/18M1167152Benveniste A, M´etivier M, Priouret P（1990）自适应算法与随机逼近。Springer Verlag, https://doi.org/10.1007/978-3-642-75894-2Bonettini S, Porta F, Ruggiero V 等（2021）成像中的前向-后向方法变度量技术。计算与应用数学杂志 385:113,192。https://doi.org/https://doi.org/10.1016/j.cam.2020.113192

BorkarVS (2008) 随机逼近：动力系统视角。剑桥大学出版社，剑桥；新德里印度图书机构，https://doi.org/https://doi.org/10.1007/978-93-86279-38-5，动力系统视角

Brown L (1986) 统计指数族基础：及其在统计决策理论中的应用。讲义-专著系列《统计指数族基础》，数理统计学会，https://doi.org/10.1214/lnms/1215466757

Capp´e O, Moulines E (2009) 潜在数据模型的在线期望最大化算法。J Roy Stat Soc B Met 71(3):593–613。https://doi.org/10.1111/j.1467-9868.2009.00698.xCeleux G, Diebolt J (1985) SEM算法：一种从期望最大化算法推导出的混合问题概率教师算法。计算统计季刊 2:73–82 Chen HG, Rockafellar R (1997) 前向-后向分裂的收敛速率。SIAM J Optim 7:421–444。https://doi.org/https://doi.org/10.1137/S1052623495290179Chen J, Zhu J, Teh Y, et al (2018) 带方差缩减的随机期望最大化。在：Bengio S, Wallach H, Larochelle H, et al (编)《神经信息处理系统进展》第31卷。Curran Associates, Inc.，第7967–7977页，https://doi.org/10.5555/3327757.3327893Chen X, Liu S, Sun R, et al (2019) 一类用于非凸优化的 Adam 类算法的收敛性。载于：国际学习表征会议 Choi H, Hobert J (2013) 用于贝叶斯逻辑回归的 Polya-Gamma 吉布斯采样器具有一致遍历性。电子统计学杂志 7(none):2054–2064。https://doi.org/10.1214/13-EJS837Chouzenoux E, Pesquet JC, Repetti A (2014) 用于最小化可微函数之和的变度量前向-后向算法函数与凸函数。《优化理论与应用杂志》162(1):107–132。https://doi.org/10.1007/s10957-013-0465-7Combettes P, Pesquet J (2011) 信号处理中的近端分裂方法。载于：Bauschke HH, Burachik RS, Combettes PL 等（编）《科学与工程反问题的定点算法》。施普林格优化及其应用丛书，施普林格出版社，第185–212页，https://doi.org/10.1007/978-1-4419-9569-8Combettes P,V˜u B (2014) 具有对偶单调包含应用的变度量前向-后向分裂。《优化》63(9):1289–1318。https://doi.org/10.1080/02331934.2012.733883Combettes PL, Wajs VR (2005) 通过近端前向-后向分裂进行信号恢复。《多尺度建模与仿真》4(4):1168–1200。https://doi.org/10.1137/050626090Defazio A, Bach F, Lacoste-Julien S (2014) SAGA：一种快速增量梯度法，适用于非强凸复合目标函数。载于：第27届神经信息处理系统国际会议论文集——第1卷。麻省理工学院出版社，美国马萨诸塞州剑桥，NIPS'14，第1646–1654页 Delyon B, Lavielle M, Moulines E (1999) EM算法随机逼近版本的收敛性。《Ann Statist》27(1):94–128。https://doi.org/10.1214/aos/1018031103Dempster A, Laird N, Rubin D (1977) 通过EM算法从不完整数据中进行最大似然估计。《J Roy Stat Soc B Met》39(1):1–38 Devroye L (1986) 非均匀随机变量生成（最初与……一同出版）。施普林格出版社，https://doi.org/10.1007/978-1-4613-8643-8Eicke B (1992) 希尔伯特空间中凸约束病态问题的迭代方法。《数值泛函分析与优化》13(5-6):413–429。https://doi.org/10.1080/01630569208816489

Everitt B (1984) 潜变量模型导论。Chapman and Hall，伦敦；纽约，
h
t
t
p
s://
d
o
i.org/https://doi.org/10.1007/978-94-009-5564-6
Fang C, Li C, Lin Z 等 (2018) Spider：通过随机路径积分微分估计器实现近乎最优的非凸优化。载于：Bengio S, Wallach H, Larochelle H 等（编）《神经信息处理系统进展》，第31卷。Curran Associates, Inc. Fort G, Moulines E (2003) 曲指数族蒙特卡罗期望最大化的收敛性。《Ann Statist》31(4):1220–1259 Fort G, Moulines E (2021) 扰动近端预条件 Spider 算法：非渐近收敛界。载于：2021年 IEEE 统计信号处理研讨会（SSP），第96–100页，https://doi.org/10.1109/SSP49050.2021.9513846Fort G, Moulines E, Priouret P (2011) 自适应交互马尔可夫链蒙特卡罗算法的收敛性。《Ann Statist》39(6):3262–3289 Fort G, Risser L, Atchad´e Y 等 (2018) 随机 FISTA 算法：如此之快？载于：2018年 IEEE 统计信号处理研讨会（SSP），第796–800页，https://doi.org/10.1109/SSP.2018.8450740Fort G, Moulines E, Wai HT (2020) 随机路径积分微分估计器期望最大化算法。载于：第34届神经信息处理系统国际会议论文集。Curran Associates Inc.，美国纽约州雷德胡克，NIPS'20 Fort G, Gach P, Moulines E (2021a) 有限和优化的快速增量期望最大化：非渐近收敛性。《Stat Comput》31(4):48。https://doi.org/10.1007/s11222-021-10023-9Fort G, Moulines E, Wai HT (2021b) Geom-Spider-EM：用于非凸有限和优化的更快速方差缩减随机期望最大化算法。载于：ICASSP 2021 - 2021年 IEEE 国际声学、语音与信号处理会议 (ICASSP)，第 3135–3139 页，
h
t
t
p
s://
d
oi.org/10.1109/ICASSP39728.2021.9414271Ghadimi S, Lan G (2013) 用于非凸随机规划的随机一阶与零阶方法。SIAM Journal on Optimization 23(4):2341–2368。 https://doi.org/10.1137/120880811Ghadimi S, Lan G, Zhang H (2016) 用于非凸随机复合优化的小批量随机逼近方法。Math Program 155(1-2):267–305。https://doi.org/10.1007/s10107-014-0846-1Gower R, Goldfarb D, Richtarik P (2016) 随机块 BFGS：从数据中挖掘更多曲率。载于：Balcan MF, Weinberger KQ (编) 第 33 届国际机器学习会议论文集，机器学习研究论文集，第 48 卷。PMLR，纽约，美国纽约州，第 1869–1878 页 Hiriart-Urruty JB, Lemar´echal C (1996) 凸分析与最小化算法。Springer Verlag，海德堡，
h
t
t
p
s://
doi.org/https://doi.org/10.1007/978-3-662-02796-7，两卷本——第 2 次印刷版 Horv´ath S, Lei L,Richt´arik P, et al (2022年) 随机梯度方法对非凸优化的适应性。SIAM Journal on Mathematics of Data Science 4(2):634–648。https://doi.org/10.1137/21M1394308Johnson R, Zhang T (2013) 使用预测方差缩减加速随机梯度下降。载于：Burges C, Bottou L, Welling M, 等 (编) 神经信息处理系统进展，第 26 卷。Curran Associates, Inc. Karimi B, Wai HT, Moulines E, 等 (2019) 关于（快速）增量期望最大化方法的全局收敛。载于：Wallach H, Larochelle H, Beygelzimer A, 等 (编) 神经信息处理系统进展 32。Curran Associates, Inc.，第 2837–2847 页

Karimi H, Nutini J, Schmidt M（2016）Polyak-Lojasiewicz 条件下梯度方法与近端梯度方法的线性收敛。载于：Fras- coni P、Landwehr N、Manco G 等（编）《数据库中的机器学习与知识发现》。施普林格国际出版社，第 795–811 页

Kolte R, Erdogdu M, Ozgur A（2015）利用二阶信息加速 SVRG。载于：《神经信息处理系统进展》——OptML 研讨会，第 1–5 页

Lan G（2020）机器学习的一阶与随机优化方法。施普林格数据科学系列，施普林格国际出版社，
h
t
tps://doi.org/https://doi.org/10.1007/978-3-030-39568-1

Lee JD, Sun Y, Saunders MA（2014）用于最小化复合函数的近端牛顿型方法。《SIAM 优化杂志》24(3):1420–1443。https://doi.org/10.1137/130921428

Li Z, Li J（2018）一种用于非光滑非凸优化的简单近端随机梯度法。载于：Bengio S、Wallach H、Larochelle H 等（编），《神经信息处理系统进展》，第 31 卷。Curran Associates, Inc.

Li Z, Bao H, Zhang X 等（2021）PAGE：一种用于非凸优化的简单且最优的概率梯度估计器。载于：Meila M、Zhang T（编），《第 38 届国际机器学习会议论文集》，《机器学习研究论文集》，第 139 卷。PMLR，第 6286–6295 页

McLachlan G, Krishnan T（2008）《EM 算法及其扩展》，第 2 版。Wiley 统计学与概率论丛书，Wiley，https://doi.org/10.1002/9780470191613

Metel M，Takeda A（2021年）用于非光滑非凸约束稀疏优化的随机近端方法。《机器学习研究期刊》22(115)：1–36

Meyn S，Tweedie R（1993）《马尔可夫链与随机稳定性》。施普林格出版社，伦敦，https://doi.org/https://doi.org/10.1007/978-1-4471-3267-7

Moreau J（1965）希尔伯特空间中的邻近性与对偶性。《法国数学学会通报》93：273–299。´e et dualit´e dansunespace hilbertien. Bulletin de la Soci´et´eMath´ematique deFrance 93:273–299. https://doi.org/10.24033/bsmf.1625

Moritz P、Nishihara R、Jordan M（2016）一种线性收敛的随机 L-BFGS 算法。载于：Gretton A、Robert CC（编），第19届人工智能与统计国际会议论文集，《机器学习研究论文集》第51卷。PMLR，第249–258页

Neal RM、Hinton GE（1998）一种为增量、稀疏及其他变体提供依据的 EM 算法观点。载于：Jordan MI（编），《图模型学习》。施普林格·荷兰，多德雷赫特，第355–368页，https://doi.org/10.1007/978-94-011-5014-9 12

Ng SK, McLachlan GJ (2003) 关于使用增量期望最大化算法拟合正态混合模型时分块数量的选择。Stat Comput 13(1):45–55.https://doi.org/10.1023/A:1021987710829

Nguyen L, Liu J, Scheinberg K, 等 (2017) SARAH：一种使用随机递归梯度解决机器学习问题的新方法。载于：Precup D, Teh YW（编）《第34届国际机器学习会议论文集》，第 2613–2621 页

NhanHP, Lam MN, Dzung TP, 等 (2020) Prox-SARAH：一种用于随机复合非凸优化的高效算法框架。机器学习研究期刊 21(110):1–48

Park Y, Dhar S, Boyd S, 等 (2019) 变度量近端梯度法与对角 Barzilai-Borwein 步长。https://doi.org/10.48550/ARXIV.1910.07056

Polson NG、Scott J、Windle J（2013）使用 P'olya–Gamma 潜变量对逻辑回归模型进行贝叶斯推断。期刊

《美国统计学会期刊》108(504):1339–1349。
https://doi.org/https://doi.org/10.1080/01621459.2013.829001

Reddi SJ, Hefny A, Sra S, 等（2016）非凸优化的随机方差缩减。载于：Balcan MF、Weinberger KQ（编），《第33届国际机器学习会议论文集》，《机器学习研究论文集》第48卷。PMLR，纽约，纽约州，美国，第314–323页

Repetti A, Wiaux Y（2021）用于复合最小化问题的变度量前向-后向算法。SIAM J Optim 31：1215–1241。
https://doi.org/https://doi.org/10.1137/19M1277552

Repetti A, Chouzenoux E, Pesquet JC（2014）一种预条件前向-后向方法及其在大规模非凸光谱解混问题中的应用。载于：2014 IEEE 国际声学、语音与信号处理会议（ICASSP），第1498–1502页，
https://doi.org/10.1109/ICASSP.2014.6853847

Robert C, Casella G（2004）蒙特卡罗统计方法。Spri nger Verlag,
https://doi.org/https://doi.org/10.1007/978-1-4757-4145-2

Wang Z、Ji K、Zhou Y 等（2019）Spider-boost 与动量：更快的方差缩减算法。载于：Wallach HM、Larochelle H、Beygelzimer A 等（编），《神经信息处理系统进展》第32卷：2019年神经信息处理系统年度会议，NeurIPS 2019，2019年12月8日至14日，加拿大不列颠哥伦比亚省温哥华，第2403–2413页

Wei G, Tanner（1990）EM算法和简易数据增强算法的蒙特卡罗实现。《美国统计学会杂志》85(411)：699–704。https://doi.org/10.1080/01621459.1990.10474930

Wu C (1983) 关于期望最大化算法的收敛性质。Ann Statist 11(1):95–103。
https://doi.org/10.1214/aos/1176346060

Yun J、Lozano AC、Yang E（2021年）：面向结构化神经网络的自适应近端梯度方法。载于：Ranzato M、Beygelzimer A、Dauphin Y 等（编）：《神经信息处理系统进展》，第34卷。Curran Associates, Inc.，第24,365–24,378页

Zhang J, Xiao L (2019) 一种采用增量方差缩减的随机复合梯度算法。载于：WallachH、Larochelle H、Beygelz- imer A 等（编），《神经信息处理系统进展》，第32卷。Curran Associates, Inc.

Zhang Q, Huang F, Deng C 等（2022年）更快的随机拟牛顿方法。IEEE Transactions on Neural Networks and Learn- ing Systems 33(9):4388–4397。
https://doi.org/10.1109/TNNLS.2021.3056947

Zhou D, Xu P, Gu Q（2020）非凸优化的随机嵌套方差缩减。机器学习研究期刊 21(103):1–63