arXiv:2303.00937v1 [math.OC] 2 Mar 2023

非精确在线优化方法的跟踪误差界与动态遗憾界：基于序贯半定规划的一般性分析 *

Usman Syed† Emiliano Dall'Anese‡ Bin Hu§

2023年3月3日

摘要

本文提出一个统一框架，用于分析多种设定下非精确在线优化方法的跟踪误差和动态遗憾。具体而言，我们借助控制理论中的二次约束方法，构建序贯半定规划（SDP），其可行点自然对应于多种非精确在线优化方法的跟踪误差界，包括非精确在线梯度下降（OGD）方法、在线梯度下降-上升方法、在线随机梯度方法以及非精确近端在线梯度方法。我们为所提出的序贯半定规划（SDP）提供精确解析解，并由此得到本文所研究在线算法的细粒度跟踪误差界。我们还提供一种简单的程序，将所得跟踪误差界转换为动态遗憾界。本文分析的主要创新之处在于，我们以统一的方式，在各种非精确预言机假设下推导出所提出序贯半定规划（SDP）的精确解析解。

1 引言

优化与控制理论之间有着深刻而根本的联系。近年来，研究者开始利用控制理论中的二次约束方法 [26]，对迭代优化算法进行统一分析和计算机辅助设计 [1–4,6,11–14, 16–19, 24, 27, 28, 30, 31, 34, 35, 40, 42]。在这些研究中，研究者将李雅普诺夫论证与二次约束相结合，推导出规模较小的半定规划（SDP），其可行点对应于迭代优化算法的收敛或鲁棒性保证（近期综述见 [23]）。基于 SDP 的方法有一个优点：通常可以通过模块化方式修改由此得到的半定规划条件，以处理关于梯度预言机的不同假设（例如，预言机含噪声 [1] 或非精确 [12, 17, 24]）。在目标

*资助：Usman Syed 和 Bin Hu 的工作得到了美国国家科学基金会 CAREER 奖项（资助编号 2048168）的支持。

†U. Syed 任职于伊利诺伊大学厄巴纳-香槟分校协同科学实验室和电气与计算机工程系，电子邮箱：usyed3@illinois.edu

‡E. Dall'Anese 任职于科罗拉多大学博尔德分校电气、计算机与能源工程系，电子邮箱：emiliano.dallanese@colorado.edu§B. Hu 任职于伊利诺伊大学厄巴纳-香槟分校协同科学实验室和电气与计算机工程系，电子邮箱：binhu7@illinois.edu

函数为静态且不随时间变化的情形下，二次约束方法导出了新的半定规划条件，自然涵盖了基于动量的加速方法 [2,3,11,16,24,30,35,40,42], 、近端／算子分裂方法 [6,14,28,31], 、分布式梯度方法 [4, 13, 34, 35], 以及随机有限和方法 [18, 19]的复杂度分析。然而，在许多实际情形中，目标函数会随时间变化，而上述二次约束框架与当前时变在线优化理论之间存在显著差距 [7, 15, 25, 32, 33]。本文通过提出一种序贯半定规划方法，简化对非精确在线优化方法的分析，从而弥合这一差距。

为便于说明，考虑时变代价函数 $f_t\colon\ \mathbb{R}^p \to \mathbb{R}$，其中 $t \in \{0, 1, 2\cdots\}$是时间索引。广受欢迎的在线梯度下降（OGD）方法的迭代形式为

$$x_{t+1} = x_t - \alpha_t \nabla f_t(x_t), \tag{1}$$

其中，在每个时间步，决策变量根据时间 t处的目标函数梯度进行更新。这类算法充分利用了可用的流数据，并可用于利用动态环境的时变特性。值得注意的是，遗憾和跟踪误差一直被用作在线优化算法的两项主要性能指标。遗憾已在在线学习文献中得到广泛研究，并为包括在线投资组合选择和网页排序在内的决策任务中的探索-利用权衡提供了精确刻画 [15,32], 而跟踪误差则是衡量在线算法跟踪能力的常用指标，应用涉及目标跟踪、信号检测和路径估计 [7, 10, 25, 33]。这两项指标相辅相成，对于理解在线优化方法的性能至关重要。

本文尤其关注非精确在线优化方法的跟踪误差和动态遗憾分析。目前，在线优化方法的跟踪误差分析仍是逐个情形进行的，并且依赖专家的深入见解 [7, 10,25, 33]。当预言机的底层假设发生变化时，这类分析可能无法推广。例如， [25] 中对在线梯度下降的跟踪误差分析无法直接用于处理在线梯度预言机存在相对误差的情况。因此，开发更为系统的分析方法大有裨益。本文提出了一种统一的基于半定规划的分析方法，涵盖确定性和随机情形下的一大类非精确在线优化方法。我们的贡献总结如下。

1. 我们利用控制理论构造序贯半定规划条件，其可行点自然对应于多种非精确在线优化方法跟踪误差的上界[1] ，包括非精确在线梯度下降、随机在线梯度下降以及非精确近端在线梯度下降（IP-OGD）方法。

2. 我们为所提出的序贯半定规划给出了精确解析解，从而得到了通过这些半定规划能够获得的最紧跟踪误差界。这些新的跟踪误差界有助于理解存在预言机不精确性时在线优化方法的性能。此外，我们的分析首次给出了 IP-OGD 的均方跟踪误差界，补充了 [7] 中提出的原始 L_1 误差界。

[1]假设 x_t^*是 f_t的全局最小值，且 $\{x_t\}$由某种在线算法生成。本文将确定性情形下的跟踪误差定义为$\|x_t - x_t^*\|^2$，将随机情形下的跟踪误差定义为 $\mathbb{E}\,\|x_t - x_t^*\|^2$。

3. 我们提供了一个简单的流程，可将跟踪误差界转换为动态遗憾界。由此，我们为本文研究的不精确在线算法得到了新的动态遗憾界。

相关工作。此前，研究者利用二次约束构造序贯半定规划，以分析使用不精确预言机 [17] 的有偏随机梯度下降（SGD）。从概念上看，我们提出的序贯半定规划方法可视为将 [17] 中的工作扩展到在线设置。然而，本文的一项独特创新在于，我们通过严谨的代数论证，得到了所提出的序贯半定规划条件的精确解析解。相比之下， [17] 中的序贯半定规划尚未以解析方式精确求解。具体而言， [17] 中有偏 SGD 的界是通过数值模拟的指导选取一个特定可行点得到的，而该可行点是否为原始序贯半定规划的最优解，目前仍不清楚（详见 [17, 第 2.2] 节的讨论）。因此，本文得到的精确半定规划解补充了 [17] 中的原始分析，并进一步加深了我们对二次约束方法的理解。另一类使用半定规划分析优化方法的相关工作，以性能估计问题（PEP）的构造为基础 [9, 37, 38]。当目标函数是静态的，性能估计问题框架可用于在各种设置下，获得一阶优化方法最坏情况性能的紧上界和下界 [5, 8, 20–22, 29, 36–38]。目前，性能估计问题框架尚未扩展到在线设置。鉴于二次约束方法与性能估计问题框架之间存在根本联系 [39], ，将性能估计问题框架扩展用于在线优化方法的紧跟踪误差分析是可行的。这是一个值得探索的未来方向。

1.1 符号约定

$p \times p$ 单位矩阵和 $p \times p$ 零矩阵分别记作 I_p 和 0_p。两个矩阵 A 和 B 的克罗内克积记作 $A \otimes B$。当矩阵 P 为负半定（负定）时，我们使用记号 $P \preceq$ （$\prec$）0。可微函数 $f\colon \mathbb{R}^p \to \mathbb{R}$ 若满足 $\|\nabla f（x）- \nabla f（y）\| \leq L\|x - y\|$ 对所有 x、$y \in \mathbb{R}^p$ 成立，则称其为 L-光滑；若满足 $f（x） \geq f（y） + \nabla f（y）^T（x - y） + \frac{m}{2}\|x - y\|^2$ 对所有 x、$y \in \mathbb{R}^p$ 成立，则称其为 m-强凸。令 $\mathcal{F}（m, L）$ 表示由可微函数 $f\colon \mathbb{R}^p \to \mathbb{R}$ 组成的函数集合，其中这些函数均为 L-光滑且 m-强凸。对于任意 $f \in \mathcal{F}（m, L）$ 且满足 $m > 0$ 的情形，存在唯一的 x^* 满足 $\nabla f（x^*） = 0$，并且对于任意 $x \in \mathbb{R}^p$ [24, , 以下不等式成立。命题 5]:

$$\begin{bmatrix} x - x^* \\ \nabla f(x) \end{bmatrix}^T \begin{bmatrix} -2mLI_p & (L+m)I_p \\ (L+m)I_p & -2I_p \end{bmatrix} \begin{bmatrix} x - x^* \\ \nabla f(x) \end{bmatrix} \geq 0. \tag{2}$$

然而，满足上述不等式的函数不一定是凸函数。满足式 (2) 且具有某个唯一全局最小值 x^* 的可微函数集合记为 $\mathcal{S}(m, L)$。这一类函数具有扇区有界梯度。条件 (2) 与信号处理文献 [42, 附录 A] 中的正则性条件密切相关。显然，我们有 $\mathcal{F}(m, L) \subset \mathcal{S}(m, L)$。

1.2 问题设定

本文研究时变目标函数 $f_t : \mathbb{R}^p \to \mathbb{R}$，其中回顾一下， t 是时间索引。根据所分析的算法，我们将假设 $f_t \in \mathcal{F}(m_t, L_t)$

或 $f_t \in \mathcal{S}(m_t, L_t)$。时间 t 处目标函数的全局最小值记为 x_t^*。我们进一步假设序列 $\{\sigma_t\}$ 对所有 t 均满足以下不等式：

$$\left\|x_{t+1}^* - x_t^*\right\| \leq \sigma_t. \tag{3}$$

这里，$\{\sigma_t\}$ 可以视为预先规定的界限，用于刻画全局最小值随时间变化的速度。最后，为便于表述，我们采用以下记号：

$$\begin{aligned} Y_t &:= \begin{bmatrix} 2L_t m_t & -(L_t + m_t) \\ -(L_t + m_t) & 2 \end{bmatrix}, \quad M_t = \begin{bmatrix} 2m_t & -1 \\ -1 & 0 \end{bmatrix}, \\ \mu_t &:= \begin{cases} 1 - m_t \alpha_t & \text{if } 0 \leq \alpha_t \leq \frac{2}{m_t + L_t} \\ \alpha_t L_t - 1 & \text{if } \frac{2}{m_t + L_t} \leq \alpha_t \leq \frac{2}{L_t} \end{cases} \end{aligned} \tag{4}$$

其中，α_t 是待分析的在线优化方法的步长。

2 预备知识：用于算法分析的序贯半定规划框架

基于 [17], 的工作，我们首先提出一种通用流程，可用于构造在线优化算法跟踪误差分析中的序贯半定规划。为便于表述，我们采用 [16]中关于二次约束方法的耗散性视角。首先，我们以线性状态空间形式引入以下离散时间动力系统：

$$\xi_{t+1} = \xi_t + (B_t \otimes I_p)\, w_t, \tag{5}$$

其中，$\xi_t \in \mathbb{R}^p$ 是状态，$B_t \in \mathbb{R}^{1\times q}$ 是某个输入矩阵，而 $w_t \in \mathbb{R}^{n_w}$ 是输入。为使维度相容，我们有 $n_w = pq$。我们可以将 w_t 解释为驱动力，而耗散性理论中的一个关键概念是供给率条件，它刻画了由于输入 w_t而导致的状态 ξ_t 中的能量变化。在本文中，我们考虑二次约束形式的供给率条件：

$$\begin{bmatrix} \xi_t \\ w_t \end{bmatrix}^T (X_t^{(j)} \otimes I_p) \begin{bmatrix} \xi_t \\ w_t \end{bmatrix} \leq \Lambda_t^{(j)}, \tag{6}$$

其中，$X_t^{(j)} \in \mathbb{R}^{(1+q)\times(1+q)}$ 和 $\Lambda_t^{(j)} \in \mathbb{R}$ 是针对 $j = 1$、2、.....、J预先指定的。(6) 左侧可视为驱动力 w_t 向内部状态 ξ_t输送的能量，而 $\Lambda_t^{(j)}$ 仅为所输送的能量提供一个上界。稍后我们将证明，许多在线优化算法都可以改写为 (5) 的形式（其中 ξ_t 是 t处的跟踪误差），并满足某个供给率条件 (6)。跟踪误差分析归结为求出$\|\xi_t\|^2$（在确定性情形下）或 $\mathbb{E}\,\|\xi_t\|^2$（在随机情形下）的紧上界。我们的分析依赖于耗散性理论中的以下标准结果 $[16, 41]$。

命题 2.1. 考虑满足供给率条件 (6) 的 (5)。若存在非负标量 $\{\lambda_t^{(j)}\}_{j=1}^J$ 和 ρ_t，使得以下条件成立

$$\begin{bmatrix} 1 - \rho_t^2 & B_t \\ B_t^T & B_t^T B_t \end{bmatrix} \preceq \sum_{j=1}^{J} \lambda_t^{(j)} X_t^{(j)}, \tag{7}$$

则 (5) 的迭代满足 $\|\xi_{t+1}\|^2 \le \rho_t^2\|\xi_t\|^2 + \sum_{j=1}^J \lambda_t^{(j)}\Lambda_t^{(j)}$。此外，若 (7) 对所有 t 均成立，则可得到以下界：

$$\|\xi_t\|^2 \le \left(\prod_{k=0}^{t-1}\rho_k^2\right)\|\xi_0\|^2 + \sum_{k=0}^{t-1}\left(\left(\prod_{l=k+1}^{t-1}\rho_l^2\right)\left(\sum_{j=1}^{J}\lambda_k^{(j)}\Lambda_k^{(j)}\right)\right). \tag{8}$$

证明。为完整起见，我们简要概述证明过程。由（7）可直接得到

$$\begin{bmatrix}\xi_t\\ w_t\end{bmatrix}^T\left(\begin{bmatrix}1-\rho_t^2 & B_t\\ B_t^T & B_t^TB_t\end{bmatrix}\otimes I_p\right)\begin{bmatrix}\xi_t\\ w_t\end{bmatrix} \le \sum_{j=1}^{J}\lambda_t^{(j)}\begin{bmatrix}\xi_t\\ w_t\end{bmatrix}^T\left(X_t^{(j)}\otimes I_p\right)\begin{bmatrix}\xi_t\\ w_t\end{bmatrix}. \tag{9}$$

由（5）容易验证，（9）的左侧等于 $\|\xi_{t+1}\|^2 - \rho_t^2\|\xi_t\|^2$。因此，我们可以利用供给率条件（6）以及 $\{\lambda_t^{(j)}\}_{j=1}^J$ 的非负性来证明 $\|\xi_{t+1}\|^2 \le \rho_t^2\|\xi_t\|^2 + \sum_{j=1}^J \lambda_t^{(j)}\Lambda_t^{(j)}$。随后，我们可以利用数学归纳法证明 $\|\xi_t\|^2 \le \left(\prod_{k=0}^{t-1}\rho_k^2\right)\|\xi_0\|^2 + C_t$，其中 $\{C_t\}$ 根据以 $C_0 = 0$ 为初始值的递推关系 $C_{t+1} = \rho_t^2 C_t + \sum_{j=1}^J \lambda_t^{(j)}\Lambda_t^{(j)}$ 计算得到。直接验证可知，C_t 等于（8）右侧的第二项。证明完毕。

□

正如我们将在后文展示的，命题 2.1 可用于获得在线优化方法的各种跟踪误差界。固定 B_t 和 $\{X_t^{(j)}\}$后，条件 (7) 关于 ρ_t^2 和 $\lambda_t^{(j)}$ 是线性的，因此成为一个线性矩阵不等式（LMI），其可行性可使用半定规划求解器进行验证。令 $\mathcal{T}_t \subseteq \mathbb{R}_+^{J+1}$表示满足线性矩阵不等式条件 (7) 的元组 $\left(\rho_t, \lambda_t^{(1)}, \ldots, \lambda_t^{(J)}\right)$ 所组成的集合。根据 (8)，笛卡尔积 $\prod_{k=0}^{t-1}\mathcal{T}_k$ 中的任何可行点都将给出 $\|\xi_t\|^2$的一个上界。计算有用的跟踪误差界时，一个关键问题是如何选择 $\{\rho_t\}$ 和 $\{\lambda_t^{(j)}\}$，使得 (8) 的右侧最小。通过线性矩阵不等式 (8) 可为 $\|\xi_t\|^2$ 获得的最佳界，应当是以下最小化问题的解：

$$\underset{(\rho_k,\lambda_k^{(j)})\in\mathcal{T}_k, 0\le k\le t-1}{\text{minimize}} \quad \left(\prod_{k=0}^{t-1}\rho_k^2\right)\|\xi_0\|^2 + \sum_{k=0}^{t-1}\left(\left(\prod_{l=k+1}^{t-1}\rho_l^2\right)\left(\sum_{j=1}^{J}\lambda_k^{(j)}\Lambda_k^{(j)}\right)\right). \tag{10}$$

应用 [17, 第 2.2], 节中的论证，可以证明上述问题能够通过下面形式化的序贯半定规划方法精确求解。

命题 2.2。将（10）的全局最优解记为 $\hat{} U_t$。令 $\hat{}U_0 = U_0 = \|\xi_0\|^2$。则 $\hat{}$ 序列 $\{U_t\}$ 满足以下递推关系

$$\hat{U}_{t+1} := \underset{(\rho_t,\lambda_t^{(j)})\in\mathcal{T}_t}{\text{minimize}} \quad \left(\rho_t^2\hat{U}_t + \sum_{j=1}^{J}\lambda_t^{(j)}\Lambda_t^{(j)}\right). \tag{11}$$

证明。将 (10) 中的代价函数记为 U_t。我们可以验证，$\{U_t\}$ 可以通过 $U_{t+1} = \rho_t^2 U_t + \sum_{j=1}^J \lambda_t^{(j)}\Lambda_t^{(j)}$ 递归计算，并以 $U_0 = \|\xi_0\|^2$ 为初始化条件。因此，我们可以重写 (10)

as

$$\hat{U}_t := \underset{(\rho_k,\lambda_k^{(j)})\}_{k=0}^{t-1}}{\text{minimize}} \quad U_t$$
$$\text{subject to} \quad U_{k+1} = \rho_k^2 U_k + \sum_{j=1}^{J} \lambda_k^{(j)} \Lambda_k^{(j)},\ \left(\rho_k, \lambda_k^{(j)}\right) \in \mathcal{T}_k,\ 0 \le k \le t-1$$

其余证明与 [17] 中命题 1 的证明几乎相同。我们可以采用类似的动态规划论证，具体细节从略。□

对于每个 t，子问题 (11) 都会化简为一个半定规划，因为代价函数关于 ρ_t^2 和 $\lambda_t^{(j)}$ 是线性的。因此，我们可以采用统一的分析流程，其中包括以下两个步骤。

1. 将待分析的在线算法改写为形式 (5)，并满足某个供给率条件 (6)。应将状态 ξ_t 设为跟踪误差。

2. 应用命题 2.1 得到线性矩阵不等式条件 (7)，并求解序贯半定规划问题 (11)，以界定 $\|\xi_t\|^2$。

为了推导闭式跟踪误差界，我们需要得到序贯半定规划问题 (11) 的精确解析解。这并非易事。事实上，[17] 中的序贯半定规划尚未得到解析求解。为说明这一问题，我们考虑精确在线梯度下降更新规则 (1)。记 $e_t = x_t^* - x_{t+1}^*$，则在线梯度下降迭代 (1) 可以

可改写为 $x_{t+1} - x_{t+1}^* = x_t - x_t^* - \alpha_t \nabla f_t(x_t) + e_t$。若设 $\xi_t = x_t - x_t^*$, $w_t = \begin{bmatrix}\nabla f_t(x_t)^T & e_t^T\end{bmatrix}^T$，且 $B_t = \begin{bmatrix}-\alpha_t & 1\end{bmatrix}$，则在线梯度下降更新规则 (1) 可视为 (5) 的一个特例。供给率条件取决于对 $\{f_t\}$ 的假设。若 $f_t \in \mathcal{S}(m_t, L_t)$ 且 $\left\|x_{t+1}^* - x_t^*\right\| \le \sigma_t$，则有

$$\begin{bmatrix} x_t - x_t^* \\ \nabla f_t(x_t) \end{bmatrix}^T \begin{bmatrix} -2L_t m_t I_p & (L_t + m_t) I_p \\ (L_t + m_t) I_p & -2 I_p \end{bmatrix} \begin{bmatrix} x_t - x_t^* \\ \nabla f_t(x_t) \end{bmatrix} \ge 0, \quad \|e_t\|^2 \le \sigma_t^2.$$

显然，上述两个不等式可改写为供给率条件 (6) 的形式（对于 $j = 1,\ 2$），其中 $X_t^{(1)} = \mathsf{diag}(Y_t, 0)$、$\Lambda_t^{(1)} = 0$, $X_t^{(2)} = \mathsf{diag}(0, 0, 1)$，且 $\Lambda_t^{(2)} = \sigma_t^2$（回忆一下，$Y_t$ 定义见 (4)）。根据命题 2.1，若存在非负标量 $(\rho_t, \lambda_t^{(1)}, \lambda_t^{(2)})$ 使得：

$$\begin{bmatrix} 1-\rho_t^2 & -\alpha_t & 1 \\ -\alpha_t & \alpha_t^2 & -\alpha_t \\ 1 & -\alpha_t & 1-\lambda_t^{(2)} \end{bmatrix} + \lambda_t^{(1)} \begin{bmatrix} -2L_t m_t & L_t + m_t & 0 \\ L_t + m_t & -2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \preceq 0, \tag{12}$$

则在线梯度下降方法 (1) 满足 $\left\|x_{t+1} - x_{t+1}^*\right\|^2 \le \rho_t^2 \left\|x_t - x_t^*\right\|^2 + \lambda_t^{(2)} \sigma_t^2$。根据命题 2.2，针对精确在线梯度下降，由线性矩阵不等式 (12) 所能得到的最紧跟踪误差界由以下序贯半定规划问题给出：

$$\hat{U}_{t+1} = \underset{(\rho_t, \lambda_t^{(1)}, \lambda_t^{(2)}) \in \mathcal{T}_t}{\text{minimize}} \ \rho_t^2 \hat{U}_t + \lambda_t^{(2)} \sigma_t^2, \tag{13}$$

其中， $\mathcal{T}_t \subseteq \mathbb{R}_+^3$ 表示满足条件 (12) 的元组集合 $\left(\rho_t, \lambda_t^{(1)}, \lambda_t^{(2)}\right)$。要得到 U_t 的精确公式并不容易。要获得 U_t 的上界则较为简单，因为只需选取一个特定的可行点，而无需证明其最优性。具体而言，我们可以令 $\nu_t = \frac{\sigma_t}{\mu_t\sqrt{\hat{U}_t}}$，然后验证采用以下决策变量取值 [2] 时，(12) 是可行的：

$$\rho_t^2 = \mu_t^2\left(1+\nu_t\right),\ \lambda_t^{(1)} = \frac{\mu_t\alpha_t}{L_t-m_t}\left(1+\nu_t\right),\ \lambda_t^{(2)} = 1+\frac{1}{\nu_t}.$$

因此，我们有 $\hat{U}_{t+1} \leq (\mu_t\sqrt{\hat{U}_t}+\sigma_t)^2$。[17] 中的原始序贯半定规划分析正是基于选择特定的可行点（未对最优性进行论证）。然而，要获得紧性则更加困难。接下来，我们将建立新的代数论证，为针对各种非精确在线优化方法构造的序贯半定规划问题（11）推导精确解析解。我们针对非精确在线梯度下降方法所得出的结果之一实际上将精确在线梯度下降分析作为特例涵盖在内，并指出（13）的精确解实际上由 $\hat{U}_{t+1} = (\mu_t\sqrt{\hat{U}_t}+\sigma_t)^2$ 给出（参见第 3.1 节中的注 3.1）。

3 非精确在线梯度下降的跟踪误差界

我们首先应用序贯半定规划方法，为非精确在线梯度下降方法推导跟踪误差界。

3.1 具有加性绝对误差的非精确在线梯度下降分析

不精确OGD方法的更新规则如下：

$$x_{t+1} = x_t - \alpha_t(\nabla f_t(x_t) + v_t), \tag{14}$$

其中， v_t 是用于刻画梯度不精确性的误差项。我们首先假设 v_t 是满足界限 $\|v_t\| \leq c_t$的绝对误差。我们有如下结果。

引理 3.1。考虑不精确OGD方法 (14)。对于所有 t，我们假设：i) $f_t \in \mathcal{S}(m_t, L_t)$；ii)$\|x_{t+1}^* - x_t^*\| \leq \sigma_t$；iii) $\|v_t\| \leq c_t$。如果存在非负标量 $(\rho_t, \lambda_t^{(1)}, \lambda_t^{(2)}, \lambda_t^{(3)})$，使得以下矩阵不等式成立

$$\begin{bmatrix} 1-\rho_t^2-2\lambda_t^{(1)}L_tm_t & -\alpha_t+\lambda_t^{(1)}(L_t+m_t) & 1 & -\alpha_t \\ -\alpha_t+\lambda_t^{(1)}(L_t+m_t) & \alpha_t^2-2\lambda_t^{(1)} & -\alpha_t & \alpha_t^2 \\ 1 & -\alpha_t & 1-\lambda_t^{(2)} & -\alpha_t \\ -\alpha_t & \alpha_t^2 & -\alpha_t & \alpha_t^2-\lambda_t^{(3)} \end{bmatrix} \preceq 0, \tag{15}$$

那么，由 (14) 得到的迭代序列满足 $\left\|x_{t+1}-x_{t+1}^*\right\|^2 \leq \rho_t^2\left\|x_t-x_t^*\right\|^2 + \lambda_t^{(2)}\sigma_t^2 + \lambda_t^{(3)}c_t^2$。

[2]若 $0 \leq \alpha_t \leq \frac{2}{L_t+m_t}$，将上述参数代入 (12) 式的左侧可得

$$\frac{(1+\nu_t)\left((m_t+L_t)\alpha_2^2-2\alpha_t\right)}{L_t-m_t}\begin{bmatrix} m_t^2 & -m_t & 0 \\ -m_t & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} - \nu_t\begin{bmatrix} 1 \\ -\alpha_t \\ -\frac{1}{\nu_t} \end{bmatrix}\begin{bmatrix} 1 & -\alpha_t & -\frac{1}{\nu_t} \end{bmatrix} \preceq 0.$$

对于 $\frac{2}{L_t+m_t} \leq \alpha_t \leq \frac{2}{L_t}$ 的情形，验证过程类似。

证明。我们可以将 (14) 改写为 $x_{t+1}-x_{t+1}^*=x_t-x_t^*-\alpha\nabla f_t(x_t)+e_t-\alpha v_t$，其中 $e_t=x_t^*-x_{t+1}^*$。若令 $\xi_t=x_t-x_t^*$, $w_t=\begin{bmatrix}\nabla f_t(x_t)^T & e_t^T & v_t^T\end{bmatrix}^T$，且 $B_t=\begin{bmatrix}-\alpha_t & 1 & -\alpha_t\end{bmatrix}$，则 (14) 成为 (5) 的一个特例。假设 $f_t\in\mathcal{S}(m_t,L_t)$ 确保了一个供给率条件，其中 $X_t^{(1)}=\mathsf{diag}(Y_t,0,0)$，且 $\Lambda_t^{(1)}=0$（回忆一下，Y_t 由 (4) 定义）。界 $\|x_{t+1}^*-x_t^*\|\le\sigma_t$ 导出了第二个供给率条件，其中 $X_t^{(2)}=\mathsf{diag}(0,0,1,0)$，且 $\Lambda_t^{(2)}=\sigma_t^2$。根据 $\|v_t\|\le c_t$，我们有$X_t^{(3)}=\mathsf{diag}(0,0,0,1)$，且 $\Lambda_t^{(3)}=c_t^2$。由此得到所需的线性矩阵不等式。 □

接下来，我们精确求解序贯半定规划 (11)，并得到以下界。

定理 3.2. 假设对于所有 t，$\alpha_t\le\frac{2}{L_t}$。令 $\mathcal{T}_t\subseteq\mathbb{R}_+^4$ 表示满足条件 (15) 的元组集合 $\left(\rho_t,\lambda_t^{(1)},\lambda_t^{(2)},\lambda_t^{(3)}\right)$。假设引理 3.1 中的假设对所有 t 均成立。那么，(11) 中定义的 ˆ界 U_t 在 $J=3$ 时为

$$\hat{U}_{t+1}=\left(\mu_t\sqrt{\hat{U}_t}+\sigma_t+\alpha_t c_t\right)^2. \tag{16}$$

证明. 根据命题 2.2，我们需要求解以下问题：

$$\hat{U}_{t+1}=\underset{(\rho_t,\lambda_t^{(1)},\lambda_t^{(2)},\lambda_t^{(3)})\in\mathcal{T}_t}{\text{minimize}}\ \rho_t^2\hat{U}_t+\lambda_t^{(2)}\sigma_t^2+\lambda_t^{(3)}c_t^2. \tag{17}$$

由于线性矩阵不等式约束，精确求解 (17) 并不容易。我们将把 (17) 转化为一个约束更简单的等价问题。令 $\lambda_t^{(2)}=1+\frac{1}{\nu_t}$ 和 $\lambda_t^{(3)}=\alpha_t^2(1+\nu_t)\left(1+\frac{1}{\zeta_t}\right)$。根据舒尔补，(15) 等价于 $\zeta_t>0$、$\nu_t>0$ 以及

$$\begin{bmatrix}(1+\nu_t)(1+\zeta_t)-\rho_t^2-2\lambda_t^{(1)}L_tm_t & \lambda_t^{(1)}(L_t+m_t)-\alpha_t(1+\nu_t)(1+\zeta_t)\\ \lambda_t^{(1)}(L_t+m_t)-\alpha_t(1+\nu_t)(1+\zeta_t) & \alpha_t^2(1+\nu_t)(1+\zeta_t)-2\lambda_t^{(1)}\end{bmatrix}\preceq 0.$$

令 $\tau_t:=\frac{2\lambda_t^{(1)}}{(1+\nu_t)(1+\zeta_t)}-\alpha_t^2$。我们有 $\lambda_t^{(1)}=\frac{1}{2}\left(\alpha_t^2+\tau_t\right)(1+\nu_t)(1+\zeta_t)$。定义 $h(\tau_t)$ 为

$$h(\tau_t)=1-\alpha_t(L_t+m_t)+\frac{\alpha_t^2(m_t^2+L_t^2)}{2}+\frac{\tau_t(L_t-m_t)^2}{4}+\frac{\alpha_t^2(\alpha_t(L_t+m_t)-2)^2}{4\tau_t}. \tag{18}$$

那么，上述矩阵不等式条件成立，当且仅当 $\zeta_t>0$、$\nu_t>0$、$\tau_t\ge 0$ 和 $(1+\nu_t)(1+\zeta_t)h(\tau_t)\le\rho_t^2$。容易验证，对于所有 $\tau_t\ge 0$，$h(\tau_t)\ge 0$。此外，我们可以证明，使 $h(\tau_t)$ 最小化的 τ_t 的最优选择必须满足 $\frac{\tau_t}{4}(L_t-m_t)^2=\frac{\alpha_t^2(\alpha_t(L_t+m_t)-2)^2}{4\tau_t}$，最终得到 $\min_{\tau_t\ge 0}h(\tau_t)=\mu_t^2$，其中 μ_t 由 (4) 定义。因此我们有

$$\begin{aligned}\hat{U}_{t+1}&=\min_{\nu_t>0}\left(\min_{\zeta_t>0}\left((1+\nu_t)\left(\min_{\tau_t\ge 0}\left((1+\zeta_t)h(\tau_t)\hat{U}_t+\alpha_t^2c_t^2\left(1+\frac{1}{\zeta_t}\right)+\frac{\sigma_t^2}{\nu_t}\right)\right)\right)\right)\\&=\min_{\nu_t>0}\left(\min_{\zeta_t>0}\left((1+\nu_t)\left((1+\zeta_t)\left(\min_{\tau_t\ge 0}h(\tau_t)\right)\hat{U}_t+\alpha_t^2c_t^2\left(1+\frac{1}{\zeta_t}\right)+\frac{\sigma_t^2}{\nu_t}\right)\right)\right)\\&=\min_{\nu_t>0}\left(\min_{\zeta_t>0}\left((1+\nu_t)\left((1+\zeta_t)\mu_t^2\hat{U}_t+\left(1+\frac{1}{\zeta_t}\right)\alpha_t^2c_t^2+\frac{\sigma_t^2}{\nu_t}\right)\right)\right)\end{aligned}$$

现在我们可以将 (17) 求解为

$$\hat{U}_{t+1} = \min_{\nu_t>0}\left((1+\nu_t)\left(\mu_t\sqrt{\hat{U}_t}+\alpha_t c_t\right)^2 + \alpha_t^2\left(1+\frac{1}{\nu_t}\right)\sigma_t^2\right) = \left(\mu_t\sqrt{\hat{U}_t}+\sigma_t+\alpha_t c_t\right)^2.$$

至此，证明完成。 □

ˆ ˆ 我们可以看到，(16) 给出了描述 $\{U_t\}$ 递推关系的简单解析公式；根据 $\sqrt{\hat{}}U_{t+1} = \mu_t\sqrt{\hat{}}U_t+\sigma_t+\alpha_t c_t$ 这一事实， U_t 也可以很容易地改写为关于 ˆ U_0 和 $\{\mu_k,\sigma_k\}_{k=0}^{t-1}$ 的函数。在第 6 节中，我们将说明如何将这一递归跟踪误差界转换为动态遗憾界。还可以很容易地验证，使 (16) 中迭代界最小化的 α_t 取值为 ˆ$\alpha_t = \frac{2}{L_t+m_t}$。若对所有 t 取 $L_t = L$、$m_t = m$ 和 $\alpha_t = \alpha$，则 (16) 中的递归界还可以进一步简化。此时，我们有 $\mu_t = \mu\ \forall t$，其中 μ 为

$$\mu := \begin{cases} 1-m\alpha & \text{if } 0\le\alpha\le\frac{2}{m+L} \\ \alpha L-1 & \text{if } \frac{2}{m+L}\le\alpha\le\frac{2}{L} \end{cases},$$

于是，(16) 可简化为 $\sqrt{\hat{}}U_{t+1} = \mu\sqrt{\hat{}}U_t+\sigma+\alpha c$，等价地可改写为 $\sqrt{\hat{}}U_t = \mu^t\left(\sqrt{\hat{}}U_0 - \frac{\sigma+\alpha c}{1-\mu}\right)+\frac{\sigma+\alpha c}{1-\mu}$。显然， $\sqrt{\hat{}}U_t$ 以 μ 指定的速率线性收敛至稳态值 $(\sigma+\alpha c)/(1-\mu)$。

注 3.1（与精确在线梯度下降结果的联系）。如果 $c_t = 0$，我们可以修改上述论证，证明式 (13) 的精确解析解为 ˆ $U_{t+1} = (\mu_t\sqrt{\hat{}}U_t+\sigma_t)^2$。相比之下，第 2 节中的可行点论证只能证明 ˆ $U_{t+1}\le(\mu_t\sqrt{\hat{}}U_t+\sigma_t)^2$。如果 $m_t = m$、$L_t = L$，并且对所有 t都有 $\alpha_t = \alpha$，则 ˆ U_t的公式变为 $\sqrt{\hat{}}U_t = \mu^t\left(\sqrt{\hat{}}U_0 - \frac{\sigma}{1-\mu}\right)+\frac{\sigma}{1-\mu}$。显然，$\sqrt{\hat{}}U_t$ 以由 μ给出的线性速率收敛到其稳态值（即 $\sigma/(1-\mu)$）。由于我们有 $\|x_t - x_t^*\|\le\sqrt{U_t}$，可以看出，我们的分析在 *[25,* 定理 3.*1]* 中恢复了先前的精确在线梯度下降界。如果 $\sigma\equiv 0$，我们还可以恢复静态优化情形下的收敛界，即 $\sqrt{\hat{}}U_t\le\mu^t\sqrt{\hat{}}U_0$。这为我们的统一分析提供了良好的合理性检验。

3.2 带加性相对误差的非精确在线梯度下降分析

接下来，我们考虑具有更复杂误差模型的非精确在线梯度下降方法 (14)。我们假设偏差项 v_t 满足 $\|v_t\|\le\delta_t\|\nabla f_t(x_t)\|$，其中 δ_t 是某个已知常数。这种相对误差界已在静态代价设定中得到广泛研究 [5, 12, 17, 24]。正如 [5], 中所述，该假设意味着 $|\sin(\theta)|\le\delta_t$，其中 θ 是 $\nabla f_t(x_t)$ 与真实更新方向 $(\nabla f_t(x_t)+v_t)$ 之间的夹角。现在，我们讨论这种相对误差项对在线梯度下降跟踪能力的影响。我们有如下线性矩阵不等式条件。

引理 3.3. 考虑非精确OGD方法 (14)。对于所有 t，我们假设：i) $f_t\in\mathcal{S}(m_t,L_t)$；ii) $\|x_{t+1}^*-x_t^*\|\le\sigma_t$；iii) $\|v_t\|\le\delta_t\|\nabla f_t(x_t)\|$ 对于 $\delta_t>0$。若存在非负标量 $(\rho_t,\lambda_t^{(1)},\lambda_t^{(2)},\lambda_t^{(3)})$，使得以下矩阵不等式成立

$$\begin{bmatrix} 1-\rho_t^2-2\lambda_t^{(1)}L_t m_t & -\alpha_t+\lambda_t^{(1)}(L_t+m_t) & 1 & -\alpha_t \\ -\alpha_t+\lambda_t^{(1)}(L_t+m_t) & \alpha_t^2-2\lambda_t^{(1)}+\delta_t^2\lambda_t^{(3)} & -\alpha_t & \alpha_t^2 \\ 1 & -\alpha_t & 1-\lambda_t^{(2)} & -\alpha_t \\ -\alpha_t & \alpha_t^2 & -\alpha_t & \alpha_t^2-\lambda_t^{(3)} \end{bmatrix}\preceq 0, \tag{19}$$

则非精确在线梯度下降更新 (14) 满足 $\left\|x_{t+1}-x_{t+1}^*\right\|^2 \leq \rho_t^2\left\|x_t-x_t^*\right\|^2+\lambda_t^{(2)}\sigma_t^2$。

证明。我们可以采用引理 3.1 证明中的（$X_t^{(1)}$，$\Lambda_t^{(1)}$，$X_t^{(2)}$，$\Lambda_t^{(2)}$）。不等式 $\|v_t\| \leq \delta_t\left\|\nabla f_t\left(x_t\right)\right\|$ 可以改写为第三个供给率条件，其中 $X_t^{(3)}=\operatorname{diag}(0,-\delta_t^2,0,1)$ 且$\Lambda_t^{(3)}=0$。由此可得所需结论。

□

需要注意的是，只有在对 δ_t 施加某个合理的界时，引理 3.3 中的线性矩阵不等式才是可行的。我们将在下一个结果中给出这样的界。可以看出，（19）与（15）非常相似。唯一的区别在于，线性矩阵不等式（19）的第（2，2）个元素中多了一个额外项 $\delta_t^2\lambda_t^{(3)}$。然而，这一微小变化会使所得半定规划问题（11）的解析解变得复杂得多。我们将这一结果形式化如下。

定理 3.4。假设对于所有 t，$0 \leq \alpha_t \leq \frac{2}{(1+\delta_t)L_t}$ 和 $\delta_t \in [0, \frac{2m_t}{L_t+m_t})$。令 $\mathcal{T}_t \subseteq \mathbb{R}_+^4$ 表示满足条件 (19) 的元组集合 $\left(\rho_t, \lambda_t^{(1)}, \lambda_t^{(2)}, \lambda_t^{(3)}\right)$。假设引理 3.3 中的条件对所有 t 均成立。那么，(11) 中定义的递归界 U_t 在 $J=3$ 下为 $\hat{U}_{t+1}=\left(\hat{\rho}_t \sqrt{\hat{U}_t}+\sigma_t\right)^2$，其中 $\hat{\rho}_t$ 定义为

$$\hat{\rho}_t:=\begin{cases}1-\alpha_t m_t(1-\delta_t) & \text{if } 0 \leq \alpha_t \leq \alpha_{t_-} \\ \left(1-\frac{2\alpha_t L_t m_t}{L_t+m_t}+\frac{\alpha_t\delta_t^2(L_t+m_t-2\alpha_t L_t m_t)}{2-\alpha_t(L_t+m_t)}\right)^{\frac{1}{2}} & \text{if } \alpha_{t_-} \leq \alpha_t \leq \alpha_{t_+} \\ (1+\delta_t)\alpha_t L_t-1 & \text{if } \alpha_{t_+} \leq \alpha_t \leq \frac{2}{(1+\delta_t)L_t}\end{cases} \tag{20}$$

对于 $\alpha_{t_-}:=\frac{1}{1-\delta_t}(\frac{2}{L_t+m_t}-\frac{\delta_t}{m_t})$ 和 $\alpha_{t_+}:=\frac{1}{1+\delta_t}(\frac{2}{L_t+m_t}+\frac{\delta_t}{L_t})$。

证明。根据命题 2.2，我们需要求解以下问题：

$$\hat{U}_{t+1}=\underset{(\rho_t,\lambda_t^{(1)},\lambda_t^{(2)},\lambda_t^{(3)})\in\mathcal{T}_t}{\text{minimize}} \rho_t^2\hat{U}_t+\lambda_t^{(2)}\sigma_t^2. \tag{21}$$

为了对 (21) 进行解析求解，我们将再次进行变量代换。令 $\lambda_t^{(2)}=1+\frac{1}{\nu_t}$ 和 $\lambda_t^{(3)}=\alpha_t^2(1+\nu_t)\left(1+\frac{1}{\zeta_t}\right)$。然后，我们可以应用舒尔补，将 (19) 转换为以下等价形式：

$$\begin{bmatrix}\psi\left(\zeta_t,\nu_t\right)-\rho_t^2-2L_t m_t\lambda_t^{(1)} & \lambda_t^{(1)}\left(L_t+m_t\right)-\alpha_t\psi\left(\zeta_t,\nu_t\right) \\ \lambda_t^{(1)}\left(L_t+m_t\right)-\alpha_t\psi\left(\zeta_t,\nu_t\right) & \alpha_t^2\psi\left(\zeta_t,\nu_t\right)\left(1+\frac{\delta_t^2}{\zeta_t}\right)-2\lambda_t^{(1)}\end{bmatrix} \preceq 0,\ \zeta_t>0,\ \nu_t>0,$$

其中，$\psi(\zeta_t,\nu_t):=(1+\zeta_t)(1+\nu_t)$。然后，我们设 $\tau_t:=2\lambda_t^{(1)}-\alpha_t^2\psi(\zeta_t,\nu_t)\left(1+\frac{\delta_t^2}{\zeta_t}\right)$。因此，有 $\lambda_t^{(1)}=\frac{1}{2}\left(\tau_t+\alpha_t^2\psi(\zeta_t,\nu_t)\left(1+\frac{\delta_t^2}{\zeta_t}\right)\right)$。通过应用舒尔补并设$\chi_t:=\alpha_t\left(1+\frac{\delta_t^2}{\zeta_t}\right)$，可以证明，上述线性矩阵不等式成立，当且仅当 $\zeta_t>0$、$\nu_t>0$、$\tau_t \geq 0$ 和$h(\zeta_t,\nu_t,\tau_t) \leq \rho_t^2$，其中 $h(\zeta_t,\nu_t,\tau_t)$定义为

$$\begin{aligned}h\left(\zeta_t,\nu_t,\tau_t\right)=-\psi(\zeta_t,\nu_t)\left(\alpha_t L_t m_t\chi_t-1\right)+\frac{\alpha_t^2\psi(\zeta_t,\nu_t)^2}{4\tau_t}\left(2-\chi_t\left(L_t+m_t\right)\right)^2+\frac{\tau_t}{4} \\ \left(L_t-m_t\right)^2-\frac{\alpha_t\psi(\zeta_t,\nu_t)}{2}\left(L_t+m_t\right)\left(2-\chi_t\left(L_t+m_t\right)\right).\end{aligned}$$

因此，(21)等价于以下优化问题：

$$\hat{U}_{t+1} = \min_{\nu_t>0} \left(\min_{\zeta_t>0} \left(\min_{\tau_t\geq 0} \left(\min_{h(\zeta_t,\nu_t,\tau_t)\leq \rho_t^2} \left(\rho_t^2 \hat{U}_t + \left(1+\frac{1}{\nu_t}\right)\sigma_t^2 \right) \right) \right) \right). \tag{22}$$

注意到 $1-\alpha_t(m_t+L_t) + \frac{\alpha_t^2}{(m_t^2+L_t^2)_2} \geq 0$，因此对于所有 $\nu_t>0$、$\zeta_t>0$ 和 $\tau_t\geq 0$，我们有 $h(\zeta_t,\nu_t,\tau_t)\geq 0$。因此，我们可以将（22）简化为

$$\begin{aligned}
\hat{U}_{t+1} &= \min_{\nu_t>0} \left(\min_{\zeta_t>0} \left(\min_{\tau_t\geq 0} \left(h(\zeta_t,\nu_t,\tau_t)\hat{U}_t + \left(1+\frac{1}{\nu_t}\right)\sigma_t^2 \right) \right) \right) \\
&= \min_{\nu_t>0} \left(\min_{\zeta_t>0} \left(\hat{h}(\zeta_t,\nu_t)\hat{U}_t + \left(1+\frac{1}{\nu_t}\right)\sigma_t^2 \right) \right) \\
&= \min_{\nu_t>0} \left(\left(\min_{\zeta_t>0} \hat{h}(\zeta_t,\nu_t) \right)\hat{U}_t + \left(1+\frac{1}{\nu_t}\right)\sigma_t^2 \right),
\end{aligned}$$

ˆ其中 $h(\zeta_t,\nu_t) := \min_{\tau_t\geq 0} h(\zeta_t,\nu_t,\tau_t)$ 得出以下显式公式：

$$\begin{aligned}
\hat{h}(\zeta_t,\nu_t) := \psi(\zeta_t,\nu_t)(1-\alpha_t L_t m_t \chi_t) + \frac{\alpha_t\psi(\zeta_t,\nu_t)(L_t-m_t)}{2}\Bigg| \alpha_t(L_t+m_t) \\
\left(1+\frac{\delta_t^2}{\zeta_t}\right) - 2 \Bigg| - \frac{\alpha_t\psi(\zeta_t,\nu_t)(L_t+m_t)}{2}(2-\chi_t(L_t+m_t)).
\end{aligned} \tag{23}$$

显然，上述最优 ˆh 在 $\tau_t = {}^{\alpha_t\psi(}_{L_t-m_t}{}^{\zeta_t\ \nu_t}\Big|\alpha_t{}^{,)}(L_t+m_t)\left(1+\frac{\delta_t^2}{\zeta_t}\right)-2\Big|$ 处取得。证明的其余部分依赖于以下关键观察：

$$\min_{\zeta_t>0} \hat{h}(\zeta_t,\nu_t) = (1+\nu_t)\hat{\rho}_t^2, \tag{24}$$

其中，ˆρ_t 由式 (20) 定义。一旦证明了式 (24)，便可轻易验证

$$\hat{U}_{t+1} = \min_{\nu_t>0} \left((1+\nu_t)\hat{\rho}_t^2\hat{U}_t + \left(1+\frac{1}{\nu_t}\right)\sigma_t^2 \right) = \left(\hat{\rho}_t\sqrt{\hat{U}_t} + \sigma_t \right)^2,$$

这将直接得到所需的结论。

ˆ 为验证式 (24)，我们将利用式 (23) 推导 $\min_{\zeta_t>0} h(\zeta_t,\nu_t)$ 的解析形式，这在很大程度上取决于 $\bar{\zeta}_t$ 的符号：$= {}^{\delta_t^2\alpha_t}_{(L_t+m_t)2-\alpha_t(L_t+m_t)}$ 如果 $\alpha_t \geq \frac{2}{L_t+m_t}$，则有 $\bar{\zeta}_t \leq 0$。在这种情况下，我们可以应用式 (23) 证明，对于所有 $\zeta_t>0$，$\hat{h}(\zeta_t,\nu_t) = (1+\zeta_t)(1+\nu_t)\left((\alpha_t L_t-1)^2 + \frac{\alpha_t^2\delta_t^2 L_t^2}{\zeta_t}\right)$。然后，我们可以通过选择 $\zeta_t = \frac{\alpha_t\delta_t L_t}{|\alpha_t L_t - 1|}$ 来最小化 ˆh。因此，对于 $\alpha \geq \frac{2}{m_t+L_t}$，式 (24) 成立。注意，我们要求 ˆ$\rho_t<1$，以使迭代界有意义。这对步长 α_t 构成了一个上界。具体而言，我们需要 $\alpha_t \leq {}^{2}{}_{(1+\delta_t)L_t}$。

如果 $\alpha_t < \frac{2}{m+L}$，则有 $\bar{\zeta}_t > 0$。于是，由（23）可得

$$\hat{h}(\zeta_t,\nu_t) = \begin{cases} (1+\zeta_t)(1+\nu_t)\left((\alpha_t L_t-1)^2 + \frac{\alpha_t^2\delta_t^2 L_t^2}{\zeta_t}\right) & \text{if } 0<\zeta_t\leq\bar{\zeta}_t \\ (1+\zeta_t)(1+\nu_t)\left((1-\alpha_t m_t)^2 + \frac{\alpha_t^2\delta_t^2 m_t^2}{\zeta_t}\right) & \text{if } \zeta_t>\bar{\zeta}_t \end{cases} \tag{25}$$

为得到 $\min_{\zeta_t>0} h(\zeta_t, \nu_t)$ 的解析形式，我们定义 $\zeta_{tL} := \frac{\alpha_t\delta_t L_t}{|\alpha_t L_t - 1|}$ 和 $\zeta_{tm} := \frac{\alpha_t\delta_t m_t}{1-m_t\alpha_t}$。由于事实 $\alpha_t < \frac{2}{m_t+L_t}$，我们可以证明 $\zeta_{tm} \le \zeta_{tL}$。$\min_{\zeta_t>0} h(\zeta_t, \nu_t)$ 的解析形式取决于 ζ_t 与 (ζ_{tm}, ζ_{tL}) 之间的排序关系，具体如下。

- 若 $\zeta_t \le \zeta_{tm} \le \zeta_{tL}$，我们可以固定 ν_t，并证明 h 在 $0 < \zeta_t \le \zeta_t$时单调递减。由于 $\zeta_{tm} \ge \zeta_t$，我们知道 h 在 $\zeta_t \le \zeta_{tm}$ 时单调递减，并在 $\zeta_t \ge \zeta_{tm}$时单调递增。因此，h 在 $\zeta_t = \zeta_{tm}$处取得最小值。关系式 $\zeta_t \le \zeta_{tm}$可以等价地转换为步长界 $0 \le \alpha_t \le \alpha_{t-}$。
- 若 $\zeta_{tm} \le \zeta_t \le \zeta_{tL}$，则 h 在 $0 < \zeta_t \le \zeta_t$ 时单调递减，并在 $\zeta_t \ge \zeta_t$时单调递增。因此，h 在 $\zeta_t = \zeta_t$处取得全局最小值。关系式 $\zeta_{tm} \le \zeta_t \le \zeta_{tL}$ 可以转换为步长界 $\alpha_{t-} \le \alpha_t \le \alpha_{t+}$。
- 若 $\zeta_{tm} < \zeta_{tL} \le \zeta_t$，我们可以应用类似的论证，证明 h 对于 $\zeta_t \le \zeta_{tL}$ 单调递减，而对于 $\zeta_t \ge \zeta_{tL}$ 单调递增。因此，h 在 $\zeta_t = \zeta_{tL}$处达到最小值，且界 $\zeta_{tm} < \zeta_{tL} \le \zeta_t$ 等价于步长界 $\alpha_t \ge \alpha_{t+}$。

总之，对于任意 $\nu_t > 0$，通过选择 $\zeta_t = \frac{\alpha_t\delta_t m_t}{1-m_t\alpha_t}$，函数 $h(\zeta_t, \nu_t)$ 在 $0 \le \alpha_t \le \alpha_{t-}$，$\zeta_t = \frac{\delta_t^2\alpha_t(L_t+m_t)}{2-\alpha_t(L_t+m_t)}$处取得最小值；对于 $\alpha_{t-} \le \alpha_t \le \alpha_{t+}$，则在 $\zeta_t = \frac{\alpha_t\delta_t L_t}{\alpha_t L_t - 1}$处取得最小值；对于 $\alpha_{t+} \le \alpha_t \le \frac{2}{(1+\delta_t)L_t}$，则在相应位置取得最小值。因此，(24) 如期成立，证明完毕。 □

上述定理刻画了跟踪误差对 δ_t 的依赖关系。显然，增加 δ_t 会导致收敛速度变慢以及稳态误差增大。如果对所有 t 都有 $\delta_t = 0$，则我们的界约化为精确在线梯度下降界，即 $\sqrt{U_{t+1}} = \mu_t\sqrt{U_t} + \sigma_t$。如果对所有 t 都有 $\sigma_t = 0$（即 $x_{t+1}^* = x_t^*$），并且（m_t，L_t，δ_t）在 t 上不发生变化，则我们的界约化为 [12, 命题 1.3 和 1.4]，它们给出了静态代价设定下具有相对误差的非精确梯度法的收敛速率。

备注 3.2。若对所有 t 取 $L_t = L$、$m_t = m$、$\alpha_t = \alpha$ 和 $\delta \ge \delta_t$，则递归界 $U_{t+1} = (\rho_t\sqrt{U_t} + \sigma_t)^2$ 可进一步简化。在这种情况下，我们有 $\rho_t = \rho$。简化后的收敛界为 $\sqrt{U_{t+1}} = \rho\sqrt{U_t} + \sigma_t$，由此得到渐近界 $\sqrt{U_t} = \rho^t\left(\sqrt{U_0} - \frac{\sigma}{1-\rho}\right) + \frac{\sigma}{1-\rho}$。显然，$\sqrt{U_t}$ 以由 ρ 量化的收敛速率线性收敛至稳态值 $\sigma/(1-\rho)$。

3.3 非精确变分不等式设定下的分析

现在我们考虑一种不同的广义设定，将不精确OGD方法扩展到变分不等式（VI）问题。在每个时间步 t，VI 问题都要求找到点 x_t^*，使得 $F_t(x_t^*)^T(x - x_t^*) \ge 0\ \forall x \in \mathcal{X}$，其中 $F_t\colon \mathbb{R}^p \to \mathbb{R}^p$ 是一个向量场，而 $\mathcal{X}$是一个给定的凸集。为简单起见，我们考虑 $\mathcal{X} = \mathbb{R}^p$，因此 x_t^* 满足 $F_t(x_t^*) = 0$。以下算法可视为将不精确OGD方法扩展到 VI 设定：

$$x_{t+1} = x_t - \alpha_t(F_t(x_t) + v_t). \tag{26}$$

我们假设 v_t 是一个满足一般误差界 $\|v_t\|^2 \le \delta_t^2\|F_t(x_t)\|^2 + c_t^2$ 的偏差项，其中δ_t 和 c_t。如果 F_t 恰好是某个函数 f_t的梯度，那么上述算法就退化为

不精确OGD方法。然而，上述 VI 设定更为一般，可以涵盖极小极大博弈问题，此时 (26) 实际上变为梯度下降-上升方法的在线版本 [43]。

我们感兴趣的是分析由 (26) 生成的迭代序列 $\{x_t\}$ 对序列 $\{x_t^*\}$ 的跟踪精度。我们作出标准假设 [43]，即 F_t 是强单调的[3]，且满足利普希茨条件。这些假设导出了新的供给率条件，并据此构造以下线性矩阵不等式条件。

引理 3.5。考虑更新规则 (26)。对于所有 t，我们假设：i) F_t 是 m_t-强单调的；ii) F_t 满足 L_t-利普希茨条件，即对于所有 x、$\tilde{x} \in \mathbb{R}^p$，$\|F_t(x) - F_t(\tilde{x})\| \le L_t \|x - \tilde{x}\|$；iii) $\|v_t\|^2 \le \delta_t^2 \|F_t(x_t)\|^2 + c_t^2$；以及 iv) $\|x_{t+1}^* - x_t^*\| \le \sigma_t$。若存在非负标量 $(\rho_t, \lambda_t^{(1)}, \lambda_t^{(2)}, \lambda_t^{(3)}, \lambda_t^{(4)})$，使得

$$\begin{bmatrix} 1-\rho_t^2+\lambda_t^{(4)}L_t^2-2\lambda_t^{(1)}m_t & \lambda_t^{(1)}-\alpha_t & 1 & -\alpha_t \\ \lambda_t^{(1)}-\alpha_t & \alpha_t^2-\lambda_t^{(4)}+\delta_t^2\lambda_t^{(3)} & -\alpha_t & \alpha_t^2 \\ 1 & -\alpha_t & 1-\lambda_t^{(2)} & -\alpha_t \\ -\alpha_t & \alpha_t^2 & -\alpha_t & \alpha_t^2-\lambda_t^{(3)} \end{bmatrix} \preceq 0, \tag{27}$$

那么，(26) 的迭代序列满足 $\left\|x_{t+1} - x_{t+1}^*\right\|^2 \le \rho_t^2 \|x_t - x_t^*\|^2 + \lambda_t^{(2)}\sigma_t^2 + \lambda_t^{(3)}c_t^2$。

证明。记 $e_t = x_t^* - x_{t+1}^*$。于是，(26) 可改写为 $x_{t+1} - x_{t+1}^* = x_t - x_t^* - \alpha_t F_t(x_t) + e_t - \alpha_t v_t$，这是 (5) 在 $\xi_t = x_t - x_t^*$、$B_t = \begin{bmatrix} -\alpha_t & 1 & -\alpha_t \end{bmatrix}$ 和 $w_t = \begin{bmatrix} F_t(x_t)^T & e_t^T & v_t^T \end{bmatrix}^T$ 下的一个特例。由于假设 F_t 是 m_t-强单调的，我们知道 $F_t(x_t)^T(x_t - x_t^*) \ge m_t \|x_t - x_t^*\|^2$。记 $M_t = \begin{bmatrix} 2m_t & -1 \\ -1 & 0 \end{bmatrix}$，可得满足 $\left(X_t^{(1)}, \Lambda_t^{(1)}\right) = \left(\begin{bmatrix} M_t & 0_{2\times 2} \\ 0_{2\times 2} & 0_{2\times 2} \end{bmatrix}, 0\right)$ 的供给率 (6)。假设 $\|x_{t+1}^* - x_t^*\| \le \sigma_t$ 导出第二个供给率条件，其中 $X_t^{(2)} = \mathrm{diag}(0, 0, 1, 0)$，且 $\Lambda_t^{(2)} = \sigma_t^2$。关于 v_t 的误差界可改写为第三个供给率条件，其中 $X_t^{(3)} = \mathrm{diag}(0, -\delta_t^2, 0, 1)$，且 $\Lambda_t^{(3)} = c_t^2$。最后，假设 F_t 是 L_t-利普希茨的，由此得到第四个供给率条件，其中 $X_t^{(4)} = \mathrm{diag}(-L_t^2, 1, 0, 0)$，且 $\Lambda_t^{(4)} = 0$。然后，我们可以直接应用命题 2.1，得到线性矩阵不等式条件 (27)。

□

同样，我们可以求解由此得到的序贯半定规划，并得到以下界。

定理 3.6. 假设 $\alpha_t \le {}^{2}\!/_{(m_t - \delta_t L_t)\frac{L_t^2(1-\delta_t^2)}{L_t}}$ 且对所有 t，$\delta_t \le \frac{m_t}{L_t}$。令 $\mathcal{T}_t \subseteq \mathbb{R}_+^5$ 为满足 (27) 的元组 $(\rho_t, \lambda_t^{(1)}, \lambda_t^{(2)}, \lambda_t^{(3)}, \lambda_t^{(4)})$ 的集合。假设引理 3.5 中的条件成立。则在 $J = 4$ 下，式 (11) 中定义的ˆ界 U_t 为

$$\hat{U}_{t+1} = \left(\sqrt{(1 - 2m_t\alpha_t + \alpha_t^2 L_t^2)\hat{U}_t} + \alpha_t\sqrt{c_t^2 + \delta_t^2 L_t^2 \hat{U}_t} + \sigma_t\right)^2. \tag{28}$$

证明。证明仍依赖于舒尔补引理和一些代数运算。详细证明见附录。

□

[3]如果对于给定的 $m_t > 0$，对所有 $x, \tilde{x} \in \mathbb{R}^p$ 均有 $F_t(x) - F_t(\tilde{x})^T(x - \tilde{x}) \ge m_t \|x - \tilde{x}\|^2$，则称映射 F_t 为 m_t-强单调。

由 (28) 可得 $\sqrt{U_{t+1}} \leq (\sqrt{1-2m_t\alpha_t+\alpha_t^2L_t^2}+\alpha_tL_t\delta_t)\sqrt{\hat{}}U_t+\alpha_t\hat{c}_t+\sigma_t$。条件 $\alpha_t \leq \frac{2(m_t-\delta_tL_t)}{L_t^2(1-\delta_t^2)}$ 以及 $\delta_t \leq \frac{m_t}{L_t}$ 确保了 $\sqrt{1-2m_t\alpha_t+\alpha_t^2L_t^2}+\alpha_tL_t\delta_t \leq 1$。我们还可以进一步推导出动态遗憾界。该结果将在第 6 节中给出。如果对于所有 t，$L_t=L$、$m_t=m$、$\delta_t=\delta$、$\hat{c}_t=c$、$\sigma_t=\sigma$和 $\alpha_t=\alpha$ 均成立，那么很容易证明 U_t 线性收敛到一个小球，如下所示：

$$\hat{U}_t \leq (\sqrt{1-2m\alpha+\alpha^2L^2}+\alpha L\delta)^t\hat{U}_0+\frac{\alpha c+\sigma}{1-\sqrt{1-2m\alpha+\alpha^2L^2}+\alpha L\delta}.$$

注 3.3. 我们对定理 3.6 的证明稍作修改，即可证明 *[17, 注 3]* 中的递归界实际上给出了 *[17, 命题 1]*中序贯半定规划的精确解。因此，我们的证明技术弥补了 *[17]*中的分析空缺。更多说明见附录。

4 在线随机梯度方法的跟踪误差界

本节考虑在线梯度预言机中的不精确性具有随机性的情形。我们将针对这类随机在线算法的均方误差分析，对序贯半定规划方法进行相应调整。

4.1 随机动态系统的耗散性理论

第 2 节提出的耗散性框架可用于分析受到随机噪声 [17, 19]影响的动态系统。考虑输入序列 $\{w_t\}$ 为随机过程的动态系统 (5)。我们可以将命题 2.1 修改如下。

命题 4.1. 考虑系统 (5)，其中 $\{w_t\}$ 为随机过程。假设 $\{\xi_t, w_t\}$满足以下关于 $j=1,\cdots,J$的期望供给率条件：

$$\mathbb{E}\left(\begin{bmatrix}\xi_t\\ w_t\end{bmatrix}^T(X_t^{(j)}\otimes I_p)\begin{bmatrix}\xi_t\\ w_t\end{bmatrix}\right) \leq \Lambda_t^{(j)}. \tag{29}$$

如果存在非负标量 $\{\lambda_t^{(j)}\}_{j=1}^J$ 和 ρ_t 使线性矩阵不等式条件（7）成立，则必有 $\mathbb{E}\|\xi_{t+1}\|^2 \leq \rho_t^2\mathbb{E}\|\xi_t\|^2+\sum_{j=1}^J\lambda_t^{(j)}\Lambda_t^{(j)}$。此外，如果（29）对于 $j\in\Omega$ 中的 Ω 取等号，其中 Ω 是 $\{1,2,\ldots,J\}$ 的子集，则允许 $\lambda_t^{(j)}$ 对于任意 $j\in\Omega$ 取负值，且线性矩阵不等式条件仍然成立。

证明。由（7）可知，（9）几乎必然成立。于是，我们可以对（9）两侧取期望，并应用期望供给率条件（29），从而得到所需的结论[4]。 □

基于上述结果，我们仍然可以应用命题 2.2 中的序贯半定规划方法，为受随机噪声影响的在线优化方法获得均方跟踪误差界。接下来，我们将介绍两个案例研究。

[4]考虑（29）对于 $j\in\Omega$ 取等号的情形。对于任意 $j\in\Omega$，我们可以允许 $\lambda_t^{(j)}$ 取负值，并且不难验证所需的结论仍然成立。

4.2 加性 IID 噪声下随机在线梯度下降的分析

现在，在另一种假设下，我们研究非精确 OGD 方案（14）的性能：序列 $\{v_t\}$ 是满足均方误差界 $\mathbb{E}\|v_t\|^2 \le c_t^2$ 的零均值 IID 过程。我们可以应用命题 29 得到以下结果。

引理 4.1。考虑递推关系 (14)。对于所有 t，我们假设：i) $f_t \in \mathcal{S}(m_t, L_t)$，ii) $\|x_{t+1}^* - x_t^*\| \le \sigma_t$，以及 iii) $\{v_t\}$ 是满足 $\mathbb{E}\|v_t\|^2 \le c_t^2$ 的零均值 IID 过程。如果存在非负标量 $(\rho_t, \lambda_t^{(1)}, \lambda_t^{(2)}, \lambda_t^{(3)})$ 和实标量 $(\lambda_t^{(4)}, \lambda_t^{(5)}, \lambda_t^{(6)})$，使得以下矩阵不等式成立（ Y_t 的定义见 (4)）

$$\begin{bmatrix} 1-\rho^2 & -\alpha_t & 1 & -\alpha_t - \lambda_t^{(4)} \\ -\alpha_t & \alpha_t^2 & -\alpha_t & \alpha_t^2 - \lambda_t^{(5)} \\ 1 & -\alpha_t & 1-\lambda_t^{(2)} & -\alpha_t - \lambda_t^{(6)} \\ -\alpha_t - \lambda_t^{(4)} & \alpha_t^2 - \lambda_t^{(5)} & -\alpha_t - \lambda_t^{(6)} & \alpha_t^2 - \lambda_t^{(3)} \end{bmatrix} - \lambda_t^{(1)} \begin{bmatrix} Y_t & 0_{2\times 2} \\ 0_{2\times 2} & 0_{2\times 2} \end{bmatrix} \preceq 0, \tag{30}$$

那么必有 $\mathbb{E}\left\|x_{t+1} - x_{t+1}^*\right\|^2 \le \rho_t^2 \mathbb{E}\|x_t - x_t^*\|^2 + \lambda_t^{(2)}\sigma_t^2 + \lambda_t^{(3)} c_t^2$。

证明。令 $\xi_t = x_t - x_t^*$、$w_t = \begin{bmatrix} \nabla f_t(x_t)^T & e_t^T & v_t^T \end{bmatrix}^T$ 以及 $B_t = \begin{bmatrix} -\alpha_t & 1 & -\alpha_t \end{bmatrix}$。假设 (i) 和 (ii) 使我们能够用 $(X_t^{(1)}, \Lambda_t^{(1)}) = (\mathrm{diag}(Y_t, 0, 0), 0)$ 和 $(X_t^{(2)}, \Lambda_t^{(2)}) = (\mathrm{diag}(0, 0, 1, 0)\, \sigma_t^2)$ 指定前两个期望供给率条件。均方误差界 $\mathbb{E}\|v_t\|^2 \le c_t^2$ 可以改写为另一个期望供给率条件，即 $(X_t^{(3)}, \Lambda_t^{(3)}) = (\mathrm{diag}(0, 0, 0, 1)\, c_t^2)$。基于 $\{v_t\}$ 是一个零均值 IID 过程这一事实，我们有 $\mathbb{E}\left(v_t^T (x_t - x_t^*)\right) = 0$、$\mathbb{E}\left(v_t^T \nabla f_t(x_t)\right) = 0$ 以及 $\mathbb{E}\left(v_t^T e_t\right) = 0$。这三个条件给出了等式形式的期望供给率条件，并可与前三个供给率条件结合，以推导出所需的线性矩阵不等式条件。注意，由于相关的供给率条件采用等式形式，$(\lambda_t^{(4)}, \lambda_t^{(5)}, \lambda_t^{(6)})$ 可以为负。

□

接下来，我们可以应用序贯半定规划方法，得到以下递归均方跟踪误差界[5]。

定理 4.2. 假设 $\alpha_t \le \frac{2}{L_t}$。令 $\mathcal{T}_t \subseteq \mathbb{R}_+^4 \times \mathbb{R}^3$ 表示满足线性矩阵不等式条件 (30) 的元组集合 $\left(\rho_t, \lambda_t^{(1)}, \ldots, \lambda_t^{(6)}\right)$。如果 (4.1) 中的假设成立，那么由 (11) 定义的界 $\hat{} U_t$ 在 $J = 6$ 下恰好为

$$\hat{U}_{t+1} = \left(\mu_t \sqrt{\hat{U}_t} + \sigma_t\right)^2 + \alpha_t^2 c_t^2. \tag{31}$$

$\hat{}$ 证明。根据命题 2.2，我们知道 U_{t+1} 可以计算为

$$\hat{U}_{t+1} = \underset{(\rho_t, \lambda_t^{(1)}, \ldots, \lambda_t^{(6)}) \in \mathcal{T}_t}{\text{minimize}} \ \rho_t^2 \hat{U}_t + \lambda_t^{(2)}\sigma_t^2 + \lambda_t^{(3)} c_t^2. \tag{32}$$

[5]在这种情况下，我们将有 $\mathbb{E}\|x_t - x_t^*\|^2 \le \hat{} U_t$，其中 $\hat{} U_t$ 由命题 2.2 中的序贯半定规划问题求解得到。

对于任意的 $(\rho_t, \lambda_t^{(1)}, \cdots, \lambda_t^{(6)}) \in \mathcal{T}_t$，(30) 左侧的 3×3 左上块必须是负半定的，因此我们知道 (12) 成立。令 $\hat{\mathcal{T}}_t \subset \mathbb{R}_+^4 \times \mathbb{R}^3$ 表示满足线性矩阵不等式条件 (12) 的元组$\left(\rho_t, \lambda_t^{(1)}, \cdots, \lambda_t^{(6)}\right)$ 集合。于是我们知道 $\mathcal{T}_t \subset \hat{\mathcal{T}}_t$。注意到对于 $\mathcal{T}_t$中的任意点都有 $\lambda_t^{(3)} \geq \alpha_t^2$，我们可以作出如下断言：

$$\begin{aligned}
\hat{U}_{t+1} &\geq \underset{(\rho_t, \lambda_t^{(1)}, \cdots, \lambda_t^{(6)}) \in \mathcal{T}_t}{\text{minimize}} \rho_t^2 \hat{U}_t + \lambda_t^{(2)} \sigma_t^2 + \alpha_t^2 c_t^2 \\
&\geq \underset{(\rho_t, \lambda_t^{(1)}, \cdots, \lambda_t^{(6)}) \in \hat{\mathcal{T}}_t}{\text{minimize}} \rho_t^2 \hat{U}_t + \lambda_t^{(2)} \sigma_t^2 + \alpha_t^2 c_t^2 \\
&= \underset{(\rho_t, \lambda_t^{(1)}, \lambda_t^{(2)}) \in \hat{\mathcal{V}}_t}{\text{minimize}} \rho_t^2 \hat{U}_t + \lambda_t^{(2)} \sigma_t^2 + \alpha_t^2 c_t^2 \\
&= \left(\mu_t \sqrt{\hat{U}_t} + \sigma_t\right)^2 + \alpha_t^2 c_t^2,
\end{aligned}$$

其中，$\hat{\mathcal{V}}_t \subset \mathbb{R}_+^3$ 表示满足线性矩阵不等式条件 (12) 的元组集合 $\left(\rho_t, \lambda_t^{(1)}, \lambda_t^{(2)}\right)$，最后一步可利用第 3.1 节中的论证轻松验证。也很容易验证，$\rho_t^2 = \mu_t^2(1+\nu_t)$ $\lambda_t^{(1)} = \frac{\mu_t \alpha_t}{(1+\nu_t) L_t - m_t}$，$\lambda_t^{(2)} = \alpha_t^2$，$\lambda_t^{(3)} = 1 + \frac{1}{\nu_t}$，$\lambda_t^{(4)} = -\alpha_t$，$\lambda_t^{(5)} = \alpha_t^2$，$\lambda_t^{(6)} = -\alpha_t$是 $\mathcal{T}_t$ 中的一个可行点。对于上述可行点，我们有 $\rho_t^2 \hat{U}_t + \lambda_t^{(2)} \sigma_t^2 + \alpha_t^2 c_t^2 = \left(\mu_t \sqrt{\hat{U}_t} + \sigma_t\right)^2 + \alpha_t^2 c_t^2$。因此，我们也有 $\hat{U}_{t+1} \leq \left(\mu_t \sqrt{\hat{U}_t} + \sigma_t\right)^2 + \alpha_t^2 c_t^2$。结合 U_{t+1}的 $\hat{}$ 上下界，即可得到所需结论。

□

备注4.1。我们可以看到，(31)中的界始终小于或等于(16)中的界。显然，关于 $\{v_t\}$ 的随机性假设使我们能够推导出一个改进界。如果对于所有 t，我们都有$L_t = L$、$m_t = m$、$\alpha_t = \alpha$、$\sigma_t = \sigma$和 $c_t = c$，那么(31)就变为$\hat{} U_{t+1} = (\mu \sqrt{\hat{} U_t + \sigma})^2 + \alpha^2 c^2$。此时，(31)和(16)似乎都给出了界 $\sqrt{\hat{} U_t} \leq \mu^t \sqrt{\hat{} U_0} + \frac{\sigma + \alpha c}{1-\mu}$，这表明$\sqrt{\hat{} U_t}$ 以速率 μ线性收敛到渐近值$\frac{\alpha c}{1-\mu}$。然而，我们可以利用(31)推导出一个具有更优渐近值的不同界。注意到 $0 \leq \mu \leq 1$，(31)的不动点唯一，且满足 $U^* = (\mu \sqrt{U^*} + \sigma)^2 + \alpha^2 c^2$。我们有 $\sqrt{U^*} = \frac{\mu\sigma + \sqrt{\sigma^2 + \alpha^2 c^2 (1-\mu^2)}}{1-\mu^2}$。注意，(31)的右侧关于$\hat{} U_t$是凹函数，因此我们有$\hat{} U_{t+1} \leq U^* + \left(\mu + \frac{\sigma}{\sqrt{U^*}}\right)(\hat{} U_t - U^*) \leq \left(\mu + \frac{\sigma}{\sqrt{U^*}}\right)^{t+1} \hat{} U_0 + U^*$。由于我们知道$(\mu + \sigma/\sqrt{U^*})^2 = 1 - \alpha^2 c^2 / U^* \leq 1$，我们得到一个替代界，它能够给出真实值 U^*，代价是收敛速度较慢。

4.3 有限和情形下随机在线梯度下降的分析

对于许多机器学习任务，代价函数具有有限和结构。例如，监督学习中的总代价通常可以分解为训练集中不同数据点的损失之和。如果用于训练的数据点随时间变化，那么我们有 $f_t = \frac{1}{n} \sum_{i=1}^n f_t^{(i)}$，其中 n 是训练集的大小。我们可以不使用完整梯度信息，而是从训练集中随机抽取一个数据点，并采用以下随机在线梯度方案：

$$x_{t+1} = x_t - \alpha \nabla f_t^{(i_t)}(x_t), \tag{33}$$

其中， $i_t \in \{1, 2, \cdots, n\}$ 是以独立同分布方式抽取的索引。在这种情况下，我们仍然可以应用命题 4.1，得到以下线性矩阵不等式条件。

引理 4.3。考虑递推关系 (33)。对于所有 t，我们假设：i) f_t 是 m_t-强凸的，且 ii) $\|x_{t+1}^* - x_t^*\| \leq \sigma_t$。记 $G_t = \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_t^{(i)}(x_t^*)\right\|^2$。基于关于 $f_t^{(i)}$ 的假设，我们定义 $\tilde{X}_t \in \mathbb{R}^{2\times 2}$ 如下。

$$\tilde{X}_t := \begin{cases} \begin{bmatrix} 2L_t^2 & 0 \\ 0 & -1 \end{bmatrix} & \text{if } f_t^{(i)} \text{ is } L_t\text{-smooth} \\ \begin{bmatrix} 0 & L_t \\ L_t & -1 \end{bmatrix} & \text{if } f_t^{(i)} \text{ is } L_t\text{-smooth and convex} \\ \begin{bmatrix} -2L_t m_t & L_t + m_t \\ L_t + m_t & -1 \end{bmatrix} & \text{if } f_t^{(i)} \text{ is } L_t \in \mathcal{F}(m_t, L_t) \end{cases}. \tag{34}$$

如果存在非负标量 $(\rho_t, \lambda_t^{(1)}, \lambda_t^{(2)}, \lambda_t^{(3)})$，使得

$$\begin{bmatrix} 1-\rho_t^2 & -\alpha_t & 1 \\ -\alpha_t & \alpha_t^2 & -\alpha_t \\ 1 & -\alpha_t & 1-\lambda_t^{(2)} \end{bmatrix} + \lambda_t^{(1)} \begin{bmatrix} -2m_t & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \lambda_t^{(3)} \begin{bmatrix} \tilde{X}_t & 0_{2\times 1} \\ 0_{1\times 2} & 0 \end{bmatrix} \preceq 0, \tag{35}$$

那么 (33) 的迭代序列满足 $\mathbb{E}\left\|x_{t+1} - x_{t+1}^*\right\|^2 \leq \rho_t^2 \mathbb{E}\left\|x_t - x_t^*\right\|^2 + \lambda_t^{(2)}\sigma_t^2 + 2\lambda_t^{(3)} G_t^2$。

证明。我们可以设定 $e_t = x_t^* - x_{t+1}^*$，并将 (33) 改写为 $x_{t+1} - x_{t+1}^* = x_t - x_t^* - \alpha \nabla f_t^{(i_t)}(x_t) + e_t$。若设定 $\xi_t = x_t - x_t^*$、$B_t = \begin{bmatrix} -\alpha_t & 1 \end{bmatrix}$ 和 $w_t = \begin{bmatrix} \nabla f_t^{(i_t)}(x_t)^T & e_t^T \end{bmatrix}^T$，则这就成为 (5) 的一个特例。

f_t 的 m_t-强凸性导出一个供给率条件，其中 $(X_t^{(1)}, \Lambda_t^{(1)}) = (\text{diag}(M_t, 0), 0)$，这里 M_t 由 (4) 给出。假设 $\|e_t\| \leq \sigma_t$ 导出第二个供给率条件，其中 $X_t^{(2)} = \text{diag}(0, 0, 1)$ 和 $\Lambda_t^{(2)} = \sigma_t^2$。关于 $f_t^{(i)}$ 的假设可转化为第三个供给率条件，其中 $X_t^{(3)} = -\text{diag}(\tilde{X}_t, 0)$ 和 $\Lambda_t^{(3)} = 2G_t^2$。将这些供给率条件与命题 4.1 结合，即可立即得到所需的线性矩阵不等式。

□

将 （35）代入我们的序贯半定规划框架，可得到以下结果。

定理 4.4。假设 $\alpha_t \in [0, \bar{\alpha}_t]$，其中 $\bar{\alpha}_t$ 定义为

$$\bar{\alpha}_t := \begin{cases} \frac{m_t}{L_t^2} & \text{if } f_t^{(i)} \text{ is } L_t\text{-smooth} \\ \frac{1}{L_t} & \text{if } f_t^{(i)} \text{ is } L_t\text{-smooth and convex} \\ \frac{1}{L_t + m_t} & \text{if } f_t^{(i)} \text{ is } L_t\text{-smooth and } m_t\text{-strongly convex} \end{cases}.$$

令 $\mathcal{T}_t \subseteq \mathbb{R}_+^4$ 表示满足（35）的元组集合 $\left(\rho_t, \lambda_t^{(1)}, \lambda_t^{(2)}, \lambda_t^{(3)}\right)$。假设引理 4.3 中的条件成立。那么，带有 $J = 3$ 的（11）中的 $\hat{} U_t$ 为

$$\hat{U}_{t+1} = \left(\sqrt{\hat{\rho}_t \hat{U}_t + 2\alpha_t^2 G_t^2} + \sigma_t\right)^2, \tag{36}$$

其中 $\hat{}\rho_t = 1 - 2\alpha_t m_t + \tilde{m}\alpha_t^2$ 和 $\tilde{m}_t := \tilde{X}_t(1,1) + 2m_t \tilde{X}_t(2,1)$。

证明。根据命题 2.2，我们需要求解以下问题：

$$\hat{U}_{t+1} = \underset{(\rho_t,\lambda_t^{(1)},\lambda_t^{(2)},\lambda_t^{(3)})\in\mathcal{T}_t}{\text{minimize}} \rho_t^2\hat{U}_t + \lambda_t^{(2)}\sigma_t^2 + 2\lambda_t^{(3)}G_t^2. \tag{37}$$

令 $\lambda_t^{(2)} = 1 + \frac{1}{\nu_t}$。我们可以利用舒尔补将 (35) 转化为等价条件，即 $\nu_t > 0$, $\lambda_t^{(1)} \geq 0$, $\lambda_t^{(3)} \geq 0$ and :

$$\begin{bmatrix} (1+\nu_t) - \rho_t^2 - 2m_t\lambda_t^{(1)} + \lambda_t^{(3)}\tilde{X}_t(1,1) & \lambda_t^{(1)} + \lambda_t^{(3)}\tilde{X}_t(1,2) - \alpha_t(1+\nu_t) \\ \lambda_t^{(1)} + \lambda_t^{(3)}\tilde{X}(1,2) - \alpha_t(1+\nu_t) & \alpha_t^2(1+\nu_t) + \lambda_t^{(3)}\tilde{X}(2,2) \end{bmatrix} \preceq 0.$$

注意到根据 (34)，$\tilde{X}_t(2,2) = -1$，我们令 $\tau_t = \lambda_t^{(3)} - \alpha_t^2(1+\nu_t)$，由此得到 $\lambda_t^{(3)} = \tau_t + \alpha_t^2(1+\nu_t)$。根据舒尔补引理，我们可以令 $\phi_t := (1+\nu_t)\left(1 - \alpha_t\tilde{X}_t(2,1)\right)$，并证明上述条件等价于 $\nu_t > 0$, $\tau_t \geq 0$, $\lambda_t^{(1)} \geq 0$，以及 $h(\nu_t,\tau_t,\lambda_t^{(1)}) \leq \rho_t^2$，其中 $h(\nu_t,\tau_t,\lambda_t^{(1)})$ 定义为

$$\begin{aligned} h(\nu_t,\tau_t,\lambda_t^{(1)}) = (1+\nu_t)\left(1 + \alpha_t^2\tilde{X}_t(1,1)\right) + 2\lambda_t^{(1)}\left(\tilde{X}_t(2,1) - m_t\right) \\ + \tau_t\left(\tilde{X}_t(1,1) + (\tilde{X}_t(2,1))^2\right) + \frac{(\lambda_t^{(1)} - \alpha_t\phi_t)^2}{\tau_t} - 2\alpha_t\phi_t\tilde{X}_t(2,1). \end{aligned}$$

接下来，我们要证明 $h(\nu_t,\tau_t,\lambda_t^{(1)}) \geq 0$。我们需要考虑与 (34) 中对 $f_t^{(i)}$ 所作的不同假设相对应的三种情况。

第一种情况中，我们假设 $f_t^{(i)}$ 仅为 L_t-光滑。我们有 $\tilde{X}_t(1,1) = 2L_t^2$，且 $\tilde{X}_t(2,1) = 0$。对于任意固定的 $\nu_t > 0$ 和 $\tau_t \geq 0$，有

$$h(\nu_t,\tau_t,\lambda_t^{(1)}) = (1+\nu_t)\left(1 + \alpha_t^2\tilde{X}_t(1,1)\right) - 2\lambda_t^{(1)}m_t + 2L_t^2\tau_t + \frac{(\lambda_t^{(1)} - \alpha_t(1+\nu_t))^2}{\tau_t},$$

根据以下事实，该式非负：

$$h(\nu_t,\tau_t,\lambda_t^{(1)}) \geq h(\nu_t,\tau_t,\alpha_t\phi_t + \tau_t m_t) = (1+\nu_t)\left(1 - 2m_t\alpha_t + 2L_t^2\alpha_t^2\right) + \tau_t(2L_t^2 - m_t^2).$$

第二种情况中，我们假设 $f_t^{(i)}$ 是 L_t-光滑且为凸函数。我们有 $\tilde{X}_t(1,1) = 0$，且 $\tilde{X}_t(2,1) = L_t$。对于任意固定的 $\nu_t > 0$ 和 $\tau_t \geq 0$，有：

$$h(\nu_t,\tau_t,\lambda_t^{(1)}) = 1 + \nu_t + 2\lambda_t^{(1)}(L_t - m_t) + \tau_t L_t^2 + \frac{(\lambda_t^{(1)} - \alpha_t\phi_t)^2}{\tau_t} - 2\alpha_t\phi_t L_t.$$

我们可以证明，对于所有 $\lambda_t^{(1)}$， $h(\nu_t,\tau_t,\lambda_t^{(1)})$ 有如下所示的非负下界 $\hat{h}(\nu_t,\tau_t)$：

$$\hat{h}(\nu_t,\tau_t) := \begin{cases} (1+\nu_t)(1 + 2\alpha_t m_t(\alpha_t L_t - 1)) + \tau_t m_t(2L_t - m_t) & \text{if } 0 \leq \tau_t \leq \frac{\alpha_t\phi_t}{L_t - m_t} \\ 1 + \nu_t + \tau_t L_t^2 + \frac{\alpha_t^2\phi_t^2}{\tau_t} - 2\alpha_t\phi_t L_t & \text{if } \tau_t > \frac{\alpha_t\phi_t}{L_t - m_t} \end{cases}$$

由于我们知道，对于任意固定的 ν_t， $\frac{\alpha_t\phi_t}{L_t} \leq \frac{\alpha_t\phi_t}{L_t - m_t}$，因此很容易得出，$\hat{h}(\nu_t,\tau_t)$ 在 $0 \leq \tau_t \leq \frac{\alpha_t\phi_t}{L_t - m_t}$ 上线性递增，随后又在 $\tau_t > \frac{\alpha_t\phi_t}{L_t - m_t}$ 上单调递增。因此，$\hat{h}$ 达到其

最小值于 $\tau_t = 0$。因此，我们必须有 $h(\nu_t, \tau_t, \lambda_t^{(1)}) \geq h(\nu_t, \tau_t, \alpha_t(1+\nu_t)(1-\alpha_t L_t) - \tau_t(L_t - \hat{m}_t))$ $= h(\nu_t, \tau_t) \geq 0$。

在第三种情形中，假设 $f_t^{(i)}$ 是 L_t-光滑且 m_t-强凸的。我们有 $\tilde{X}_t(1,1) = -2L_t m_t$ 和 $X_t(2,1) = L_t + m_t$。对于任意 $\nu_t > 0$ 和 $\tau_t \geq 0$，有

$$h(\nu_t, \tau_t, \lambda_t^{(1)}) = (1+\nu_t)\left(1 - 2\alpha_t^2 L_t m_t\right) + 2\lambda_t^{(1)} L_t + \tau_t\left(L_t^2 + m_t^2\right) + \frac{(\lambda_t^{(1)} - \alpha_t\phi_t)^2}{\tau_t} - 2\alpha_t\phi_t(L_t + m_t).$$

同样，我们可以证明，对于所有 $\lambda_t^{(1)}$，$h(\nu_t, \tau_t, \lambda_t^{(1)})$ 存在一个非负下界 $\hat{h}(\nu_t, \tau_t)$。具体而言，对于 $0 \leq \tau_t \leq \frac{\alpha_t\phi_t}{L_t}$，我们令 $\hat{h}(\nu_t, \tau_t) = (1+\nu_t)\left(1 - 2\alpha_t m_t + 2\alpha_t^2 m_t^2\right) + \tau_t m_t^2$；对于 $\tau_t > \frac{\alpha_t\phi_t}{L_t}$，令 $\hat{h}(\nu_t, \tau_t) = (1+\nu_t)\left(1 - 2\alpha_t^2 L_t m_t\right) + \tau_t\left(L_t^2 + m_t^2\right) + \frac{\alpha_t^2\phi_t^2}{\tau_t} - 2\alpha_t\phi_t(L_t + m_t)$。由于对于任意固定的 ν_t，我们有 $\frac{\alpha_t\phi_t}{L_t^2+m_t^2} \leq \frac{\alpha_t\phi_t}{L_t}$，因此可以得出，$\hat{h}(\nu_t, \tau_t)$ 随 $0 \leq \tau_t \leq \frac{\alpha_t\phi_t}{L_t}$ 线性增加，随后又随 $\tau_t > \frac{\alpha_t\phi_t}{L_t}$ 单调增加。因此，$\hat{h}$ 在 $\tau_t = 0$ 处取得最小值。这使我们得出结论：$h(\nu_t, \tau_t, \lambda_t^{(1)}) \geq h(\nu_t, \tau_t, \alpha_t(1+\nu_t)(1 - \alpha_t(L_t + m_t) - \tau_t L_t) = \hat{h}(\nu_t, \tau_t) \geq 0$。

现在我们知道 $h(\nu_t, \tau_t, \lambda_t^{(1)}) \geq 0$。记 $F(\nu_t, \tau_t) := \left(1 + \frac{1}{\nu_t}\right)\sigma_t^2 + 2\left(\tau_t + \alpha_t^2(1+\nu_t)\right)G_t^2$，则优化问题 (37) 可按如下方式求解

$$\begin{aligned}
\hat{U}_{t+1} &= \min_{\nu_t > 0}\left(\min_{\tau_t \geq 0}\left(\min_{\lambda_t^{(1)} \geq 0}\left(\min_{\rho_t^2 \geq h\left(\nu_t, \tau_t, \lambda_t^{(1)}\right)}\left(\rho_t^2 \hat{U}_t + F(\nu_t, \tau_t)\right)\right)\right)\right) \\
&= \min_{\nu_t > 0}\left(\min_{\tau_t \geq 0}\left(\min_{\lambda_t^{(1)} \geq 0}\left(h(\nu_t, \tau_t, \lambda_t^{(1)})\hat{U}_t + F(\nu_t, \tau_t)\right)\right)\right) \\
&= \left(\sqrt{\hat{\rho}_t \hat{U}_t + 2\alpha_t^2 G_t^2} + \sigma_t\right)^2.
\end{aligned}$$

注意，最优的 $\lambda_t^{(1)}$ 由 $\alpha_t(1+\nu_t)(1 - \alpha_t \tilde{X}_t(2,1))$ 给出。由于 $\lambda_t^{(1)} \geq 0$ 和 $\nu_t > 0$，因此上述优化问题的推导解在 $0 \leq \alpha_t X_t(2,1) \leq 1$ 时成立，这与文献 [17]中的已有结果一致。此外，注意到 $\rho_t \leq 1$ 和 $0 \leq \alpha_t X_t(2,1) \leq 1$ 给出了使该结果成立的 α_t 所需范围。

□

若假设对所有 t 均有 $L_t = L$、$m_t = m$、$G_t = G$、$\sigma_t = \sigma$ 和 $\alpha_t = \alpha$，则可将 (36) 简化为 $\sqrt{\hat{U}_{t+1}} = \sqrt{\hat{\rho}\hat{U}_t + 2\alpha^2 G^2} + \sigma$。由此得到界 $\sqrt{\hat{U}_t} \leq \hat{\rho}^{\frac{t}{2}}\sqrt{\hat{U}_0} + \frac{\sqrt{2}\alpha G + \sigma}{1 - \sqrt{\hat{\rho}}}$，表明 $\sqrt{\hat{U}_t}$ 以线性速率 $\hat{\rho}^{\frac{1}{2}}$ 收敛至低于稳态值 $\frac{\sqrt{2}\alpha G + \sigma}{1 - \sqrt{\hat{\rho}}}$。我们还可以采用备注4.1中的论证，得到一个结合真实稳态值极限 U^* 的替代界，该极限满足 $\sqrt{U^*} = \sqrt{\hat{\rho}U^* + 2\alpha^2 G^2} + \sigma$。细节从略。

5 非精确近端在线梯度下降的新界

非精确近端在线梯度下降（IP-OGD）方法在 [7] 中提出，在解决非光滑时变优化问题（如鲁棒子空间跟踪）方面展现出巨大潜力。在这种设定下，时刻 t 的代价函数表示为总和 $f_t + g_t$，其中 f_t 是一个光滑凸

函数，而 g_t 是非光滑凸函数。对于每个 t，IP-OGD 方法（或等价地，[7] 中的算法 1）按如下方式迭代

$$x_{t+1} = \operatorname*{argmin}_{x\in\mathbb{R}^p} \left\{ \frac{1}{2\alpha} \|x - x_t + \alpha(\nabla f_t(x_t) + v_t)\|^2 + g_t(x) \right\}, \tag{38}$$

其中，项 v_t 表示梯度计算中的不精确性，并假设其满足某个给定的范数界。[7] 中采用的假设是 $\mathbb{E}\|v_t\| \le c_t$。因此，[7] 中推导出的跟踪误差界在 L_1 意义下成立，即得到了 $\mathbb{E}\|x_t - x_t^*\|$ 的上界。本文将对 v_t 采用更标准的均方误差界，即假设 $\mathbb{E}\|v_t\|^2 \le c_t^2$。我们关注的是推导均方跟踪误差 $\mathbb{E}\|x_t - x_t^*\|^2$ 的上界。我们的跟踪误差界在 L_2 意义下成立，因此自然地补充了 [7] 中的 L_1 误差界。除 v_t 的假设存在这一关键差异外，我们分析中的其他假设均与 [7] 中所用假设相同。我们可以得到以下线性矩阵不等式条件。

引理 5.1. 对所有 t，假设：i) $f_t \in \mathcal{F}(m_t, L_t)$，ii) g_t 是凸的且为 L_g-利普希茨连续，iii) $\|x_{t+1}^* - x_t^*\| \le \sigma_t$，以及 iv) $\mathbb{E}\|v_t\|^2 \le c_t^2$。若存在非负标量 ρ_t 和 $\{\lambda_t^{(j)}\}_{j=1}^5$，使得以下矩阵不等式成立

$$M(Q_t, S_t, R_t) = \begin{bmatrix} Q_t & S_t \\ S_t^T & R_t \end{bmatrix} \preceq 0,$$

其中，(Q_t, S_t, R_t) 定义为

$$Q_t = \begin{bmatrix} 1 - \rho_t^2 - 2\lambda_t^{(1)} L_t m_t & \lambda_t^{(1)}(L_t + m_t) - \alpha_t \\ \lambda_t^{(1)}(L_t + m_t) - \alpha_t & \alpha_t^2 - 2\lambda_t^{(1)} \end{bmatrix},$$

$$S_t = \begin{bmatrix} \lambda_t^{(5)} - \alpha_t & -\alpha_t & -\alpha_t & 1 \\ \alpha_t^2 - \alpha_t\lambda_t^{(5)} & \alpha_t^2 & \alpha_t^2 & -\alpha_t \end{bmatrix},$$

$$R_t = \begin{bmatrix} \alpha_t^2 - 2\alpha_t\lambda_t^{(5)} & \alpha_t^2 - \alpha_t\lambda_t^{(5)} & \alpha_t^2 - \alpha_t\lambda_t^{(5)} & \lambda_t^{(5)} - \alpha_t \\ \alpha_t^2 - \alpha_t\lambda_t^{(5)} & \alpha_t^2 - \lambda_t^{(4)} & \alpha_t^2 & -\alpha_t \\ \alpha_t^2 - \alpha_t\lambda_t^{(5)} & \alpha_t^2 & \alpha_t^2 - \lambda_t^{(3)} & -\alpha_t \\ \lambda_t^{(5)} - \alpha_t & -\alpha_t & -\alpha_t & 1 - \lambda_t^{(2)} \end{bmatrix},$$

*the*IP-OGD 方法满足以下均方跟踪误差界限 *d:*

$$\mathbb{E}\left\|x_{t+1} - x_{t+1}^*\right\|^2 \le \rho_t^2 \mathbb{E}\|x_t - x_t^*\|^2 + \lambda_t^{(2)}\sigma_t^2 + \lambda_t^{(3)} c_t^2 + 4\lambda_t^{(4)} L_g^2.$$

证明。IP-OGD 递推关系可改写为 $x_{t+1} = x_t - \alpha(\nabla f_t(x_t) + v_t) - \alpha r_t$，其中$r_t \in \partial g_t(x_{t+1})$。记 $e_t = x_t^* - x_{t+1}^*$，则（38）可改写为 $x_{t+1} - x_{t+1}^* = x_t - x_t^* - \alpha(\nabla f_t(x_t) - \nabla f_t(x_t^*)) - \alpha(r_t - r_{t+1}^*) - \alpha(r_{t+1}^* - r_t^*) - \alpha v_t + e_t$，其中 r_t^*是 g_t在x_t^* 处取值的一个次梯度，并满足 $r_t^* = -\nabla f_t(x_t^*)$； r_{t+1}^* 是 g_t 在 x_{t+1}^*处取值的任意次梯度。令 $\xi_t := x_t - x_t^*$。于是，IP-OGD 方法是（5）的一个特例，其中 B_t和 w_t分别为 $B_t = \begin{bmatrix} -\alpha_t & -\alpha_t & -\alpha_t & -\alpha_t & 1 \end{bmatrix}$，且 $w_t = \begin{bmatrix} \nabla f_t(x_t)^T - \nabla f_t(x_t^*)^T & r_t^T - r_{t+1}^{*T} & r_{t+1}^{*T} - r_t^{*T} & v_t^T & e_t^T \end{bmatrix}^T$。接下来，我们给出五种不同的期望供给率条件，以刻画上述引理陈述中的假设。对于 $j = 1$，我们需要指定（$X_t^{(j)}, \Lambda_t^{(j)}$）、2、$\cdots$和 5。该假设

$f_t \in \mathcal{F}(m_t, L_t)$ 导出一个形式为（29）的期望供给率条件，其中（$X_t^{(1)}, \Lambda_t^{(1)}$） =(diag($Y_t$, 0, 0, 0, 0), 0)，其中 Y_t 由（4）给出。条件 $\|e_t\| \leq \sigma_t$ 可改写为二次型（29），其中（$X_t^{(2)}, \Lambda_t^{(2)}$） = (diag(0, 0, 0, 0, 0, 1) σ_t^2)。同样，可以直接将 $\mathbb{E}\|v_t\|^2 \leq c_t^2$ 改写为（29）的形式，其中（$X_t^{(3)}, \Lambda_t^{(3)}$） = $\left(\text{diag}(0, 0, 0, 0, 1, 0)\, c_t^2\right)$。由于对于所有 t， g_t 都是 L_g- 利普希茨的，因此我们有 $\left\|r_t^* - r_{t+1}^*\right\| \leq 2L_g$，它可改写为（29）的形式，其中（$X_t^{(4)}, \Lambda_t^{(4)}$） = $\left(\text{diag}(0, 0, 0, 1, 0, 0)\, 4L_g^2\right)$。最后一个供给率条件直接由 g_t的凸性推出，而条件 $(r_t - r_{t+1}^*)^T (x_{t+1} - x_{t+1}^*) \geq 0$ 可用于导出具有以下参数的供给率条件:

$$X_t^{(5)} = \begin{bmatrix} 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & \alpha_t & 0 & 0 & 0 \\ -1 & \alpha_t & 2\alpha_t & \alpha_t & \alpha_t & -1 \\ 0 & 0 & \alpha_t & 0 & 0 & 0 \\ 0 & 0 & \alpha_t & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \end{bmatrix}, \text{ and } \Lambda_t^{(5)} = 0$$

然后，我们可以应用命题 4.1 来构造所需的线性矩阵不等式。 □

应用我们的序贯半定规划方法，可以得到 IP-OGD 的如下均方跟踪误差界。

定理 5.2. 假设对于所有 t， $\alpha_t \leq \frac{2}{L_t}$。令 $\mathcal{T}_t \subseteq \mathbb{R}_+^6$ 表示满足 $M(Q_t, S_t, R_t) \preceq 0$ 的元组集合 $\left(\rho_t, \lambda_t^{(1)}, \ldots, \lambda_t^{(5)}\right)$。假设引理 5.1 中的条件成立，则可按如下方式计算式 (11) 中定义的界 $\hat{} U_t$:

$$\hat{U}_{t+1} = \left(\mu_t \sqrt{\hat{U}_t} + \alpha_t \sigma_t + c_t + 2\alpha_t L_g\right)^2. \tag{39}$$

证明。我们需要求解以下半定规划:

$$\hat{U}_{t+1} = \underset{(\rho_t, \lambda_t^{(1)}, \ldots, \lambda_t^{(5)}) \in \mathcal{T}_t}{\text{minimize}} \rho_t^2 \hat{U}_t + \lambda_t^{(2)} \sigma_t^2 + \lambda_t^{(3)} c_t^2 + 4\lambda_t^{(4)} L_g^2. \tag{40}$$

我们可以修改前面的论证，并应用舒尔补引理，将上述优化问题转化为以下等价形式:

$$\hat{U}_{t+1} = \min_{\chi_t > 0} \left(\min_{\nu_t > 0} \left(\min_{\zeta_t > 0} \left(\min_{\upsilon_t > 0} \left(\min_{\tau_t \geq 0} \left(\min_{\phi_t h(\tau_t) \leq \rho_t^2} \left(\rho_t^2 \hat{U}_t + F(\chi_t, \nu_t, \zeta_t, \upsilon_t) \right) \right) \right) \right) \right) \right) \tag{41}$$

其中， $F(\chi_t, \nu_t, \zeta_t, \upsilon_t)$ 定义为

$$F(\chi_t, \nu_t, \zeta_t, \upsilon_t) := \psi_t(\chi_t)^{-1} (1 + \nu_t) \left(\frac{c_t^2}{\nu_{t_{-1}}} + \alpha_t^2 (1 + \zeta_t) \left(\frac{\sigma_t^2}{\zeta_t} + 4 \left(1 + \frac{1}{\upsilon_t} \right) L_g^2 \right) \right).$$

这里我们使用记号 $\psi_t(\chi_t) := \left(1 + \frac{1}{\chi_t} \left(\frac{\chi_t - 1}{2}\right)^2\right)^{-1}$。然后我们只需按

以下逐层方式求解 (41)：

$$\begin{aligned}
\hat{U}_{t+1} &= \min_{\chi_t>0}\left(\min_{\nu_t>0}\left(\min_{\zeta_t>0}\left(\min_{\upsilon_t>0}\left(\min_{\tau_t\geq 0}\left(\phi_t h(\tau_t)\hat{U}_t + F(\chi_t,\nu_t,\zeta_t,\upsilon_t)\right)\right)\right)\right)\right) \\
&= \min_{\chi_t>0}\left(\min_{\nu_t>0}\left(\min_{\zeta_t>0}\left(\min_{\upsilon_t>0}\left(\phi_t\left(\min_{\tau_t\geq 0} h(\tau_t)\right)\hat{U}_t + F(\chi_t,\nu_t,\zeta_t,\upsilon_t)\right)\right)\right)\right) \\
&= \min_{\chi_t>0}\left(\min_{\nu_t>0}\left(\min_{\zeta_t>0}\left(\min_{\upsilon_t>0}\left(\phi_t\mu_t^2\hat{U}_t + F(\chi_t,\nu_t,\zeta_t,\upsilon_t)\right)\right)\right)\right) \\
&= \min_{\nu_t>0}\left((1+\nu_t)\left(\min_{\zeta_t>0}\left(\frac{c_t^2}{\nu_t} + (1+\zeta_t)\left(\frac{\alpha_t^2\sigma_t^2}{\zeta_t} + \left(\mu_t\sqrt{\hat{U}_t} + 2\alpha_t L_g\right)^2\right)\right)\right)\right) \\
&= \min_{\nu_t>0}\left((1+\nu_t)\left(\frac{c_t^2}{\nu_t} + \left(\mu_t\sqrt{\hat{U}_t} + \alpha_t\left(\sigma_t + 2L_g\right)\right)^2\right)\right) \\
&= \left(\mu_t\sqrt{\hat{U}_t} + \alpha_t\sigma_t + c_t + 2\alpha_t L_g\right)^2.
\end{aligned}$$

我们的证明完成了。 □

注 5.1. 通过对所有 t 取 $L_t = L$、 $m_t = m$、 $\alpha_t = \alpha$、 $c_t = c$ 和 $\sigma_t = \sigma$，可以将式 (39) 中的递归界进一步简化为 $\hat{U}_{t+1} = \left(\mu\sqrt{\hat{U}_t} + \alpha\sigma + c + 2\alpha L_g\right)^2$。简化后的迭代界导出了以下渐近界：

$$\sqrt{\hat{U}_t} = \mu^t\left(\sqrt{\hat{U}_0} - \frac{\alpha\sigma + c + 2\alpha L_g}{1-\mu}\right) + \frac{\alpha\sigma + c + 2\alpha L_g}{1-\mu}$$

显然， $\sqrt{\hat{U}_t}$ 以 μ 所量化的速率线性收敛至其稳态值。

6 动态遗憾分析

有一种常规方法可以将前几节推导出的跟踪误差界转化为动态遗憾的新上界。具体而言，以下引理很有用。

引理 6.1. 假设 f_t 是 L-光滑的， $\nabla f_t(x_t^*) = 0$，且 $\|x_t - x_t^*\|^2 \leq \hat{U}_t\ \forall t$。如果跟踪误差界 U_t 满足以下不等式，其中 $0 \leq \gamma < 1$ 和 $u_t \geq 0$ 为某些量：

$$\sqrt{\hat{U}_{t+1}} \leq \gamma\sqrt{\hat{U}_t} + u_t, \tag{42}$$

那么，该在线优化方法的动态遗憾可以界定为

$$\sum_{t=0}^{T}(f_t(x_t) - f_t(x_t^*)) \leq \frac{L}{(1-\gamma)^2}\hat{U}_0 + \frac{L}{(1-\gamma)^2}\left(\sum_{t=0}^{T-1} u_t\right)^2. \tag{43}$$

证明。由于 f_t 是 L-光滑的，且 $\nabla f_t(x_t^*) = 0$，我们有 $f_t(x_t) \leq f_t(x_t^*) + \frac{L}{2}\|x_t - x_t^*\|^2$。因此，动态遗憾满足 $\sum_{t=0}^{T}(f_t(x_t) - f_t(x_t^*)) \leq \frac{L}{2}\sum_{t=0}^{T}\|x_t - x_t^*\|^2 \leq \frac{L}{2}\sum_{t=0}^{T}\hat{U}_t \leq$

$\frac{L}{2}\left(\sum_{t=0}^{T}\sqrt{\hat{U}_t}\right)^2$。根据 (42)，我们有 $\sqrt{\hat{U}_t} \le \gamma^t\sqrt{\hat{U}_0} + \gamma^{t-1}u_0 + \gamma^{t-2}u_1 + \cdots + u_{t-1}$。由此得到以下界

$$\sum_{t=0}^{T}\sqrt{\hat{U}_t} \le \sum_{t=0}^{T}\gamma^t\sqrt{\hat{U}_0} + \frac{1}{1-\gamma}\sum_{t=0}^{T-1}u_t \le \frac{1}{1-\gamma}\sqrt{\hat{U}_0} + \frac{1}{1-\gamma}\sum_{t=0}^{T-1}u_t.$$

因此，我们有 $\left(\sum_{t=0}^{T}\sqrt{\hat{U}_t}\right)^2 \le \left(2\hat{U}_0 + 2(\sum_{t=0}^{T-1}u_t)^2\right)/(1-\gamma)^2$。这直接导出了所需的结论。

□

ˆ 根据上述结果，将跟踪误差界 U_t转换为动态遗憾界几乎是轻而易举的。接下来，我们简要总结这些结果。

• 具有绝对误差的非精确在线梯度下降。若对所有 t，$\alpha_t = \alpha$ 且 $\mu_t = \mu$，则有 $\sqrt{\hat{U}_{t+1}} \le \mu\sqrt{\hat{U}_t} + \sigma_t + \alpha c_t$。根据引理 6.1，以下界成立

$$\sum_{t=0}^{T}(f_t(x_t) - f_t(x_t^*)) \le \frac{L}{(1-\mu)^2}\hat{U}_0 + \frac{2L}{(1-\mu)^2}\left(\sum_{t=0}^{T-1}\sigma_t\right)^2 + \frac{2L\alpha^2}{(1-\mu)^2}\left(\sum_{t=0}^{T-1}c_t\right)^2.$$

• 具有相对误差的非精确在线梯度下降。若对所有 t，$\alpha_t = \alpha$ 且 $\hat{}\rho_t = \hat{}\rho$，则有 $\sqrt{\hat{U}_{t+1}} \le \hat{}\rho\sqrt{\hat{U}_t} + \sigma_t$。根据引理 6.1，以下界成立

$$\sum_{t=0}^{T}(f_t(x_t) - f_t(x_t^*)) \le \frac{L}{(1-\hat{\rho})^2}\hat{U}_0 + \frac{L}{(1-\hat{\rho})^2}\left(\sum_{t=0}^{T-1}\sigma_t\right)^2.$$

•带加性噪声的随机在线梯度下降。我们可以利用引理 6.1 的随机变体得到随机动态遗憾界。具体而言，我们有 $\mathbb{E}\|x_t - x_t^*\|^2 \le \hat{}$ U_t 以及 $\mathbb{E}\ f_t\ x_t\ -f_t\ x_t^*\ \le \frac{L}{2}\mathbb{E}\|x_t - x_t^*\|^2$ 4 2, $\sqrt{U_{t+1}} \le \mu\sqrt{U_t} + (\ (\ )\ (\ ))$。根据定理 ., 我 $\sigma_t + \alpha c_t$ 6 1 们得到界 ˆ ˆ。然后，我们可以稍微修改引理 . 的证明来说明

$$\sum_{t=0}^{T}\mathbb{E}(f_t(x_t) - f_t(x_t^*)) \le \frac{L}{(1-\mu)^2}\hat{U}_0 + \frac{2L}{(1-\mu)^2}\left(\sum_{t=0}^{T-1}\sigma_t\right)^2 + \frac{2L\alpha^2}{(1-\mu)^2}\left(\sum_{t=0}^{T-1}c_t\right)^2.$$

•有限和情形下的随机在线梯度下降。根据定理 4.4，我们得到跟踪误差界 $\sqrt{\hat{U}_{t+1}} \le \sqrt{\hat{}\rho}\hat{}$ $U_t + \sqrt{2}\alpha G_t + \sigma_t$。类似地，我们可以立即得到随机动态遗憾界：

$$\sum_{t=0}^{T}\mathbb{E}(f_t(x_t) - f_t(x_t^*)) \le \frac{L\hat{U}_0}{\left(1-\sqrt{\hat{\rho}}\right)^2} + \frac{2L}{\left(1-\sqrt{\hat{\rho}}\right)^2}\left(\sum_{t=0}^{T-1}\sigma_t\right)^2 + \frac{4L\alpha^2}{\left(1-\sqrt{\hat{\rho}}\right)^2}\left(\sum_{t=0}^{T-1}G_t\right)^2$$

•IP-OGD。在这种情况下，我们需要稍微修改分析。具体而言，我们希望推导出 $\sum_{t=0}^{T}\mathbb{E}(h_t(x_t) - h_t(x_t^*))$ 的一个上界。我们有

$$\sum_{t=0}^{T}\mathbb{E}(h_t(x_t) - h_t(x_t^*)) \le \sum_{t=0}^{T}\mathbb{E}(f_t(x_t) - f_t(x_t^*)) + \sum_{t=0}^{T}\mathbb{E}(g_t(x_t) - g_t(x_t^*))$$

注意，现在我们有 $\nabla f_t(x_t^*) \neq 0$。由于 x_t^* 是 $f_t + g_t$的全局最小值，因此必有$-\nabla f_t(x_t^*) \in \partial g_t(x_t^*)$。由于 g_t 是 L_g-利普希茨的，因此我们有 $\|\nabla f_t(x_t^*)\| \leq L_g$。现在我们有

$$\sum_{t=0}^{T} \mathbb{E}(h_t(x_t) - h_t(x_t^*)) \leq \frac{L}{2} \sum_{t=0}^{T} \hat{U}_t + 2L_g \sum_{t=0}^{T} \mathbb{E}\,\|x_t - x_t^*\| \leq \frac{L}{2} \sum_{t=0}^{T} \hat{U}_t + 2L_g \sum_{t=0}^{T} \sqrt{\hat{U}_t}.$$

假设 (42) 成立。我们可以轻松修改引理 6.1 的证明，从而证明

$$\sum_{t=0}^{T} \mathbb{E}(h_t(x_t) - h_t(x_t^*)) \leq \frac{L\hat{U}_0}{(1-\gamma)^2} + \frac{2L_g\sqrt{\hat{U}_0}}{1-\gamma} + \frac{L}{(1-\gamma)^2} \left(\sum_{t=0}^{T-1} u_t \right)^2 + \frac{2L_g}{1-\gamma} \sum_{t=0}^{T-1} u_t.$$

对于 IP-OGD，若对所有 t 都有 $\alpha_t = \alpha$ 和 $\mu_t = \mu$，则有 $\gamma = \mu$ 和 $u_t = \alpha\sigma_t + c_t + 2\alpha L_t$。因此，我们得到上述随机动态遗憾界，其中 $\gamma = \mu$ 和 $u_t = \alpha\sigma_t + c_t + 2\alpha L_t$。

7 结论

本文针对一大类非精确在线优化方法，推导了序贯半定规划的解析解，从而得到其跟踪误差和动态遗憾的上界。我们的分析以统一的方式处理确定性和随机情形下的非精确在线优化方法，并灵活地应对预言机不精确性。未来，研究非精确在线镜像下降方法的分析将颇有意义；这需要使用布雷格曼散度作为跟踪误差的距离度量。

定理 3.6 的证明

证明。根据命题 2.2，我们需要求解以下问题:

$$\hat{U}_{t+1} = \underset{(\rho_t, \lambda_t^{(1)}, \lambda_t^{(2)}, \lambda_t^{(3)}, \lambda_t^{(4)}) \in \mathcal{T}_t}{\text{minimize}} \rho_t^2 \hat{U}_t + \lambda_t^{(2)} \sigma_t^2 + \lambda_t^{(3)} c_t^2. \tag{44}$$

令 $\lambda_t^{(2)} = 1 + \frac{1}{\nu_t}$和 $\lambda_t^{(3)} = \alpha_t^2(1+\nu_t)\left(1 + \frac{1}{\zeta_t}\right)$。然后，我们可以两次应用舒尔补，证明 27 等价于 $\zeta_t > 0$, $\nu_t > 0$ 和

$$\begin{bmatrix} \psi(\zeta_t, \nu_t) - \rho_t^2 + \lambda_t^{(4)} L_t^2 - 2\lambda_t^{(1)} m_t & \lambda_t^{(1)} - \alpha_t \psi(\zeta_t, \nu_t) \\ \lambda_t^{(1)} - \alpha_t \psi(\zeta_t, \nu_t) & \alpha_t^2 \psi(\zeta_t, \nu_t)(1 + \frac{\delta_t^2}{\zeta_t}) - \lambda_t^{(4)} \end{bmatrix} \preceq 0$$

其中 $\psi(\zeta_t, \nu_t) := (1+\zeta_t)(1+\nu_t)$。令 $\tau_t = \lambda_t^{(4)} - \alpha_t^2 \psi(\zeta_t, \nu_t)(1 + \frac{\delta_t^2}{\zeta_t})$。我们有 $\lambda_t^{(4)} = \tau_t + \alpha_t^2 \psi(\zeta_t, \nu_t)(1 + \frac{\delta_t^2}{\zeta_t})$。应用舒尔补，可以证明上述条件成立，当且仅当 $\nu_t > 0$, $\zeta_t > 0$, $\tau_t \geq 0$, 且 $h(\nu_t, \zeta_t, \tau_t, \lambda_t^{(1)}) \leq \rho_t^2$, 其中 $h(\nu_t, \zeta_t, \tau_t, \lambda_t^{(1)}) := \psi_t(\nu_t, \zeta_t)\left(1 + \alpha_t^2 L_t^2 (1 + \frac{\delta_t^2}{\zeta_t}) - \frac{2\alpha_t \lambda_t^{(1)}}{\tau_t}\right) + \tau_t L_t^2 - 2\lambda_t^{(1)} m_t + \frac{\alpha_t^2 \psi_t (\nu_t, \zeta_t)}{\tau_t} \lambda_t^{(1)\,2}$。对于任意 $\nu_t > 0, \zeta_t > 0$和 $\tau_t \geq 0$, 使 h 最小的~~$\lambda_t^{(1)}$ 的最优取值~~为 $\lambda_t^{(1)} = \tau_t m_t + \alpha_t \psi_t \geq 0$。我们有

$$h(\nu_t, \zeta_t, \tau_t, \lambda_t^{(1)}) \geq h(\nu_t, \zeta_t, \tau_t, \tau_t m_t + \alpha_t \psi_t) = \psi_t \bar{F}(\zeta_t) + \tau_t (L_t^2 - m_t^2) \geq 0,$$

其中 $\bar{F}(\zeta_t) := 1 - 2\alpha_t m_t + \alpha^2 L_t^2 \left(1 + \frac{\delta_t^2}{\zeta_t}\right)$。那么我们可以记 $F(\nu_t, \zeta_t) := \left(1 + \frac{1}{\nu_t}\right)\sigma_t^2 + \alpha_t^2(1+\nu_t)\left(1 + \frac{1}{\zeta_t}\right)c_t^2$，并如下求解（44）

$$\begin{aligned}
\hat{U}_{t+1} &= \min_{\nu_t>0}\left(\min_{\zeta_t>0}\left(\min_{\tau_t\geq 0}\left(\min_{\lambda_t^{(1)}\geq 0}\left(\min_{\rho_t^2\geq h\left(\nu_t,\zeta_t,\tau_t,\lambda_t^{(1)}\right)}\left(\rho_t^2\hat{U}_t + F(\nu_t,\zeta_t)\right)\right)\right)\right)\right) \\
&= \min_{\nu_t>0}\left(\min_{\zeta_t>0}\left(\min_{\tau_t\geq 0}\left(\min_{\lambda_t^{(1)}\geq 0}\left(h(\nu_t,\zeta_t,\tau_t,\lambda_t^{(1)})\hat{U}_t + F(\nu_t,\zeta_t)\right)\right)\right)\right) \\
&= \min_{\nu_t>0}\left(\min_{\zeta_t>0}\left(\min_{\tau_t\geq 0}\left(\psi_t(\nu_t,\zeta_t)\bar{F}(\zeta_t)\hat{U}_t + \tau_t(L_t^2 - m_t^2)\hat{U}_t + F(\nu_t,\zeta_t)\right)\right)\right) \\
&= \min_{\nu_t>0}\left(\min_{\zeta_t>0}\left((1+\nu_t)\left((1+\zeta_t)\left(\bar{F}(\zeta_t)\hat{U}_t + \frac{\alpha_t^2 c_t^2}{\zeta_t}\right) + \frac{\sigma_t^2}{\nu_t}\right)\right)\right) \\
&= \min_{\nu_t>0}\left((1+\nu_t)\left(\left(\sqrt{\left(1 - 2\alpha_t m_t + \alpha_t^2 L_t^2\right)\hat{U}_t} + \alpha_t\sqrt{c_t^2 + \delta_t^2 L_t^2 \hat{U}_t}\right)^2 + \frac{\sigma_t^2}{\nu_t}\right)\right) \\
&= \left(\sqrt{(1 - 2m_t\alpha_t + \alpha_t^2 L_t^2)\hat{U}_t} + \alpha_t\sqrt{c_t^2 + \delta_t^2 L_t^2 \hat{U}_t} + \sigma_t\right)^2.
\end{aligned}$$

□

B 注 3.3 的解释

[17, 命题 1] 中的序贯优化问题等价于以下序贯半定规划：

$$\hat{U}_{t+1} = \underset{(\rho_t,\lambda_t^{(1)},\lambda_t^{(3)},\lambda_t^{(4)})\in\mathcal{T}_t}{\text{minimize}} \rho_t^2\hat{U}_t + 2\lambda_t^{(4)}G^2 + \lambda_t^{(3)}c^2, \tag{45}$$

其中，$\mathcal{T}_t \subseteq \mathbb{R}_+^4$ 是满足条件 (27) 且具有 $\lambda_t^{(2)} = 0$ 的元组 $(\rho_t, \lambda_t^{(1)}, \lambda_t^{(3)}, \lambda_t^{(4)})$ 的集合。值得一提的是，上述优化问题与式 (44) 中具有 $\sigma_t \equiv 0$ 且代价函数略有修改的序贯半定规划类似。接下来，我们可以解析地求解 (45)，以证明 [17, 注 3] 中的递归界确实是 [17, 命题1]中序贯半定规划的精确解。令 $\lambda_t^{(3)} = \alpha_t^2\left(1 + \frac{1}{\zeta_t}\right)$，并重复附录 A 中为证明定理 3.6 所进行的步骤，可直接得到以下界：

$$\begin{aligned}
\hat{U}_{t+1} &= \min_{\zeta_t>0}\left((1+\zeta_t)\left(\left(1 - 2\alpha_t m + 2\alpha_t^2 L^2\right)\hat{U}_t + 2\alpha_t^2 G^2\left(1 + \frac{\delta^2}{\zeta_t}\right) + \frac{\alpha_t^2 c^2}{\zeta_t}\right)\right) \\
&= \left(\alpha_t\sqrt{c^2 + 2\delta^2 G^2 + 2\delta^2 L^2 \hat{U}_t} + \sqrt{\left(1 - 2\alpha_t m + 2\alpha_t^2 L^2\right)\hat{U}_t + 2G^2\alpha_t^2}\right)^2.
\end{aligned}$$

参考文献

[1] N. S. Aybat、A. Fallah、M. Gurbuzbalaban 和 A. Ozdaglar。《用于光滑强凸函数的鲁棒加速梯度方法》。SIAM Journal on Optimization，30(1)：717–751，2020。

[2] A. Badithela 和 P. Seiler。《利用积分二次约束分析重球算法》。载于：2019年美国控制会议（ACC），第 4081–4085 页，2019 年。[3] S. Cyrus、B. Hu、B. Van Scoy 和 L. Lessard。《用于强凸函数的鲁棒加速优化算法》。载于：美国控制会议，第 1376–1381 页，2018 年。[4] A. Datar 和 H. Werner。《利用积分二次约束对寻源动力学进行鲁棒性能分析》。载于：2022年美国控制会议（ACC），第 5229–5234 页，2022 年。[5] E. De Klerk、F. Glineur 和 A. Taylor。《关于采用精确线搜索的梯度法在光滑强凸函数上的最坏情况复杂度》。Optimization Letters，11(7)：1185–1199，2017 年。[6] N. K. Dhingra、S. Z. Khong 和 M. R. Jovanovi´c。《用于非光滑复合优化的近端增广拉格朗日方法》。IEEE Transactions on Automatic Control，64(7)：2861–2868，2018 年。[7] R. Dixit、A. S. Bedi、R. Tripathi 和 K. Rajawat。《采用非精确近端在线梯度下降算法的在线学习》。IEEE Transactions on Signal Processing，67(5)：1338–1352，2019 年。[8] R.-A. Dragomir、A. B. Taylor、A. d'Aspremont 和 J. Bolte。《Breg man一阶方法的最优复杂度与认证》。《数学规划》，194(1)：41–83，2022 年。[9] Y. Drori 和 M. Teboulle。《光滑凸最小化中一阶方法的性能：一种新方法》。《数学规划》，145(1-2)：451–482，2014 年。[10] M. Fazlyab、S. Paternain、V. M. Preciado 和 A. Ribeiro。《用于时变凸优化的预测-校正内点法》。IEEE Transactions on Automatic Control，63(7)：1973–1986，2017 年。[11] M. Fazlyab、A. Ribeiro、M. Morari 和 V. M. Preciado。《通过积分二次约束分析优化算法：非强凸问题》。SIAM Journal on Optimization，28(3)：2654–2689，2018 年。[12] O. Gannot。《非精确梯度法的频域分析》。《数学规划》，第 1–42 页，2021 年。[13] S. Han。《一致性优化去中心化算法的系统设计》。arXiv 预印本 arXiv:1903.01023，2019 年。[14] S. Hassan-Moghaddam 和 M. R. Jovanovi´c。《近端梯度流与Douglas–Rachford分裂动力学：通过积分二次约束实现全局指数稳定性》。Automatica，123：109311，2021 年。[15] E. Hazan。《在线凸优化导论》。Foundations and Trends® in Optimization，2(3-4)：157–325，2016 年。[16] B. Hu 和 L. Lessard。《Nesterov加速法的耗散性理论》。载于：第34届国际机器学习会议论文集，第 70 卷，第 1549–1557 页，2017 年。

[17] B. Hu、P. Seiler 和 L. Lessard. 使用序贯半定规划分析有偏随机梯度下降. 《数学规划》，187(1):383–408，2021.[18] B. Hu、P. Seiler 和 A. Rantzer. 运用跳跃系统理论和二次约束统一分析随机优化方法. 载于《2017年学习理论会议论文集》，第65卷，第1157–1189页，2017.[19] B. Hu、S. Wright 和 L. Lessard. 加速随机方差缩减的耗散性理论：运用半定规划统一分析随机方差缩减梯度和 Katyusha. 载于《国际机器学习会议》，第2043–2052页，2018.[20] D. Kim 和 J. Fessler. 光滑凸最小化的优化一阶方法. 《数学规划》，159(1-2):81–107，2016.[21] D. Kim 和 J. A. Fessler. 再看快速迭代收缩阈值算法（快速迭代收缩阈值算法）. 《SIAM Journal on Optimization》，28(1):223–250，2018.[22] D. Kim 和 J. A. Fessler. 推广光滑凸最小化的优化梯度法. 《SIAM Journal on Optimization》，28(2):1920–1950，2018.[23] L. Lessard. 优化算法分析：一种耗散性方法. 《IEEE 控制系统杂志》，42(3):58–72，2022.[24] L. Lessard、B. Recht 和 A. Packard. 通过积分二次约束分析和设计优化算法. 《SIAM Journal on Optimization》，26(1):57–95，2016.[25] L. Madden、S. Becker 和 E. Dall'Anese. 一阶在线优化方法跟踪误差的界. 《优化理论与应用杂志》，189(2):437–457，2021.[26] A. Megretski 和 A. Rantzer. 通过积分二次约束进行系统分析. IEEE《自动控制汇刊》，42:819–830，1997.[27] H. Mohammadi、M. Razaviyayn 和 M. R. Jovanovic. 加速一阶算法在强凸优化问题中的鲁棒性. IEEE《自动控制汇刊》，2020.[28] R. Nishihara、L. Lessard、B. Recht、A. Packard 和 M. Jordan. ADMM 收敛性的一般分析. 载于《第32届国际机器学习会议论文集》，第343–352页，2015.[29] E. K. Ryu、A. B. Taylor、C. Bergeling 和 P. Giselsson. 算子分裂性能估计：紧致收缩因子与最优参数选择. 《SIAM Journal on Optimization》，30(3):2251–2271，2020.[30] C. Scherer 和 C. Ebenbauer. 加速梯度算法的凸综合. 《SIAM 控制与优化杂志》，59(6):4615–4645，2021.[31] J. H. Seidman、M. Fazlyab、V. M. Preciado 和 G. J. Pappas. 用于 Douglas–Rachford 分裂分析和参数选择的控制理论方法. 《IEEE 控制系统快报》，4(1):199–204，2019。

[32] S. Shalev-Shwartz. 在线学习与在线凸优化.《机器学习基础与趋势》，4(2):107–194, 2011.[33] A. Simonetto、E. Dall'Anese、S. Paternain、G. Leus 和 G. B. Giannakis. 时变凸优化：时间结构化算法及其应用.《IEEE会刊》，108(11):2032–2048，2020.[34] A. Sundararajan、B. Hu 和 L. Lessard. 分布式优化算法的鲁棒收敛分析. 载于年度Allerton通信、控制与计算会议，第1206–1212页，2017.[35] A. Sundararajan、B. Van Scoy 和 L. Lessard. 时变图上的一阶分布式优化算法分析与设计.《IEEE网络系统控制汇刊》，7(4):1597–1608, 2020.[36] A. Taylor 和 F. Bach. 随机一阶方法：通过势函数进行非渐近分析与计算机辅助分析. 载于《2019年学习理论会议论文集》，第2934–2992页，2019.[37] A. Taylor、J. Hendrickx 和 F. Glineur. 光滑强凸插值与一阶方法的精确最坏情况性能.《数学规划》，161(1-2):307–345, 2017.[38] A. Taylor、J. M. Hendrickx 和 F. Glineur. 复合凸优化的一阶方法的精确最坏情况性能.《SIAM Journal on Optimization》，27(3):1283–1313，2017.[39] A. Taylor、B. Van Scoy 和 L. Lessard. 一阶方法的 Lyapunov函数：紧致的自动化收敛保证. 载于《第35届国际机器学习会议论文集》，第4897–4906页，2018.[40] B. Van Scoy、R. Freeman 和 K. Lynch. 已知最快的全局收敛一阶强凸函数最小化方法.《IEEE 控制系统快报》，2(1):49–54，2017.[41] J. Willems. 耗散动力系统（第一部分）：一般理论.《Archive for Rational Mech. and Analysis》，45(5):321–351，1972.[42] H. Xiong、Y. Chi、B. Hu 和 W. Zhang. 正则性条件下非凸优化中加速梯度下降的解析收敛区域. Automatica，113，2020.[43] G. Zhang、X. Bao、L. Lessard 和 R. Grosse. 通过积分二次约束统一分析光滑博弈的一阶方法.《机器学习研究杂志》，22(103), 2021。