arXiv:2602.08372v1 [cs.LG] 9 Feb 2026

通过从折扣遗憾到动态遗憾的约化研究动态遗憾：及其在曲线损失和 Adam 优化器中的应用

Yan-Feng Xie[1,2] Yu-Jie Zhang[3] Peng Zhao[1,2] Zhi-Hua Zhou[1,2]

[1] 新型软件技术国家重点实验室，南京大学，中国[2] 人工智能学院，南京大学，中国
[3] 华盛顿大学，美国

摘要

我们研究非平稳在线学习中的动态遗憾最小化，重点关注跟随正则化领导者（FTRL）方法。FTRL 对曲线损失以及理解 Adam 等自适应优化器至关重要，但现有研究对 FTRL 的动态遗憾分析较少。为此，我们以从折扣遗憾到动态遗憾的约化为基础，提出一种模块化方法，用于获得与 FTRL 相关问题的动态遗憾界。具体而言，我们关注两种具有代表性的曲线损失：线性回归和逻辑回归。我们的方法不仅简化了在线线性回归最优动态遗憾的现有证明，还为在线逻辑回归带来了新的动态遗憾保证。除在线凸优化外，我们还将该约化应用于分析 Adam 优化器，并在随机、非凸和非光滑设定下获得最优收敛速率。该约化还使我们能够更细致地分析具有两个折扣参数的 Adam（β_1，β_2），并为 Adam 优化器的裁剪和无裁剪变体均得到了新结果。

1 引言

在线学习为机器学习和序贯决策问题提供了一个通用建模框架，并已得到广泛研究 [Cesa-Bianchi 和 Lugosi,2006, Shalev-Shwartz,2012, Lattimore 和 Szepesv´ari, 2020]. 本文研究在线凸优化（OCO）框架下的非平稳在线学习 [Zinkevich, 2003, Hazan, 2016, Orabona, 2019], 其中在每一轮 $t \in [T]$，学习者从紧凸集 $\mathcal{X} \subseteq \mathbb{R}^d$ 中选择一个决策 $\mathbf{x}_t \in \mathcal{X}$，随后承受损失 $f_t(\mathbf{x}_t)$，其中 $f_t : \mathcal{X} \to \mathbb{R}$ 是凸的。非平稳在线学习中一种广泛使用的性能指标是动态遗憾 [Zhang 等人, 2018], 它将学习者的表现与随时间变化的比较器序列 $\mathbf{u}_1, \ldots, \mathbf{u}_T \in \mathcal{X}$进行比较:

$$\text{D-Reg}_T(\mathbf{u}_{1:T}) = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t), \tag{1}$$

当比较器固定时，即对所有 t，$\mathbf{u}_t = \mathbf{u}$，这就简化为标准静态遗憾。动态遗憾能更好地反映算法在非平稳环境中的表现。

在线凸优化（OCO）中的一个经典算法框架是在线镜像下降（OMD） [Beck 和 Teboulle, 2003]。OMD 会进行局部且贪心的更新，即从当前迭代点沿梯度方向迈出一步$\mathbf{x}_t$，因此直观而言，它受过去信息的影响较小，能够快速响应环境变化。因此，推导动态遗憾保证的一种常见方法是采用 OMD 类型的方法，再利用两层在线集成 [Zhao 等人, 2024] 以适应未知的非平稳性。

在线凸优化中的另一个基础框架是跟随正则化领导者（FTRL） [Abernethy 等人, 2008]。与在线镜像下降不同，FTRL 通过最小化一个$\mathbf{x}_t$ 全局目标来进行选择，通常是累积损失加上一个正则化项。这一更新规则使其成为优化许多在线学习问题的静态遗憾的便捷工具，尤其适用于必须仔细利用损失结构才能获得先进保证的问题。一个典型例子是具有曲线损失的在线学习，例如在线

线性回归、在线逻辑回归 [J´ez´equel 等人，2020，Agarwal 等人，2022，Mayo 等人，2022，Jacobsen 和 Cutkosky， 2024], 以及遗憾最优性和计算效率都至关重要的场景 [Zimmert 等人，2022，Mhammedi 和 Rakhlin，2022，Mhammedi 和 Gatmiry，2023，J´ez´equel 等人， 2025]。

然而，跟随正则化领导者（FTRL）的这一优势也伴随着取舍：使用所有过去的信息可能会减缓其在非平稳环境中的适应速度。事实上，已有负面结果表明，即使环境变化轻微，FTRL 也可能遭受线性动态遗憾 [Jacobsen 和 Cutkosky，2022，定理 2]。近期研究旨在保留 FTRL 的核心框架，同时引入改进，使其能够更好地追踪不断变化的比较器。一项值得关注的成果来自Mhaisen 和 Iosifidis [2025], ，他们提出加入校正项以有效剪枝历史信息，从而为凸损失实现最优动态遗憾保证。然而，现有结果未能涵盖若干重要的曲线损失，而在这些损失上，FTRL 相较于在线镜像下降（OMD）具有优势。

近期，Ahn 等人 [2024] 的一项深刻研究揭示了 FTRL 的动态遗憾与 Adam 优化器之间一种自然却引人入胜的联系 [Kingma 和 Ba， 2015], 尤其有助于理解其动量分量的作用。具体而言，Ahn 等人 [2024]表明，Adam 可以被视为一种优化折扣遗憾的 FTRL 实例，其定义如下：

$$\textsc{Reg}_{t;\beta}(\mathbf{u}) = \sum_{s=1}^{t} \beta^{t-s}\big(f_s(\mathbf{x}_s) - f_s(\mathbf{u})\big). \tag{2}$$

Ahn 等人 [2024] 建立了一种将折扣遗憾最小化与动态遗憾联系起来的归约。这一联系表明，Adam 可以被视为在隐式最小化动态遗憾，从而有助于处理非凸和非光滑设定下的非平稳性。后续工作 [Ahn 和 Cutkosky，2024] 利用这一视角分析了略作修改的裁剪版 Adam 的折扣遗憾，并推导出收敛保证。尽管取得了成功，他们的分析集中于耦合选择 $\beta_2 = \beta_1^2$，这比实践中使用的许多参数设置更为狭窄。要在更一般的参数选择下分析收敛保证（特别是 $\beta_2 \neq \beta_1^2$），需要一种细粒度的方法来执行折扣到动态（D2D）归约。

基于 D2D 归约 [Ahn 等人的工作， 2024], 我们提出了一种模块化的折扣到动态遗憾分析，适用于分析 FTRL 的动态遗憾：我们并非在确定一个经过调参的折扣遗憾闭式界之后再应用 D2D 归约，而是在重缩放损失的未调参上界模板层面进行表述。这种模板级视角使得促成保证的关键项在整个归约过程中始终保持明确，并支持更加灵活的调参，同时还能在不同的算法实例化和优化情形之间复用。

我们的结果 I：曲面损失的动态遗憾。 我们将模块化归约实例化于两个代表性回归任务。对于在线线性回归 [Vovk，2001，Azoury 和 Warmuth， 2001],它简化了对 [Jacobsen 和 Cutkosky， 2024] 的分析，避免了特定于算法的布雷格曼散度论证，同时恢复了匹配的最优界。对于在线逻辑回归，它为 AIOLI 的折扣变体给出了动态遗憾保证 [J´ez´equel 等人， 2020]。该界避免了逻辑回归分析中通常出现的指数依赖，并提高了非平稳环境中的稳健性。

我们的结果 II：Adam 优化器。 除在线凸优化之外，同一 D2D 分析还可通过在线到非凸（O2NC）转换，用于分析 Adam 优化器在随机、非凸和非光滑设定中的收敛速率 [Cutkosky 等人，2023，Zhang and Cutkosky，2024]。利用这一改进方法，我们在更灵活的参数选择下获得了收敛保证，而非受限的选择 $\beta_2 = \beta_1^2$ [Ahn 和 Cutkosky， 2024], 同时保持到 (c, ϵ)-驻点的最优收敛 [Zhang and Cutkosky， 2024]。

具体而言，对于裁剪 Adam，我们给出了两个充分的收敛条件。第一个要求$\beta_2 \geq \max\{1 - \frac{\nu}{G+\sigma}, \beta_1^4\}$，其中 ν是 Adam 的分母中的数值稳定性项， G和σ 分别是 Lipschitz 常数和梯度方差。这有助于阐明裁剪的作用：它允许动量参数之间的耦合更弱，最低可至 $\beta_2 \geq \beta_1^4$。第二个给出了一个满足 $\beta_2 > \beta_1^2$ 的边际型条件，并量化了间隙 $\beta_2 - \beta_1^2$ 对收敛率的影响。我们还研究了非光滑和非凸情形下的无裁剪 *Adam*，并得到了相匹配的最优收敛速率

采用类似的分析。最后，我们的结果还可以推出非凸光滑和二阶光滑设定下的最优速率。

技术。 我们在重缩放损失的未调优上界层面构建 D2D 归约，并给出明确的动态遗憾界。 *(i)* 首先，这一视角使动态遗憾分析更易复用：它不仅能简洁地证明无约束设定下的折扣 VAW 预测器，还提供了一种直接方法，可为另一种具有代表性的曲线损失——逻辑损失——推导保证。更广泛地说，同一框架也适用于本文示例之外的情形，可作为处理非平稳环境的一般工具。 *(ii)* 其次，在非凸和非光滑设定下应用 O2NC 转换分析 Adam 时，一个主要难点出现在 $\beta_2 \neq \beta_1^2$。在这种情况下，在线问题中诱导的学习率不再单调，从而使标准论证失效。我们的归约有助于分离决定最终收敛率的各项，使我们能够分别加以控制。这对于获得 Adam 的细粒度保证尤为有用。具体而言，我们提出了一个自信调参引理，用于处理分母为历史梯度指数加权和的学习率。我们还改进了对诱导的在线算法的分析，消除了对数值稳定性项 ν的依赖。关键步骤是采用尺度无关风格的论证，界定学习率在不同轮次间的变化幅度 [Orabona 和 P´等人， 2018]。由此得到 Adam 的一个改进收敛条件，并使其对间隔的依赖 $\beta_2 - \beta_1^2$ 更加明确。

组织结构。第2节介绍折扣到动态模块化分析。第3 节利用模块化归约研究曲线损失的动态遗憾。第4 节通过 O2NC 转换研究 Adam 的裁剪变体和无裁剪变体的收敛条件。第5 节总结全文。相关工作的进一步讨论、省略的证明以及补充细节均见附录。

2 折扣到动态模块化归约

本节首先回顾折扣遗憾与动态遗憾优化之间的联系，然后介绍折扣到动态模块化分析并作进一步讨论。所有证明均放在附录B中。

2.1 动态遗憾与折扣遗憾的联系

Ahn 等人 [2024]揭示了动态遗憾与折扣遗憾之间的联系，以阐释动量项在 Adam 中的作用。下面总结他们的结果。

引理 1(改编自 Ahn 等人 [2024]中的定理 B.3)。以下陈述对任意$T > 0$ 以及任意比较序列 $\mathbf{u}_1, \dots, \mathbf{u}_T \in \mathcal{X} \subseteq \mathbb{R}^d$ 均成立：

$$\text{D-Reg}_T(\mathbf{u}_{1:T}) = \beta \sum_{t=1}^{T-1} (\text{Reg}_{t;\beta}(\mathbf{u}_t) - \text{Reg}_{t;\beta}(\mathbf{u}_{t+1})) + (1-\beta) \sum_{t=1}^{T} \text{Reg}_{t;\beta}(\mathbf{u}_t) + \beta \text{Reg}_{T;\beta}(\mathbf{u}_T),$$

其中 $\text{Reg}_{t;\beta}(\mathbf{u})$在公式 (2)中定义。

为了解释这一结果，请注意，左侧是动态遗憾，它以变化的比较序列为基准。右侧则通过折扣遗憾来表示：它取决于折扣因子 β，但在每个时刻都只与一个固定的比较器进行比较。这种转换在现有分析工具主要针对固定比较器设计的场景中很有帮助，FTRL 方法 [Orabona， 2019]。

利用引理 1 的一种直接方法是，先为 $\text{Reg}_{t;\beta}$ 建立折扣遗憾界，再将其代入恒等式以得到动态遗憾界。这种方法具有普适性，并已成功用于分析裁剪 Adam [Ahn 和 Cutkosky， 2024]。然而，所得表达式通常包含多重求和与多项，因此对其进行化简可能会掩盖最终界对关键量（例如非平稳性和稳定性）的依赖关系。受此启发，我们提出一种互补且模块化的归约方法：在未调参上界的层面表述 D2D 归约，从而在整个推导过程中保持主要项明确，并便于进行调参和权衡，以获得最优且信息量更丰富的保证。

2.2 模块化D2D归约

基于 D2D 归约，我们给出以下定理，为 FTRL 在一系列设定下提供模块化分析。我们的分析以未调参上界模板的形式表述，不预先确定具体的调参方式。这种模板层面的形式保留了学习率选择的灵活性，也使这种转换能够复用于不同的算法族。

定理 1（模块化D2D归约）。假设一个在线学习算法 $\mathcal{A}$ 输出决策 $\mathbf{x}_t \in \mathcal{X}$ 并满足以下重缩放遗憾界：对于任意 $t \in [T]$ 以及任意比较器 $\mathbf{u} \in \mathcal{X}$，

$$\sum_{s=1}^{t} \beta^{-s} f_s(\mathbf{x}_s) - \sum_{s=1}^{t} \beta^{-s} f_s(\mathbf{u}) \le \varphi_t(\mathbf{u}) + \sum_{s=1}^{t} \Lambda_s, \tag{3}$$

其中 $\beta \in (0, 1]$，$\Lambda_s \ge 0$ 为稳定性项，而 $\varphi_t(\cdot) \ge 0$ 是分析中出现的依赖于比较器的量。于是，对于任意 $\mathbf{u}_1$、……、$\mathbf{u}_T \in \mathcal{X}$，$\mathcal{A}$ 满足

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t) \le \beta\varphi_1(\mathbf{u}_1) + \sum_{t=1}^{T} \beta^t \Lambda_t + \beta \sum_{t=1}^{T-1} \Big(F_t^{\beta,\varphi}(\mathbf{u}_{t+1}) - F_t^{\beta,\varphi}(\mathbf{u}_t) \Big) + \beta \sum_{t=1}^{T-1} \beta^t \Big(\varphi_{t+1}(\mathbf{u}_{t+1}) - \varphi_t(\mathbf{u}_{t+1}) \Big),$$

其中 $F^{\beta,\varphi}_t(\mathbf{u}) = \beta^t \varphi_t(\mathbf{u}) + \sum_{s=1}^{t} \beta^{t-s} f_s(\mathbf{u})$。

定理 1 受到Ahn 等人 [2024], Jacobsen 和 Cutkosky [2024]工作的启发。$\sum_{t=1}^{T} \beta^t \Lambda_t$ 项汇总了稳定性贡献，并刻画迭代点随时间发生变化的程度（例如，在$\mathbf{x}_{t-1}$ 和$\mathbf{x}_t$之间）。在许多分析中，Λ_t 采取 $\Lambda_t \approx \|\beta^{-t} \nabla f_t(\mathbf{x}_t)\|^2_{t,*}$的形式，其中 $\|\cdot\|_t$ 是时变范数，而$\|\cdot\|_{t,*}$是其对偶范数。

量 $\varphi_t(\cdot)$ 取决于具体分析，不必与算法所用的正则化项一致。对于在线回归，可以取 $\varphi_t(\mathbf{u}) = \frac{\lambda}{2}\|\mathbf{u}\|_2^2$，它与时间无关；因此项$\varphi_{t+1}(\mathbf{u}_{t+1}) - \varphi_t(\mathbf{u}_{t+1})$ 消失。然而，对于折扣 FTRL，φ_t 通常随 t 变化，我们需要控制由此产生的差分项，以确保其不影响最优速率。

为理解项 $\beta \sum_{t=1}^{T-1} \big(F^{\beta,\varphi}_t(\mathbf{u}_{t+1}) - F^{\beta,\varphi}_t(\mathbf{u}_t) \big)$，为简便起见，我们假设 $\varphi_t(\mathbf{u}) = \varphi(\mathbf{u})$ 与时间无关，并定义 $f_0(\mathbf{u}) = \varphi(\mathbf{u})$。于是，该项可由 $\frac{\beta}{1-\beta} P_T^{\beta}$ 界定，其中

$$P_T^{\beta} = \sum_{t=1}^{T-1} \sum_{s=0}^{t} p_{t,s}^{\beta} \big[f_s(\mathbf{u}_{t+1}) - f_s(\mathbf{u}_t) \big]_+, \tag{4}$$

并且 $p_{t,s}^{\beta} = \frac{\beta^{t-s}}{\sum_{\tau=0}^{t} \beta^{t-\tau}}$ 是归一化几何权重。量 P_T^{β} 通过损失差度量比较器变化，而标准路径长度 $P_T = \sum_{t=1}^{T-1} \|\mathbf{u}_{t+1} - \mathbf{u}_t\|_2$ [Zinkevich, 2003]度量决策空间中的变化。此外，若每个 f_s 关于范数 $\|\cdot\|$ 是 G-利普希茨的（即 $\|\nabla f_s(\mathbf{u})\|_* \le G$），则 $P_T^{\beta} \le G \sum_{t=1}^{T-1} \|\mathbf{u}_{t+1} - \mathbf{u}_t\|$，其中 $\|\cdot\|$ 可被选择以匹配问题的几何结构。如 Jacobsen 和 Cutkosky [2024], 所示，以 P_T^{β} 表示的动态遗憾界对于无约束在线线性回归问题可以是极小极大最优的。

总而言之，定理 1 为推导显式的动态遗憾保证提供了一个模块化模板。接下来的两节将给出自成一体的应用，以进一步展示这一分析的有效性。

3 曲面损失的动态遗憾

我们考虑以下在线回归协议：

1: **对于** $t = 1, \ldots\ldots, T$**执行**2: 学习者接收一个特征向量 $\mathbf{z}_t \in \mathcal{Z} \subseteq \mathbb{R}^d$;3: 学习者提交一个决策 $\mathbf{x}_t \in \mathcal{X} \subseteq \mathbb{R}^d$;

4: 环境揭示标签 $y_t \in \mathcal{Y} \subseteq \mathbb{R}$；5: 学习者遭受损失 f_t $(\mathbf{x}_t) = \ell(\mathbf{x}_t^\top \mathbf{z}_t, y_t)$。
6: **结束循环**

其中 $\ell(\cdot,\cdot)$: $\mathbb{R} \times \mathbb{R} \to \mathbb{R}$ 是损失函数。我们称该协议为非正规的，因为学习者在看到 $\mathbf{x}_t$ 后，才选择 $\mathbf{z}_t$，并使用对实例 $\mathbf{z}_t$可能并非线性的预测。我们注意到，在许多机器学习问题中，这是一个较弱的假设 [Orabona, 2019]。接下来，我们将该协议具体化，以研究两种典型曲线损失的动态遗憾：平方损失（第 3.1 节）和逻辑损失（第 3.2 节）。证明见附录 C。

3.1 在线线性回归

我们首先研究采用平方损失 $\ell(\hat{y}, y) = \frac{1}{2}(\hat{y}-y)^2$, 的无约束在线线性回归，其中 $\mathcal{X} = \mathbb{R}^d$、 $\mathcal{Z} = \mathbb{R}^d$ $\mathcal{Y} = \mathbb{R}$ t $\hat{y}_t = {}_t^\top {}_t$ f_t ${}_t$ $=$，并遵循该协议。在第 轮中，学习者预测 $\mathbf{xz}$，并遭受损失 $(\mathbf{x})$ $_1$ $_2(\mathbf{x}_t^\top \mathbf{z}_t - y_t)^2$。

在有界域上，平方损失具有指数凹性，ONS 等二阶方法能够实现对数静态遗憾 [Hazan, 2016, Hazan 等人, 2007]。在我们的设定 $\mathcal{X} = \mathbb{R}^d$中，这种一致有界性（因而全局指数凹性）通常并不存在。然而，一旦我们能够利用特征 $\mathbf{z}_t$提供的侧信息，VAW预测器 [Vovk, 2001, Azoury 和 Warmuth, 2001]即使在 $\mathcal{X} = \mathbb{R}^d$的情况下也能实现对数遗憾。其更新可写为 FTRL解：

$$\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \left\{ \frac{\lambda}{2}\|\mathbf{x}\|_2^2 + \frac{1}{2}(\mathbf{x}^\top \mathbf{z}_t)^2 + \frac{1}{2}\sum_{s=1}^{t-1}(\mathbf{x}^\top \mathbf{z}_s - y_s)^2 \right\}.$$

VAW预测器确保如下保证。

引理2. 在无约束在线线性回归中，*VAW*预测器对任意 $t \in [T]$确保如下界：

$$\sum_{s=1}^{t} f_s(\mathbf{x}_s) - f_s(\mathbf{u}) \le \frac{\lambda}{2}\|\mathbf{u}\|_2^2 + \frac{1}{2}\sum_{s=1}^{t-1} y_s^2 \mathbf{z}_s^\top A_s^{-1} \mathbf{z}_s,$$

对任意$\mathbf{u} \in \mathbb{R}^d$，其中$A_t = \lambda I + \sum_{s=1}^{t} \mathbf{z}_s \mathbf{z}_s^\top$且 $I \in \mathbb{R}^{d\times d}$ 为单位矩阵。

引理2是针对固定比较器陈述的。然而，在非平稳情形中，最优预测器可能随时间漂移，因此我们使用动态遗憾（1）来评估性能，并通过第2节中的模块化D2D归约对其进行分析。

遵循Jacobsen 和 Cutkosky [2024], ，我们考虑带有因子 $\beta \in (0, 1]$, 的VAW的折扣变体，该变体强调近期数据。在第 t轮，它通过以下方式选择决策

$$\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \left\{ \frac{\lambda\beta^t}{2}\|\mathbf{x}\|_2^2 + \frac{1}{2}(\mathbf{x}^\top \mathbf{z}_t)^2 + \frac{1}{2}\sum_{s=1}^{t-1} \beta^{t-s}(\mathbf{x}^\top \mathbf{z}_s - y_s)^2 \right\}.$$

利用重缩放技巧 [Zhang 等人, 2024], 定义 $\tilde{}\mathbf{z}_t = \beta^{-t/2}\mathbf{z}_t$ 和 $\tilde{y}_t = \beta^{-t/2} y_t$, 使得 $\ell(\mathbf{x}^\top \tilde{}\mathbf{z}_t, \tilde{y}_t) = \beta^{-t}\ell(\mathbf{x}^\top \mathbf{z}_t, y_t)$。将引理2 应用于重缩放后的序列，可得如下形式的遗憾界，对任意$t \in [T]$:

$$\sum_{s=1}^{t} \beta^{-s}\Big(\ell(\mathbf{x}_s^\top \mathbf{z}_s, y_s) - \ell(\mathbf{u}^\top \mathbf{z}_s, y_s)\Big) \le \frac{\lambda}{2}\|\mathbf{u}\|_2^2 + \sum_{s=1}^{t} \beta^{-2s} y_s^2 \mathbf{z}_s^\top (\lambda I + \sum_{\tau=1}^{s} \beta^{-\tau}\mathbf{z}_\tau \mathbf{z}_\tau^\top)^{-1}\mathbf{z}_s,$$

因此，上述表达式符合定理 1 的条件，其中 $\varphi_t(\mathbf{u}) = \frac{\lambda}{2}\|\mathbf{u}\|_2^2$ 以及 $\Lambda_t = \beta^{-2t} y_t^2 \mathbf{z}_t^\top (\lambda I + \sum_{s=1}^{t} \beta^{-s}\mathbf{z}_s \mathbf{z}_s^\top)^{-1}\mathbf{z}_t$, 由此得到以下动态遗憾保证。

定理 2。 在无约束在线线性回归中，折扣 *VAW* 预测器满足以下动态遗憾界：对于任意 $\mathbf{u}_1, \ldots, \mathbf{u}_T \in \mathbb{R}^d$,

$$\frac{\beta\lambda}{2}\|\mathbf{u}_1\|_2^2 + \frac{d}{2}\big(\max_{t\in[T]} y_t^2\big) \cdot \ln\left(1 + \frac{\sum_{t=1}^{T} \beta^{T-t}\|\mathbf{z}_t\|_2^2}{\lambda d}\right) + \frac{\beta}{1-\beta} \cdot P_T^\beta + \frac{1-\beta}{\beta} \cdot \frac{d}{2}\sum_{t=1}^{T} y_t^2,$$

其中 $\beta \in (0,1)$，$\lambda > 0$，且 P_T^β 在公式 (4) 中定义。此外，存在一种双层集成算法，可确保动态遗憾为 $\mathcal{O}\big(d \log T + \sqrt{dTP_T^\beta}\big)$。

定理 2 达到了 Jacobsen 和 Cutkosky [2024] 的最优速率。我们的推导更加直接：它将模块化 D2D 归约应用于重缩放遗憾模板，避免了将更新重写为在线镜像下降以及随之而来的 Bregman 散度论证 [Jacobsen 和 Cutkosky。 2024] 与折扣在线牛顿步（ONS）分析相比 [Hazan 等人，2007，Yuan 和 Lamperski， 2020]，我们的结果不依赖指数凹性，并允许无约束决策。通过利用 Jacobsen 和 Cutkosky [2024]，提出的双层算法，我们可以得到一个 $\mathcal{O}(d$ 对数 $T + \sqrt{dTP_T^\beta}$ 界，与该无约束设定下的下界相匹配。

3.2 在线逻辑回归

接下来，我们研究具有 $\ell(\widehat{y}, y) = \ln\big(1+\exp(-y\widehat{y})\big)$ 的在线逻辑回归。在第 t 轮，学习者预测 $\widehat{y}_t = \mathbf{x}_t^\top \mathbf{z}_t$，并遭受 $f_t(\mathbf{x}_t) = \ln\big(1 + \exp(-y_t \mathbf{x}_t^\top \mathbf{z}_t)\big)$。我们假设标签为二元标签 $\mathcal{Y} = \{+1, -1\}$，且特征有界 $\mathcal{Z} = \{\mathbf{z} \in \mathbb{R}^d : \|\mathbf{z}\|_2 \le R\}$。学习者可以选择 $\mathbf{x}_t \in \mathbb{R}^d$，而比较序列对所有 t 都满足 $\|\mathbf{u}_t\|_2 \le B$。我们遵循相关约定来设定在线逻辑回归中的有界性假设 [Hazan 等人, 2007, Foster 等人, 2018, J´ez´equel 等人，2020]。

在有界比较器集合 $\|\mathbf{u}\|_2 \le B$ 上，逻辑损失是指数凹的，因此应用 ONS 可得静态遗憾上界 $\mathcal{O}(de^B \log T)$：它关于 T 呈对数增长，但关于 B 呈指数增长。对于适当在线凸优化算法而言，这种对 B 的指数依赖通常无法避免 [Hazan 等人， 2014]。然而，Foster 等人[2018] 展示了非适当性的重要性，并且可以通过一种高效方法将对 B 的依赖从 e^B 降至 BJ [ez´equel 等人提出的方法， 2020]。

为了应对回归目标可能发生的漂移，我们还旨在高效地获得动态遗憾保证，同时避免对 B的指数依赖。为此，我们提出 AIOLI 算法的一种折扣变体 [J´ez´equel 等人， 2020]，其形式化定义如下：

$$\mathbf{x}_t = \arg\min_{\mathbf{x} \in \mathbb{R}^d} \left\{ \frac{\beta^t \lambda}{2} \|\mathbf{x}\|_2^2 + h_t(\mathbf{x}) + \sum_{s=1}^{t-1} \beta^{t-s} \widehat{f}_s(\mathbf{x}) \right\},$$

其中 $h_t(\mathbf{x}) = \ell(\mathbf{x}^\top \mathbf{z}_t, +1) + \ell(\mathbf{x}^\top \mathbf{z}_t, -1)$ 使用观测到的特征 $\mathbf{z}_t$，可视为一个乐观项 [Chiang 等人，2012，Rakhlin 和 Sridharan， 2013]，它通过对即将到来的损失函数进行猜测来实现。这里，$\widehat{f}_t(\mathbf{x})$ 是 $f_t(\mathbf{x})$ 的替代函数，定义为 $\widehat{f}_t(\mathbf{x}) = f_t(\mathbf{x}_t) + \langle \nabla f_t(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle + \frac{\eta_t}{2} \langle \nabla f_t(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle^2$，其中$\eta_t = \frac{\exp(y_t \widehat{y}_t)}{1+BR}$。

在引理 3中，我们从乐观跟随正则化领导者 [Orabona 的视角，建立了重缩放遗憾的任意时刻保证，2019]，满足定理 1的条件。通过应用所提出的模块化分析，我们得出以下结果。

定理 3。 假设 $\|\mathbf{z}_t\|_2 \le R$ 对所有 $t \in [T]$均成立。对于任意比较序列 $\mathbf{u}_1$、.....、$\mathbf{u}_T$，满足$\|\mathbf{u}_t\|_2 \le B$，带有参数的折扣 *AIOLI* 算法的动态遗憾界为：

$$\beta\lambda\|\mathbf{u}_1\|_2^2 + d(1+BR)\log\left(1 + \frac{R^2 \sum_{t=1}^T \beta^{T-t}}{d\lambda(1+BR)}\right) + \frac{\beta}{1-\beta} \cdot P_T^\beta + \frac{1-\beta}{\beta} \cdot d(1+BR)T,$$

其中 $\beta \in (0,1)$，$\lambda > 0$，且 P_T^β 定义见式 (4)。

证明推迟至附录 C.2。我们指出，折扣 AIOLI 更新位于 $\mathbb{R}^d$ 中，且不需要投影。额外的计算成本来自涉及乐观项 $h_t(\mathbf{x})$ 的隐式更新，该更新可通过 $\mathcal{O}(\log T)$ 次迭代 [J´ez´equel 等人， 2020]实现。

利用逻辑损失的可混合性，我们提出了一种双层在线集成算法来调整折扣因子 β [Mhammedi 和 Rakhlin， 2022，Zhao 等人， 2024]，总结于算法1中。下面的定理给出了算法1相应的理论保证，证明及更多相关细节见附录C.3。

定理 4. 假设 $\|\mathbf{z}_t\|_2 \le R$ 对所有 $t \in [T]$ 均成立，并定义凸损失 $f_t(\mathbf{x}) = \ell(\mathbf{x}^\top \mathbf{z}_t, y_t)$，其中$\ell(\widehat{y}, y) = \ln(1 + \exp(-y\widehat{y}))$。令 $\lambda = 1/B^2$。令 $C = \max\{1, 2R\}$，并定义参数范围

算法 1 用于学习逻辑回归中 β 的集成算法

输入：基学习器的数量 N、折扣因子池 $\{\beta_i\}_{i=1}^N$，以及参数 (B, R, λ)。1：初始化：$q_{1,i} = 1$ 以及基学习器 $\mathcal{B}_i$（折扣 AIOLI），其参数为 β_i, B, R, λ，适用于 $i \in [N]$。2：对于 $t = 1$ 到 T，执行 3：接收特征 $\mathbf{z}_t \in \mathbb{R}^d$；4：每个基学习器 $\mathcal{B}_i$ 输出 $\mathbf{x}_{t,i} \in \mathbb{R}^d$，并预测 $\widehat{y}_{t,i} = \mathbf{x}_{t,i}^\top \mathbf{z}_t$；5：设置 $p_{t,i} = q_{t,i} / \sum_{j=1}^N q_{t,j}$；6：预测

$$\widehat{y}_t = \ln\left(\frac{\sum_{i=1}^N p_{t,i} \frac{\exp(\widehat{y}_{t,i})}{1+\exp(\widehat{y}_{t,i})}}{\sum_{i=1}^N p_{t,i} \frac{1}{1+\exp(\widehat{y}_{t,i})}}\right);$$

7: 观察标签 $y_t \in \{-1, +1\}$ 并更新 $q_{t+1,\,i} = q_{t,\,i} \exp\left(-\ell(\widehat{y}_{t,\,i},\ y_t)\right)$ 对所有 $i \in [N]$。8: **结束循环**

$\eta_{\min} = \sqrt{\frac{d(1+BR)}{CB}}$，$\eta_{\max} = dT$，以及几何网格 $\{\eta_i = 2^{i-1}\eta_{\min} : \ \beta_i = \frac{\eta_i}{1+\eta_i} \in (0, 1) \ i \in [N]\}$，其中 $N = \left\lceil \log_2\left(\frac{\eta_{\max}}{\eta_{\min}}\right)\right\rceil + 1$。使用折扣因子池 $\{\beta_i = \frac{\eta_i}{1+\eta_i}\}_{i=1}^N$ 运行算法 1。记作 $P_T^\beta = \sum_{t=1}^{T-1} \sum_{s=0}^t p_t^\beta$，$_s\left[f_s(u_{t+1}) - f_s(u_t)\right]_+$，其中 $f_0(u) = \frac{\lambda}{2}\|u\|_2^2$，并令 $\beta_\star \in (0, 1]$ 满足

$$\beta_\star = \frac{\sqrt{d(1+BR)T}}{\sqrt{d(1+BR)T} + \sqrt{P_T^{\beta_\star}}}.$$

于是，算法1保证，对于任意比较序列$\mathbf{u}_1$、……、$\mathbf{u}_T$，满足 $\|\mathbf{u}_t\|_2 \le B$,

$$\sum_{t=1}^T \ell(\widehat{y}_t, y_t) - \sum_{t=1}^T \ell(\mathbf{u}_t^\top \mathbf{z}_t, y_t) \ \le \ \mathcal{O}\left(dB\log(BT) + \sqrt{dBTP_T^{\beta_\star}}\right).$$

由于在线逻辑回归中的比较器有界，无约束线性回归的下界并不能直接适用 [Jacobsen 和 Cutkosky, 2024]。已有研究针对指数凹损失，在不同的算法设计下，实现了对 T 和标准路径长度 P_T 的最优依赖，并给出了 $\widetilde{\mathcal{O}}(T^{1/3}P_T^{2/3})$ 动态遗憾界 [Baby and Wang, 2021, Zhang 等人，2025]，其中我们用 $\mathcal{O}(\cdot)$ 隐去 T 中的对数因子。这些方法要么依赖更复杂的黑箱分析 [Baby and Wang, 2021, 2022]，要么在某些情况下会产生较高的计算开销 [Zhang 等人，2025]。我们的方法提供了互补的权衡：通过一种高效算法和透明、模块化的分析，我们可以获得 $\mathcal{O}(\sqrt{TP_T})$ 界；此外，我们的框架也适用于无界决策集。

4 Adam 的收敛条件

本节将指数化 O2NC 框架 [Cutkosky 等人，2023，Zhang andCutkosky， 2024] 与模块化D2D归约（定理1）相结合，以分析随机、非凸和非光滑设定下的 Adam 优化器。我们首先介绍问题设定和 O2NC 转换，然后按照Ahn 和 Cutkosky [2024], 的模板研究裁剪 Adam，最后讨论无裁剪变体。所有证明均放在附录D中。

4.1 问题设定

我们研究以下随机、非凸且非光滑优化问题，其中 F： $\mathbb{R}^d \to \mathbb{R}$是定义如下的目标函数：

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \triangleq \mathbb{E}_\xi\left[f(\mathbf{x};\xi)\right].$$

我们作出以下假设。

算法 2 指数化 O2NC [Zhang andCutkosky, 2024]

输入：在线凸优化算法 $\mathcal{A}$，初始点 $\mathbf{x}_0$，参数 $T \in \mathbb{N}$, $\beta \in (0,1)$，损失函数 $\ell_t(\Delta)$。1：初始化：$\bar{\mathbf{x}}_0 = \mathbf{x}_0$ 2：对于 $t = 1, \ldots, T$，执行 3：接收来自 $\mathcal{A}$ 的 Δ_t； 4：更新 $\mathbf{x}_t = \mathbf{x}_{t-1} + s_t \cdot \Delta_t$，$s_t$ i.i.d.$\sim$ Exp(1)； 5：计算 $\mathbf{g}_t = \nabla f(\mathbf{x}_t, \xi_t)$； 6：将损失 $\ell_t(\Delta)$ 发送给 $\mathcal{A}$； 7：$\bar{\mathbf{x}}_t = \frac{\beta - \beta^t}{1-\beta^t} \bar{\mathbf{x}}_{t-1} + \frac{1-\beta}{1-\beta^t} \mathbf{x}_t$； 8：循环结束 9：返回：$\bar{\mathbf{x}} \sim \mathrm{Unif}(\{\bar{\mathbf{x}}_t : t \in [T]\})$。

假设 1。 F 可微且为 G-利普希茨，即对于所有 $\mathbf{x} \in \mathbb{R}^d$，$\|\nabla F(\mathbf{x})\|_2 \le G$。

假设 2。 给定一个初始点 $\mathbf{x}_0$, $F(\mathbf{x}_0) - \inf_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) \le F^*$ 对于某个已知的 F *.

假设 3。 假设 F 满足：对任意 $\mathbf{x}$、$\mathbf{y}$，$F(\mathbf{y}) - F(\mathbf{x}) = \int_0^1 \langle \nabla F(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle \mathrm{d}t$。

假设 4。 随机梯度是无偏的，$\mathbb{E}_\xi[\nabla f(\mathbf{x}; \xi)] = \nabla F(\mathbf{x})$，具有有界方差，$\mathbb{E}_\xi \|\nabla f(\mathbf{x}; \xi) - \nabla F(\mathbf{x})\|_2^2 \le \sigma^2$，且几乎必然有界，即 $\|\nabla f(\mathbf{x}; \xi)\|_2 \le G$。

假设 1 – 3 在 O2NC 框架中是标准的 [Cutkosky 等人，2023，Zhang and Cutkosky，2024]。为了在灵活的（β_1，β_2）条件下获得更细粒度的保证，我们还假设随机梯度有界，作为一种权衡；这是相关分析中常见的技术条件 [Cutkosky 和 Orabona，2019，Guo 等人，2021，D´efossez 等人，2022]。这一假设可以通过对噪声施加适当的尾部条件来放宽 [van Handel，2016]，但代价是收敛速率中会出现额外的对数因子。

定义 1. 假设 F： $\mathbb{R}^d \to \mathbb{R}$ 是一个可微函数。我们称$\mathbf{x}$ 为一个（c, ϵ）驻点 t 如果 $\|\nabla F(\mathbf{x})\|_c \le \epsilon$，其中

$$\|\nabla F(\mathbf{x})\|_c = \inf_{\substack{\mathbf{y} \sim P \in \mathcal{P}(\mathbb{R}^d) \\ \mathbb{E}[\mathbf{y}] = \mathbf{x}}} \|\mathbb{E}[\nabla F(\mathbf{y})]\|_2 + c \cdot \mathbb{E}\|\mathbf{y} - \mathbf{x}\|_2^2,$$

其中 $\mathcal{P}(\mathbb{R}^d)$ 表示 $\mathbb{R}^d$ 上所有分布的集合。

这一概念相当一般。通过适当选择 c，它可以涵盖若干标准概念 [ZhangandCutkosky，2024]，包括当目标函数光滑或二阶光滑时的一阶驻点。如果 F是 G-利普希茨的，那么一个（c，ϵ）驻点也意味着一个（δ，$(1 + \frac{2G}{c\delta^2})\epsilon$Goldstein 驻点，这一概念在非凸非光滑优化中被广泛使用 [Goldstein，1977，Zhang 等人，2020，Tian 等人，2022，Davis 等人，2022]。

Cutkosky 等人的开创性工作 [2023] 提出了 O2NC 转换。它将随机非凸和非光滑优化归约为在线学习问题，并且在标准假设下能够达到最优收敛速率。我们关注 O2NC 的一个指数化变体 [Zhang and Cutkosky，2024]，它与我们的 D2D 归约更为兼容。

算法2 概述了指数化 O2NC 转换 [Zhang and Cutkosky，2024]。在第4行中，Exp(a) 表示参数为 a的指数分布。不同的替代损失 $\ell_t(\Delta)$ 在第6行中会产生 Adam 的裁剪变体或无裁剪变体，具体将在后文说明。

为理解 O2NC，根据引理 6 和更新规则，我们有 $\mathbb{E}_{s_t}[F(\mathbf{x}_t) - F(\mathbf{x}_{t-1})] = \mathbb{E}_{s_t}[\langle \nabla F(\mathbf{x}_t)\ \Delta_t \rangle] = \mathbb{E}_{s_t,\ \xi_t}[\langle \mathbf{g}_t,\ \Delta_t \rangle]$。这表明可以使用在线学习器，根据由 $\langle \mathbf{g}_t,\ \Delta \rangle$（及其变体）构造的损失来选择更新方向 Δ_t，从而将优化问题转化为在线学习问题。

4.2 通过跟随正则化领导者理解 Adam

Ahn 等人 [2024]证明，Adam 可以理解为折扣 FTRL，这将 Adam 与动态遗憾联系起来，并凸显了引入折扣因子的重要性。基于这一

联系，我们可以使用模块化归约来分析折扣 FTRL 的动态遗憾，再将结果推广到 Adam。我们首先对折扣 FTRL 进行如下形式化定义。

$$\Delta_{t+1} = \arg\min_{\Delta \in \mathcal{D}} \left\{ \frac{\|\Delta\|_2^2}{2\eta_t} + \sum_{s=1}^{t} \beta^{-s} \ell_s(\Delta) \right\}, \tag{5}$$

其中， $\eta_t > 0$ 和 $\mathcal{D} \subseteq \mathbb{R}^d$ （一个闭凸集）将在后文指定。等价地，根据式（5）的一阶最优性条件， Δ_{t+1} 满足以下形式

$$\Delta_{t+1} = \Pi_{\mathcal{D}} \left[-\eta_t \sum_{s=1}^{t} \beta^{-s} \nabla \ell_s(\Delta_{t+1}) \right]. \tag{6}$$

在本节余下部分，我们将选择 $\ell_t(\cdot)$、 η_t和 $\mathcal{D}$ ，使得得到的折扣 FTRL 更新在 O2NC 框架下与（裁剪或无裁剪的）Adam 更新相匹配，并通过归约给出保证和分析。

4.3 裁剪 Adam 的收敛条件

我们首先遵循Ahn 和 Cutkosky [2024] 的算法模板来研究裁剪 Adam，并设定 $\ell_t(\Delta) = \langle \mathbf{g}_t, \Delta \rangle$ 以及 $\mathcal{D} = \{\Delta : \|\Delta\|_2 \le D\}$。我们选择

$$\eta_t = \gamma \cdot \frac{(1-\beta_1)\beta_1^t}{\nu + \sqrt{(1-\beta_2)\sum_{s=1}^{t} \beta_2^{t-s} \|\mathbf{g}_s\|_2^2}}, \tag{7}$$

并在式（ $\beta = \beta_1$ 5）中设定。作出这些选择后，式（6）中的更新规则得到:

$$\Delta_{t+1} = \mathrm{Clip}_D \left[-\gamma \cdot \frac{(1-\beta_1)\sum_{s=1}^{t} \beta_1^{t-s} \mathbf{g}_s}{\nu + \sqrt{(1-\beta_2)\sum_{s=1}^{t} \beta_2^{t-s} \|\mathbf{g}_s\|_2^2}} \right], \tag{8}$$

其中， $\mathrm{Clip}_D[\mathbf{a}] = \min\{\|\mathbf{a}\|_2, \ D\} \cdot \mathbf{a}/\|\mathbf{a}\|_2$ 是裁剪算子。与Ahn 和 Cutkosky[2024], 相比，我们的算法和分析不需要 $\beta_2 = \beta_1^2$。分子和分母采用 Adam 中使用的标准动量形式。与原始 Adam [Kingma 和 Ba相比， 2015], 我们提议省略校正项，使分析聚焦于动量和自适应缩放的主要影响 [D´efossez 等人，2022，Wang等人，2023，Ahn 和 Cutkosky， 2024]。这些校正因子主要影响早期迭代，并随时间推移呈指数衰减。

4.3.1 关于 β_2 的宽松条件

对于裁剪 Adam，我们首先给出以下定理，证明见附录 D.6。

定理 *5*。在假设 *1–4* 下，在 $\mathcal{D} = \{\Delta : \|\Delta\|_2 \le D\}$ 上使用 $\ell_t(\Delta) = \langle g_t, \Delta \rangle$ 和 η_t ，并采用式 *(7)* 中的 $\beta = \beta_1$ 和式 *(5)* 中的折扣 *FTRL* 更新运行算法 *2*。这意味着式 *(8)* 中的裁剪 *Adam* 更新规则。令 $1 - (16$ ϵ $(G+\sigma))^2 \le \beta_1 < 1$, $D = (1-\beta_1)\sqrt{\epsilon}/\sqrt{48c}$, $\gamma = \beta_1 D/\sqrt{1-\beta_1}$, $0 < \nu \le G + \sigma$ 和 $\max\{1 - \nu/(G+\sigma), \beta_1^4\} \le \beta_2 < 1$。若 T 满足

$$T \ge \max\left\{ \frac{1}{1-\beta_1} \max\left\{ \frac{16F^*\sqrt{48c}}{\epsilon^{3/2}}, \ \frac{16(G+\sigma)}{\epsilon} \right\}, \frac{\ln 2}{1-\beta_2} \right\},$$

则算法 $\bar{\mathbf{x}}$ 的输出 *2* 满足 $\mathbb{E}[\|\nabla F(\bar{\mathbf{x}})\|_c] \le \epsilon$。

该定理为指数移动平均（指数移动平均）迭代点提供了收敛保证$\bar{\mathbf{x}}$ [Ruppert，1988年，Polyak 和 Juditsky，1992年，Ahn 和 Cutkosky， 2024] 当更新方向由裁剪 Adam 生成时，指数化 O2NC 返回的迭代点。该定理保证迭代复杂度为

$$\mathcal{O}(\max\{(G+\sigma)^2 F^* c^{1/2} \epsilon^{-7/2}, \ (G+\sigma)^3 \epsilon^{-3}, \ (G+\sigma)\nu^{-1}\}).$$

只有当 ν 足够小时，最后一项才占主导，此时 $\nu \le \Theta(\min\{\epsilon^{7/2}/(GF^*c^{1/2}),\ \epsilon^3/G^2\})$。在大多数情形下，首项与寻找一个（$F^*G^2c^{1/2}\epsilon^{-7/2}$）驻点的下界 $\Omega(c,\ \epsilon)$ 相匹配 [Zhang and Cutkosky, 2024]。

我们分析的一个关键启示是，对（β_1, β_2）的选择可以有更精细的理解。一些现有的收敛分析采用标准条件 $\beta_2 \ge \beta_1^2$ [Luo 等人，2019，Zhang 等人，2022b，D´efossez 等人，2022]。在裁剪下，我们的分析表明，这种耦合条件可以放宽：只需取 $\beta_2 \ge \beta_1^4$（至多再满足一个涉及 ν 的约束）即可。这个放宽是通过利用裁剪带来的益处推导得出的，即 $\|\Delta_t\|_2 \le D$。如果将 ν, G, σ 视为常数，也可以选择 $\beta_1 = \Theta(1 - (\epsilon\nu)^2/(G+\sigma)^3)$，这会在复杂度中引入一个额外的（$G+\sigma$）$\nu^{-2}$ 因子，同时使条件 $\beta_2 \ge \beta_1^4$保持不变，且不含 ν。

引理 8 （O2NC 转换）将收敛分析归结为控制动态遗憾。结合模块化 D2D 转换（定理 1）与折扣 FTRL 遗憾界，关键的动态遗憾项可以分解为三个明确的组成部分：

$$\frac{1}{DT}\mathbb{E}\Bigg[\frac{1-\beta_1}{\beta_1}\gamma\sum_{t=1}^{T}\alpha_{t-1}\|\mathbf{g}_t\|_2^2 + \frac{1}{\gamma(1-\beta_1)}\sum_{t=1}^{T}\left(\frac{\beta_1}{2\alpha_{t-1}}\|\Delta_{t+1}\|_2^2 - \frac{1}{2\alpha_t}\|\Delta_{t+1}\|_2^2\right) + \frac{\beta_1}{\gamma(1-\beta_1)}\sum_{t=1}^{T-1}\left(\frac{1}{2\alpha_t}\|\mathbf{u}_{t+1}\|_2^2 - \frac{1}{2\alpha_{t-1}}\|\mathbf{u}_t\|_2^2\right)\Bigg], \tag{9}$$

其中，我们用 $\alpha_t = 1/\big(\nu + \sqrt{(1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\|\mathbf{g}_s\|_2^2}\big)$ 表示，而 $\mathbf{u}_t \in \mathbb{R}^d$ 是引入分析的比较序列，其中包含 $\|\mathbf{u}_t\|_2 = D$ 。

第一项称为在线学习中的稳定性项。我们使用一种新的自信调参引理（引理 9）对其进行分析，得到以下上界：

$$\mathcal{O}\left(\sqrt{1-\beta_1}(1-\beta_2)\frac{(G+\sigma)^2}{\nu}\right).$$

To k要使该项保持理想的阶 $\mathcal{O}(\sqrt{1-\beta_1}(G+\sigma))$，需要对 $\beta_2 \ge 1 - \frac{\nu}{G+\sigma}$ 进行调参。

对于（9）中的第二项，关键观察是，裁剪保证了 $\|\Delta_t\|_2 \le D$，并且在调参过程中 D 很小，因此对 β_2 的要求可以放宽，而不必对 $\beta_2 \ge \beta_1^2$ 进行调参并舍弃所有项。根据引理 11，该项可由 $\frac{[\beta_1 - \sqrt{\beta_2}]_+}{\beta_1\sqrt{1-\beta_1}}(G+\sigma)$ 上界控制。要匹配目标阶，只需要求 $[\beta_1 - \sqrt{\beta_2}]_+ \le \beta_1(1-\beta_1)$，这意味着 $\beta_2 \ge \beta_1^4$。

对于（9）中的最后一项，利用 $\|\mathbf{u}_t\|_2 = D$ 对所有 t，可以使求和形成望远镜消去，并可利用引理 12进行界定。

4.3.2 关于 β_2 的边际条件

定理 5 在选择 ν 时涉及参数 β_2。直观而言，这源于对等式（9）中第一项进行最坏情况分析。该项的形式为 $\alpha_{t-1}\|\mathbf{g}_t\|_2^2$。由于学习率 α_{t-1} 滞后一步，分析无法直接将其与当前梯度 $\mathbf{g}_t$配对；在最坏情况下，我们只能对 $\alpha_{t-1}\|\mathbf{g}_t\|_2^2 \le \|\mathbf{g}_t\|_2^2/\nu$作上界估计，因此需要调节 β_2 以吸收额外因子 G/ν，从而确保达到最优收敛速率。

借鉴无尺度在线学习中的思路 [Orabona 和 P´al, 2018], 我们可以通过将界改写为关于 $\alpha_t\|\mathbf{g}_t\|_2^2$ 的形式，并加上一个取决于步长变化的校正项 $\mathcal{O}(|\eta_t - \eta_{t-1}| \cdot \|\sum_{s=1}^{t-1}\beta_1^{t-1-s}\mathbf{g}_s\|_2)$，来改进这一步。这样做的好处是，$\alpha_t$ 会与 $\mathbf{g}_t$ 关联起来。直观而言，由于$\alpha_t \le 1/(\nu + \sqrt{(1-\beta_2)}\|\mathbf{g}_t\|_2^2)$，我们可以得到 $\alpha_t\|\mathbf{g}_t\|_2^2 \le \|\mathbf{g}_t\|_2/\sqrt{1-\beta_2}$，从而避免 ν 占据主导地位。这种改进并非没有代价：必须仔细分析额外的校正项，而且它会使最优收敛率中出现一个边际型条件。

定理 6. 在假设 *1 – 4*成立的条件下，运行算法 *2*，使用 $\beta = \beta_1$ 以及式 (5) 中的折扣 *FTRL* 更新，作用于 $\mathcal{D} = \{\Delta :\ \|\Delta\|_2 \le D\}$，使用 $\ell_t(\Delta) = \langle \mathbf{g}_t,\ \Delta\rangle$ 以及 η_t 式 (7)。这意味着裁剪后的式（(8)）中的 *Adam* 更新规则。任取 $\rho \in [0,1)$ 以及 $\nu \in (0, G+\sigma]$。调整 $1 - \left(\frac{\epsilon\sqrt{1-\rho^2}}{64\ G+\sigma}()\right) \le$ 2

$\beta_1 < 1$。定义间隔 $m = \frac{1-\rho}{2}(1-\beta_1^2)$，并选择 β_2 使得 $\beta_2 \in [\beta_1^2 + m,\ 1-m]$。设定 $D = \frac{(1-\beta_1)\sqrt{\epsilon}}{\sqrt{48c}}$，$\gamma = \frac{\beta_1 D}{\sqrt{1-\beta_1}}$。如果 T 满足

$$T \geq \max\left\{\frac{1}{1-\beta_1}\cdot\max\left\{\frac{32F^*\sqrt{c}}{\epsilon^{3/2}},\ \frac{16(G+\sigma)}{\epsilon}\right\}, \frac{32G}{\epsilon\sqrt{1-\beta_1}\sqrt{1-\rho^2}}\cdot\ln\left(1+\frac{G}{\nu}\right), \frac{\ln 2}{1-\beta_2},\right\},$$

那么，算法 $\bar{\mathbf{x}}_T$ 的输出满足 $\mathbb{E}[\|\nabla F(\bar{\mathbf{x}}_T)\|_c] \leq \epsilon$。

证明见附录 D.7。定理 6 表明，我们要求 β_2 与 β_1^2 和 1 都保持至少 m 的间隔。具体而言，条件 $\beta_2 \in [\beta_1^2 + m, 1-m]$ 意味着 $\beta_2 > \beta_1^2$，且其中心位于 $(1+\beta_1^2)/2$。参数 ρ 控制有效范围的宽度：较小的 ρ 会迫使 β_2 更接近 $(1+\beta_1^2)/2$，而 $\rho = 0$ 给出了最紧但限制也最严格的选择 $\beta_2 = (1+\beta_1^2)/2$。对于 β_2 的其他取值，定理通过 ρ 量化其影响，明确展现了较宽的可行范围与较松的界之间的权衡。在附录 D.2 中，我们列出 (β_1, β_2) 的常见取值，以说明不同设置如何影响 ρ。

定理 6 的收敛速率为

$$\mathcal{O}\left(\max\left\{\frac{(G+\sigma)^2F^*c^{1/2}}{(1-\rho)\epsilon^{7/2}}, \frac{(G+\sigma)^3}{(1-\rho)\epsilon^3}, \frac{G^2}{(1-\rho)\epsilon^2}\cdot\ln\left(\frac{G}{\nu}\right)\right\}\right),$$

其中，差距 $\beta_2 - \beta_1^2$ 可以通过最终速率中 ρ 的选择来量化。将 $1-\rho$ 视为常数，只要满足 定理 6 就能达到先前的最优收敛速率 $G/\nu \leq \mathcal{O}(\exp(G/\epsilon))$，这是一个温和的条件。这类结果是遵循无尺度在线学习的思路，通过精细分析得到的 [Orabona 和 P´等， 2018]，其中我们在引理 21中总结了这一适用于一般时变学习率的跟随正则化领导者技术。关键在于仔细控制偏差 $|\eta_t - \eta_{t-1}| \cdot \|\sum_{s=1}^{t-1}\beta_1^{t-1-s}\mathbf{g}_s\|_2$的速率，从而在最终结果中得到不被支配的速率。

4.4 无裁剪 Adam 的收敛条件

对于裁剪 Adam，使用裁剪操作有助于推导关于 (β_1, β_2) 的更精细收敛条件，即 $\beta_2 \geq \beta_1^4$。虽然裁剪 Adam 在实践中并非总是首选，我们进一步分析无裁剪变体。

在指数化 O2NC 框架中，可以通过让在线学习器最小化复合损失来处理无裁剪情形 [Zhang 等人， 2024]。核心思路是，O2NC 分析得出$\mathbb{E}[\|\nabla F(\bar{\mathbf{x}}_T)\|_c] \lesssim \mathbb{E}[\sum_{t=1}^T\langle\mathbf{g}_t, \Delta_t - \mathbf{u}_t\rangle] + c\mu\mathbb{E}[\sum_{t=1}^T\|\Delta_t\|_2^2]$。存在裁剪时，$\|\Delta_t\|_2 \leq D$，第二项可立即得到界。若没有裁剪，Zhang 等人 [2024] 提出通过标准的测度变换技术，将这一二次项吸收到遗憾中 [Foster 等人，2020，Chen 等人， 2021]: $\mathbb{E}[\sum_{t=1}^T\langle\mathbf{g}_t, \Delta_t - \mathbf{u}_t\rangle] + c\mu\mathbb{E}[\sum_{t=1}^T\|\Delta_t\|_2^2] = \mathbb{E}[\text{D-Reg}_T'(\mathbf{u}_{1:T})] + c\mu\mathbb{E}[\sum_{t=1}^T\|\mathbf{u}_t\|_2^2]$，其中$\text{D-Reg}_T'$由复合损失 $\langle\mathbf{g}_t, \Delta\rangle + \mu\|\Delta\|_2^2$定义。由于 O2NC 分析中的比较器选取方式满足 $\|\mathbf{u}_t\|_2 = D$，因此项 $c\mu\mathbb{E}[\sum_{t=1}^T\|\mathbf{u}_t\|_2^2]$ 有界于 $\mathcal{O}\left(c\mu TD^2\right)$。

受这一观点的启发，我们将算法 2 中的替代损失设为 $\ell_t(\Delta) = \langle\mathbf{g}_t, \Delta\rangle + \frac{\mu}{2}\|\Delta\|_2^2$其中 $\mu > 0$待调节，并取 $\mathcal{D} = \mathbb{R}^d$。将这一选择代入式（5），并将式（7）中的 η_t代入，引理 4给出了无裁剪 Adam 的闭式更新:

$$\Delta_{t+1} = -\frac{\gamma(1-\beta_1)\sum_{s=1}^t\beta_1^{t-s}\mathbf{g}_s}{\nu + \gamma\mu(1-\beta_1^t) + \sqrt{(1-\beta_2)\sum_{s=1}^t\beta_2^{t-s}\|\mathbf{g}_s\|_2^2}}. \tag{10}$$

所得算法是原始 Adam 的一种近似，其分母中增加了阻尼项 $\gamma\mu(1-\beta_1^t)$。随着 t 增大， $\gamma\mu(1-\beta_1^t)$会迅速达到 $\mathcal{O}(\gamma\mu)$。在我们的调节下， $\gamma\mu = \mathcal{O}(G+\sigma)$，因此在最坏情况下，该项与二阶矩项处于同一数量级。

The 无裁剪 Adam 满足以下定理，证明见附录D。 8.

定理 7。 在假设 *1 – 4*下，运行算法 *2*，采用 $\beta = \beta_1$ 以及式 (5) 中的折扣 *FTRL* 更新规则，应用于 $\mathcal{D} = \mathbb{R}^d$，使用 $\ell_t(\Delta) = \langle\mathbf{g}_t, \Delta\rangle + \frac{\mu}{2}\|\Delta\|_2^2$和 $\bar{\eta}_t$ 式 (7)中的内容。这意味着无裁剪 *Adam*

式 (10)中的更新规则。设$1-\left(16\ \frac{\epsilon}{(G+\sigma)}\right)^2 \le \beta_1 < 1$，$D=\frac{(1-\beta_1)}{\sqrt{96c}}\sqrt{\epsilon}$，$\gamma=\frac{\beta_1 D}{\sqrt{1-\beta_1}}$，$0<\nu\le G+\sigma$，以及最大值$\{1-\frac{\nu}{G+\sigma},\beta_1^2\} \le \beta_2 < 1$。如果 T满足

$$T \ge \max\left\{\frac{1}{1-\beta_1}\max\left\{\frac{16F^*\sqrt{96c}}{\epsilon^{3/2}},\ \frac{48(G+\sigma)}{\epsilon}\right\},\frac{\ln 2}{1-\beta_2}\right\},$$

那么，算法 $\bar{\mathbf{x}}$ *2* 的输出 满足 $\mathbb{E}[\|\nabla F(\bar{\mathbf{x}})\|_c] \le \epsilon$。

定理 7 的证明采用了与定理 5相同的主要步骤。关键区别在于式（9）中的第二项：在不进行裁剪的情况下，$\|\Delta_t\|_2^2$ 无法按照所需的阶数进行界定，因此我们要求 $\frac{\beta_1}{2\alpha_{t-1}} - \frac{1}{2\alpha_t} \le 0$ 使该项保持非正，这就导出了条件 $\beta_2 \ge \beta_1^2$。

此外，得益于模块化 D2D 分析的灵活性，我们还为无裁剪 Adam 提供了间隔型收敛条件。证明详见附录 D.9。

定理 *8*。在假设 *1–4* 下，使用 $\ell_t(\Delta) = \langle g_t, \Delta\rangle + \frac{\mu}{2}\|\Delta\|_2^2$和 η_t，在 $\mathcal{D}=\mathbb{R}^d$上运行算法 *2*，采用 $\beta=\beta_1$ 以及式（*5*）中的折扣 *FTRL* 更新，并令其满足式（*7*）。这意味着式（*10*）中的无裁剪 *Adam* 更新规则。任取 $\rho \in [0,1)$和任意 $\nu \in (0, G+\sigma]$。选择 β_1 使得$1-\left(\frac{\epsilon\sqrt{1-\rho^2}}{64\ G+\sigma}\right) \le \beta_1 < 1$。定义间隔 $m=\frac{1-\rho}{2}$，$1-\beta_1^2$，并选择 $\beta_2 \in [\beta_1^2+m,\ 1-m]$。设定$D=\frac{(1-\beta_1)}{\sqrt{96c}}\sqrt{\epsilon}$,$\gamma=\frac{\beta_1 D}{\sqrt{1-\beta_1}}$,$\mu=\frac{24cD}{(1-\beta_1)^2}$。若 T满足

$$T \ge \max\left\{\frac{1}{1-\beta_1}\max\left\{\frac{32F^*\sqrt{96c}}{\epsilon^{3/2}},\ \frac{48(G+\sigma)}{\epsilon}\right\},\ \frac{\ln 2}{1-\beta_2},\ \frac{32G}{\epsilon\sqrt{1-\beta_1}\sqrt{1-\rho^2}}\cdot\ln\left(1+\frac{\gamma\mu+G}{\nu}\right)\right\},$$

那么输出 $\bar{\mathbf{x}}$ 算法 *2* 满足 $\mathbb{E}\left[\|\nabla F(\bar{})\|_c\right] \le \epsilon$。

4.5 更多推论

由于 (c,ϵ)-平稳性的普适性， [Zhang and Cutkosky， 2024], 在附录 D.3中展示了我们的定理如何应用于随机、非凸且光滑的设定。通过适当设置 c 光滑性常数，我们的结果可以推导出非凸光滑优化中一阶驻点的最优收敛速率。

此外，我们的结果可以推导出以 ℓ_1-范数衡量的收敛速率，利用其内在的优势坐标稀疏性，使用Cutkosky 等人在附录 G 中的转换 [2023]。

5 结论

我们通过模块化 D2D 归约研究动态遗憾。我们简化了线性回归的分析，并为逻辑回归提供了新的保证。除了在线凸优化之外，将这一归约与指数化 O2NC 框架相结合，可在随机、非凸且非光滑的设定下，为裁剪版和无裁剪 Adam 带来最优收敛速率，并允许更灵活的参数选择（β_1，β_2）。未来的研究方向包括将该方法扩展到更复杂的跟随正则化领导者变体 [J´ez´equel 等人，2025] 并在放宽假设的情况下获得 Adam 的保证。

参考文献

Jacob Abernethy、Elad Hazan 和 Alexander Rakhlin。暗中竞争：一种用于多臂赌博机线性优化的高效算法。载于第 *21* 届年度学习理论会议（*COLT*）论文集，第 263–274 页，2008 年。

Naman Agarwal、Satyen Kale 和 Julian Zimmert。在线多类别逻辑回归的高效方法。见第 *33* 届国际算法学习理论会议（*ALT*）论文集，第 3–33 页，2022。

Kwangjun Ahn 和 Ashok Cutkosky。带模型指数移动平均的 Adam 对非凸优化有效。见 神经信息处理系统进展 *37*（*NeurIPS*），第 94909–94933 页，2024年。Kwangjun Ahn、Zhiyu Zhang、Yunbum Kook 和 Yan Dai。通过更新的在线学习理解 Adam 优化器：Adam 实为伪装的跟随正则化领导者。见第 *41* 届国际机器学习会议（*ICML*）论文集，第 619–640 页，2024年。Ahmet Alacaoglu、Yura Malitsky、Panayotis Mertikopoulos 和 Volkan Cevher。Adam类算法的一种新遗憾分析。见 第 *37* 届国际机器学习会议（*ICML*）论文集，第 202–210 页，2020。Yossi Arjevani、Yair Carmon、John C Duchi、Dylan J Foster、Nathan Srebro 和 Blake Woodworth。非凸随机优化的下界。*Mathematical Programming*，199(1):165–214，2023。Peter Auer、Nicol`o Cesa-Bianchi 和 Claudio Gentile。自适应且自信的在线学习算法。*Journal of Computer and System Sciences*，64(1):48–75，2002。Katy S Azoury 和 Manfred K Warmuth。指数族分布在线密度估计的相对损失界。*Machine Learning*，43:211–246，2001。Dheeraj Baby 和 Yu-Xiang Wang。指数凹在线学习中的最优动态遗憾。见 第 *34* 届学习理论会议（*COLT*）论文集，第 359–409 页，2021。Dheeraj Baby 和 Yu-Xiang Wang。适当在线学习中强凸损失及其他情形的最优动态遗憾。见第 *25* 届国际人工智能与统计会议（*AISTATS*）论文集，第 1805–1845 页，2022。Yong Bai、Yu-Jie Zhang、Peng Zhao、Masashi Sugiyama 和 Zhi-Hua Zhou。以可证明保证适应在线标签偏移。见 神经信息处理系统进展 *35*（*NeurIPS*），第 29960–29974 页，2022。Amir Beck 和 Marc Teboulle。凸优化的镜像下降与非线性投影次梯度方法。*Operations Research Letters*，31(3):167–175，2003。Nicol`o Cesa-Bianchi 和 G´abor Lugosi。*Prediction, Learning, and Games*。剑桥大学出版社，2006。Nicol`o Cesa-Bianchi、Pierre Gaillard、G´abor Lugosi 和 Gilles Stoltz。镜像下降遇上固定份额（且毫无遗憾）。见 神经信息处理系统进展 *25*（*NIPS*），第 989–997 页，2012。Liyu Chen、Haipeng Luo 和 Chen-Yu Wei。化不可能的调参为可能：一种新的专家算法及其应用。见 第 *34* 届学习理论会议（*COLT*）论文集，第 1216–1259 页，2021。Alexey Chernov 和 Fedor Zhdanov。折扣损失下基于专家建议的预测。见 第 *21* 届国际算法学习理论会议（*ALT*）论文集，第 255–269 页，2010。Chao-Kai Chiang、Tianbao Yang、Chia-Jung Lee、Mehrdad Mahdavi、Chi-Jen Lu、Rong Jin 和 Shenghuo Zhu。具有渐进变化的在线优化。见 第 *25* 届学习理论会议（*COLT*）论文集，第 6.1–6.20 页，2012。Ashok Cutkosky。无参数、动态且强自适应在线学习。见 第 *37* 届国际机器学习会议（*ICML*）论文集，第 2250–2259 页，2020。Ashok Cutkosky 和 Francesco Orabona。非凸随机梯度下降中基于动量的方差缩减。见神经信息处理系统进展 *32*（*NeurIPS*），第 15210–15219 页，2019。Ashok Cutkosky、Harsh Mehta 和 Francesco Orabona。通过在线到非凸转换实现的最优随机非光滑非凸优化。见第 *40* 届国际机器学习会议（*ICML*）论文集，第 6643–6670 页，2023。

Amit Daniely、Alon Gonen 和 Shai Shalev-Shwartz。强自适应在线学习。载于第 *32* 届国际机器学习会议（*ICML*）论文集，第 1405–1411 页，2015 年。Damek Davis、Dmitriy Drusvyatskiy、Yin Tat Lee、Swati Padmanabhan 和叶光豪。用于高维和低维 Lipschitz函数的、具有复杂度保证的梯度采样方法。载于神经信息处理系统进展第 *35* 卷（*NeurIPS*），第 6692–6703 页，2022 年。Alexandre D´efossez、L´eon Bottou、Francis R. Bach 和 Nicolas Usunier。Adam 和 Adagrad 的一个简单收敛性证明。机器学习研究汇刊，2022 年。Dylan J. Foster、Satyen Kale、Haipeng Luo、Mehryar Mohri 和 Karthik Sridharan。逻辑回归：不拘泥于正确设定的重要性。载于第 *31* 届学习理论会议（*COLT*）论文集，第 167–208 页，2018 年。Dylan J. Foster、Claudio Gentile、Mehryar Mohri 和 Julian Zimmert。上下文赌博机中的模型误设适应。载于神经信息处理系统进展第 *33* 卷（*NeurIPS*），第 11478–11489 页，2020 年。Yoav Freund 和 Daniel Hsu。一种新的对冲算法及其在推断潜在随机变量中的应用。*ArXiv* 预印本，arXiv:0806.4802，2008 年。Pierre Gaillard、Gilles Stoltz 和 Tim van Erven。基于超额损失的二阶界。载于第 *27* 届学习理论会议（*COLT*）论文集，第 176–196 页，2014 年。Allen A. Goldstein。利普希茨连续函数的优化。*Mathematical Programming*，13(1)：14–22，1977。郭志帅、徐毅、印卧涛、Rong Jin 和 Tianbao Yang。Adam族算法的一种新型收敛性分析。*ArXiv* 预印本，arXiv:2112.03459，2021 年。Andr´as Gy¨orgy 和 Csaba Szepesv´ari。漂移遗憾、镜像下降与矩阵。载于第 *33* 届国际机器学习会议（*ICML*）论文集，第 2943–2951 页，2016 年。Eric C. Hall 和 Rebecca Willett。在线凸规划中的动态模型与跟踪遗憾。载于第 *30* 届国际机器学习会议（*ICML*）论文集，第 579–587 页，2013 年。Elad Hazan。《在线凸优化导论》。优化基础与趋势，2 (3-4)：157–325，2016 年。Elad Hazan 和 C. Seshadhri。在线决策问题的自适应算法。计算复杂性电子讨论会（*ECCC*），14(088)，2007 年。Elad Hazan、Amit Agarwal 和 Satyen Kale。在线凸优化的对数遗憾算法。*Machine Learning*，69(2-3)：169–192，2007 年。Elad Hazan、Tomer Koren 和 Kfir Y. Levy。逻辑回归：随机优化和在线优化的紧界。载于第 *27* 届学习理论会议（*COLT*）论文集，第 197–209 页，2014 年。Mark Herbster 和 Manfred K. Warmuth。跟踪最佳专家。*Machine Learning*，32(2)：151–178，1998 年。胡子豪、Guanghui Wang 和 Jacob D. Abernethy。测地度量空间上的动态遗憾最小化。载于第 *36* 届年度学习理论会议（*COLT*），第 195 卷，第 4336–4383 页，2023 年。Andrew Jacobsen 和 Ashok Cutkosky。无参数镜像下降。载于第 *35* 届学习理论会议（*COLT*）论文集，第 4160–4211 页，2022 年。Andrew Jacobsen 和 Ashok Cutkosky。通过折扣处理动态环境中的在线线性回归。载于第 *41* 届国际机器学习会议（*ICML*）论文集，第 21083–21120 页，2024 年。

Andrew Jacobsen 和 Francesco Orabona。静态遗憾最小化与动态遗憾最小化之间的等价性。载于神经信息处理系统进展第 *37* 卷（*NeurIPS*），第 98835–98870 页，2024 年。Andrew Jacobsen、Alessandro Rudi、Francesco Orabona 和 Nicol`o Cesa-Bianchi。动态遗憾可归约为核化静态遗憾。载于神经信息处理系统进展第 *38* 卷（*NeurIPS*），即将出版，2025 年。R´emi J´ez´equel、Pierre Gaillard 和 Alessandro Rudi。在线逻辑回归的高效非适定学习。载于第 *33* 届学习理论会议（*COLT*）论文集，第 2085–2108 页，2020 年。R´emi J´ez´equel、Dmitrii Ostrovskii 和 Pierre Gaillard。高效且近乎最优的在线投资组合选择。运筹学数学，2025 年。Michael I. Jordan、Tianyi Lin 和 Manolis Zampetakis。确定性非光滑非凸优化的复杂度问题。*ArXiv* 预印本，arXiv:2209.12463，2022 年。Kwang-Sung Jun、Francesco Orabona、Stephen Wright 和 Rebecca Willett。利用抛硬币投注适应变化环境的在线学习。电子统计学杂志，11(2)：5282–5310，2017 年。

Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。载于第*3*届国际学习表征会议（*ICLR*）论文集，2015年。

Walid Krichene、Benjamin Drigh`es 和 Alexandre Bayen。自私路由中无悔学习的收敛性。载于第*31*届国际机器学习会议（*ICML*）论文集，第163–171页，2014。

Tor Lattimore 和 Csaba Szepesv´ari。多臂老虎机算法。剑桥大学出版社，2020。

Haochuan Li、Ali Jadbabaie 和 Alexander Rakhlin。宽松假设下 Adam 的收敛性。载于神经信息处理系统进展第*36*卷（*NeurIPS*），第52166–52196页，2023。

Long-Fei Li、Peng Zhao 和 Zhi-Hua Zhou。对抗性线性混合马尔可夫决策过程的近最优动态遗憾。载于神经信息处理系统进展第*37*卷（*NeurIPS*），第55858–55883页，2024。

Xinghan Li、Haodong Wen 和 Kaifeng Lyu。Adam 可降低一种独特形式的尖锐度：极小化器流形附近的理论洞见。发表于神经信息处理系统进展第 *38* 卷（*NeurIPS*），即将刊载，2025年。

Haipeng Luo 和 Robert E. Schapire。无需参数即可实现一切：AdaNormalHedge。发表于第 *28* 届年度计算学习理论会议（*COLT*）论文集，第 1286–1304 页，2015年。

Liangchen Luo、Yuanhao Xiong、Yan Liu 和 XuSun。具有动态学习率界的自适应梯度方法。发表于第 *7* 届国际学习表征会议（*ICLR*）论文集，2019。

Jack JMayo、H´edi Hadiji 和 TimvanErven。适用于曲线损失的无尺度无约束在线学习。发表于第 *35* 届学习理论会议（*COLT*）论文集，第 4464–4497 页，2022。

Naram Mhaisen 和 George Iosifidis. 关于正则化领导者跟随法的动态遗憾：采用历史剪枝的乐观策略。载于第*42*届国际机器学习会议（*ICML*）论文集，第43990–44016页，2025年。Zakaria Mhammedi 和 Khashayar Gatmiry. 用于高效在线指数凹优化的拟牛顿步。载于第*36*届学习理论会议（*COLT*）论文集，第4473–4503页，2023年。Zakaria Mhammedi 和 Alexander Rakhlin. 用于投资组合选择的阻尼在线牛顿步。载于第*35*届学习理论会议（*COLT*）论文集，第5561–5595页，2022年。

Quan Nguyen. 如何在 Adam 中设置 (β_1, β_2)：在线学习视角。载于第*37*届国际算法学习理论会议（*ALT*）论文集，即将出版，2026年。Francesco Orabona.《在线学习现代导论》。*ArXiv*预印本，arXiv:1912.13213v8，2019年。Francesco Orabona 和 D´avid P´al. 无尺度在线学习。理论计算机科学，716:50–69，2018年。Antonio Orvieto 和 Robert M Gower. 探寻 Adam 的秘密诀窍。载于神经信息处理系统进展第*38*卷（*NeurIPS*），即将出版，2025年。Boris T. Polyak 和 Anatoli B. Juditsky. 通过平均加速随机逼近。*SIAM*控制与优化杂志，30(4):838–855，1992年。Yu-Yang Qian、Peng Zhao、Yu-Jie Zhang、Masashi Sugiyama 和 Zhi-Hua Zhou. 利用小波实现高效的非平稳在线学习及其在在线分布偏移适应中的应用。载于第*41*届国际机器学习会议（*ICML*）论文集，第41383–41415页，2024年。Alexander Rakhlin 和 Karthik Sridharan. 基于可预测序列的在线学习。载于第*26*届学习理论会议（*COLT*）论文集，第993–1019页，2013年。Sashank J. Reddi、Satyen Kale 和 Sanjiv Kumar. Adam 及其后续方法的收敛性。载于第*6*届国际学习表征会议（*ICLR*）论文集，2018年。David Ruppert. 从缓慢收敛的 Robbins–Monro 过程中进行高效估计。技术报告，康奈尔大学运筹学与工业工程系，1988年。Shai Shalev-Shwartz. 在线学习与在线凸优化。机器学习基础与趋势，4(2):107–194，2012年。Lai Tian、Kaiwen Zhou 和 Anthony Man-Cho So. 利普希茨函数近似平稳性概念的有限时间复杂度与实际计算。载于第*39*届国际机器学习会议（*ICML*）论文集，第21360–21379页，2022年。Tim van Erven、Peter D. Gr¨unwald、Mark D. Reid 和 Robert C. Williamson. 统计学习中的可混合性。载于第*26*届神经信息处理系统会议（*NeurIPS*）论文集，第25卷，第1691–1699页，2012年。Ramon van Handel. 高维概率，2016年。讲义，普林斯顿大学。Volodya Vovk. 竞争性在线统计。国际统计评论，69(2):213–248，2001年。Bohan Wang、Jingwen Fu、Huishuai Zhang、Nanning Zheng 和 Wei Chen. 缩小 Adam 迭代复杂度上界与下界之间的差距。载于神经信息处理系统进展第*36*卷（*NeurIPS*），第39006–39032页，2023年。Wenhao Yang、Sifan Yang 和 Lijun Zhang. 折扣在线凸优化：在连续区间上的一致遗憾。载于第*14*届国际学习表征会议（*ICLR*）论文集，2026年。Jianjun Yuan 和 Andrew G. Lamperski. 在线最小二乘及其扩展中静态遗憾与动态遗憾之间的权衡。载于第*34*届*AAAI*人工智能会议论文集，第6712–6719页，2020年。Jingzhao Zhang、Hongzhou Lin、Stefanie Jegelka、Suvrit Sra 和 Ali Jadbabaie. 求解非凸非光滑函数驻点的复杂度。载于第*37*届国际机器学习会议（*ICML*）论文集，第11173–11182页，2020年。Lijun Zhang、Shiyin Lu 和 Zhi-Hua Zhou. 动态环境中的自适应在线学习。载于神经信息处理系统进展第*31*卷（*NeurIPS*），第1330–1340页，2018年。

Mengxiao Zhang、Peng Zhao、Haipeng Luo 和 Zhi-Hua Zhou。时变零和博弈中的无悔学习。载于第*39*届国际机器学习会议（*ICML*）论文集，第26772–26808页，2022a。Qinzi Zhang 和 Ashok Cutkosky。非光滑非凸优化中的随机缩放与动量。载于第*41*届国际机器学习会议（*ICML*）论文集，第58780–58799页，2024年。Yu-Jie Zhang、Peng Zhao 和 Masashi Sugiyama。曲线损失下的非平稳在线学习：通过可混合性改进动态遗憾。载于第*42*届国际机器学习会议（*ICML*）论文集，第77044–77068页，2025。Yushun Zhang、Congliang Chen、Naichen Shi、Ruoyu Sun 和 Zhi-Quan Luo。无需修改更新规则，Adam 也能收敛。载于神经信息处理系统进展第*35*卷（*NeurIPS*），第28386–28399页，2022b。Zhiyu Zhang、David Bombara 和 Heng Yang。折扣自适应在线学习：迈向更好的正则化。载于第*41*届国际机器学习会议（*ICML*）论文集，第58631–58661页，2024年。Peng Zhao 和 Lijun Zhang。强凸光滑函数动态遗憾的改进分析。载于第*3*届动力学与控制学习会议（*L4DC*）论文集，第48–59页，2021。Peng Zhao、Guanghui Wang、Lijun Zhang 和 Zhi-Hua Zhou。非平稳环境中的老虎机凸优化。载于第*23*届国际人工智能与统计会议（*AISTATS*）论文集，第1508–1518页，2020。Peng Zhao、Yu-Jie Zhang、Lijun Zhang 和 Zhi-Hua Zhou。自适应性与非平稳性：在线凸优化中问题依赖的动态遗憾。机器学习研究期刊，25 (98):1–52，2024年。Peng Zhao、Yan-Feng Xie、Lijun Zhang 和 Zhi-Hua Zhou。非平稳在线学习的高效方法。机器学习研究期刊，26(208):1–66，2025。Zhiming Zhou、Qingru Zhang、Guansong Lu、Hongwei Wang、Weinan Zhang 和 Yong Yu。Adashift：自适应学习率方法的去相关与收敛。载于第*7*届国际表征学习会议（*ICLR*）论文集，2019。Julian Zimmert、Naman Agarwal 和 Satyen Kale。推动在线学习投资组合与量子态的效率—遗憾帕累托前沿。载于第*35*届学习理论会议（*COLT*）论文集，第182–226页，2022。Martin Zinkevich。在线凸规划与广义无穷小梯度上升。载于第*20*届国际机器学习会议（*ICML*）论文集，第928–936页，2003。Fangyu Zou、Li Shen、Zequn Jie、Weizhong Zhang 和 Wei Liu。Adam 和 RMSProp 收敛的一个充分条件。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第11127–11135页，2019。

A 相关工作

非平稳在线学习。非平稳性是在线凸优化（OCO）固有的，其中损失序列可能随时间漂移。对这种漂移建模的一种常见方法，是将基准从静态遗憾加以强化，从而引入区间遗憾 [Hazan 和 Seshadhri，2007，Daniely 等人，2015，Jun 等人， 2017]，切换遗憾 [Herbster 和 Warmuth，1998，Gy¨orgy 和 Szepesv´ari， 2016]，以及动态遗憾 [Zinkevich，2003，Cesa-Bianchi 等人，2012，Hall 和 Willett，2013]等概念。本文聚焦于动态遗憾，近年来，在各种假设和算法族下，这一问题已得到广泛研究 [Zhang 等人，2018，Zhao 等人，2020，Cutkosky，2020，Zhao 和 Zhang，2021，Zhang 等人，2022a，Baby and Wang，2021，Bai 等人，2022，Baby and Wang，2022，Jacobsen 和 Cutkosky，2022，Hu 等人，2023，Qian 等人，2024，Li 等人，2024，Zhao 等人，2024, 2025]。

对于曲线损失，获得动态遗憾保证会变得更为微妙：即便已有路径长度等漂移度量，要在高效调节学习率的同时保持对（T， P_T）的尖锐依赖，仍可能颇具挑战 [Zhang 等人， 2025]。一种通用方法是将区间遗憾保证转化为动态遗憾界 [BabyandWang， 2021, 2022]。这些分析通常依赖于比较器的 Karush–Kuhn–Tucker 条件，技术上可能较为繁琐，从而在推广到更广泛的设定时可能限制灵活性。Zhang 等人 [2025]利用可混合性 [Vovk，2001，van Erven 等人， 2012]以及指数权重方法，以更高的计算成本为代价获得尖锐的权衡。 Yuan 和 Lamperski [2020]使用折扣算法分析并给出动态遗憾界，但其方法仅限于有界域。与我们的线性回归应用最为接近的是，Jacobsen 和 Cutkosky [2024] 分析了折扣 VAW，并借助基于 Bregman 散度的论证，在无约束设定下给出了匹配的下界。相对于这些工作，我们的贡献是一种模块化的从折扣遗憾到动态遗憾的约化，它给出更直接的分析并能恢复最优保证，同时也为无界决策域上的曲线损失提供新的保证。

折扣在线学习与 D2D 归约。折扣遗憾已成为强调近期损失并适应漂移环境的重要方式之一 [Freund 和 Hsu, 2008, Chernov 和 Zhdanov, 2010, Cesa-Bianchi 等人 , 2012, Krichene 等人 , 2014]。在在线凸优化中，Zhang 等人 [2024]针对给定的折扣因子提出了梯度自适应保证，后续工作则研究折扣因子未知时的自适应问题 [Yang 等人，2026]。最近，Nguyen [2026] 通过对在线凸优化中一种由简化版 Adam 变体推导出的跟随正则化领导者方法进行折扣遗憾分析，考察了 (β_1, β_2) 的选择。与这项工作相比，我们着重于通过 O2NC 进行随机、非凸和非光滑优化，并分析采用标准指数移动平均动量形式的 Adam，在灵活的 (β_1, β_2) 机制下推导收敛保证。另一个相关方向是通过剪枝历史信息，让跟随正则化领导者更能及时响应近期变化。具体而言，Mhaisen 和 Iosifidis [2025]保留了跟随正则化领导者的核心框架，同时在输入梯度中加入校正项，这可以视为剪除历史信息的影响，并由此实现最优动态遗憾保证。与他们的工作相比，我们的分析适用于曲线损失，即使在无约束设定下也成立。作为一项局限，目前尚不清楚如何利用我们的方法为一般凸损失恢复最优动态遗憾界。

在折扣遗憾与动态遗憾之间架起桥梁，Ahn 等人 [2024] 明确建立了这两种度量之间的联系。另有研究采用不同方法，通过在扩展空间中建立动态遗憾与静态遗憾之间的约化 [Jacobsen 和 Orabona, 2024年，Jacobsen 等人，2025]。基于从折扣遗憾到动态遗憾的约化，我们提出了一种模块化分析方法，将关键项明确保留，并支持在多种场景中复用分析，包括凸问题，以及通过 O2NC 转换处理非凸优化。

Adam 的收敛性。Kingma 和 Ba [2015] 提出了 Adam，并给出了其在凸设定下的保证；此后，Adam 在凸且光滑的目标以及非凸目标上的收敛条件得到了广泛研究 [Reddi 等人，2018，Zhou 等人，2019，Zou 等人，2019，Alacaoglu 等人，2020，Guo 等人，2021，Zhang 等人，2022b，D´efossez 等人，2022，Li 等人，2023，Wang等人， 2023]。最近，Li 等人 [2025]从更具几何性的视角研究了 Adam。我们的工作则通过 O2NC 框架研究 Adam，以推导其在非光滑设定下的保证。Ahn 和 Cutkosky [2024] 分析了随机、非凸、非光滑设定下针对指数移动平均（EMA）迭代的裁剪 Adam。

他们的更新采用了略有不同的动量归一化方式（例如，$m_t = \beta m_{t-1} + \mathbf{g}_t$ 而不是$m_t = \beta m_{t-1} + (1-\beta)\mathbf{g}_t$），其理论分析聚焦于受限的参数选择 $\beta_2 = \beta_1^2$。在本工作中，我们在标准动量设计和更灵活的参数选择下，研究随机、非凸、非光滑情形中的 Adam。我们的结果涵盖裁剪版和无裁剪版 Adam 优化器，给出了收敛到 (c, ϵ)-驻点的最优收敛速率；通过恰当选择c，还可恢复在光滑和二阶光滑设定下收敛到一阶驻点的最优速率。

O2NC 转换。 受非光滑目标函数中易于处理的平稳性概念启发，[Zhang 等人，2020]，非光滑非凸优化受到越来越多的关注，[Davis 等人，2022，Tian 等人，2022，Jordan 等人，2022]。由 Cutkosky 等人 [2023] 提出的开创性工作构建了在线到非凸（O2NC）框架，将随机非凸非光滑优化转化为在线学习问题，并在标准假设下实现了最优速率。Zhang and Cutkosky[2024] 通过引入（c，ϵ）平稳性，并允许通过最小化复合损失来处理无界方向，从而扩展了 O2NC。我们以这一指数化 O2NC 框架为基础，通过模块化 D2D 分析，为 Adam 的指数移动平均迭代序列推导出收敛保证。

B 第 2 节证明

本部分将给出第 2 节中所述理论结果的证明。

B.1 引理1的证明

为完整起见，我们给出引理 1的证明，该证明改编自 Ahn 等人 [2024]。

证明。 回顾截至时间 t 相对于固定比较器 $\mathbf{u} \in \mathcal{X}$的折扣遗憾:

$$\textsc{Reg}_{t;\beta}(\mathbf{u}) = \sum_{s=1}^{t} \beta^{t-s}\big(f_s(\mathbf{x}_s) - f_s(\mathbf{u})\big), \qquad \text{with the convention } \textsc{Reg}_{0;\beta}(\mathbf{u}) = 0.$$

一个关键恒等式是，对于每个 $t \geq 1$ 以及任意 $\mathbf{u} \in \mathcal{X}$,

$$\begin{aligned} f_t(\mathbf{x}_t) - f_t(\mathbf{u}) &= \textsc{Reg}_{t;\beta}(\mathbf{u}) - \beta \textsc{Reg}_{t-1;\beta}(\mathbf{u}) \\ &= (1-\beta)\textsc{Reg}_{t;\beta}(\mathbf{u}) + \beta\big(\textsc{Reg}_{t;\beta}(\mathbf{u}) - \textsc{Reg}_{t-1;\beta}(\mathbf{u})\big). \end{aligned} \tag{11}$$

事实上，展开定义可得

$$\textsc{Reg}_{t;\beta}(\mathbf{u}) - \beta \textsc{Reg}_{t-1;\beta}(\mathbf{u}) = \sum_{s=1}^{t} \beta^{t-s}(f_s(\mathbf{x}_s) - f_s(\mathbf{u})) - \sum_{s=1}^{t-1} \beta^{t-s}(f_s(\mathbf{x}_s) - f_s(\mathbf{u})) = f_t(\mathbf{x}_t) - f_t(\mathbf{u}).$$

现在，对 11式求和，令 $t = 1, \ldots, T$，并取随时间变化的比较器 $\mathbf{u} = \mathbf{u}_t$:

$$\sum_{t=1}^{T}\big(f_t(\mathbf{x}_t) - f_t(\mathbf{u}_t)\big) = (1-\beta)\sum_{t=1}^{T}\textsc{Reg}_{t;\beta}(\mathbf{u}_t) + \beta\sum_{t=1}^{T}\big(\textsc{Reg}_{t;\beta}(\mathbf{u}_t) - \textsc{Reg}_{t-1;\beta}(\mathbf{u}_t)\big).$$

接下来，我们通过望远镜求和来改写第二个求和式。注意到

$$\sum_{t=1}^{T}\big(\textsc{Reg}_{t;\beta}(\mathbf{u}_t) - \textsc{Reg}_{t-1;\beta}(\mathbf{u}_t)\big) = \textsc{Reg}_{T;\beta}(\mathbf{u}_T) + \sum_{t=1}^{T-1}\big(\textsc{Reg}_{t;\beta}(\mathbf{u}_t) - \textsc{Reg}_{t;\beta}(\mathbf{u}_{t+1})\big),$$

其中我们使用 $\textsc{Reg}_{0;\beta}(\mathbf{u}_1) = 0$。将其代入前一个等式即可完成证明 . □

B.2 定理1的证明

证明。 根据引理 1，可得

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t) = (1-\beta)\sum_{t=1}^{T} \textsc{Reg}_{t;\beta}(\mathbf{u}_t) + \beta \sum_{t=1}^{T-1} \Big(\textsc{Reg}_{t;\beta}(\mathbf{u}_t) - \textsc{Reg}_{t;\beta}(\mathbf{u}_{t+1})\Big) + \beta \textsc{Reg}_{T;\beta}(\mathbf{u}_T).$$

对于任意 $t \in [T]$ 以及任意固定的 $\mathbf{u} \in \mathcal{X}$，将假设的缩放后的界乘以 β^t，得到 s

$$\textsc{Reg}_{t;\beta}(\mathbf{u}) = \sum_{s=1}^{t} \beta^{t-s}\big(f_s(\mathbf{x}_s) - f_s(\mathbf{u})\big) \le \beta^t \varphi_t(\mathbf{u}) + \sum_{s=1}^{t} \beta^t \Lambda_s.$$

将其应用于 $\mathbf{u}$ $=\mathbf{u}_t$ 并求和，可得

$$(1-\beta)\sum_{t=1}^{T} \textsc{Reg}_{t;\beta}(\mathbf{u}_t) \le (1-\beta)\sum_{t=1}^{T} \Big(\beta^t \varphi_t(\mathbf{u}_t) + \sum_{s=1}^{t} \beta^t \Lambda_s\Big)$$

$$= (1-\beta)\sum_{t=1}^{T} \beta^t \varphi_t(\mathbf{u}_t) + (1-\beta)\sum_{t=1}^{T}\sum_{s=1}^{t} \beta^t \Lambda_s = (1-\beta)\sum_{t=1}^{T} \beta^t \varphi_t(\mathbf{u}_t) + (1-\beta)\sum_{s=1}^{T} \Lambda_s \sum_{t=s}^{T} \beta^t$$

$$= (1-\beta)\sum_{t=1}^{T} \beta^t \varphi_t(\mathbf{u}_t) + (1-\beta)\sum_{s=1}^{T} \Lambda_s \cdot \frac{\beta^s(1-\beta^{T-s+1})}{1-\beta} = (1-\beta)\sum_{t=1}^{T} \beta^t \varphi_t(\mathbf{u}_t) + \sum_{s=1}^{T} \beta^s \Lambda_s - \sum_{s=1}^{T} \beta^{T+1} \Lambda_s,$$

其中，负项 $-\beta^{T+1}\sum_{s=1}^{T} \Lambda_s$ 留作后用。接下来，根据 $\textsc{Reg}_{t;\beta}(\cdot)$ 的定义，

$$\beta \sum_{t=1}^{T-1} \Big(\textsc{Reg}_{t;\beta}(\mathbf{u}_t) - \textsc{Reg}_{t;\beta}(\mathbf{u}_{t+1})\Big) = \beta \sum_{t=1}^{T-1}\sum_{s=1}^{t} \beta^{t-s}\Big(f_s(\mathbf{u}_{t+1}) - f_s(\mathbf{u}_t)\Big).$$

现在，我们将上述等式与项 $(1-\beta)\sum_{t=1}^{T} \beta^t \varphi_t(\mathbf{u}_t)$ 合并。利用 $(1-\beta)\beta^t = \beta^t - \beta^{t+1}$，可得

$$(1-\beta)\sum_{t=1}^{T} \beta^t \varphi_t(\mathbf{u}_t) = \beta\varphi_1(\mathbf{u}_1) - \beta^{T+1}\varphi_T(\mathbf{u}_T) + \sum_{t=1}^{T-1} \Big(\beta^{t+1}\varphi_{t+1}(\mathbf{u}_{t+1}) - \beta^{t+1}\varphi_t(\mathbf{u}_t)\Big)$$

$$= \beta\varphi_1(\mathbf{u}_1) - \beta^{T+1}\varphi_T(\mathbf{u}_T) + \beta \sum_{t=1}^{T-1} \Big(\beta^t \varphi_t(\mathbf{u}_{t+1}) - \beta^t \varphi_t(\mathbf{u}_t)\Big) + \beta \sum_{t=1}^{T-1} \beta^t \Big(\varphi_{t+1}(\mathbf{u}_{t+1}) - \varphi_t(\mathbf{u}_{t+1})\Big),$$

其中，负项 $-\beta^{T+1}\varphi_T(\mathbf{u}_T)$ 很有用，并且我们应用恒等式

$$\beta^{t+1}\varphi_{t+1}(\mathbf{u}_{t+1}) = \beta \cdot \beta^t \varphi_t(\mathbf{u}_{t+1}) + \beta \cdot \beta^t \Big(\varphi_{t+1}(\mathbf{u}_{t+1}) - \varphi_t(\mathbf{u}_{t+1})\Big).$$

定义 $F^{\beta,\varphi}_t(\mathbf{u}) = \beta^t \varphi_t(\mathbf{u}) + \sum_{s=1}^{t} \beta^{t-s} f_s(\mathbf{u})$。那么

$$\beta \sum_{t=1}^{T-1} \Big(\beta^t \varphi_t(\mathbf{u}_{t+1}) - \beta^t \varphi_t(\mathbf{u}_t)\Big) + \beta \sum_{t=1}^{T-1}\sum_{s=1}^{t} \beta^{t-s}\Big(f_s(\mathbf{u}_{t+1}) - f_s(\mathbf{u}_t)\Big) = \beta \sum_{t=1}^{T-1} \Big(F^{\beta,\varphi}_t(\mathbf{u}_{t+1}) - F^{\beta,\varphi}_t(\mathbf{u}_t)\Big).$$

最后，应用关于 $\textsc{Reg}_{T;\beta}(\mathbf{u}_T)$ 的界可得

$$\beta \textsc{Reg}_{T;\beta}(\mathbf{u}_T) \le \beta^{T+1}\varphi_T(\mathbf{u}_T) + \beta^{T+1}\sum_{s=1}^{T} \Lambda_s,$$

这可以由上文的负项 $-\beta^{T+1}\varphi_T(\mathbf{u}_T)$ 和 $-\beta^{T+1}\sum_{s=1}^{T} \Lambda_s$ 抵消。综上所述，证明完成：

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t) \le \beta\varphi_1(\mathbf{u}_1) + \sum_{t=1}^{T} \beta^t \Lambda_t + \beta \sum_{t=1}^{T-1} \Big(F^{\beta,\varphi}_t(\mathbf{u}_{t+1}) - F^{\beta,\varphi}_t(\mathbf{u}_t)\Big)$$
$$+ \beta \sum_{t=1}^{T-1} \beta^t \Big(\varphi_{t+1}(\mathbf{u}_{t+1}) - \varphi_t(\mathbf{u}_{t+1})\Big).$$

□

C 第 3节中省略的细节

我们预 第3节补充了省略的细节，包括若干省略的关键引理和讨论 .

C.1 在线线性回归的证明

我们首先给出定理 2 的正式表述。

定理 （定理 2 的扩展版本）。 对于无约束在线线性回归，折扣 *VAW* 预测器满足以下动态遗憾界：对于任意 $\mathbf{u}_1, \ldots, \mathbf{u}_T \in \mathbb{R}^d$,

$$\begin{aligned}
&\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t) \\
&\le \frac{\beta\lambda}{2}\|\mathbf{u}_1\|_2^2 + \frac{d}{2}\big(\max_{t\in[T]} y_t^2\big) \cdot \ln\left(1 + \frac{\sum_{t=1}^{T} \beta^{T-t}\|\mathbf{z}_t\|_2^2}{\lambda d}\right) + \beta \sum_{t=1}^{T-1} \left(F_t^{\beta,\varphi}(\mathbf{u}_{t+1}) - F_t^{\beta,\varphi}(\mathbf{u}_t)\right) + \frac{1-\beta}{\beta} \cdot \frac{d}{2} \sum_{t=1}^{T} y_t^2 \\
&\le \frac{\beta\lambda}{2}\|\mathbf{u}_1\|_2^2 + \frac{d}{2}\big(\max_{t\in[T]} y_t^2\big) \cdot \ln\left(1 + \frac{\sum_{t=1}^{T} \beta^{T-t}\|\mathbf{z}_t\|_2^2}{\lambda d}\right) + \frac{\gamma}{1-\gamma} P_T^{\gamma} + \frac{1-\beta}{\beta} \cdot \frac{d}{2} \sum_{t=1}^{T} y_t^2,
\end{aligned}$$

其中我们假设 $0 < \beta \le \gamma < 1$，且 P_T^{γ} 由公式 (4) 定义。此外，存在一种双层集成算法，可确保一个 $\mathcal{O}\big(d \log T + \sqrt{dTP_T^{\beta_\star}}\big)$ 动态遗憾界，其中

$$\beta_\star = \frac{\sqrt{\frac{d}{2}\sum_{t=1}^{T} y_t^2}}{\sqrt{\frac{d}{2}\sum_{t=1}^{T} y_t^2} + \sqrt{P_T^{\beta_\star}}}.$$

定理 2 的证明。 通过重新参数化为 $\widetilde{\mathbf{z}}_t = \beta^{-t/2}\mathbf{z}_t$ ，并 $\widetilde{y}_t = \beta^{-t/2} y_t$，使用样本（$\widetilde{\mathbf{z}}_t$，$\widetilde{y}_t$）向 VAW预测器提供数据，引理 2 保证对于任意 $\mathbf{u} \in \mathbb{R}^d$,

$$\begin{aligned}
\sum_{s=1}^{t} \beta^{-s}\big(f_s(\mathbf{x}_s) - f_s(\mathbf{u})\big) &= \sum_{s=1}^{t} \big(\ell(\mathbf{x}_s^\top \widetilde{\mathbf{z}}_s, \widetilde{y}_s) - \ell(\mathbf{u}^\top \widetilde{\mathbf{z}}_s, \widetilde{y}_s)\big) \\
&\le \frac{\lambda}{2}\|\mathbf{u}\|_2^2 + \sum_{s=1}^{t} \beta^{-2s} y_s^2 \mathbf{z}_s^\top \left(\lambda I + \sum_{\tau=1}^{s} \beta^{-\tau}\mathbf{z}_\tau \mathbf{z}_\tau^\top\right)^{-1} \mathbf{z}_s.
\end{aligned}$$

选择 $\varphi_t(\mathbf{u}) = \frac{\lambda}{2}\|\mathbf{u}\|_2^2$ 和 $\Lambda_t = \beta^{-2t} y_t^2 \mathbf{z}_t^\top \left(\lambda I + \sum_{s=1}^{t} \beta^{-s}\mathbf{z}_s\mathbf{z}_s^\top\right)^{-1} \mathbf{z}_t$，定理 1 给出:

$$\begin{aligned}
&\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{u}_t) \\
&\le \frac{\beta\lambda}{2}\|\mathbf{u}_1\|_2^2 + \sum_{t=1}^{T} \beta^{-t} y_t^2 \mathbf{z}_t^\top \left(\lambda I + \sum_{s=1}^{t} \beta^{-s}\mathbf{z}_s\mathbf{z}_s^\top\right)^{-1} \mathbf{z}_t + \beta \sum_{t=1}^{T-1} \left(F_t^{\beta,\varphi}(\mathbf{u}_{t+1}) - F_t^{\beta,\varphi}(\mathbf{u}_t)\right) \\
&= \frac{\beta\lambda}{2}\|\mathbf{u}_1\|_2^2 + \sum_{t=1}^{T} y_t^2 \mathbf{z}_t^\top \left(\lambda\beta^t I + \sum_{s=1}^{t} \beta^{t-s}\mathbf{z}_s\mathbf{z}_s^\top\right)^{-1} \mathbf{z}_t + \beta \sum_{t=1}^{T-1} \left(F_t^{\beta,\varphi}(\mathbf{u}_{t+1}) - F_t^{\beta,\varphi}(\mathbf{u}_t)\right), \qquad (12)
\end{aligned}$$

其中 $F_t^{\beta,\varphi}(\mathbf{u}) = \beta^t \varphi_t(\mathbf{u}) + \sum_{s=1}^{t} \beta^{t-s} f_s(\mathbf{u})$。

对于式（12）中的第二项，引理25 表明

$$\begin{aligned}
\sum_{t=1}^{T} y_t^2 \mathbf{z}_t^\top \left(\lambda\beta^t I + \sum_{s=1}^{t} \beta^{t-s}\mathbf{z}_s\mathbf{z}_s^\top\right)^{-1} \mathbf{z}_t &\le d \ln\left(\tfrac{1}{\beta}\right) \sum_{t=1}^{T} y_t^2 + \big(\max_t y_t^2\big) d \ln\left(1 + \tfrac{\sum_{t=1}^{T} \beta^{T-t}\|\mathbf{z}_t\|_2^2}{\lambda d}\right) \\
&\le d\frac{1-\beta}{\beta} \sum_{t=1}^{T} y_t^2 + \big(\max_t y_t^2\big) d \ln\left(1 + \tfrac{\sum_{t=1}^{T} \beta^{T-t}\|\mathbf{z}_t\|_2^2}{\lambda d}\right),
\end{aligned}$$

其中我们使用 $\ln(1/\beta) \le (1-\beta)/\beta$。

对于（12）中的第三项，令 $f_0(\mathbf{u}) = \frac{\lambda}{2}\|\mathbf{u}\|_2^2$，根据引理 18，我们有

$$\beta \sum_{t=1}^{T-1} \left(F_t^{\beta,\varphi}(\mathbf{u}_{t+1}) - F_t^{\beta,\varphi}(\mathbf{u}_t) \right) \le \frac{\gamma}{1-\gamma} P_T^{\gamma},$$

结合上述界，第一项断言的证明即告完成。至于最终的动态遗憾保证，注意到未调参的动态遗憾界恰好与 Jacobsen 和 Cutkosky [2024]中的定理 3.1 相符。将 Jacobsen 和 Cutkosky [2024] 中的定理 4.2 应用于未调参的保证，即可完成证明。

□

C.2 在线逻辑回归的证明

在本部分中，我们证明定理 3。我们先建立一个中间结果，然后应用归约来完成证明。我们的主要证明步骤遵循J´ez´equel 等人 [2020], 的思路，但我们从乐观跟随正则化领导者的视角来解读这一论证。

C.2.1 关键引理

引理 3。考虑在线逻辑回归，其中 $f_t(x) = \ell(x^\top z_t, y_t) = \ln(1 + \exp(-y_t x^\top z_t))$，其中$y_t \in \{+1, -1\}$ 且 $\|z_t\|_2 \le R$ 对所有 $t \in [T]$均成立。假设比较器满足 $\|u\|_2 \le B$。令 $\beta \in (0, 1)$，且 $\lambda > 0$。定义乐观函数 $h_t(x) = \ell(x^\top z_t, +1) + \ell(x^\top z_t, -1)$，梯度 $g_t = \nabla f_t(x_t)$，以及学习率

$$\eta_t = \frac{\exp(y_t \mathbf{x}_t^\top \mathbf{z}_t)}{1 + BR}. \tag{13}$$

定义替代损失 $\widehat{f}_t(\mathbf{x}) = f_t(\mathbf{x}_t) + \langle \mathbf{g}_t, \mathbf{x} - \mathbf{x}_t \rangle + \frac{\eta_t}{2}\langle \mathbf{g}_t, \mathbf{x} - \mathbf{x}_t \rangle^2$，并运行折扣 *AIOLI* 更新（如第 3.2 节所述）：

$$\mathbf{x}_t \in \arg\min_{\mathbf{x} \in \mathbb{R}^d} \left\{ \frac{\lambda \beta^t}{2}\|\mathbf{x}\|_2^2 + h_t(\mathbf{x}) + \sum_{s=1}^{t-1} \beta^{t-s} \widehat{f}_s(\mathbf{x}) \right\}.$$

Let

$$A_t = \lambda I + \sum_{s=1}^{t} \beta^{-s} \eta_s \mathbf{g}_s \mathbf{g}_s^\top, \qquad t \in [T]. \tag{14}$$

那么对于 任意 $t \in [T]$ 以及任意满足 $\|\mathbf{u}\|_2 \le B$ 的 $\mathbf{u} \in \mathbb{R}^d$，折扣 *AIOLI* 满足以下界：

$$\sum_{s=1}^{t} \beta^{-s} \big(f_s(\mathbf{x}_s) - f_s(\mathbf{u}) \big) \le \frac{\lambda}{2}\|\mathbf{u}\|_2^2 + (1 + BR) \sum_{s=1}^{t} \beta^{-2s} \eta_s \mathbf{g}_s^\top A_s^{-1} \mathbf{g}_s.$$

证明。 根据引理 17，对于任意满足 $\mathbf{u}$ $\|\mathbf{u}\|_2 \le B$ 的，替代损失满足

$$f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \le \widehat{f}_t(\mathbf{x}_t) - \widehat{f}_t(\mathbf{u}) = \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{u} \rangle - \frac{\eta_t}{2}\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{u} \rangle^2.$$

将上式乘以 β^{-t} 并对 $t \in [s]$ 求和，得到

$$\sum_{s=1}^{t} \beta^{-s} \big(f_s(\mathbf{x}_s) - f_s(\mathbf{u}) \big) \le \sum_{s=1}^{t} \beta^{-s} \langle \mathbf{g}_s, \mathbf{x}_s - \mathbf{u} \rangle - \sum_{s=1}^{t} \frac{\beta^{-s} \eta_s}{2} \langle \mathbf{g}_s, \mathbf{x}_s - \mathbf{u} \rangle^2. \tag{15}$$

接下来，将目标函数乘以 β^{-t}，把折扣 AIOLI 改写为重新缩放形式：

$$\mathbf{x}_t \in \arg\min_{\mathbf{x} \in \mathbb{R}^d} \left\{ \sum_{s=1}^{t-1} \langle \beta^{-s} \mathbf{g}_s, \mathbf{x} \rangle + \beta^{-t} h_t(\mathbf{x}) + \psi_t(\mathbf{x}) \right\}, \tag{16}$$

其中正则化项为

$$\psi_t(\mathbf{x}) = \frac{\lambda}{2}\|\mathbf{x}\|_2^2 + \sum_{s=1}^{t-1} \frac{\beta^{-s}\eta_s}{2}\langle \mathbf{g}_s, \mathbf{x} - \mathbf{x}_s\rangle^2.$$

因此，式（16）恰好是采用线性损失 $\langle \beta^{-t}\mathbf{g}_t, \mathbf{x}\rangle$ 和乐观函数 $\beta^{-t}h_t(\mathbf{x})$ 的乐观跟随正则化领导者（引理 22）。令 $\hat{\mathbf{x}}_t$ 表示不采用乐观策略时的决策：

$$\widehat{\mathbf{x}}_t \in \arg\min_{\mathbf{x}\in\mathbb{R}^d} \Big\{ F_t(\mathbf{x}) \Big\}, \qquad F_t(\mathbf{x}) = \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} \langle \beta^{-s}\mathbf{g}_s, \mathbf{x}\rangle.$$

将引理 22 应用于 $\{\langle \beta^{-t}\mathbf{g}_t, \mathbf{x}\rangle\}_{t=1}^{s}$，对于任意满足 $\mathbf{u}$ 使得 $\|\mathbf{u}\|_2 \le B$ 的

$$\sum_{s=1}^{t} \langle \beta^{-s}\mathbf{g}_s, \mathbf{x}_s - \mathbf{u}\rangle \le \psi_{t+1}(\mathbf{u}) - \min_{\mathbf{x}} \psi_1(\mathbf{x}) + \sum_{s=1}^{t} \Big(\langle \beta^{-s}\mathbf{g}_s, \mathbf{x}_s\rangle + F_s(\widehat{\mathbf{x}}_s) - F_{s+1}(\widehat{\mathbf{x}}_{s+1}) \Big), \tag{17}$$

这里我们使用 $F_{t+1}(\hat{\mathbf{x}}_{t+1}) \le F_{t+1}(\mathbf{u})$，以去掉引理 22中的最后一项。由于 $\psi_1(\mathbf{x}) = \frac{\lambda}{2}\|\mathbf{x}\|_2^2$，因此有 $\min_{\mathbf{x}} \psi_1(\mathbf{x}) = 0$。从式（17）两边减去 $\sum_{s=1}^{t} \frac{\beta^{-s}\eta_s}{2}\langle \mathbf{g}_s, \mathbf{x}_s - \mathbf{u}\rangle^2$，并与式（15）结合，得到

$$\sum_{s=1}^{t} \beta^{-s}\big(f_s(\mathbf{x}_s) - f_s(\mathbf{u})\big) \le \frac{\lambda}{2}\|\mathbf{u}\|_2^2 + \sum_{s=1}^{t} \Big(\langle \beta^{-s}\mathbf{g}_s, \mathbf{x}_s\rangle + F_s(\widehat{\mathbf{x}}_s) - F_{s+1}(\widehat{\mathbf{x}}_{s+1}) \Big). \tag{18}$$

接下来需要界定右侧的稳定性项。根据定义，$\psi_{s+1}(\cdot)$ 是 Orabona 的近端正则化项 [, 2019] 并满足 $\psi_{s+1}(\mathbf{x}) - \psi_s(\mathbf{x}) = \frac{\beta^{-s}\eta_s}{2}\langle \mathbf{g}\{\text{v34>}, \mathbf{x} - \mathbf{x}_s\rangle^2$。特别地，——

$$\psi_{s+1}(\mathbf{x}_s) - \psi_s(\mathbf{x}_s) = 0, \quad \nabla\psi_{s+1}(\mathbf{x}_s) - \nabla\psi_s(\mathbf{x}_s) = \mathbf{0}. \tag{19}$$

首先，式（19）意味着

$$\begin{aligned}
\langle \beta^{-s}\mathbf{g}_s, \mathbf{x}_s\rangle + F_s(\widehat{\mathbf{x}}_s) - F_{s+1}(\widehat{\mathbf{x}}_{s+1}) &= F_{s+1}(\mathbf{x}_s) - F_s(\mathbf{x}_s) + \psi_s(\mathbf{x}_s) - \psi_{s+1}(\mathbf{x}_s) + F_s(\widehat{\mathbf{x}}_s) - F_{s+1}(\widehat{\mathbf{x}}_{s+1}) \\
&= F_{s+1}(\mathbf{x}_s) - F_s(\mathbf{x}_s) + F_s(\widehat{\mathbf{x}}_s) - F_{s+1}(\widehat{\mathbf{x}}_{s+1}) \\
&= (F_{s+1}(\mathbf{x}_s) - F_{s+1}(\widehat{\mathbf{x}}_{s+1})) - (F_s(\mathbf{x}_s) - F_s(\widehat{\mathbf{x}}_s)).
\end{aligned}$$

其次，根据一阶最优条件，

$$\begin{aligned}
&\nabla F_s(\mathbf{x}_s) + \beta^{-s}\nabla h_s(\mathbf{x}_s) = \mathbf{0}, \\
&\nabla F_{s+1}(\mathbf{x}_s) = \nabla F_s(\mathbf{x}_s) + \beta^{-s}\mathbf{g}_s + \nabla\psi_{s+1}(\mathbf{x}_s) - \nabla\psi_s(\mathbf{x}_s) = \beta^{-s}\mathbf{g}_s - \beta^{-s}\nabla h_s(\mathbf{x}_s),
\end{aligned}$$

其中，最后一个等号利用了式（19）以及 $\nabla F_s(\mathbf{x}_s) = -\beta^{-s}\nabla h_s(\mathbf{x}_s)$。由于 $F_{s+1}(\cdot)$ 是一个海森矩阵为 A_s（定义见式（14））的二次函数，我们有如下标准恒等式

$$F_{s+1}(\mathbf{x}) - F_{s+1}(\widehat{\mathbf{x}}_{s+1}) = \frac{1}{2}\big\|\nabla F_{s+1}(\mathbf{x})\big\|_{A_s^{-1}}^2 \qquad \text{for all } \mathbf{x}\in\mathbb{R}^d.$$

利用该恒等式和 $A_s \succeq A_{s-1}$，我们得到

$$\begin{aligned}
&\langle \beta^{-s}\mathbf{g}_s, \mathbf{x}_s\rangle + F_s(\widehat{\mathbf{x}}_s) - F_{s+1}(\widehat{\mathbf{x}}_{s+1}) \\
&= (F_{s+1}(\mathbf{x}_s) - F_{s+1}(\widehat{\mathbf{x}}_{s+1})) - (F_s(\mathbf{x}_s) - F_s(\widehat{\mathbf{x}}_s)) = \frac{1}{2}\big\|\nabla F_{s+1}(\mathbf{x}_s)\big\|_{A_s^{-1}}^2 - \frac{1}{2}\big\|\nabla F_s(\mathbf{x}_s)\big\|_{A_{s-1}^{-1}}^2 \\
&\le \frac{\beta^{-2s}}{2}\Big(\big\|\mathbf{g}_s - \nabla h_s(\mathbf{x}_s)\big\|_{A_s^{-1}}^2 - \big\|\nabla h_s(\mathbf{x}_s)\big\|_{A_s^{-1}}^2\Big) = \frac{\beta^{-2s}}{2}\langle \mathbf{g}_s, \mathbf{g}_s - 2\nabla h_s(\mathbf{x}_s)\rangle_{A_s^{-1}}.
\end{aligned} \tag{20}$$

对于逻辑损失，令 $\mathbf{g}_s^{-y_s} = \nabla\ell(\mathbf{x}_s^\top \mathbf{z}_s, -y_s)$ 表示相对于相反标签的梯度。于是 $\nabla h_s(\mathbf{x}_s) = \mathbf{g}_s + \mathbf{g}_s^{-y_s}$，并且

$$\frac{1}{2}\langle \mathbf{g}_s, \mathbf{g}_s - 2\nabla h_s(\mathbf{x}_s)\rangle_{A_s^{-1}} = \frac{1}{2}\langle \mathbf{g}_s, -\mathbf{g}_s - 2\mathbf{g}_s^{-y_s}\rangle_{A_s^{-1}} \le -\langle \mathbf{g}_s, \mathbf{g}_s^{-y_s}\rangle_{A_s^{-1}}.$$

此外，根据逻辑梯度的显式形式，$\mathbf{g}_s^{-y_s} = -\exp(y_s \mathbf{x}_s^\top \mathbf{z}_s)\mathbf{g}_s = -(1+BR)\eta_s \mathbf{g}_s$ 根据（13）。因此

$$-\langle \mathbf{g}_s, \mathbf{g}_s^{-y_s} \rangle_{A_s^{-1}} = (1+BR)\eta_s \mathbf{g}_s^\top A_s^{-1} \mathbf{g}_s.$$

将其与（20）结合可得

$$\langle \beta^{-s}\mathbf{g}_s, \mathbf{x}_s \rangle + F_s(\widehat{\mathbf{x}}_s) - F_{s+1}(\widehat{\mathbf{x}}_{s+1}) \le (1+BR)\beta^{-2s}\eta_s \mathbf{g}_s^\top A_s^{-1}\mathbf{g}_s.$$

将上述不等式代入（18）即可完成证明。 □

C.2.2 定理3的证明

本节将证明定理3 的完整版本。

定理 （定理 3 的扩展版本）。 假设 $\|\mathbf{z}_t\|_2 \le R$ 对所有 $t \in [T]$均成立。对于满足 $\|\mathbf{u}_t\|_2 \le B$的任意比较序列 $\mathbf{u}_1, \dots, \mathbf{u}_T$，参数为 $\lambda > 0$ 和 $\beta \in (0,1)$ 的折扣 *AIOLI* 算法的动态遗憾界为：

$$\begin{aligned}\sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}_t) &\le \frac{\beta\lambda}{2}\|\mathbf{u}_1\|_2^2 + d(1+BR)\log\left(1 + \frac{R^2\sum_{t=1}^T \beta^{T-t}}{d\lambda(1+BR)}\right) \\ &\quad + \beta\sum_{t=1}^{T-1}\left(F_t^{\beta,\varphi}(\mathbf{u}_{t+1}) - F_t^{\beta,\varphi}(\mathbf{u}_t)\right) + \frac{1-\beta}{\beta}\cdot d(1+BR)T \\ &\le \frac{\beta\lambda}{2}\|\mathbf{u}_1\|_2^2 + d(1+BR)\log\left(1 + \frac{R^2\sum_{t=1}^T \beta^{T-t}}{d\lambda(1+BR)}\right) + \frac{\gamma}{1-\gamma}P_T^\gamma + \frac{1-\beta}{\beta}\cdot d(1+BR)T\end{aligned}$$

其中假设 $0 < \beta \le \gamma < 1$，且 P_T^γ 由公式 （4）给出。

证明。 我们应用引理 3 和 D2D 归约。令 $\mathbf{u}_1, \dots, \mathbf{u}_T \in \mathbb{R}^d$ 满足 $\|\mathbf{u}_t\|_2 \le B$。引理 3给出，对每个 $t \in [T]$ 以及任意固定的 $\mathbf{u}$，

$$\sum_{s=1}^t \beta^{-s}\big(f_s(\mathbf{x}_s) - f_s(\mathbf{u})\big) \le \frac{\lambda}{2}\|\mathbf{u}\|_2^2 + \sum_{s=1}^t (1+BR)\beta^{-2s}\eta_s \mathbf{g}_s^\top A_s^{-1}\mathbf{g}_s,$$

其中 A_s 定义于 (14)。因此，定理 1 的假设在 $\varphi_t(\mathbf{u}) = \frac{\lambda}{2}\|\mathbf{u}\|_2^2$ 以及 $\Lambda_s = (1+BR)\beta^{-2s}\eta_s \mathbf{g}_s^\top A_s^{-1}\mathbf{g}_s$下成立。由于 φ_t 与时间无关，定理 1 中的最后一项为零，因此我们得到

$$\begin{aligned}\sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{u}_t) &\le \frac{\beta\lambda}{2}\|\mathbf{u}_1\|_2^2 + (1+BR)\sum_{t=1}^T \beta^t \cdot \beta^{-2t}\eta_t \mathbf{g}_t^\top A_t^{-1}\mathbf{g}_t + \beta\sum_{t=1}^{T-1}\left(F_t^{\beta,\varphi}(\mathbf{u}_{t+1}) - F_t^{\beta,\varphi}(\mathbf{u}_t)\right) \\ &= \frac{\beta\lambda}{2}\|\mathbf{u}_1\|_2^2 + (1+BR)\sum_{t=1}^T \eta_t \mathbf{g}_t^\top \widetilde{A}_t^{-1}\mathbf{g}_t + \beta\sum_{t=1}^{T-1}\left(F_t^{\beta,\varphi}(\mathbf{u}_{t+1}) - F_t^{\beta,\varphi}(\mathbf{u}_t)\right),\end{aligned}$$

其中我们用了 $\widetilde{A}_t = \beta^t A_t = \lambda\beta^t I + \sum_{s=1}^t \beta^{t-s}\eta_s \mathbf{g}_s \mathbf{g}_s^\top$，从而有 $A_t^{-1} = \beta^t \widetilde{A}_t^{-1}$。我们通过将折扣势引理（引理 25）应用于 $\tilde{\mathbf{g}}_t = \sqrt{\eta_t}\mathbf{g}_t$，来界定稳定性求和 $\sum_{t=1}^T \eta_t \mathbf{g}_t^\top \widetilde{A}_t^{-1}\mathbf{g}_t$：

$$\sum_{t=1}^T \eta_t \mathbf{g}_t^\top \widetilde{A}_t^{-1}\mathbf{g}_t \le dT\ln\frac{1}{\beta} + d\ln\left(1 + \frac{\sum_{t=1}^T \beta^{T-t}\eta_t\|\mathbf{g}_t\|_2^2}{d\lambda}\right) \le dT\frac{1-\beta}{\beta} + d\ln\left(1 + \frac{R^2\sum_{t=1}^T \beta^{T-t}}{d\lambda(1+BR)}\right),$$

其中，在第二行我们使用 $\ln(1/\beta) \le (1-\beta)/\beta$，并且 $\eta_t\|\mathbf{g}_t\|_2^2 = \|\eta_t\mathbf{g}_t\|_2\|\mathbf{g}_t\|_2 = \frac{1}{1+BR}\|\mathbf{g}_t^{-y_t}\|_2\|\mathbf{g}_t\|_2 \le \frac{R^2}{1+BR}$。对于路径长度项，令 $f_0(\mathbf{u}) = \frac{\lambda}{2}\|\mathbf{u}\|_2^2$，根据引理 18，我们得出：

$$\beta\sum_{t=1}^{T-1}\left(F_t^{\beta,\varphi}(\mathbf{u}_{t+1}) - F_t^{\beta,\varphi}(\mathbf{u}_t)\right) \le \frac{\gamma}{1-\gamma}P_T^\gamma.$$

□

C.3 定理4的证明

在本小节中，我们将展示如何为在线逻辑回归学习合适的折扣因子。核心思路是采用两层在线集成 [Mhammedi 和 Rakhlin，2022，Jacobsen 和 Cutkosky，2024年，Zhao 等人， 2024] 来处理选择 β时的不确定性。

具体来说，我们维护 N 个基学习器。基学习器 i 运行折扣 AIOLI，使用指定的折扣因子 $\beta_i \in (0,1]$, , 并在第 t 轮输出决策 $\mathbf{x}_{t,\,i} \in \mathbb{R}^d$ 。令 $\widehat{y}_{t,\,i} = \mathbf{x}_{t,\,i}^{\top}\mathbf{z}_t$ 为其预测。元学习器将这些基学习器组合起来，生成最终预测 $\widehat{y}_t \in \mathbb{R}$，随后学习者承受逻辑损失 $\ell(\widehat{y}_t,\ y_t)$。

对于任意固定的索引 $i \in [N]$，区间动态遗憾可以分解为

$$\sum_{t=1}^{T} \ell(\widehat{y}_t, y_t) - \sum_{t=1}^{T} \ell(\mathbf{u}_t^{\top}\mathbf{z}_t, y_t) = \underbrace{\sum_{t=1}^{T} \Big(\ell(\widehat{y}_t, y_t) - \ell(\widehat{y}_{t,i}, y_t)\Big)}_{\text{Meta Regret}} + \underbrace{\sum_{t=1}^{T} \Big(\ell(\widehat{y}_{t,i}, y_t) - \ell(\mathbf{u}_t^{\top}\mathbf{z}_t, y_t)\Big)}_{\text{Base Regret}}. \tag{21}$$

如果存在一个 β_i 接近理想选择的值，那么基遗憾具有所需的阶数。剩下的主要问题是控制元遗憾，同时避免引入对 B 的指数依赖。为此，我们利用逻辑损失的可混合性 [Vovk，2001，Cesa-Bianchi 和 Lugosi，2006，van Erven 等人，2012，Foster 等人，2018，Mhammedi 和 Rakhlin，2022，Zhang 等人， 2025]。我们先给出可混合损失的定义。

定义 2 （可混合损失）。 若损失 $\ell(\cdot,\cdot)$：$\mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ 满足以下条件，则称其为 η-可混合的：对于任意预测 $\widehat{y}_1, \ldots, \widehat{y}_N \in \mathbb{R}$，以及任意权重 $p_1, \ldots, p_N \geq 0$ （满足 $\sum_{i=1}^{N} p_i = 1$），存在一个混合预测 $\widehat{y}_{\text{mix}} \in \mathbb{R}$，使得对于任意 $y \in \mathcal{Y}$，

$$\ell(\widehat{y}_{\text{mix}}, y) \leq -\frac{1}{\eta} \ln\Big(\sum_{i=1}^{N} p_i e^{-\eta \ell(\widehat{y}_i, y)}\Big).$$

逻辑损失是 1-可混合的，利用这一性质，我们可以避免元遗憾中的指数因子。

命题 1 (Foster 等人 [2018])。 逻辑损失 $\ell(\widehat{y}, y) = \ln(1+\exp(-y\widehat{y}))$ 在 $\mathbb{R}\times\{-1,+1\}$ 上是1-可混合的。

此外，混合预测可以写成以下闭式形式： [Foster 等人， 2018]。令 $\sigma(a) = \exp(a)/(1+\exp(a))$。给定 $\widehat{y}_1, \ldots, \widehat{y}_N$ 以及权重 p_i，一种有效的取法是

$$\widehat{y}_{\text{mix}} = \ln\left(\frac{\sum_{i=1}^{N} p_i \sigma(\widehat{y}_i)}{\sum_{i=1}^{N} p_i (1-\sigma(\widehat{y}_i))}\right).$$

定理4的证明。固定任意基学习器索引 $i \in [N]$。回顾（21）中的分解：

$$\sum_{t=1}^{T} \ell(\widehat{y}_t, y_t) - \sum_{t=1}^{T} \ell(\mathbf{u}_t^{\top}\mathbf{z}_t, y_t) = \underbrace{\sum_{t=1}^{T} \Big(\ell(\widehat{y}_t, y_t) - \ell(\widehat{y}_{t,i}, y_t)\Big)}_{\text{Meta Regret}} + \underbrace{\sum_{t=1}^{T} \Big(f_t(\mathbf{x}_{t,i}) - f_t(\mathbf{u}_t)\Big)}_{\text{Base Regret}}.$$

元遗憾。由逻辑损失的1-可混合性（命题1）以及算法1中的特定混合预测，对每个 t,

$$\ell(\widehat{y}_t, y_t) \leq -\ln\left(\sum_{j=1}^{N} p_{t,j} \exp\big(-\ell(\widehat{y}_{t,j}, y_t)\big)\right).$$

利用 $p_{t,j} = q_{t,j}/\sum_{k=1}^{N} q_{t,k}$ 以及更新规则 $q_{t+1,j} = q_{t,j}\exp(-\ell(\widehat{y}_{t,j}, y_t))$，我们得到

$$\ell(\widehat{y}_t, y_t) \leq -\ln\left(\sum_{j=1}^{N} \frac{q_{t,j}}{\sum_{k=1}^{N} q_{t,k}} \exp\big(-\ell(\widehat{y}_{t,j}, y_t)\big)\right) = -\ln\left(\frac{\sum_{j=1}^{N} q_{t+1,j}}{\sum_{k=1}^{N} q_{t,k}}\right).$$

对 $t=1,\dots,T$ 求和，得到望远镜求和界

$$\sum_{t=1}^{T}\ell(\widehat{y}_t,y_t)\le-\ln\left(\frac{\sum_{j=1}^{N}q_{T+1,j}}{\sum_{j=1}^{N}q_{1,j}}\right)=\ln\left(\sum_{j=1}^{N}q_{1,j}\right)-\ln\left(\sum_{j=1}^{N}q_{T+1,j}\right).$$

由于 $q_{1,\,j}=1$，我们有 $\sum_{j=1}^{N}q_{1,\,j}=N$，并且还有 $\sum_{j=1}^{N}q_{T+1,\,j}\ge q_{T+1,\,i}$。因此，

$$\sum_{t=1}^{T}\ell(\widehat{y}_t,y_t)\le\ln N-\ln q_{T+1,i}=\ln N+\sum_{t=1}^{T}\ell(\widehat{y}_{t,i},y_t),$$

这意味着 Meta Regret $\le\ln N$。注意，$N=\mathcal{O}(\log T)$，因此 $\ln N=\mathcal{O}(\log\log T)$，我们将其视为 t 常数 按照前述约定 [Gaillard 等人，2014，Luo 和 Schapire，2015]。

对于任意 $s\in[T]$，损失 $f_s(\mathbf{x})=\ell(\mathbf{x}^\top\mathbf{z}_s,\ y_s)$ 关于 $\|\cdot\|_2$ 是 R-利普希茨的，因为 $\|\nabla f_s(\mathbf{x})\|_2\le\|\mathbf{z}_s\|_2\le R$。因此，对于任意 $t\in[T-1]$ 和任意 $s\le t$，

$$\left[f_s(\mathbf{u}_{t+1})-f_s(\mathbf{u}_t)\right]_+\le\left|f_s(\mathbf{u}_{t+1})-f_s(\mathbf{u}_t)\right|\le\max\left\{R,\frac{\lambda}{2}\|\mathbf{u}_{t+1}\|_2^2\right\}\|\mathbf{u}_{t+1}-\mathbf{u}_t\|_2\le\max\{2R,1\}B,$$

其中我们使用了 $\|\mathbf{u}_t\|_2\le B$ 和 $\lambda=\frac{1}{B^2}$。由于 $\sum_{s=0}^{t}p_{t,\,s}^{\beta}=1$，我们对所有 $\beta\in(0,1]$，

$$P_T^{\beta}=\sum_{t=1}^{T-1}\sum_{s=0}^{t}p_{t,s}^{\beta}\left[f_s(\mathbf{u}_{t+1})-f_s(\mathbf{u}_t)\right]_+\le CBT,$$

其中 $C=$最大值为$\{2R,1\}$。特别是，对于定理陈述中的 $\beta_\star$，$P_T^{\beta_\star}\le CBT$，因此

$$\eta_\star=\frac{\beta_\star}{1-\beta_\star}=\sqrt{\frac{d(1+BR)T}{P_T^{\beta_\star}}}\ge\sqrt{\frac{d(1+BR)T}{CBT}}=\sqrt{\frac{d(1+BR)}{CB}}=\eta_{\min}.$$

基准遗憾。对于折扣率为 $\beta_i\in(0,1)$ 且 $\lambda=1/B^2$ 的基学习器，折扣 AIOLI 保证（定理，采用将路径长度项转换为 $\frac{\beta}{1-\beta}P_T^{\beta}$ 之前的形式）给出

$$\sum_{t=1}^{T}\left(f_t(\mathbf{x}_{t,i})-f_t(\mathbf{u}_t)\right)\le\underbrace{\beta_i\cdot\frac{\lambda}{2}\|\mathbf{u}_1\|_2^2}_{\le1/2}+\underbrace{d(1+BR)\log\left(1+\frac{R^2B^2T}{d(1+BR)}\right)}_{\text{Term-A}}+\underbrace{d(1+BR)T\log\left(\frac{1}{\beta_i}\right)}_{\text{Term-B}}$$
$$+\underbrace{\beta_i\sum_{t=1}^{T-1}\left(F_t^{\beta_i,\varphi}(\mathbf{u}_{t+1})-F_t^{\beta_i,\varphi}(\mathbf{u}_t)\right)}_{\text{Term-C}}.\tag{22}$$

现在，我们通过从候选池中选择合适的 β_i，来对 Term-B 和 Term-C 进行上界估计。定义 $\eta_i=\beta_i/(1-\beta_i)$。利用 $\ln(1/x)\le(1/x)-1$，我们有 $\ln(\frac{1}{\beta_i})\le(1-\beta_i)/\beta_i$，因此

$$\text{Term-B}\le\frac{d(1+BR)T}{\eta_i}.$$

接下来，根据引理 18，对于任意 $0<\beta\le\gamma<1$，

$$\beta\sum_{t=1}^{T-1}\left(F_t^{\beta,\varphi}(\mathbf{u}_{t+1})-F_t^{\beta,\varphi}(\mathbf{u}_t)\right)\le\frac{\gamma}{1-\gamma}P_T^{\gamma}.\tag{23}$$

令 $\eta_\star=\beta_\star/(1-\beta_\star)$。由于 $\eta_\star\ge\eta_{\min}$，根据学习率池的构造，存在一个索引$i\in[N]$，使得 $\eta_i\le\min\{\eta_\star,\eta_{\max}\}\le2\eta_i$。我们考虑两种情况。

情况 1：如果 $\eta_\star\le\eta_{\max}$。那么 $\eta_i\le\eta_\star$ 因此 $\beta_i\le\beta_\star$ 根据 $\beta=\eta/(1+\eta)$ 在 η中的单调性。应用 (23) 和 $(\beta,\gamma)=(\beta_i,\beta_\star)$ 得到

$$\text{Term-C}\le\frac{\beta_\star}{1-\beta_\star}P_T^{\beta_\star}=\eta_\star P_T^{\beta_\star}.$$

此外， $\eta_\star \le 2\eta_i$ 意味着 $\eta_i \ge \eta_\star/2$，因此

$$\textsc{Term-B} \le \frac{d(1+BR)T}{\eta_i} \le \frac{2d(1+BR)T}{\eta_\star}.$$

将这些界代入（22）可得

$$\sum_{t=1}^{T}\big(f_t(\mathbf{x}_{t,i}) - f_t(\mathbf{u}_t)\big) \le \frac{1}{2} + d(1+BR)\log\Big(1 + \frac{R^2B^2T}{d(1+BR)}\Big) + \eta_\star P_T^{\beta_\star} + \frac{2d(1+BR)T}{\eta_\star}.$$

最后，由于最优值为 $\eta_\star = \sqrt{d(1+BR)T/P_T^{\beta_\star}}$，我们有 $\eta_\star P_T^{\beta_\star} = \sqrt{d(1+BR)TP_T^{\beta_\star}}$ 以及 $\frac{d(1+BR)T}{\eta_\star} = \sqrt{d(1+BR)TP_T^{\beta_\star}}$，因此最后两项的上界为 $3\sqrt{d(1+BR)TP_T^{\beta_\star}}$。

情况 2：若 $\eta_\star > \eta_{\max}$。我们取 i，使得 $\frac{1}{2}\eta_{\max} \le \eta_i \le \eta_{\max}$ 且仍有 $\eta_i \le \eta_\star$，因此 $\beta_i \le \beta_\star$。同样的论证可得 $\textsc{Term-C} \le \eta_\star P_T^{\beta_\star} = \sqrt{d(1+BR)TP_T^{\beta_\star}}$，而根据 $\eta_{\max} = dT$，$\textsc{Term-B} \le 2d(1+BR)T/\eta_{\max} = 2(1+BR)$。因此

$$\sum_{t=1}^{T}\big(f_t(\mathbf{x}_{t,N}) - f_t(\mathbf{u}_t)\big) \le \frac{1}{2} + d(1+BR)\log\Big(1 + \frac{R^2B^2T}{d(1+BR)}\Big) + \sqrt{d(1+BR)TP_T^{\beta_\star}} + 2(1+BR).$$

结合元遗憾界和基础遗憾界，我们完成证明。 □

D 第 4 节中省略的细节

在本部分中，我们给出第4节中省略的细节，包括若干关键引理和讨论，以及关于 Adam 优化器结果的详细证明。

D.1 无裁剪 Adam 的闭式更新规则

在本部分中，我们推导无裁剪 Adam 的闭式更新规则。

引理 4。固定 β_1、$\beta_2 \in (0, 1)$ 以及参数 $\gamma > 0$、$\nu > 0$和 $\mu > 0$。考虑带折扣的跟随正则化领导者更新（5），其中 $\beta = \beta_1$ 以及 $\mathcal{D} = \mathbb{R}^d$，替代损失为

$$\ell_t(\Delta) = \langle \mathbf{g}_t, \Delta\rangle + \frac{\mu}{2}\|\Delta\|_2^2,$$

，且基础步长 η_t 按照 (7) 中的设定选取，即：

$$\eta_t = \gamma \cdot \frac{(1-\beta_1)\beta_1^t}{\nu + \sqrt{(1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\|\mathbf{g}_s\|_2^2}}.$$

那么，该更新可写成闭式形式

$$\Delta_{t+1} = -\frac{\gamma(1-\beta_1)\sum_{s=1}^{t}\beta_1^{t-s}\mathbf{g}_s}{\nu + \gamma\mu(1-\beta_1^t) + \sqrt{(1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\|\mathbf{g}_s\|_2^2}}.$$

证明。代入 $\ell_s(\Delta) = \langle \mathbf{g}_s, \Delta\rangle + \frac{\mu}{2}\|\Delta\|_2^2$ 和 $\beta = \beta_1$ 到式（5）中，更新为

$$\begin{aligned}\Delta_{t+1} &= \arg\min_{\Delta\in\mathbb{R}^d}\Big\{\frac{1}{2\eta_t}\|\Delta\|_2^2 + \sum_{s=1}^{t}\beta_1^{-s}\langle \mathbf{g}_s, \Delta\rangle + \sum_{s=1}^{t}\beta_1^{-s}\frac{\mu}{2}\|\Delta\|_2^2\Big\}\\ &= \arg\min_{\Delta\in\mathbb{R}^d}\Big\{\langle\sum_{s=1}^{t}\beta_1^{-s}\mathbf{g}_s, \Delta\rangle + \frac{1}{2}\Big(\frac{1}{\eta_t} + \mu\sum_{s=1}^{t}\beta_1^{-s}\Big)\|\Delta\|_2^2\Big\}.\end{aligned}$$

T目标函数关于 Δ 是一个严格凸二次函数，根据一阶最优性，极小点为：

$$0 = \sum_{s=1}^{t} \beta_1^{-s} \mathbf{g}_s + \Big(\frac{1}{\eta_t} + \mu \sum_{s=1}^{t} \beta_1^{-s}\Big) \Delta_{t+1},$$

由此可得

$$\Delta_{t+1} = -\eta_t' \sum_{s=1}^{t} \beta_1^{-s} \mathbf{g}_s, \qquad \eta_t' = \frac{1}{\frac{1}{\eta_t} + \mu \sum_{s=1}^{t} \beta_1^{-s}}. \tag{24}$$

接下来，计算几何级数 $\sum_{s=1}^{t} \beta_1^{-s} = \frac{\beta_1^{-t}-1}{1-\beta_1}$。根据 η_t 在（7）中的定义，我们有

$$\frac{1}{\eta_t} = \frac{\nu + \sqrt{(1-\beta_2)\sum_{s=1}^{t} \beta_2^{t-s} \|\mathbf{g}_s\|_2^2}}{\gamma(1-\beta_1)\beta_1^t}.$$

将这些恒等式代入（ η_t' ）中24的定义，得到

$$\eta_t' = \frac{1}{\frac{\nu + \sqrt{(1-\beta_2)\sum_{s=1}^{t} \beta_2^{t-s} \|\mathbf{g}_s\|_2^2}}{\gamma(1-\beta_1)\beta_1^t} + \mu \cdot \frac{\beta_1^{-t}-1}{1-\beta_1}} = \frac{\gamma(1-\beta_1)\beta_1^t}{\nu + \gamma\mu(1-\beta_1^t) + \sqrt{(1-\beta_2)\sum_{s=1}^{t} \beta_2^{t-s} \|\mathbf{g}_s\|_2^2}}.$$

将该表达式与 $\sum_{s=1}^{t} \beta_1^{-s} \mathbf{g}_s$ 相乘，即完成证明。 □

D.2 通过 ρ 考察 (β_1, β_2) 的选择效果

为体现定理6中的数量级，我们列出以下几种具有代表性的（β_1, β_2）取值：

- PyTorch 中的默认设置（0.9,0.999）： $\rho \approx 0.989$ 和 $1/\sqrt{1-\rho^2} \approx 6.9$;
- 大语言模型中的常见选择（0.9,0.95）： $\rho \approx 0.473$ 和 $1/\sqrt{1-\rho^2} \approx 1.13$;
- Orvieto 和 Gower 推荐的 [2025] (0.95,0.95)： $\rho \approx 0.025$ 和 $1/\sqrt{1-\rho^2} \approx 1.0$。

对于 β_2 接近 1 的典型取值，所得常数适中。我们的理论允许 β_2 取值范围较广，只要满足 $\beta_2 > \beta_1^2$；并且在适当选择 β_1 且时域 T 足够长的情况下，该理论或可为实践中 Adam 的收敛行为提供理论支持，尤其是在模型包含非光滑组件的场景中。

D.3 非凸与光滑设定的影响

与（ϵ, δ）-Goldstein 驻点类似，（c, ϵ）-驻点也可以通过以下引理约化为一阶驻点。

引理 5 （Zhang and Cutkosky [2024] 中的引理 2.3）。 对于（c, ϵ）-驻点性，对于非凸光滑函数可转化为一阶驻点性：

- 假设 F 是 L-光滑的。如果 $\|\nabla F(\mathbf{x})\|_c \le \epsilon$，其中 $c = L^2\epsilon^{-1}$，那么 $\|\nabla F(\mathbf{x})\|_2 \le 2\epsilon$。
- 假设 F 是 ρ-二阶光滑的。如果 $\|\nabla F(\mathbf{x})\|_c \le \epsilon$，其中 $c = \rho/2$，那么 $\|\nabla F(\mathbf{x})\|_2 \le 2\ \epsilon$.

我们的结果（定理 5–8）达到了 $\mathcal{O}(\max\{c^{1/2}\epsilon^{-7/2}, \epsilon^{-3}\}) = \mathcal{O}(c^{1/2}\epsilon^{-7/2})$ 速率，可收敛至一个 (c, ϵ)-驻点。通过设置 $c = \mathcal{O}(\epsilon^{-1})$，我们为非凸和光滑函数 [Arjevani 等人， 2023]达到了 $\mathcal{O}(\epsilon^{-4})$ 最优速率。通过设置 $c = \mathcal{O}(1)$，我们为二阶光滑函数 [Arjevani 等人， 2023]获得了 $\mathcal{O}(\epsilon^{-7/2})$ 最优速率。

D.4 O2NC 转换的引理

本节汇总了用于 O2NC 转换的引理。

引理 6 (Zhang and Cutkosky [2024]中的引理 3.1。在假设 3下，令 $s \sim \mathrm{Exp}(\lambda)$ 其中$\lambda > 0$，则

$$\mathbb{E}_s\left[F(\mathbf{x}+s\cdot\Delta)-F(\mathbf{x})\right]=\mathbb{E}_s\left[\langle\nabla F(\mathbf{x}+s\cdot\Delta),\Delta\rangle\right].$$

引理 7 (改编自Zhang and Cutkosky [2024]中的定理 C.1。在假设 1 – 4下，将算法 2 中的损失设为 $\ell_t(\Delta) = \langle \mathbf{g}_t, \Delta\rangle + \frac{\mu}{2}\|\Delta\|_2^2$，定义比较器 $\mathbf{u}_t$ 和动态遗憾 $\text{D-REG}_T(\mathbf{u}_{1:T})$ 如下：

$$\mathbf{u}_t = -D\cdot\frac{\sum_{s=1}^t \beta^{t-s}\nabla F(\mathbf{x}_s)}{\|\sum_{s=1}^t \beta^{t-s}\nabla F(\mathbf{x}_s)\|_2},\quad \text{D-REG}_T(\mathbf{u}_{1:T}) = \sum_{t=1}^T \langle \mathbf{g}_t, \Delta_t - \mathbf{u}_t\rangle + \frac{\mu}{2}\|\Delta_t\|_2^2 - \frac{\mu}{2}\|\mathbf{u}_t\|_2^2.$$

设定 $\mu = \frac{24cD}{(1-\beta)^2}$，则算法 2 **中返回的决策** x 满足：

$$\begin{aligned}\mathbb{E}\|\nabla F(\bar{\mathbf{x}})\|_c \le{}& \frac{F^*}{DT} + \frac{2G+\sigma}{(1-\beta)T} + \sigma\sqrt{1-\beta} + \frac{12cD^2}{(1-\beta)^2}\\ &+ \frac{1}{DT}\mathbb{E}\left[\text{D-REG}_T(\mathbf{u}_{1:T}) + \beta\sum_{t=1}^{T-1}\sum_{s=1}^{t}\beta^{t-s}\left(\ell_s(\mathbf{u}_t) - \ell_s(\mathbf{u}_{t+1})\right)\right].\end{aligned}$$

证明。根据Zhangand Cutkosky [2024],中的定理 C.1，同样的设定可得以下保证：

$$\mathbb{E}\|\nabla F(\bar{\mathbf{x}})\|_c \le \frac{F^*}{DT} + \frac{2G+\sigma}{(1-\beta)T} + \sigma\sqrt{1-\beta} + \frac{12cD^2}{(1-\beta)^2} + \frac{1}{DT}\mathbb{E}\left[\beta\text{REG}_{T;\beta}(\mathbf{u}_T) + (1-\beta)\sum_{t=1}^T \text{REG}_{t;\beta}(\mathbf{u}_t)\right],$$

其中 $\text{REG}_{t;\beta}(\mathbf{u}) = \sum_{s=1}^t \left(\beta^{t-s}\langle \mathbf{g}_s, \Delta_s - \mathbf{u}\rangle + \frac{\mu\beta^{t-s}}{2}(\|\Delta_s\|_2^2 - \|\mathbf{u}\|_2^2)\right)$。根据引理 1以及 $\ell_t(\Delta) = \langle \mathbf{g}_t, \Delta\rangle + \frac{\mu}{2}\|\Delta\|_2^2$的定义，可得：

$$\beta\text{REG}_{T;\beta}(\mathbf{u}_T) + (1-\beta)\sum_{t=1}^T \text{REG}_{t;\beta}(\mathbf{u}_t) = \text{D-REG}_T(\mathbf{u}_{1:T}) + \beta\sum_{t=1}^{T-1}\sum_{s=1}^{t}\beta^{t-s}\left(\ell_s(\mathbf{u}_t) - \ell_s(\mathbf{u}_{t+1})\right),$$

至此，证明完成。 □

引理 8 (改编自 Ahn 和 Cutkosky [2024])。 在假设 1 – 4下，将算法 2 中的损失设为 $\ell_t(\Delta) = \langle \mathbf{g}_t, \Delta\rangle$，定义比较器 $\mathbf{u}_t$ 以及动态遗憾 $\text{D-REG}_T(\mathbf{u}_{1:T})$如下：

$$\mathbf{u}_t = -D\cdot\frac{\sum_{s=1}^t \beta^{t-s}\nabla F(\mathbf{x}_s)}{\|\sum_{s=1}^t \beta^{t-s}\nabla F(\mathbf{x}_s)\|_2},\quad \text{D-REG}_T(\mathbf{u}_{1:T}) = \sum_{t=1}^T \langle \mathbf{g}_t, \Delta_t - \mathbf{u}_t\rangle.$$

假设 $\|\Delta_t\|_2 \le D$ 对任意 t均成立，则算法 $\bar{\mathbf{x}}$ 在算法 2 中返回的决策满足：

$$\begin{aligned}\mathbb{E}\|\nabla F(\bar{\mathbf{x}})\|_c \le{}& \frac{F^*}{DT} + \frac{2G+\sigma}{(1-\beta)T} + \sigma\sqrt{1-\beta} + \frac{12cD^2}{(1-\beta)^2}\\ &+ \frac{1}{DT}\mathbb{E}\left[\text{D-REG}_T(\mathbf{u}_{1:T}) + \beta\sum_{t=1}^{T-1}\sum_{s=1}^{t}\beta^{t-s}\left(\ell_s(\mathbf{u}_t) - \ell_s(\mathbf{u}_{t+1})\right)\right].\end{aligned}$$

该引理的证明与引理7的证明相同，故略去。

D.5 Adam 分析的技术引理

本节将介绍并证明在 O2NC 转换下分析 Adam 时所需的若干有用技术引理。

引理 9. 令 $\beta \in (0,\ 1)$，$\epsilon > 0$，并令 $\{a_t\}_{t=1}^T$ 满足 $0 \le a_t \le A$，对所有 $t \in [T]$ 均成立。定义指数移动平均 $V_0 = 0, V_t = \beta V_{t-1} + (1-\beta)a_t$，其中 $t \ge 1$。那么，

$$\sum_{t=1}^{T} \frac{a_t}{\epsilon + \sqrt{V_{t-1}}} \le \left(\frac{T(1-\beta)}{\ln 2} + 1\right)\frac{A}{\epsilon} + \frac{2}{\sqrt{(1-\beta)2^{-1/\beta}}}\sqrt{\left(\frac{T(1-\beta)}{\ln 2} + 1\right)\sum_{t=1}^{T} a_t}.$$

证明。 展开递推关系可得，对任意 $t \ge 1$,

$$V_{t-1} = (1-\beta)\sum_{s=1}^{t-1} \beta^{t-1-s} a_s.$$

固定 $\tau \ge 1$（稍后确定），并将 $[T]$ 划分为 $K = \lceil T/\tau \rceil$ 个连续分段 $I_k = [r_k, s_k]$，每段长度至多为 τ，其中 $r_k = (k-1)\tau + 1$ 且 $s_k = \min\{k\tau, T\}$。对于任意 $t \in [r_k, s_k]$ 和任意 $s \in [r_k, t-1]$，有 $t - 1 - s \le \tau - 1$，因此 $\beta^{t-1-s} \ge \beta^{\tau-1}$。所以,

$$V_{t-1} = (1-\beta)\sum_{s=1}^{t-1} \beta^{t-1-s} a_s \ \ge\ (1-\beta)\sum_{s=r_k}^{t-1} \beta^{t-1-s} a_s \ \ge\ (1-\beta)\beta^{\tau-1}\sum_{s=r_k}^{t-1} a_s.$$

令 $c = (1-\beta)\beta^{\tau-1}$。利用上述不等式并对各分段求和，我们得到

$$\sum_{t=1}^{T} \frac{a_t}{\epsilon + \sqrt{V_{t-1}}} = \sum_{k=1}^{K}\sum_{t=r_k}^{s_k} \frac{a_t}{\epsilon + \sqrt{V_{t-1}}} \le \sum_{k=1}^{K}\sum_{t=r_k}^{s_k} \frac{a_t}{\epsilon + \sqrt{c\sum_{s=r_k}^{t-1} a_s}}. \tag{25}$$

现在我们对每个内层求和进行上界估计。固定一个分段 I_k 并定义 $S_{t-1} = \sum_{s=r_k}^{t-1} a_s$（因此 $S_{r_k-1} = 0$）。对于任意 $x \ge 0$，有 $(\epsilon + \sqrt{cx})^2 = \epsilon^2 + cx + 2\epsilon\sqrt{cx} \ge \epsilon^2 + cx$，于是

$$\frac{1}{\epsilon + \sqrt{cx}} \le \frac{1}{\sqrt{\epsilon^2 + cx}} = \frac{1}{\sqrt{c}} \cdot \frac{1}{\sqrt{\epsilon^2/c + x}}.$$

令 $\delta = \epsilon^2/c > 0$。将 $x = S_{t-1}$ 代入上述等式，得到

$$\sum_{t=r_k}^{s_k} \frac{a_t}{\epsilon + \sqrt{cS_{t-1}}} \le \frac{1}{\sqrt{c}}\sum_{t=r_k}^{s_k} \frac{a_t}{\sqrt{\delta + S_{t-1}}}.$$

我们应用引理 23，取 $a_0 = \delta$、$B = A$和 $f(u) = u^{-1/2}$，得到

$$\begin{aligned}\sum_{t=r_k}^{s_k} \frac{a_t}{\sqrt{\delta + S_{t-1}}} &\le A \cdot \frac{1}{\sqrt{\delta}} + \int_{\delta}^{\delta + \sum_{t=r_k}^{s_k} a_t} u^{-1/2}\mathrm{d}u \\ &= A \cdot \frac{1}{\sqrt{\delta}} + 2\left(\sqrt{\delta + \sum_{t=r_k}^{s_k} a_t} - \sqrt{\delta}\right) \le A \cdot \frac{1}{\sqrt{\delta}} + 2\sqrt{\sum_{t=r_k}^{s_k} a_t},\end{aligned}$$

其中我们在 $S \ge 0$处使用了 $\sqrt{\delta + S} - \sqrt{\delta} \le \sqrt{S}$。由于 $\sqrt{\delta} = \epsilon/\sqrt{c}$，我们对每个 I_k 得出结论 t

$$\sum_{t=r_k}^{s_k} \frac{a_t}{\epsilon + \sqrt{cS_{t-1}}} \le \frac{A}{\epsilon} + \frac{2}{\sqrt{c}}\sqrt{\sum_{t=r_k}^{s_k} a_t}. \tag{26}$$

结合（25）和（26），再应用柯西–施瓦茨不等式，可得

$$\sum_{t=1}^{T} \frac{a_t}{\epsilon + \sqrt{V_{t-1}}} \le K \cdot \frac{A}{\epsilon} + \frac{2}{\sqrt{c}}\sum_{k=1}^{K}\sqrt{\sum_{t=r_k}^{s_k} a_t} \le K \cdot \frac{A}{\epsilon} + \frac{2}{\sqrt{c}}\sqrt{K}\sqrt{\sum_{t=1}^{T} a_t}. \tag{27}$$

接下来需要选取 τ。令 $\tau = \left\lceil \frac{\ln 2}{1-\beta} \right\rceil$。则 $\tau - 1 < \frac{\ln 2}{1-\beta}$，因此 $\beta^{\tau-1} > \beta^{\frac{\ln 2}{1-\beta}}$。此外，对于 $\beta \in (0,1)$，有 $\ln \beta \geq -\frac{1-\beta}{\beta}$，这意味着

$$\beta^{\frac{\ln 2}{1-\beta}} = \exp\left(\frac{\ln 2}{1-\beta}\ln\beta\right) \geq \exp\left(\frac{\ln 2}{1-\beta}\cdot\left(-\frac{1-\beta}{\beta}\right)\right) = 2^{-1/\beta}.$$

因此 $\beta^{\tau-1} \geq 2^{-1/\beta}$，从而 $c = (1-\beta)\beta^{\tau-1} \geq (1-\beta)2^{-1/\beta}$。此外，由于 $K = \lceil T/\tau \rceil \leq T/\tau + 1$ 且 $\tau \geq \ln 2/(1-\beta)$，我们有

$$K \leq \frac{T(1-\beta)}{\ln 2} + 1.$$

将这些关于 c 和 K 的界代入式（27）即可完成证明。 □

引理 10. 设 $\beta \in (0,1)$，$\epsilon > 0$，并设 $\{a_t\}_{t=1}^T$ 满足对所有 $t \in [T]$ 均有 $0 \leq a_t \leq A$。定义指数移动平均 $V_0 = 0, V_t = \beta V_{t-1} + (1-\beta)a_t$，其中 $t \geq 1$。则，

$$\sum_{t=1}^{T} \frac{a_t}{\epsilon + \sqrt{V_t}} \leq \frac{2}{\sqrt{(1-\beta)2^{-1/\beta}}} \sqrt{\left(\frac{T(1-\beta)}{\ln 2} + 1\right)\sum_{t=1}^{T} a_t}.$$

证明概要。论证沿用引理 9的证明，但有一处改动：我们将引理 24 应用于式（25），而非使用引理 23。因此，我们省略细节。 □

引理 11. 令 $\beta_1, \beta_2 \in (0,1)$。令 $\{\epsilon_t\}_{t\geq 0}$ 为非负单调非减序列，并令 $\{\mathbf{g}_t\}_{t\geq 1} \subseteq \mathbb{R}^d$ 任意。定义

$$V_0 = 0, \qquad V_t = (1-\beta_2)\sum_{s=1}^{t} \beta_2^{t-s}\|\mathbf{g}_s\|_2^2 \quad (t \geq 1), \qquad \alpha_t = \frac{1}{\epsilon_t + \sqrt{V_t}} \quad (t \geq 0).$$

于是，对每个 $t \geq 1$，

$$\left[\frac{\beta_1}{\alpha_{t-1}} - \frac{1}{\alpha_t}\right]_+ \leq [\beta_1 - \sqrt{\beta_2}]_+ \sqrt{V_{t-1}},$$

其中 $[x]_+ = \max\{x, 0\}$。

证明。根据 α_t 的定义，我们有

$$\begin{aligned}\frac{\beta_1}{\alpha_{t-1}} - \frac{1}{\alpha_t} &= \beta_1\left(\epsilon_{t-1} + \sqrt{V_{t-1}}\right) - \left(\epsilon_t + \sqrt{V_t}\right) \\ &= \underbrace{\left(\beta_1\epsilon_{t-1} - \epsilon_t\right)}_{\leq 0} + \left(\beta_1\sqrt{V_{t-1}} - \sqrt{V_t}\right) \leq \beta_1\sqrt{V_{t-1}} - \sqrt{V_t},\end{aligned}$$

其中该不等式用到了 $\epsilon_{t-1} \leq \epsilon_t$ 和 $\beta_1 \leq 1$。

接下来，$V_t = \beta_2 V_{t-1} + (1-\beta_2)\|\mathbf{g}_t\|_2^2 \geq \beta_2 V_{t-1}$ 意味着 $\sqrt{V_t} \geq \sqrt{\beta_2}\sqrt{V_{t-1}}$。因此，

$$\frac{\beta_1}{\alpha_{t-1}} - \frac{1}{\alpha_t} \leq (\beta_1 - \sqrt{\beta_2})\sqrt{V_{t-1}}.$$

在两边取 $[\cdot]_+$ 即可证明该引理。 □

引理 12。 设 β_1、$\beta_2 \in$（0，1）、$\gamma > 0$、$\epsilon > 0$ 和 $\mu \geq 0$。令 $\{\mathbf{g}_t\}_{t=1}^{T-1} \subseteq \mathbb{R}^d$ 为任意函数，并定义

$$V_0 = 0, \qquad V_t = (1-\beta_2)\sum_{s=1}^{t} \beta_2^{t-s}\|\mathbf{g}_s\|_2^2 \quad (t \geq 1).$$

对于 $t \geq 0$，定义

$$\alpha_t = \frac{1}{\epsilon + \gamma\mu(1-\beta_1^t) + \sqrt{V_t}}, \qquad \eta_t = \gamma(1-\beta_1)\beta_1^t\alpha_t.$$

那么，对于任意 $T \geq 2$,

$$\sum_{t=1}^{T-1} \beta_1^t \Big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \Big) \leq \frac{\sqrt{V_{T-1}}}{\gamma(1-\beta_1)} + \frac{1}{\gamma} \sum_{t=1}^{T-2} \sqrt{V_t} + \frac{(T-1)\epsilon}{\gamma} + \mu(T-1).$$

证明。 令 $b_t = 1/\eta_t$。 通过直接重排求和,

$$\sum_{t=1}^{T-1} \beta_1^t (b_t - b_{t-1}) = \beta_1^{T-1} b_{T-1} - \beta_1 b_0 + (1-\beta_1) \sum_{t=1}^{T-2} \beta_1^t b_t, \tag{28}$$

这由下式推出

$$\sum_{t=1}^{T-1} \beta_1^t (b_t - b_{t-1}) = \sum_{t=1}^{T-1} \beta_1^t b_t - \sum_{t=1}^{T-1} \beta_1^t b_{t-1} = \sum_{t=1}^{T-1} \beta_1^t b_t - \sum_{t=0}^{T-2} \beta_1^{t+1} b_t.$$

根据 η_t 和 α_t的定义，对于每个 $t \geq 0$,

$$\beta_1^t b_t = \beta_1^t \cdot \frac{1}{\gamma(1-\beta_1)\beta_1^t \alpha_t} = \frac{1}{\gamma(1-\beta_1)} \cdot \frac{1}{\alpha_t} = \frac{\epsilon + \gamma\mu(1-\beta_1^t) + \sqrt{V_t}}{\gamma(1-\beta_1)}.$$

特别地， $V_0 = 0$ 和 $\beta_1^0 = 1$ 给出 $b_0 = \epsilon/(\gamma(1-\beta_1))$。代入式（28）得到

$$\sum_{t=1}^{T-1} \beta_1^t (b_t - b_{t-1}) = \frac{\sqrt{V_{T-1}}}{\gamma(1-\beta_1)} + \frac{1}{\gamma} \sum_{t=1}^{T-2} \sqrt{V_t} + \frac{(T-1)\epsilon}{\gamma} + \mu \Big(\frac{1-\beta_1^{T-1}}{1-\beta_1} + \sum_{t=1}^{T-2} (1-\beta_1^t) \Big). \tag{29}$$

注意到以下事实，我们可以简化最后一项

$$\frac{1-\beta_1^{T-1}}{1-\beta_1} + \sum_{t=1}^{T-2} (1-\beta_1^t) = \sum_{t=0}^{T-2} \beta_1^t + \sum_{t=1}^{T-2} 1 - \sum_{t=1}^{T-2} \beta_1^t = 1 + (T-2) = T-1.$$

将此结果与式（29）结合，即得引理。 □

引理 13. 设 $\beta_1, \beta_2 \in (0,1)$ 满足 $\beta_2 > \beta_1^2$，且设 $\gamma > 0$、 $\nu > 0$ 和 $\mu \geq 0$。令 $\{\mathbf{g}_t\}_{t \geq 1} \subseteq \mathbb{R}^d$ 为任意值，并定义 $t \geq 1$

$$V_t = (1-\beta_2) \sum_{s=1}^{t} \beta_2^{t-s} \|\mathbf{g}_s\|_2^2, \qquad A_t = \nu + \gamma\mu(1-\beta_1^t) + \sqrt{V_t}, \qquad \eta_t = \gamma(1-\beta_1) \frac{\beta_1^t}{A_t}.$$

那么，对任意 $t \geq 2$,

$$\big|\eta_{t-1} - \eta_t\big| \cdot \Big\| \sum_{s=1}^{t-1} \beta_1^{t-1-s} \mathbf{g}_s \Big\|_2 \leq \gamma(1-\beta_1)\beta_1^{t-1} \cdot \frac{A_t - \beta_1 A_{t-1}}{A_t} \cdot \sqrt{\frac{\beta_2}{(\beta_2 - \beta_1^2)(1-\beta_2)}}.$$

证明。 我们首先改写步长之差:

$$\eta_{t-1} - \eta_t = \gamma(1-\beta_1)\beta_1^{t-1} \Big(\frac{1}{A_{t-1}} - \frac{\beta_1}{A_t} \Big) = \gamma(1-\beta_1)\beta_1^{t-1} \cdot \frac{A_t - \beta_1 A_{t-1}}{A_{t-1} A_t}.$$

在 $\beta_2 \geq \beta_1^2$ 条件下，我们有$\sqrt{V_t} \geq \sqrt{\beta_2}\sqrt{V_{t-1}} \geq \beta_1 \sqrt{V_{t-1}}$。因此

$$\begin{aligned} A_t - \beta_1 A_{t-1} &= \big(\nu + \gamma\mu(1-\beta_1^t) + \sqrt{V_t}\big) - \beta_1 \big(\nu + \gamma\mu(1-\beta_1^{t-1}) + \sqrt{V_{t-1}}\big) \\ &= (1-\beta_1)(\nu + \gamma\mu) + \big(\sqrt{V_t} - \beta_1 \sqrt{V_{t-1}}\big) \geq 0, \end{aligned}$$

这意味着 $\eta_{t-1} \geq \eta_t$ ，从而 $|\eta_{t-1} - \eta_t| = \eta_{t-1} - \eta_t$。

接下来，根据柯西–施瓦茨不等式，令 $k = t-1-s$,

$$\Big\| \sum_{s=1}^{t-1} \beta_1^{t-1-s} \mathbf{g}_s \Big\|_2 = \Big\| \sum_{k=0}^{t-2} \beta_1^k \mathbf{g}_{t-1-k} \Big\|_2 = \Big\| \sum_{k=0}^{t-2} \Big(\frac{\beta_1^2}{\beta_2} \Big)^{k/2} \cdot \beta_2^{k/2} \mathbf{g}_{t-1-k} \Big\|_2$$

$$\leq \sqrt{\sum_{k=0}^{t-2}\Big(\frac{\beta_1^2}{\beta_2}\Big)^k}\cdot\sqrt{\sum_{k=0}^{t-2}\beta_2^k\|\mathbf{g}_{t-1-k}\|_2^2} = \sqrt{\sum_{k=0}^{t-2}\Big(\frac{\beta_1^2}{\beta_2}\Big)^k}\cdot\sqrt{\frac{V_{t-1}}{1-\beta_2}}.$$

将上一个公式与 $|\eta_{t-1}-\eta_t|$ 的表达式结合，并利用 $\sqrt{V_{t-1}} \leq A_{t-1}$，得到

$$\begin{aligned}
|\eta_{t-1}-\eta_t|\cdot\Big\|\sum_{s=1}^{t-1}\beta_1^{t-1-s}\mathbf{g}_s\Big\|_2 &\leq \gamma(1-\beta_1)\beta_1^{t-1}\cdot\frac{A_t-\beta_1 A_{t-1}}{A_{t-1}A_t}\cdot\sqrt{\frac{V_{t-1}}{1-\beta_2}}\cdot\sqrt{\sum_{k=0}^{t-2}\Big(\frac{\beta_1^2}{\beta_2}\Big)^k} \\
&\leq \gamma(1-\beta_1)\beta_1^{t-1}\cdot\frac{A_t-\beta_1 A_{t-1}}{A_t}\cdot\sqrt{\frac{1}{1-\beta_2}\sum_{k=0}^{t-2}\Big(\frac{\beta_1^2}{\beta_2}\Big)^k} \\
&\leq \gamma(1-\beta_1)\beta_1^{t-1}\cdot\frac{A_t-\beta_1 A_{t-1}}{A_t}\cdot\sqrt{\frac{\beta_2}{(\beta_2-\beta_1^2)(1-\beta_2)}},
\end{aligned}$$

从而完成证明。 □

引理 14。 令 $\beta \in (0,1)$，并令 $\mathcal{X} \subseteq \mathbb{R}^d$ 为非空闭凸集。考虑线性损失 $f_t(\mathbf{x}) = \langle \mathbf{g}_t, \mathbf{x}\rangle$，其中 $\mathbf{g}_t \in \mathbb{R}^d$。运行缩放的 折扣 *FTRL*

$$\mathbf{x}_{t+1} \in \arg\min_{\mathbf{x}\in\mathcal{X}}\left\{\sum_{s=1}^{t}\langle\beta^{-s}\mathbf{g}_s,\mathbf{x}\rangle+\psi_{t+1}(\mathbf{x})\right\}, \qquad \psi_{t+1}(\mathbf{x}) = \frac{1}{2\eta_t}\|\mathbf{x}\|_2^2,$$

其中 $\{\eta_t\}_{t\geq 0}$ 是任意正步长序列。令 $\mathbf{u}_1$、……、$\mathbf{u}_T \in \mathcal{X}$ 为任意比较序列。定义直径 $D_{\mathcal{X}} = \max_{\mathbf{x},\mathbf{y}\in\mathcal{X}}\|\mathbf{x}-\mathbf{y}\|_2$ 以及缩放后的累积梯度

$$\widetilde{\mathbf{g}}_{1:t} = \sum_{s=1}^{t}\beta^{-s}\mathbf{g}_s.$$

则以下动态遗憾分解成立：

$$\begin{aligned}
\sum_{t=1}^{T}\langle\mathbf{g}_t,\mathbf{x}_t-\mathbf{u}_t\rangle \leq \frac{\beta}{2\eta_0}\|\mathbf{u}_1\|_2^2 &+ \sum_{t=1}^{T}\beta^t\Lambda_t + \sum_{t=1}^{T}\beta^t\big(\psi_t(\mathbf{x}_{t+1})-\psi_{t+1}(\mathbf{x}_{t+1})\big) \\
&+ \beta\sum_{t=1}^{T-1}\sum_{s=1}^{t}\beta^{t-s}\langle\mathbf{g}_s,\mathbf{u}_{t+1}-\mathbf{u}_t\rangle + \beta\sum_{t=1}^{T-1}\beta^t\big(\psi_{t+1}(\mathbf{u}_{t+1})-\psi_t(\mathbf{u}_t)\big).
\end{aligned}$$

此外，单步稳定性项 Λ_t 可以通过以下两种方式中的<style id="2">任意一种 实例化：

选项 ***I***：对于每个 $t \in [T]$，$\Lambda_t = \frac{\eta_{t-1}}{2}\|\beta^{-t}\mathbf{g}_t\|_2^2 = \frac{1}{2}\eta_{t-1}\beta^{-2t}\|\mathbf{g}_t\|_2^2$

选项 ***II***：对于每个 $t \in [T]$，

$$\Lambda_t = \eta_t\|\beta^{-t}\mathbf{g}_t\|_2^2 + \min\left\{\frac{\eta_{t-1}}{2}\|\beta^{-t}\mathbf{g}_t\|_2^2,\ \min\big\{D_{\mathcal{X}},\ |\eta_{t-1}-\eta_t|\|\widetilde{\mathbf{g}}_{1:t-1}\|_2\big\}\cdot\|\beta^{-t}\mathbf{g}_t\|_2\right\}.$$

证明。 将引理 19 应用于缩放后的线性损失

$$\widetilde{f}_t(\mathbf{x}) = \langle\beta^{-t}\mathbf{g}_t,\mathbf{x}\rangle \qquad (t\in[T]),$$

其正则化项为 $\psi_{t+1}(\mathbf{x}) = \frac{1}{2\eta_t}\|\mathbf{x}\|_2^2$。这给出了关于单步量的标准分解

$$F_t(\mathbf{x}_t)-F_{t+1}(\mathbf{x}_{t+1})+\widetilde{f}_t(\mathbf{x}_t) \quad\text{and}\quad \psi_t(\mathbf{x}_{t+1})-\psi_{t+1}(\mathbf{x}_{t+1}).$$

接下来，我们将以两种不同的方式对单步量进行上界估计：

选项 I： 根据引理 20可得：

$$F_t(\mathbf{x}_t)-F_{t+1}(\mathbf{x}_{t+1})+\widetilde{f}_t(\mathbf{x}_t) \leq \frac{\eta_{t-1}}{2}\|\beta^{-t}\mathbf{g}_t\|_2^2+\psi_t(\mathbf{x}_{t+1})-\psi_{t+1}(\mathbf{x}_{t+1}).$$

选项 II： 或者，将引理 21 应用于重新缩放后的梯度 $\{\beta^{-t}\mathbf{g}_t\}_{t=1}^T$ 可得

$$F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + \widetilde{f}_t(\mathbf{x}_t) \le \Lambda_t + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}).$$

将上述分析与定理 1 结合，即完成证明。 □

引理 15. 固定 $\beta \in (0,1)$ 和 $\mu \ge 0$。设 $\mathcal{X} \subseteq \mathbb{R}^d$ 是一个非空闭凸集，且 $\{\mathbf{g}_t\}_{t\ge1} \subseteq \mathbb{R}^d$任意。定义复合代理损失 $\ell_t(\mathbf{x}) = \langle \mathbf{g}_t, \mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{x}\|_2^2$。考虑折扣 *FTRL* 更新

$$\mathbf{x}_{t+1} \in \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \sum_{s=1}^{t} \beta^{-s}\ell_s(\mathbf{x}) + \frac{1}{2\eta_t}\|\mathbf{x}\|_2^2 \right\},$$

这与下式等价

$$\mathbf{x}_{t+1} \in \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \sum_{s=1}^{t} \langle \beta^{-s}\mathbf{g}_s, \mathbf{x}\rangle + \frac{1}{2\eta'_t}\|\mathbf{x}\|_2^2, \right\}.$$

其中 $\frac{1}{\eta'_t} = \frac{1}{\eta_t} + \mu\sum_{s=1}^{t}\beta^{-s}$。对于任意比较序列 $\mathbf{u}_1, \dots, \mathbf{u}_T \in \mathcal{X}$，它保证：

$$\sum_{t=1}^{T}\ell_t(\mathbf{x}_t) - \sum_{t=1}^{T}\ell_t(\mathbf{u}_t) \le \frac{\beta}{2\eta'_0}\|\mathbf{u}_1\|_2^2 + \sum_{t=1}^{T}\frac{\beta^t\eta'_{t-1}}{2}\|\beta^{-t}\mathbf{g}_t\|_2^2 + \sum_{t=1}^{T}\beta^t\Big(\frac{1}{2\eta'_{t-1}}\|\mathbf{x}_{t+1}\|_2^2 - \frac{1}{2\eta'_t}\|\mathbf{x}_{t+1}\|_2^2\Big)$$
$$+ \beta\sum_{t=1}^{T-1}\sum_{s=1}^{t}\beta^{t-s}(\ell_s(\mathbf{u}_{t+1}) - \ell_s(\mathbf{u}_t)) + \beta\sum_{t=1}^{T-1}\beta^t\Big(\frac{1}{2\eta'_t}\|\mathbf{u}_{t+1}\|_2^2 - \frac{1}{2\eta'_{t-1}}\|\mathbf{u}_t\|_2^2\Big).$$

证明。 证明过程与引理 14相同步骤。我们从 D2D 归约（定理 1）出发，它引入路径项 $\beta\sum_{t=1}^{T-1}\sum_{s=1}^{t}\beta^{t-s}\big(\ell_s(\mathbf{u}^{t+1}) - \ell_s(\mathbf{u}_t)\big)$。其余各项利用引理20 以及定义进行界定，并注意到正则化项是 $1/(2\eta'_{t-1})$-强凸的。至此，证明完成。 □

D.6 定理5的证明

证明。 在整个证明中，我们使用 β_1 作为 Adam 中的一阶矩折扣因子以及算法 2 中的折扣因子，并使用 β_2 作为二阶矩项。

O2NC 归约。 根据引理8，我们有

$$\mathbb{E}\|\nabla F(\bar{\mathbf{x}})\|_c \le \frac{F^*}{DT} + \frac{2G+\sigma}{(1-\beta_1)T} + \sigma\sqrt{1-\beta_1} + \frac{12cD^2}{(1-\beta_1)^2}$$
$$+ \frac{1}{DT}\mathbb{E}\Big[\text{D-Reg}_T(\mathbf{u}_{1:T}) + \beta_1\sum_{t=1}^{T-1}\sum_{s=1}^{t}\beta_1^{t-s}\big(\ell_s(\mathbf{u}_t) - \ell_s(\mathbf{u}_{t+1})\big)\Big], \tag{30}$$

其中 $\ell_t(\Delta) = \langle \mathbf{g}_t, \Delta\rangle$ 且 $\text{D-Reg}_T(\mathbf{u}_{1:T}) = \sum_{t=1}^{T}\langle \mathbf{g}_t, \Delta_t - \mathbf{u}_t\rangle$。

使用 D2D 归约。 我们将引理 14应用于在 $\mathcal{D} = \{\Delta$ 上运行的折扣 FTRL：$\|\Delta\|_2 \le D\}$，决策为 Δ_t，梯度为$\mathbf{g}_t$，比较序列为$\mathbf{u}_t$ （见引理 8）。回顾步长（算法 2 ，采用 clippedAdam）：

$$V_0 = 0, \qquad V_t = (1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\|\mathbf{g}_s\|_2^2, \qquad \alpha_t = \frac{1}{\nu + \sqrt{V_t}}, \qquad \eta_t = \gamma(1-\beta_1)\beta_1^t\alpha_t.$$

由于 $\|\mathbf{u}_t\|_2 = D$ 对所有 t 根据构造，（30）中的路径项抵消了 $\beta_1\sum_{t=1}^{T-1}\sum_{s=1}^{t}\beta_1^{t-s}\langle \mathbf{g}_s, \mathbf{u}_{t+1} - \mathbf{u}_t\rangle$（由引理14引入），从而得到该分解

$$\text{D-Reg}_T(\mathbf{u}_{1:T}) + \beta_1\sum_{t=1}^{T-1}\sum_{s=1}^{t}\beta_1^{t-s}\big(\ell_s(\mathbf{u}_t) - \ell_s(\mathbf{u}_{t+1})\big)$$

$$\leq \underbrace{\frac{1-\beta_1}{2\beta_1}\gamma\sum_{t=1}^{T}\alpha_{t-1}\|\mathbf{g}_t\|_2^2+\frac{\beta_1}{2\gamma(1-\beta_1)\alpha_0}\|\mathbf{u}_1\|_2^2}_{\text{TERM-A}}+\underbrace{\beta_1\sum_{t=1}^{T-1}\Big(\frac{\beta_1^t}{2\eta_t}\|\mathbf{u}_{t+1}\|_2^2-\frac{\beta_1^t}{2\eta_{t-1}}\|\mathbf{u}_t\|_2^2\Big)}_{\text{TERM-B}}$$
$$+\underbrace{\sum_{t=1}^{T}\Big(\frac{\beta_1^t}{2\eta_{t-1}}\|\Delta_{t+1}\|_2^2-\frac{\beta_1^t}{2\eta_t}\|\Delta_{t+1}\|_2^2\Big)}_{\text{TERM-C}}. \tag{31}$$

项 A。 我们首先对 $\sum_{t=1}^{T}\alpha_{t-1}\|\mathbf{g}_t\|_2^2$ 给出上界，借助引理 9。对引理 9 取

$$a_t=\|\mathbf{g}_t\|_2^2,\qquad \epsilon=\nu,\qquad \beta=\beta_2,\qquad V_t=\beta_2 V_{t-1}+(1-\beta_2)a_t.$$

令 $K\geq \frac{T}{(1-\beta_2)\ln 2}+1$，则

$$\sum_{t=1}^{T}\alpha_{t-1}\|\mathbf{g}_t\|_2^2=\sum_{t=1}^{T}\frac{\|\mathbf{g}_t\|_2^2}{\nu+\sqrt{V_{t-1}}}\leq K\frac{A}{\nu}+\frac{2}{\sqrt{(1-\beta_2)2^{-1/\beta_2}}}\sqrt{K\sum_{t=1}^{T}\|\mathbf{g}_t\|_2^2}, \tag{32}$$

其中 A 是 $\|\mathbf{g}_t\|_2^2$的任意几乎处处上界。在假设 1 – 4下，取 $A=G^2$。对 (32) 取期望并使用詹森不等式，$\mathbb{E}\big[\sqrt{\sum_{t=1}^{T}\|\mathbf{g}_t\|_2^2}\big]\leq\sqrt{\mathbb{E}\sum_{t=1}^{T}\|\mathbf{g}_t\|_2^2}$，可得

$$\mathbb{E}\sum_{t=1}^{T}\alpha_{t-1}\|\mathbf{g}_t\|_2^2\leq K\frac{G^2}{\nu}+\frac{2}{\sqrt{(1-\beta_2)2^{-1/\beta_2}}}\sqrt{KT}G,$$

其中 $K=\frac{T}{(1-\beta_2)\ln 2}+1$。令 T满足 $\frac{T}{(1-\beta_2)\ln 2}\geq 1$，则 $K\leq\frac{2T}{(1-\beta_2)\ln 2}$，因此，

$$\begin{aligned}\mathbb{E}\sum_{t=1}^{T}\alpha_{t-1}\|\mathbf{g}_t\|_2^2&\leq\frac{2T(1-\beta_2)}{\ln 2}\cdot\frac{G^2}{\nu}+\frac{2}{\sqrt{(1-\beta_2)2^{-1/\beta_2}}}\cdot\sqrt{\frac{2T^2(1-\beta_2)}{\ln 2}}G\\&=\frac{2T(1-\beta_2)}{\ln 2}\cdot\frac{G^2}{\nu}+\frac{2\sqrt{2/\ln 2}}{\sqrt{2^{-1/\beta_2}}}TG.\end{aligned}$$

由于 $\|\mathbf{u}_1\|_2=D$ 且 $\alpha_0=\frac{1}{\nu}$，将上述不等式代入 TERM-A 可得

$$\mathbb{E}[\text{TERM-A}]\leq\frac{1-\beta_1}{2\beta_1}\gamma\left(\frac{2T(1-\beta_2)}{\ln 2}\cdot\frac{G^2}{\nu}+\frac{2\sqrt{2/\ln 2}}{\sqrt{2^{-1/\beta_2}}}TG\right)+\frac{\beta_1\nu}{2\gamma(1-\beta_1)}D^2. \tag{33}$$

项 B。 由于对所有 t，$\|\mathbf{u}_t\|_2=D$，

$$\text{TERM-B}=\beta_1\sum_{t=1}^{T-1}\Big(\frac{\beta_1^t}{2\eta_t}D^2-\frac{\beta_1^t}{2\eta_{t-1}}D^2\Big)=\frac{\beta_1 D^2}{2}\sum_{t=1}^{T-1}\beta_1^t\Big(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\Big).$$

根据引理12，取 $\mu=0$ 和 $\epsilon=\nu$，

$$\text{TERM-B}\leq\frac{\beta_1 D^2}{2}\left(\frac{\sqrt{V_{T-1}}}{\gamma(1-\beta_1)}+\frac{1}{\gamma}\sum_{t=1}^{T-2}\sqrt{V_t}+\frac{(T-1)\nu}{\gamma}\right). \tag{34}$$

接下来，我们对指数移动平均项的期望进行界定。首先，根据詹森，$\mathbb{E}[\sqrt{V_t}]\leq\sqrt{\mathbb{E}[V_t]}$。此外，

$$\mathbb{E}[V_t]\leq(1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\mathbb{E}\|\mathbf{g}_s\|_2^2\leq G^2,$$

因此，$\mathbb{E}[\sqrt{V_t}] \le G$ 对所有 t。此外，利用柯西–施瓦茨不等式和引理 16，

$$\mathbb{E}\sum_{t=1}^{T-2}\sqrt{V_t} \le \sqrt{T}\cdot\sqrt{\mathbb{E}\sum_{t=1}^{T}V_t} = \sqrt{T}\cdot\sqrt{\mathbb{E}\sum_{s=1}^{T}(1-\beta_2^{T+1-s})\|\mathbf{g}_s\|_2^2} \le \sqrt{T}\cdot\sqrt{\sum_{s=1}^{T}\mathbb{E}\|\mathbf{g}_s\|_2^2} \le TG.$$

对（34）取期望可得

$$\mathbb{E}[\textsc{Term-B}] \le \frac{\beta_1 D^2}{2}\left(\frac{G}{\gamma(1-\beta_1)} + \frac{TG}{\gamma} + \frac{(T-1)\nu}{\gamma}\right).$$

项 C。 利用 $\eta_t = \gamma(1-\beta_1)\beta_1^t\alpha_t$ 和 $\|\Delta_{t+1}\|_2 \le D$，

$$\begin{aligned}\textsc{Term-C} &= \sum_{t=1}^{T}\Big(\frac{\beta_1^t}{2\eta_{t-1}} - \frac{\beta_1^t}{2\eta_t}\Big)\|\Delta_{t+1}\|_2^2 = \frac{1}{2\gamma(1-\beta_1)}\sum_{t=1}^{T}\Big(\frac{\beta_1}{\alpha_{t-1}} - \frac{1}{\alpha_t}\Big)\|\Delta_{t+1}\|_2^2 \\ &\le \frac{D^2}{2\gamma(1-\beta_1)}\sum_{t=1}^{T}\Big[\frac{\beta_1}{\alpha_{t-1}} - \frac{1}{\alpha_t}\Big]_+.\end{aligned}$$

由引理 11，$\big[\frac{\beta_1}{\alpha_{t-1}} - \frac{1}{\alpha_t}\big]_+ \le [\beta_1 - \sqrt{\beta_2}]_+\sqrt{V_{t-1}}$，因此

$$\textsc{Term-C} \le \frac{D^2}{2\gamma(1-\beta_1)}[\beta_1 - \sqrt{\beta_2}]_+\sum_{t=1}^{T}\sqrt{V_{t-1}}.$$

取期望并利用 $\mathbb{E}\sum_{t=1}^{T}\sqrt{V_{t-1}} \le TG$ 可得

$$\mathbb{E}[\textsc{Term-C}] \le \frac{D^2}{2\gamma(1-\beta_1)}[\beta_1 - \sqrt{\beta_2}]_+ TG.$$

在条件 $\beta_2 \ge \beta_1^4$ 下，我们有 $\sqrt{\beta_2} \ge \beta_1^2$，因此

$$[\beta_1 - \sqrt{\beta_2}]_+ \le \beta_1 - \beta_1^2 = \beta_1(1-\beta_1).$$

代入上文 $\mathbb{E}[\textsc{Term-C}]$ 的半步可得

$$\mathbb{E}[\textsc{Term-C}] \le \frac{\beta_1 D^2}{2\gamma}TG.$$

组合并调整参数。 将（30）与（31）合并，得到

$$\begin{aligned}\mathbb{E}\|\nabla F(\bar{\mathbf{x}})\|_c \le{}& \frac{F^*}{DT} + \frac{2G+\sigma}{(1-\beta_1)T} + \sigma\sqrt{1-\beta_1} + \frac{12cD^2}{(1-\beta_1)^2} \\ &+ \frac{1}{DT}\mathbb{E}[\textsc{Term-A} + \textsc{Term-B} + \textsc{Term-C}].\end{aligned} \tag{35}$$

现在按照定理中的方式调整参数。设

$$\beta_1 = 1 - \Big(\frac{\epsilon}{16(G+\sigma)}\Big)^2, \qquad D = \frac{(1-\beta_1)\sqrt{\epsilon}}{\sqrt{48c}}, \qquad \gamma = \frac{\beta_1 D}{\sqrt{1-\beta_1}}, \qquad \beta_2 \ge \max\Big\{1 - \frac{\nu}{G+\sigma}, \beta_1^4\Big\}.$$

然后 $\frac{12cD^2}{(1-\beta_1)^2} = \epsilon/4$ 和 $\sigma\sqrt{1-\beta_1} \le (G+\sigma)\sqrt{1-\beta_1} = \epsilon/16$。接下来，我们对 FTRL 的贡献项 $\frac{1}{DT}\mathbb{E}[\textsc{Term-A} + \textsc{Term-B} + \textsc{Term-C}]$ 进行上界估计。首先，将（33）除以 DT，并代入 $\gamma = \beta_1 D/\sqrt{1-\beta_1}$：

$$\frac{1}{DT}\mathbb{E}[\textsc{Term-A}] \le \frac{\sqrt{1-\beta_1}}{4}\cdot\left(\frac{2(1-\beta_2)}{\ln 2}\cdot\frac{G^2}{\nu} + \frac{2\sqrt{2/\ln 2}}{\sqrt{2^{-1/\beta_2}}}G\right) + \frac{\nu}{2T\sqrt{1-\beta_1}}.$$

利用 $\beta_2 \geq 1-\nu/(G+\sigma)$ 得到 $(1-\beta_2)\frac{G^2}{\nu} \leq G$，因此

$$\frac{1}{DT}\mathbb{E}[\text{TERM-A}] \leq \frac{\sqrt{1-\beta_1}}{4} G\left(\frac{2}{\ln 2}+\frac{2\sqrt{2/\ln 2}}{\sqrt{2^{-1/\beta_2}}}\right)+\frac{\nu}{2T\sqrt{1-\beta_1}}. \tag{36}$$

类似地，将 $\mathbb{E}[\text{TERM-B}]$ 除以 DT 并代入 $\gamma=\beta_1 D/\sqrt{1-\beta_1}$，可得

$$\frac{1}{DT}\mathbb{E}[\text{TERM-B}] \leq \frac{G}{2T\sqrt{1-\beta_1}}+\frac{G\sqrt{1-\beta_1}}{2}+\frac{\nu\sqrt{1-\beta_1}}{2}. \tag{37}$$

将 $\mathbb{E}[\text{TERM-C}]$ 除以 DT 并代入 $\gamma=\beta_1 D/\sqrt{1-\beta_1}$，可得

$$\frac{1}{DT}\mathbb{E}[\text{TERM-C}] \leq \frac{G\sqrt{1-\beta_1}}{2}. \tag{38}$$

此外，我们仅考虑非平凡情形 $\epsilon \leq G+\sigma$（否则 $\mathbb{E}\|\nabla F(\bar{\mathbf{x}})\|_c \leq G+\sigma \leq \epsilon$ 显然成立），因此 $\beta_1 \geq 1-(1/16)^2 = 255/256$，进而 $\beta_2 \geq \beta_1^4 \geq (255/256)^4$。因此，

$$\frac{1}{\sqrt{2^{-1/\beta_2}}} = 2^{\frac{1}{2\beta_2}} \leq 2^{\frac{1}{2\beta_1^4}} < \frac{3}{2}.$$

将其代入 (36) 并利用 $\nu \leq G+\sigma$ 可得

$$\frac{1}{DT}\mathbb{E}[\text{TERM-A}] \leq \frac{\epsilon}{64}\left(\frac{2}{\ln 2}+3\sqrt{\frac{2}{\ln 2}}\right)+\frac{\nu}{2T\sqrt{1-\beta_1}} < \frac{\epsilon}{4}+\frac{\nu}{2T\sqrt{1-\beta_1}}.$$

此外，由 (37)–(38) 以及 $\nu \leq G+\sigma$，

$$\frac{1}{DT}\mathbb{E}[\text{TERM-B}+\text{TERM-C}] \leq \frac{G+\sigma}{2T\sqrt{1-\beta_1}}+\frac{\epsilon}{16}+\frac{\epsilon}{32}.$$

结合上述两个公式可得

$$\frac{1}{DT}\mathbb{E}[\text{TERM-A}+\text{TERM-B}+\text{TERM-C}] \leq \frac{11}{32}\epsilon+\frac{G+\sigma+\nu}{2T\sqrt{1-\beta_1}}. \tag{39}$$

最后，选择 T 使其满足

$$T \geq \max\left\{(1-\beta_1)^{-1}\cdot\max\left\{\frac{16F^*\sqrt{48c}}{\epsilon^{3/2}}, \frac{16(G+\sigma)}{\epsilon}\right\}, \frac{\ln 2}{1-\beta_2}\right\}.$$

Then

$$\frac{F^*}{DT} \leq \frac{\epsilon}{16}, \qquad \frac{2G+\sigma}{(1-\beta_1)T} \leq \frac{(2G+\sigma)\epsilon}{16(G+\sigma)} \leq \frac{3\epsilon}{16},$$

并且，由于 $\sqrt{1-\beta_1}=\epsilon/(16(G+\sigma))$ 且 $\nu \leq G+\sigma$，

$$\frac{G+\sigma+\nu}{2T\sqrt{1-\beta_1}} \leq \frac{2(G+\sigma)}{2T}\cdot\frac{16(G+\sigma)}{\epsilon} = \frac{16(G+\sigma)^2}{T\epsilon} \leq \frac{\epsilon}{32}.$$

Putt将所有内容代回（35）并使用（39），我们得到 $\mathbb{E}\|\nabla F(\bar{\mathbf{x}})\|_c \leq \epsilon$，完成证明 f. □

D.7 定理 6 的证明

证明。 在整个证明过程中，我们用 β_1 表示 Adam 中的一阶矩折扣因子，以及算法 2中的折扣因子；用 β_2 表示二阶矩项。

O2NC 转换。应用引理 8，取 $\beta = \beta_1$。令 $\ell_t(\Delta) = \langle \mathbf{g}_t, \Delta \rangle$ 和 $\text{D-REG}_T(\mathbf{u}_{1:T}) = \sum_{t=1}^T \langle \mathbf{g}_t, \Delta_t - \mathbf{u}_t \rangle$，即可得到

$$\mathbb{E}\|\nabla F(\bar{\mathbf{x}})\|_c \le \frac{F^*}{DT} + \frac{2G+\sigma}{(1-\beta_1)T} + \sigma\sqrt{1-\beta_1} + \frac{12cD^2}{(1-\beta_1)^2} + \frac{1}{DT}\mathbb{E}\Big[\text{D-REG}_T(\mathbf{u}_{1:T}) + \beta_1 \sum_{t=1}^{T-1}\sum_{s=1}^{t} \beta_1^{t-s}\big(\ell_s(\mathbf{u}_t) - \ell_s(\mathbf{u}_{t+1})\big)\Big].$$

回顾一下，$\|\mathbf{u}_t\|_2 = D$ 对所有 t 根据引理 8中的定义成立。

使用 D2D 归约。我们将引理14 应用于在裁剪域上运行的折扣 FTRL$\mathcal{D} = \{\Delta\colon \|\Delta\|_2 \le D\}$，其中决策记为 Δ_t，比较器记为 $\mathbf{u}_t$。我们在引理14中选择选项 II，即

$$\Lambda_t = \eta_t \|\beta_1^{-t}\mathbf{g}_t\|_2^2 + \min\Big\{\frac{\eta_{t-1}}{2}\|\beta_1^{-t}\mathbf{g}_t\|_2^2,\ \min\{2D, |\eta_{t-1}-\eta_t|\|\widetilde{\mathbf{g}}_{1:t-1}\|_2\}\cdot\|\beta_1^{-t}\mathbf{g}_t\|_2\Big\},$$

其中 $\widetilde{\mathbf{g}}_{1:t} = \sum_{s=1}^t \beta_1^{-s}\mathbf{g}_s$。利用步长

$$V_0 = 0, \qquad V_t = (1-\beta_2)\sum_{s=1}^t \beta_2^{t-s}\|\mathbf{g}_s\|_2^2, \qquad \alpha_t = \frac{1}{\nu + \sqrt{V_t}}, \qquad \eta_t = \gamma(1-\beta_1)\beta_1^t\alpha_t,$$

O2NC 转换中的路径项与 $\beta_1 \sum_{t=1}^{T-1}\sum_{s=1}^t \beta_1^{t-s}\langle \mathbf{g}_s, \mathbf{u}_{t+1} - \mathbf{u}_t\rangle$ 在引理 14中的路径项相消。我们得到分解

$$\begin{aligned}
&\text{D-REG}_T(\mathbf{u}_{1:T}) + \beta_1 \sum_{t=1}^{T-1}\sum_{s=1}^{t} \beta_1^{t-s}\big(\ell_s(\mathbf{u}_t) - \ell_s(\mathbf{u}_{t+1})\big) \\
&\le \underbrace{\sum_{t=1}^T \beta_1^t \eta_t \|\beta_1^{-t}\mathbf{g}_t\|_2^2 + \frac{\beta_1}{2\gamma(1-\beta_1)\alpha_0}\|\mathbf{u}_1\|_2^2}_{\text{TERM-A}} \\
&\quad + \underbrace{\sum_{t=1}^T \beta_1^t \min\Big\{\frac{\eta_{t-1}}{2}\|\beta_1^{-t}\mathbf{g}_t\|_2^2,\ \min\{D, |\eta_{t-1}-\eta_t|\|\widetilde{\mathbf{g}}_{1:t-1}\|_2\}\cdot\|\beta_1^{-t}\mathbf{g}_t\|_2\Big\}}_{\text{TERM-B}} \\
&\quad + \underbrace{\sum_{t=1}^T \Big(\frac{\beta_1^t}{2\eta_{t-1}}\|\Delta_{t+1}\|_2^2 - \frac{\beta_1^t}{2\eta_t}\|\Delta_{t+1}\|_2^2\Big)}_{\text{TERM-C}} + \underbrace{\beta_1 \sum_{t=1}^{T-1}\Big(\frac{\beta_1^t}{2\eta_t}\|\mathbf{u}_{t+1}\|_2^2 - \frac{\beta_1^t}{2\eta_{t-1}}\|\mathbf{u}_t\|_2^2\Big)}_{\text{TERM-D}}.
\end{aligned} \tag{40}$$

项 A。由于我们有

$$\sum_{t=1}^T \beta_1^t\eta_t\|\beta_1^{-t}\mathbf{g}_t\|_2^2 = \gamma(1-\beta_1)\sum_{t=1}^T \alpha_t\|\mathbf{g}_t\|_2^2 = \gamma(1-\beta_1)\sum_{t=1}^T \frac{\|\mathbf{g}_t\|_2^2}{\nu+\sqrt{V_t}}.$$

应用引理 10，取 $a_t = \|\mathbf{g}_t\|_2^2$，$\beta = \beta_2$，以及 $\epsilon = \nu$，我们得到

$$\sum_{t=1}^T \alpha_t\|\mathbf{g}_t\|_2^2 \le \frac{2}{\sqrt{(1-\beta_2)2^{-1/\beta_2}}}\sqrt{\Big(\frac{T(1-\beta_2)}{\ln 2}+1\Big)\sum_{t=1}^T\|\mathbf{g}_t\|_2^2}.$$

取期望，并结合詹森不等式以及 $\mathbb{E}\sum_{t=1}^T\|\mathbf{g}_t\|_2^2 \le TG^2$，得到

$$\mathbb{E}\left[\sum_{t=1}^T \beta_1^t\eta_t\|\beta_1^{-t}\mathbf{g}_t\|_2^2\right] \le \gamma(1-\beta_1)\cdot\frac{2}{\sqrt{(1-\beta_2)2^{-1/\beta_2}}}\sqrt{\frac{T(1-\beta_2)}{\ln 2}+1}\sqrt{T}G.$$

假设 $T \geq (1-\beta_2)^{-1}\ln 2$，则 $T_{(1-\beta_2)\ln 2} + 1 \leq 2T_{(1-\beta_2)\ln 2}$，因此代入 $\gamma = \frac{\beta_1 D}{\sqrt{1-\beta_1}}$ 并利用 $\beta_1 \leq 1$ 后，上述不等式给出

$$\frac{1}{DT}\mathbb{E}\left[\sum_{t=1}^{T}\beta_1^t\eta_t\|\beta_1^{-t}\mathbf{g}_t\|_2^2\right] \leq \frac{2^{3/2+1/(2\beta_2)}}{\sqrt{\ln 2}}G\sqrt{1-\beta_1},$$

根据定义：

$$\frac{1}{DT}\mathbb{E}\left[\textsc{Term-A}\right] \leq \frac{2^{3/2+1/(2\beta_2)}}{\sqrt{\ln 2}}G\sqrt{1-\beta_1} + \frac{\nu}{2T\sqrt{1-\beta_1}}.$$

项 B。我们利用引理 13 来界定偏差部分，$\mu = 0$。令 $A_t = \nu + \sqrt{V_t}$ 且 $\|\mathbf{g}_t\| \leq G$，则有：

$$\beta_1^t|\eta_{t-1} - \eta_t| \cdot \|\widetilde{\mathbf{g}}_{1:t-1}\|_2 \cdot \|\beta_1^{-t}\mathbf{g}_t\|_2 \;\leq\; \gamma(1-\beta_1)\cdot\frac{A_t - \beta_1 A_{t-1}}{A_t}\cdot\sqrt{\frac{\beta_2}{(\beta_2-\beta_1^2)(1-\beta_2)}}\cdot G.$$

S对 t 求和，并利用不等式 $1 - x \leq \ln(1/x)$（其中 $x > 0$ 且 $x = \beta_1 A_{t-1}/A_t$），可得 s

$$\sum_{t=1}^{T}\frac{A_t - \beta_1 A_{t-1}}{A_t} = \sum_{t=1}^{T}\left(1 - \frac{\beta_1 A_{t-1}}{A_t}\right) \leq \sum_{t=1}^{T}\ln\left(\frac{A_t}{\beta_1 A_{t-1}}\right) = \ln\left(\frac{A_T}{A_0}\right) - T\ln\beta_1.$$

因此，

$$\textsc{Term-B} \leq \gamma(1-\beta_1)G\Big(\ln\left(\frac{A_T}{A_0}\right) - T\ln\beta_1\Big)\cdot\sqrt{\frac{\beta_2}{(\beta_2-\beta_1^2)(1-\beta_2)}}. \tag{41}$$

接下来，我们利用间隔条件对依赖于 β_2 的因子给出上界。令 $\beta_2^\star = \frac{1+\beta_1^2}{2}$，并回忆 $\beta_2 \in [\beta_1^2 + m, 1-m]$，其中 $m = \frac{1-\rho}{2}(1-\beta_1^2)$。于是 $|\beta_2 - \beta_2^\star| \leq \frac{\rho}{2}(1-\beta_1^2)$，因此

$$(\beta_2-\beta_1^2)(1-\beta_2) = \frac{(1-\beta_1^2)^2}{4} - (\beta_2-\beta_2^\star)^2 \geq \frac{(1-\beta_1^2)^2}{4}(1-\rho^2).$$

由于 $\beta_2 \leq 1$，我们得到

$$\sqrt{\frac{\beta_2}{(\beta_2-\beta_1^2)(1-\beta_2)}} \leq \frac{2}{1-\beta_1^2}\cdot\frac{1}{\sqrt{1-\rho^2}}. \tag{42}$$

此外，$\sqrt{V_t} \leq G$ 意味着 $A_T \leq \nu + G$ 和 $A_0 = \nu$，因此

$$\ln\left(\frac{A_T}{A_0}\right) \leq \ln\left(1 + \frac{G}{\nu}\right). \tag{43}$$

另外，$-\ln\beta_1 \leq \frac{1-\beta_1}{\beta_1}$。将（42）–（43）代入（41），并利用 $\gamma = \frac{\beta_1 D}{\sqrt{1-\beta_1}}$ 可得

$$\frac{1}{DT}\mathbb{E}[\textsc{Term-B}] \leq \frac{2G}{T\sqrt{1-\beta_1}}\cdot\frac{1}{\sqrt{1-\rho^2}}\ln\left(1+\frac{G}{\nu}\right) + \frac{2G}{\sqrt{1-\rho^2}}\sqrt{1-\beta_1}.$$

项 C。由于 $\beta_2 > \beta_1^2$，引理11 给出 $\textsc{Term-C} \leq 0$。

术语-D. 我们有 $\frac{1}{DT}\mathbb{E}[\textsc{Term-D}] \leq G\sqrt{1-\beta_1}$。

组合并调整参数. 结合上述界限，我们得到：

$$\begin{aligned}\mathbb{E}\|\nabla F(\bar{\mathbf{x}})\|_c \le \frac{F^*}{DT} + \frac{2G+\sigma}{(1-\beta_1)T} + \sigma\sqrt{1-\beta_1} + \frac{12cD^2}{(1-\beta_1)^2} \\ + \frac{2^{3/2+1/(2\beta_2)}}{\sqrt{\ln 2}} G\sqrt{1-\beta_1} + G\sqrt{1-\beta_1} + \frac{2G}{\sqrt{1-\rho^2}}\sqrt{1-\beta_1} \\ + \frac{2G}{T\sqrt{1-\beta_1}} \cdot \frac{1}{\sqrt{1-\rho^2}} \ln\left(1+\frac{G}{\nu}\right) + \frac{\nu}{2T\sqrt{1-\beta_1}}. \end{aligned} \tag{44}$$

现在令

$$D = \frac{(1-\beta_1)\sqrt{\epsilon}}{\sqrt{48c}}, \qquad \gamma = \frac{\beta_1 D}{\sqrt{1-\beta_1}}, \qquad T \ge \frac{\ln 2}{1-\beta_2}.$$

Then

$$\frac{12cD^2}{(1-\beta_1)^2} = 12c \cdot \frac{(1-\beta_1)^2\epsilon}{48c} \cdot \frac{1}{(1-\beta_1)^2} = \frac{\epsilon}{4}.$$

接下来，选择 β_1 使得

$$\sqrt{1-\beta_1} \le \frac{\epsilon\sqrt{1-\rho^2}}{64(G+\sigma)}.$$

Then

$$\sigma\sqrt{1-\beta_1} \le (G+\sigma)\sqrt{1-\beta_1} \le \frac{\epsilon}{64}, \qquad \frac{2(G+\sigma)}{\sqrt{1-\rho^2}}\sqrt{1-\beta_1} \le \frac{\epsilon}{32}, \qquad (G+\sigma)\sqrt{1-\beta_1} \le \frac{\epsilon}{64}.$$

此外，由于 $\beta_2 > \beta_1^2$ 且 $\beta_1 \ge 15/16$（只要 $\epsilon \le 16(G+\sigma)/\sqrt{1-\rho^2}$ 即可推出），我们有 $2^{1/(2\beta_2)} \le 3/2$，从而 $2^{3/2+1/(2\beta_2)}/\sqrt{\ln 2} \le 6$，因此

$$\frac{2^{3/2+1/(2\beta_2)}}{\sqrt{\ln 2}}(G+\sigma)\sqrt{1-\beta_1} \le \frac{6\epsilon}{64} = \frac{3\epsilon}{32}.$$

最后，选择 T 使得

$$T \ge \frac{1}{1-\beta_1} \cdot \max\left\{\frac{32F^*\sqrt{c}}{\epsilon^{3/2}}, \frac{16(G+\sigma)}{\epsilon}\right\}, \qquad T \ge \frac{32G}{\epsilon\sqrt{1-\beta_1}\sqrt{1-\rho^2}} \cdot \ln\left(1+\frac{G}{\nu}\right).$$

于是我们有

$$\frac{F^*}{DT} \le \frac{\sqrt{48}}{32}\epsilon < \frac{\epsilon}{4}, \qquad \frac{2G+\sigma}{(1-\beta_1)T} \le \frac{\epsilon}{8}, \qquad \frac{2(G+\sigma)}{T\sqrt{1-\beta_1}} \cdot \frac{1}{\sqrt{1-\rho^2}} \ln\left(1+\frac{G}{\nu}\right) \le \frac{\epsilon}{16}.$$

此外，利用 $\nu \le G+\sigma$ 和 $T \ge 16(G+\sigma)\epsilon(1-\beta_1)$，

$$\frac{\nu}{2T\sqrt{1-\beta_1}} \le \frac{G+\sigma}{2T\sqrt{1-\beta_1}} \le \frac{\epsilon\sqrt{1-\beta_1}}{32} \le \frac{\epsilon}{128}.$$

将式（44）中的所有界合并，得到 $\mathbb{E}\|\nabla F(\bar{\mathbf{x}})\|_c \le \epsilon$ 并完成证明。 □

D.8 定理 7的证明

定理 7 的证明。 在整个证明过程中，我们使用 β_1 表示 Adam 中的一阶矩折扣因子和算法 2 中的折扣因子，并使用 β_2 表示二阶矩项。

O2NC 转换。 应用引理 7，其中 $\beta = \beta_1$ 以及其中的比较序列 $\{\mathbf{u}_t\}_{t=1}^T$。令 $\ell_t(\Delta) = \langle \mathbf{g}_t, \Delta\rangle + \frac{\mu}{2}\|\Delta\|_2^2$ 以及 $\text{D-REG}_T(\mathbf{u}_{1:T}) = \sum_{t=1}^T \big(\ell_t(\Delta_t) - \ell_t(\mathbf{u}_t)\big)$，可得

$$\mathbb{E}\|\nabla F(\bar{\mathbf{x}})\|_c \le \frac{F^*}{DT} + \frac{2G+\sigma}{(1-\beta_1)T} + \sigma\sqrt{1-\beta_1} + \frac{12cD^2}{(1-\beta_1)^2} + \frac{1}{DT}\mathbb{E}\Big[\text{D-REG}_T(\mathbf{u}_{1:T}) + \beta_1 \sum_{t=1}^{T-1}\sum_{s=1}^{t} \beta_1^{t-s}\big(\ell_s(\mathbf{u}_t) - \ell_s(\mathbf{u}_{t+1})\big)\Big]. \tag{45}$$

使用 D2D 归约。 令

$$V_0 = 0, \qquad V_t = (1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\|\mathbf{g}_s\|_2^2, \qquad \alpha_t = \frac{1}{\nu + \gamma\mu(1-\beta_1^t) + \sqrt{V_t}}, \qquad \eta_t = \gamma(1-\beta_1)\beta_1^t\alpha_t.$$

我们将引理15 应用于上述步长 $\{\Delta_t\}_{t=1}^T$ 下的折扣 FTRL 迭代序列 $\{\eta_t\}$ 以及引理 $\{\mathbf{u}_t\}$ 中的比较序列7。引理15 中的路径项抵消了式（45）中的负路径项，由此得到分解

$$\begin{aligned}
&\text{D-REG}_T(\mathbf{u}_{1:T}) + \beta_1 \sum_{t=1}^{T-1}\sum_{s=1}^{t} \beta_1^{t-s}\big(\ell_s(\mathbf{u}_t) - \ell_s(\mathbf{u}_{t+1})\big) \\
&\le \underbrace{\frac{1-\beta_1}{2\beta_1}\gamma\sum_{t=1}^{T}\alpha_{t-1}\|\mathbf{g}_t\|_2^2 + \frac{\beta_1}{2\gamma(1-\beta_1)\alpha_0}\|\mathbf{u}_1\|_2^2}_{\text{TERM-A}} + \underbrace{\beta_1\sum_{t=1}^{T-1}\Big(\frac{\beta_1^t}{2\eta_t}\|\mathbf{u}_{t+1}\|_2^2 - \frac{\beta_1^t}{2\eta_{t-1}}\|\mathbf{u}_t\|_2^2\Big)}_{\text{TERM-B}} \\
&\quad + \underbrace{\sum_{t=1}^{T}\Big(\frac{\beta_1^t}{2\eta_{t-1}}\|\Delta_{t+1}\|_2^2 - \frac{\beta_1^t}{2\eta_t}\|\Delta_{t+1}\|_2^2\Big)}_{\text{TERM-C}}.
\end{aligned} \tag{46}$$

项 A。 由于 $\gamma\mu(1-\beta_1^t) \ge 0$，我们有 $\alpha_{t-1} \le \frac{1}{\nu+\sqrt{V_{t-1}}}$。因此

$$\sum_{t=1}^{T}\alpha_{t-1}\|\mathbf{g}_t\|_2^2 \le \sum_{t=1}^{T}\frac{\|\mathbf{g}_t\|_2^2}{\nu+\sqrt{V_{t-1}}}.$$

应用引理 9，其中 $a_t = \|\mathbf{g}_t\|_2^2$, $\epsilon = \nu$, $\beta = \beta_2$。根据定理假设，对所有 t，$\|\mathbf{g}_t\|_2 \le G$，我们取 $A = \max_t \|\mathbf{g}_t\|_2^2 \le G^2$。令 $K = \frac{T}{(1-\beta_2)\ln 2} + 1$。于是

$$\sum_{t=1}^{T}\alpha_{t-1}\|\mathbf{g}_t\|_2^2 \le K\frac{G^2}{\nu} + \frac{2}{\sqrt{(1-\beta_2)2^{-1/\beta_2}}}\sqrt{K\sum_{t=1}^{T}\|\mathbf{g}_t\|_2^2}.$$

取期望，并结合詹森不等式以及 $\mathbb{E}\sum_{t=1}^T\|\mathbf{g}_t\|_2^2 \le TG^2$，得到

$$\mathbb{E}\sum_{t=1}^{T}\alpha_{t-1}\|\mathbf{g}_t\|_2^2 \le K\frac{G^2}{\nu} + \frac{2}{\sqrt{(1-\beta_2)2^{-1/\beta_2}}}\sqrt{KT}G.$$

由于 $\|\mathbf{u}_1\|_2 = D$ 且 $\alpha_0 = \frac{1}{\nu}$，将上述不等式代入 TERM-A 可得

$$\mathbb{E}[\text{TERM-A}] \le \frac{1-\beta_1}{2\beta_1}\gamma\left(K\frac{(G+\sigma)^2}{\nu} + \frac{2}{\sqrt{(1-\beta_2)2^{-1/\beta_2}}}\sqrt{KT}(G+\sigma)\right) + \frac{\beta_1\nu}{2\gamma(1-\beta_1)}D^2.$$

项 B. 由于 $\|\mathbf{u}_t\|_2 = D$ 对所有 t,

$$\textsc{Term-B} = \frac{\beta_1 D^2}{2} \sum_{t=1}^{T-1} \beta_1^t \Big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \Big).$$

应用引理 12 ，并结合 $\epsilon = \nu$ ，可得

$$\sum_{t=1}^{T-1} \beta_1^t \Big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \Big) \le \frac{\sqrt{V_{T-1}}}{\gamma(1-\beta_1)} + \frac{1}{\gamma} \sum_{t=1}^{T-2} \sqrt{V_t} + \frac{(T-1)\nu}{\gamma} + \mu(T-1).$$

利用 $\|\mathbf{g}_t\|_2 \le G$ 可推出 $V_t \le G^2$ ，进而得到 $\sqrt{V_t} \le G$ ，对所有 t均成立。对上式取期望可得

$$\mathbb{E}\Big[\sqrt{V_{T-1}} + \sum_{t=1}^{T-2} \sqrt{V_t} \Big] \le TG.$$

因此 $\mathbb{E}[\textsc{Term-B}] \le \frac{\beta_1 D^2}{2} \Big(\frac{TG}{1-\beta_1} + (T-1)\nu + \mu(T-1) \Big)$。

项 C。 利用 $\eta_t = \gamma(1-\beta_1)\beta_1^t \alpha_t$改写 $\textsc{Term-C}$：

$$\textsc{Term-C} = \sum_{t=1}^{T} \Big(\frac{\beta_1^t}{2\eta_{t-1}} - \frac{\beta_1^t}{2\eta_t} \Big) \|\Delta_{t+1}\|_2^2 = \frac{1}{2\gamma(1-\beta_1)} \sum_{t=1}^{T} \Big(\frac{\beta_1}{\alpha_{t-1}} - \frac{1}{\alpha_t} \Big) \|\Delta_{t+1}\|_2^2.$$

该序 列 $\epsilon_t = \nu + \gamma\mu(1-\beta_1^t)$ 非负且单调不减，因此引理 11 适用，并给出 s

$$\Big[\frac{\beta_1}{\alpha_{t-1}} - \frac{1}{\alpha_t} \Big]_+ \le [\beta_1 - \sqrt{\beta_2}]_+ \sqrt{V_{t-1}}.$$

在假设 $\beta_2 \ge \beta_1^2$ 的条件下，我们有 $\sqrt{\beta_2} \ge \beta_1$ ，因此有 $[\beta_1 - \sqrt{\beta_2}]_+ = 0$，这意味着对于每个 t，都有 $\frac{\beta_1}{\alpha_{t-1}} - \frac{1}{\alpha_t} \le 0$。因此， $\textsc{Term-C} \le 0$ 几乎必然成立，从而有 $\mathbb{E}[\textsc{Term-C}] \le 0$。

组合并调整参数。 将（45）与（46）以及上述论证相结合，可得

$$\mathbb{E}\|\nabla F(\bar{\mathbf{x}})\|_c \le \frac{F^*}{DT} + \frac{2G+\sigma}{(1-\beta_1)T} + \sigma\sqrt{1-\beta_1} + \frac{12cD^2}{(1-\beta_1)^2} + \frac{1}{DT}\mathbb{E}[\textsc{Term-A} + \textsc{Term-B} + \textsc{Term-C}].$$

现在我们来调整参数。令 $\mu = \frac{24cD}{(1-\beta_1)^2}$ 按照引理 7 的要求，并令

$$\beta_1 = 1 - \Big(\frac{\epsilon}{16(G+\sigma)} \Big)^2, \qquad D = \frac{(1-\beta_1)\sqrt{\epsilon}}{\sqrt{96c}}, \qquad \gamma = \frac{\beta_1 D}{\sqrt{1-\beta_1}}, \qquad \beta_2 \ge \max\Big\{ 1 - \frac{\nu}{G+\sigma}, \beta_1^2 \Big\}.$$

经过上述调整后，我们有

$$\frac{12cD^2}{(1-\beta_1)^2} = \frac{\epsilon}{8}, \qquad \frac{D\mu}{2} = \frac{12cD^2}{(1-\beta_1)^2} = \frac{\epsilon}{8}, \qquad \sigma\sqrt{1-\beta_1} \le (G+\sigma)\sqrt{1-\beta_1} = \frac{\epsilon}{16}.$$

代入 $\gamma = \beta_1 D/\sqrt{1-\beta_1}$ 和 $\alpha_0 = 1/\nu$，并假设 $T \ge \frac{2}{\ln 1-\beta_2}$，我们得到

$$\frac{1}{DT}\mathbb{E}[\textsc{Term-A}] \le G\sqrt{1-\beta_1} \left(\frac{1}{\ln 2} + \sqrt{\frac{2}{\ln 2}} \cdot \frac{1}{\sqrt{2^{-1/\beta_2}}} \right) + \frac{\nu}{2T\sqrt{1-\beta_1}}.$$

由于 $\epsilon \le G + \sigma$ 意味着 $\beta_1 \ge 255/256$ 且 $\beta_2 \ge \beta_1^2 \ge (255/256)^2$, ，我们有 $\frac{1}{\sqrt{2^{-1/\beta_2}}} = 2^{\frac{1}{2\beta_2}} < \frac{3}{2}$。因此,

$$\frac{1}{DT}\mathbb{E}[\textsc{Term-A}] \le 4G\sqrt{1-\beta_1} + \frac{\nu}{2T\sqrt{1-\beta_1}} = \frac{\epsilon}{4} + \frac{\nu}{2T\sqrt{1-\beta_1}}.$$

同理，代入 $\gamma = \beta_1 D/\sqrt{1-\beta_1}$ 可得

$$\begin{aligned}\frac{1}{DT}\mathbb{E}[\textsc{Term-B}] &\le \frac{G}{2T\sqrt{1-\beta_1}} + \frac{G\sqrt{1-\beta_1}}{2} + \frac{\nu\sqrt{1-\beta_1}}{2} + \frac{D\mu}{2} \\ &\le \frac{G}{2T\sqrt{1-\beta_1}} + \frac{\epsilon}{8} + \frac{\epsilon}{16}\end{aligned}$$

最后，选择 T 使其满足

$$T \ge \max\Big\{\frac{1}{1-\beta_1}\max\Big\{\frac{16F^*\sqrt{96c}}{\epsilon^{3/2}}, \frac{48(G+\sigma)}{\epsilon}\Big\}, \frac{\ln 2}{1-\beta_2}\Big\}. \tag{47}$$

那么 $\frac{F^*}{DT} \le \epsilon/16$ 且 $\frac{2G+\sigma}{(1-\beta_1)T} \le \epsilon/16$。此外，根据标准假设 $\nu \le G+\sigma$ 且 $\sqrt{1-\beta_1} = \epsilon/(16(G+\sigma))$，条件(47)意味着 $\frac{G+\nu}{2T\sqrt{1-\beta_1}} \le \frac{\epsilon}{32}$。结合上述界以及 $\mathbb{E}[\textsc{Term-C}] \le 0$，得到 $\mathbb{E}\|\nabla F(\bar{\mathbf{x}})\|_c \le \epsilon$，证明完毕。

□

D.9 定理8的证明

证明。在整个证明过程中，我们用 β_1 表示 Adam 中的一阶矩折扣因子（以及算法2中的折扣因子），用 β_2 表示二阶矩的指数移动平均。

O2NC 转换。应用引理7，取 $\beta = \beta_1$。令 $\ell_t(\Delta) = \langle \mathbf{g}_t, \Delta\rangle + \frac{\mu}{2}\|\Delta\|_2^2$ 且 $\text{D-Reg}_T(\mathbf{u}_{1:T}) = \sum_{t=1}^T (\ell_t(\Delta_t) - \ell_t(\mathbf{u}_t))$，我们得到

$$\begin{aligned}\mathbb{E}\|\nabla F(\bar{\mathbf{x}})\|_c \le{}& \frac{F^*}{DT} + \frac{2G+\sigma}{(1-\beta_1)T} + \sigma\sqrt{1-\beta_1} + \frac{12cD^2}{(1-\beta_1)^2} \\ &+ \frac{1}{DT}\mathbb{E}\Big[\text{D-Reg}_T(\mathbf{u}_{1:T}) + \beta_1\sum_{t=1}^{T-1}\sum_{s=1}^{t}\beta_1^{t-s}\big(\ell_s(\mathbf{u}_t) - \ell_s(\mathbf{u}_{t+1})\big)\Big].\end{aligned} \tag{48}$$

回想一下 $\|\mathbf{u}_t\|_2 = D$ 对所有 t 根据引理 7的构造。

使用 D2D 归约。定义

$$V_t = (1-\beta_2)\sum_{s=1}^{t}\beta_2^{t-s}\|\mathbf{g}_s\|_2^2, \qquad A_t = \nu + \gamma\mu(1-\beta_1^t) + \sqrt{V_t}, \qquad \eta_t = \gamma(1-\beta_1)\frac{\beta_1^t}{A_t}.$$

我们将精细单步界（引理 21）应用于经过 重新缩放的 梯度 $\{\beta_1^{-t}\mathbf{g}_t\}_{t=1}^T$ 以及二次正则项 $\psi_{t+1}(\Delta) = \frac{1}{2\eta_t}\|\Delta\|_2^2$，然后应用折扣到动态转换（定理 1）。由此得到如下分解：

$$\begin{aligned}&\text{D-Reg}_T(\mathbf{u}_{1:T}) + \beta_1\sum_{t=1}^{T-1}\sum_{s=1}^{t}\beta_1^{t-s}\big(\ell_s(\mathbf{u}_t) - \ell_s(\mathbf{u}_{t+1})\big) \\ &\le \underbrace{\sum_{t=1}^{T}\beta_1^t\eta_t\|\beta_1^{-t}\mathbf{g}_t\|_2^2 + \frac{\beta_1}{2\eta_0}\|\mathbf{u}_1\|_2^2}_{\textsc{Term-A}} + \underbrace{\sum_{t=1}^{T}\beta_1^t\min\Big\{\frac{\eta_{t-1}}{2}\|\beta_1^{-t}\mathbf{g}_t\|_2^2, |\eta_{t-1}-\eta_t|\|\widetilde{\mathbf{g}}_{1:t-1}\|_2\cdot\|\beta_1^{-t}\mathbf{g}_t\|_2\Big\}}_{\textsc{Term-B}} \\ &+ \underbrace{\sum_{t=1}^{T}\beta_1^t\big(\psi_t(\Delta_{t+1}) - \psi_{t+1}(\Delta_{t+1})\big)}_{\textsc{Term-C}} + \underbrace{\beta_1\sum_{t=1}^{T-1}\beta_1^t\big(\psi_{t+1}(\mathbf{u}_{t+1}) - \psi_t(\mathbf{u}_t)\big)}_{\textsc{Term-D}},\end{aligned} \tag{49}$$

其中 $\widetilde{\mathbf{g}}_{1:t} = \sum_{s=1}^t \beta_1^{-s}\mathbf{g}_s$。此外，由于 $\|\mathbf{u}_t\|_2 = D$ 对所有 t，因此 $\ell_s(\mathbf{u}_t) - \ell_s(\mathbf{u}_{t+1}) = \langle \mathbf{g}_s, \mathbf{u}_t - \mathbf{u}_{t+1}\rangle$，所以 (49) 中的左侧恰好是 (48) 中括号内的项。

项 A。 首先，利用 $\eta_t = \gamma(1-\beta_1)\beta_1^t\alpha_t$ 我们改写为

$$\beta_1^t\eta_t\|\beta_1^{-t}\mathbf{g}_t\|_2^2 = \gamma(1-\beta_1)\alpha_t\|\mathbf{g}_t\|_2^2.$$

由于 $\alpha_t = \frac{1}{\nu+\gamma\mu \quad (1-\beta_1^t)+\sqrt{V_t}} \le \frac{1}{\nu+\sqrt{V_{t-}}}$ 我们有

$$\mathbb{E}\sum_{t=1}^{T}\alpha_t\|\mathbf{g}_t\|_2^2 \le \mathbb{E}\sum_{t=1}^{T}\frac{\|\mathbf{g}_t\|_2^2}{\nu+\sqrt{V_t}}. \le \frac{2^{3/2+1/(2\beta_2)}}{\sqrt{\ln 2}}G\sqrt{1-\beta_1}.$$

利用 $\alpha_0 = \frac{1}{\nu}$，因此 $\eta_0 = \gamma(1-\beta_1)\alpha_0 = \gamma(1-\beta_1)/\nu$，所以

$$\frac{1}{DT}\cdot\frac{\beta_1}{2\eta_0}\|\mathbf{u}_1\|_2^2 = \frac{\nu}{2T\sqrt{1-\beta_1}}. \tag{50}$$

项 B. 我们用偏差项给最小值取上界:

$$\beta_1^t|\eta_{t-1}-\eta_t|\|\widetilde{\mathbf{g}}_{1:t-1}\|_2\cdot\|\beta_1^{-t}\mathbf{g}_t\|_2 = |\eta_{t-1}-\eta_t|\Big\|\sum_{s=1}^{t-1}\beta_1^{t-1-s}\mathbf{g}_s\Big\|_2\cdot\|\beta_1^{1-t}\mathbf{g}_t\|_2.$$

应用引理 13 并利用 $\|\mathbf{g}_t\|_2 \le G$:

$$\beta_1^t|\eta_{t-1}-\eta_t|\|\widetilde{\mathbf{g}}_{1:t-1}\|_2\cdot\|\beta_1^{-t}\mathbf{g}_t\|_2 \le \gamma(1-\beta_1)\cdot\frac{A_t-\beta_1A_{t-1}}{A_t}\cdot\sqrt{\frac{\beta_2}{(\beta_2-\beta_1^2)(1-\beta_2)}}\cdot G.$$

对 t 上述不等式求和，并利用 $1-x \le \ln(1/x)$ 与 $x = \beta_1A_{t-1}/A_t$ 的事实，可得 $\sum_{t=1}^{T}\frac{A_t-\beta_1A_{t-1}}{A_t} \le \sum_{t=1}^{T}\ln(A_t/(\beta_1A_{t-1})) \le \ln(A_T/(\beta_1^TA_0))$，从而

$$\textsc{Term-B} \le \gamma(1-\beta_1)G\Big(\ln\Big(\frac{A_T}{A_0}\Big)-T\ln\beta_1\Big)\sqrt{\frac{\beta_2}{(\beta_2-\beta_1^2)(1-\beta_2)}}. \tag{51}$$

接下来，我们利用间隔条件对依赖于 β_2 的因子进行界定。令 $\beta_2^\star = \frac{1+\beta_1^2}{2}$。由于 $\beta_2 \in [\beta_1^2+m,\ 1-m]$ 且 $m = \frac{1-\rho}{2}(1-\beta_1^2)$，我们有 $|\beta_2-\beta_2^\star| \le \frac{\rho}{2}(1-\beta_1^2)$，因此

$$(\beta_2-\beta_1^2)(1-\beta_2) = \frac{(1-\beta_1^2)^2}{4}-(\beta_2-\beta_2^\star)^2 \ge \frac{(1-\beta_1^2)^2}{4}(1-\rho^2).$$

因此

$$\sqrt{\frac{\beta_2}{(\beta_2-\beta_1^2)(1-\beta_2)}} \le \frac{2}{1-\beta_1^2}\cdot\frac{1}{\sqrt{1-\rho^2}}. \tag{52}$$

最后，由于 $\sqrt{V_t} \le G$ 和 $1-\beta_1^t \le 1$，我们有

$$A_T \le \nu+\gamma\mu+\sqrt{V_T} \le \nu+\gamma\mu+G, \qquad A_0 = \nu,$$

因此，我们有,

$$\ln\Big(\frac{A_T}{A_0}\Big) \le \ln\Big(1+\frac{\gamma\mu+G}{\nu}\Big). \tag{53}$$

此外 $-\ln\beta_1 \le \frac{1-\beta_1}{\beta_1}$。将（52）–（53）代入（51），并利用 $\gamma = \frac{\beta_1D}{\sqrt{1-\beta_1}}$可得

$$\frac{1}{DT}\mathbb{E}[\textsc{Term-B}] \le \frac{2G}{T\sqrt{1-\beta_1}}\cdot\frac{1}{\sqrt{1-\rho^2}}\ln\Big(1+\frac{\gamma\mu+G}{\nu}\Big) + \frac{2G}{\sqrt{1-\rho^2}}\sqrt{1-\beta_1}. \tag{54}$$

项 C。 由于序列 $\epsilon_t = \nu + \gamma\mu(1-\beta_1^t)$ 非负且单调不减。根据引理 11，

$$\Big[\frac{\beta_1}{\alpha_{t-1}} - \frac{1}{\alpha_t}\Big]_+ \le [\beta_1 - \sqrt{\beta_2}]_+ \sqrt{V_{t-1}}.$$

由于 $\beta_2 > \beta_1^2$，我们有 $\sqrt{\beta_2} > \beta_1$，因此 $[\beta_1 - \sqrt{\beta_2}]_+ = 0$。所以 $\textsc{Term-C} \le 0$。

项 D。 由于 $\|\mathbf{u}_t\|_2 = D$ 对所有 t，

$$\textsc{Term-D} = \frac{\beta_1 D^2}{2} \sum_{t=1}^{T-1} \beta_1^t \Big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\Big).$$

应用引理 12，其中 $\epsilon = \nu$ 和 μ 如定理陈述中所示，并使用 $\sqrt{V_t} \le G$，我们有：

$$\frac{1}{DT}\mathbb{E}[\textsc{Term-D}] \le \frac{G}{2T\sqrt{1-\beta_1}} + \frac{G\sqrt{1-\beta_1}}{2} + \frac{\nu\sqrt{1-\beta_1}}{2} + \frac{D\mu}{2}.$$

组合并调整参数。 汇总上述不等式，我们得到

$$\begin{aligned}\mathbb{E}\|\nabla F(\bar{\mathbf{x}})\|_c \le{}& \frac{F^*}{DT} + \frac{2G+\sigma}{(1-\beta_1)T} + \frac{12cD^2}{(1-\beta_1)^2} + \frac{D\mu}{2} \\ &+ \frac{2^{3/2+1/(2\beta_2)}}{\sqrt{\ln 2}} G\sqrt{1-\beta_1} + \sigma\sqrt{1-\beta_1} + \frac{2G}{\sqrt{1-\rho^2}}\sqrt{1-\beta_1} + \frac{G+\nu}{2}\sqrt{1-\beta_1} \\ &+ \frac{G+\nu}{2T\sqrt{1-\beta_1}} + \frac{2G}{T\sqrt{1-\beta_1}\sqrt{1-\rho^2}} \ln\Big(1 + \frac{\gamma\mu+G}{\nu}\Big). \end{aligned} \tag{55}$$

现在令 $D = \frac{(1-\beta_1)\sqrt{\epsilon}}{\sqrt{96c}}$ 和 $\mu = \frac{24cD}{(1-\beta_1)^2}$。于是

$$\frac{12cD^2}{(1-\beta_1)^2} = \frac{\epsilon}{8}, \qquad \frac{D\mu}{2} = \frac{12cD^2}{(1-\beta_1)^2} = \frac{\epsilon}{8}.$$

接下来，根据 β_1，$\sqrt{1-\beta_1} \le \frac{\epsilon\sqrt{1-\rho^2}}{64\,G+\sigma}$ 的选取，我们有

$$\frac{2G}{\sqrt{1-\rho^2}}\sqrt{1-\beta_1} \le \frac{\epsilon}{32}, \quad \sigma\sqrt{1-\beta_1} \le (G+\sigma)\sqrt{1-\beta_1} \le \frac{\epsilon}{64}, \quad \frac{G+\nu}{2}\sqrt{1-\beta_1} \le (G+\sigma)\sqrt{1-\beta_1} \le \frac{\epsilon}{64}.$$

此外，在温和条件 $\epsilon \le \frac{16}{(G+\sigma)\sqrt{1-\rho^2}}$ 下，上述选取意味着 $\beta_1 \ge 15/16$。由于$\beta_2 > \beta_1^2$，我们得到 $2^{\frac{1}{2\beta_2}} < \frac{3}{2}$，从而有 $\sqrt{\frac{2^{3/2+1/(2\beta_2)}}{\ln 2}} < 6$。因此，

$$\frac{2^{3/2+1/(2\beta_2)}}{\sqrt{\ln 2}}(G+\sigma)\sqrt{1-\beta_1} \le 6(G+\sigma)\sqrt{1-\beta_1} \le \frac{3\epsilon}{32}.$$

最后，按照定理陈述中的方式选取 T，使得

$$\frac{F^*}{DT} \le \frac{\epsilon}{16}, \qquad \frac{2G+\sigma}{(1-\beta_1)T} \le \frac{\epsilon}{16}, \qquad \frac{2G}{T\sqrt{1-\beta_1}\sqrt{1-\rho^2}} \ln\Big(1 + \frac{\gamma\mu+G}{\nu}\Big) \le \frac{\epsilon}{16},$$

并且 $\frac{G+\nu}{2T\sqrt{1-\beta_1}} \le \frac{\epsilon}{16}$（这由第一个 T 下界推出，因为它按 $(1-\beta_1)^{-1}$ 的尺度变化。将 (55) 中的所有项合并，得到 $\mathbb{E}\|\nabla F(\bar{\mathbf{x}})\|_c \le \epsilon$，证明完毕。

□

E 补充引理

本部分汇集了分析中用到的实用技术引理。为确保内容自成一体，我们陈述几个关键的 FTRL 引理，主要取自Orabona [2019]。

E.1 技术引理

引理 16. 设 $\beta \in (0,1]$ 和 $\{a_t\}_{t=1}^T \subseteq \mathbb{R}$。定义 $V_t = (1-\beta)\sum_{s=1}^t \beta^{t-s} a_s, t \in [T]$。则

$$\sum_{t=1}^T V_t \;=\; \sum_{t=1}^T \big(1-\beta^{T+1-t}\big) a_t.$$

证明。 通过交换求和顺序,

$$\sum_{t=1}^T V_t = (1-\beta)\sum_{t=1}^T\sum_{s=1}^t \beta^{t-s} a_s = (1-\beta)\sum_{s=1}^T a_s \sum_{t=s}^T \beta^{t-s} = \sum_{s=1}^T a_s\big(1-\beta^{T+1-s}\big).$$

□

引理 17 (J´ez´ 等人 [2020]中的引理 5。 设 $C>0$ 和 $f:\; x \in \mathbb{R} \to \ln(1+\exp(-x))$。那么对于所有 $a \in [-C, C]$ 和 $b \in \mathbb{R}$,

$$f(a) \geq f(b) + f'(b)(a-b) + \frac{e^b}{2(1+C)} f'(b)^2 (a-b)^2.$$

引理 18 (Jacobsen 和 Cutkosky [2024]中的引理3.2。 对于任意非负函数 f_0, ..., f_T, $0 < \beta \leq \gamma < 1$ $F_t^\gamma = \sum_{s=0}^t \gamma^{t-s} f_s$, 定义 $(\mathbf{u})(\mathbf{u})$, 以及

$$P_T^\gamma = \sum_{t=1}^{T-1}\sum_{s=0}^t p_{t,s}^\gamma \left[f_s(\mathbf{u}_{t+1}) - f_s(\mathbf{u}_t)\right]_+, \; where\; p_{t,s}^\gamma = \frac{\gamma^{t-s}}{\sum_{\tau=0}^t \gamma^{t-\tau}}.$$

那么我们有:

$$\beta \sum_{t=1}^{T-1} \Big(F_t^\beta(\mathbf{u}_{t+1}) - F_t^\beta(\mathbf{u}_t)\Big) \leq \frac{\gamma}{1-\gamma} P_T^\gamma.$$

E.2 FTRL引理

为确保完整性，我们给出分析中将用到的若干关键跟随正则化领导者（FTRL）结果。

标准FTRL。 我们考虑如下形式的跟随正则化领导者（FTRL）：对于闭凸集 $\mathcal{X} \subseteq \mathbb{R}^d$,

$$\mathbf{x}_{t+1} \;=\; \arg\min_{\mathbf{x}\in\mathcal{X}} \Big\{\psi_{t+1}(\mathbf{x}) + \sum_{s=1}^t f_s(\mathbf{x})\Big\}, \tag{56}$$

其中 $f_t:\; \mathcal{X} \to \mathbb{R}$ 是凸函数, $\psi_t:\; \mathcal{X} \to \mathbb{R}\cup\{+\infty\}$ 是凸正则项。为方便起见，我们定义

$$F_t(\mathbf{x}) \;=\; \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x}), \tag{57}$$

于是 $\mathbf{x}_t \in \arg\min_{\mathbf{x}\in\mathcal{X}} F_t(\mathbf{x})$ 且 $\mathbf{x}_{t+1} \in \arg\min_{\mathbf{x}\in\mathcal{X}} \big(F_t(\mathbf{x}) + f_t(\mathbf{x}) + \psi_{t+1}(\mathbf{x}) - \psi_t(\mathbf{x})\big)$。

引理 19 （Orabona [2019]中的引理7.1。）采用式 (56) 中更新规则的跟随正则化领导者（*FTRL*）满足以下静态遗憾分解：对于任意比较器 $\mathbf{u} \in \mathcal{X}$,

$$\sum_{t=1}^T f_t(\mathbf{x}_t) - f_t(\mathbf{u}) = \psi_T(\mathbf{u}) - \min_{\mathbf{x}\in\mathcal{X}} \psi_1(\mathbf{x}) + \sum_{t=1}^T \Big(F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t)\Big) + F_{T+1}(\mathbf{x}_{T+1}) - F_{T+1}(\mathbf{u}),$$

其中 $F_t(\cdot)$ 定义于 (57)。

下一个引理是跟随正则化领导者分析中经常使用的单步“稳定性”界。它明确体现了强凸性的作用，而标准的二次正则化项结论可直接由此推得。

引理 *20*（*Orabona* [2019] 中的引理 *7.6*）。设 f_t 在 $\mathcal{X}$ 上是凸且可微的，且 ψ_t 关于范数 $\|\cdot\|$ 是 λ_t-强凸的（因此，式 *(57)* 中的 F_t 也是 λ_t-强凸的）。令 $x_t \in arg\ min_{x\in\mathcal{X}} F_t(x)$，且 $x_{t+1} \in arg\ min_{x\in\mathcal{X}} F_{t+1}(x)$，并令 $g_t \in \partial f_t(x_t)$。则

$$F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) \le \frac{\|\mathbf{g}_t\|_*^2}{2\lambda_t} + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}), \tag{58}$$

其中 $\|\cdot\|_*$ 是 $\|\cdot\|$ 的对偶范数。

特别是,别是，若 ψ_t（$\mathbf{x}$）$=\frac{1}{2\eta_{t-1}}\|\mathbf{x}\|_2^2$，则 ψ_t 是 $(1/\eta_{t-1})$-相对于 的强凸函数，且，（58）变为 *es*

$$F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) \le \frac{\eta_{t-1}}{2}\|\mathbf{g}_t\|_2^2 + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}).$$

我们需要下一状态满足一个无尺度单步界，改编自 Orabona 和 P´al [2018], 在命题中使用 of 定理 m 6.

引理 21（改编自Orabona 和 P´等人 [2018]的定理 1）。设 $\mathcal{X} \subseteq \mathbb{R}^d$ 为一个直径为 $D_{\mathcal{X}} = \max_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \|\mathbf{x}-\mathbf{y}\|_2$的闭凸集。假设损失是线性的：

$$f_t(\mathbf{x}) = \langle \mathbf{g}_t, \mathbf{x}\rangle, \qquad \mathbf{g}_t \in \mathbb{R}^d.$$

考虑在（56）中使用二次正则项 ψ_t（$\mathbf{x}$）$=\frac{1}{2\eta_{t-1}}\|\mathbf{x}\|_2^2$，步长为 $\eta_{t-1} > 0$。令F_t（·）由（57）定义，并定义累积梯度$\mathbf{g}_{1:t} = \sum_{s=1}^t \mathbf{g}_s$。则有：

$$\begin{aligned} &F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) \\ &\le \eta_t\|\mathbf{g}_t\|_2^2 + \min\left\{\frac{\eta_{t-1}}{2}\|\mathbf{g}_t\|_2^2,\ \min\left\{D_{\mathcal{X}},\ |\eta_{t-1}-\eta_t|\,\|\mathbf{g}_{1:t-1}\|_2\right\}\cdot\|\mathbf{g}_t\|_2\right\} + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}). \end{aligned}$$

证明。根据 F_t 的定义，并加上再减去 $F_t(\mathbf{x}_{t+1})$ 和 $f_t(\mathbf{x}_{t+1})$，我们得到 n

$$F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) = \underbrace{\left(F_t(\mathbf{x}_t) - F_t(\mathbf{x}_{t+1})\right)}_{\text{Term-A}} + \underbrace{\left(f_t(\mathbf{x}_t) - f_t(\mathbf{x}_{t+1})\right)}_{\text{Term-B}} + \underbrace{\left(\psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1})\right)}_{\text{Term-C}}. \tag{59}$$

由于$\mathbf{x}_t \in \arg\min_{\mathbf{x}} F_t$（$\mathbf{x}$），因此有 Term-A ≤ 0。此外，f_t 是线性的，因此 Term-B $= \langle \mathbf{g}_t,\ \mathbf{x}_t - \mathbf{x}_{t+1}\rangle \le \|\mathbf{g}_t\|_2\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_2$。因此（59）意味着

$$F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) \le \|\mathbf{g}_t\|_2\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_2 + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}). \tag{60}$$

我们通过两种方式对 $\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_2$ 进行上界估计。

首先，由于 $f_s(\mathbf{x}) = \langle \mathbf{g}_s,\ \mathbf{x}\rangle$ 以及 $\psi_t(\mathbf{x}) = \frac{1}{2\eta_{t-1}}\|\mathbf{x}\|_2^2$，跟随正则化领导者可以写成投影形式

$$\mathbf{x}_t = \Pi_{\mathcal{X}}\left[-\eta_{t-1}\mathbf{g}_{1:t-1}\right], \qquad \mathbf{x}_{t+1} = \Pi_{\mathcal{X}}\left[-\eta_t\mathbf{g}_{1:t}\right],$$

其中，$\Pi_{\mathcal{X}}$ 是到 $\mathcal{X}$ 上的欧几里得投影。根据投影的非扩张性，

$$\begin{aligned} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|_2 &= \left\|\Pi_{\mathcal{X}}\left[-\eta_{t-1}\mathbf{g}_{1:t-1}\right] - \Pi_{\mathcal{X}}\left[-\eta_t\mathbf{g}_{1:t}\right]\right\| \\ &\le \left\|-\eta_{t-1}\mathbf{g}_{1:t-1} + \eta_t\mathbf{g}_{1:t}\right\|_2 = \left\|\eta_t\mathbf{g}_t - (\eta_{t-1}-\eta_t)\mathbf{g}_{1:t-1}\right\|_2 \\ &\le \eta_t\|\mathbf{g}_t\|_2 + |\eta_{t-1}-\eta_t|\,\|\mathbf{g}_{1:t-1}\|_2. \end{aligned} \tag{61}$$

其次，直接 $\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_2 \le D_{\mathcal{X}}$。结合这一事实与（61）可得

$$\begin{aligned} \|\mathbf{g}_t\|_2\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_2 &\le \min\left\{D_{\mathcal{X}}\|\mathbf{g}_t\|_2,\ \eta_t\|\mathbf{g}_t\|_2^2 + |\eta_{t-1}-\eta_t|\,\|\mathbf{g}_{1:t-1}\|_2\|\mathbf{g}_t\|_2\right\} \\ &\le \eta_t\|\mathbf{g}_t\|_2^2 + \min\left\{D_{\mathcal{X}},\ |\eta_{t-1}-\eta_t|\,\|\mathbf{g}_{1:t-1}\|_2\right\}\cdot\|\mathbf{g}_t\|_2. \end{aligned}$$

将上述界代入（60）可得

$$\begin{aligned} &F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) \le \\ &\quad \eta_t\|\mathbf{g}_t\|_2^2 + \min\left\{D_{\mathcal{X}},\ |\eta_{t-1}-\eta_t|\,\|\mathbf{g}_{1:t-1}\|_2\right\}\cdot\|\mathbf{g}_t\|_2 + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}). \end{aligned} \tag{62}$$

另一方面，应用引理 20 并结合 $\lambda_t = 1/\eta_{t-1}$ 可得

$$F_t(\mathbf{x}_t) - F_{t+1}(\mathbf{x}_{t+1}) + f_t(\mathbf{x}_t) \le \frac{\eta_{t-1}}{2}\|\mathbf{g}_t\|_2^2 + \psi_t(\mathbf{x}_{t+1}) - \psi_{t+1}(\mathbf{x}_{t+1}). \tag{63}$$

取（62）和（63）中的较小值，并对 $\{a,\ c+b\} \le c + \min\{a,b\}$ 使用 $c \ge 0$，即可完成证明。

□

乐观跟随正则化领导者。 我们考虑跟随正则化领导者的一种乐观变体 [Rakhlin 和 Sridharan，2013，Orabona，2019]，其中，在提供 h_t（$\mathbf{x}$）时使用乐观函数 $\mathbf{x}_t$。此函数可以视为对即将到来的函数的"猜测"。

$$\mathbf{x}_t \;=\; \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \psi_t(\mathbf{x}) + h_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x}) \right\}. \tag{64}$$

引理 22（改编自 J´ez´equel 等人的定理 7 [2020]）。定义不使用乐观函数的决策为 $\widehat{\mathbf{x}}_t$，使得：

$$\widehat{\mathbf{x}}_t \in \arg\min_{\mathbf{x}\in\mathcal{X}} F_t(\mathbf{x}), \qquad F_t(\mathbf{x}) \;=\; \psi_t(\mathbf{x}) + \sum_{s=1}^{t-1} f_s(\mathbf{x}).$$

对于公式（64）中的乐观跟随正则化领导者，对于任意比较器 $\mathbf{u} \in \mathcal{X}$，有

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{u}) = \psi_{T+1}(\mathbf{u}) - \min_{\mathbf{x}\in\mathcal{X}} \psi_1(\mathbf{x}) + \sum_{t=1}^{T} \Big(f_t(\mathbf{x}_t) + F_t(\widehat{\mathbf{x}}_t) - F_{t+1}(\widehat{\mathbf{x}}_{t+1}) \Big) + F_{T+1}(\widehat{\mathbf{x}}_{T+1}) - F_{T+1}(\mathbf{u}).$$

特别地，由于 F_{T+1}（$\widehat{\mathbf{x}}_{T+1}$）$\le F_{T+1}$（$\mathbf{u}$），我们得到上界

$$\sum_{t=1}^{T} \big(f_t(\mathbf{x}_t) - f_t(\mathbf{u}) \big) \le \psi_{T+1}(\mathbf{u}) - \min_{\mathbf{x}\in\mathcal{X}} \psi_1(\mathbf{x}) + \sum_{t=1}^{T} \Big(f_t(\mathbf{x}_t) + F_t(\widehat{\mathbf{x}}_t) - F_{t+1}(\widehat{\mathbf{x}}_{t+1}) \Big).$$

E.3 自信调节引理

引理 23（引理 14，见 Gaillard 等人 [2014]）。令 $a_0 > 0$ 和 $a_t \in [0, B]$ 对所有$t \in [T]$ 均为实数，并令 f: $(0,\ +\infty) \to [0, +\infty)$ 为非增函数。则

$$\sum_{t=1}^{T} a_t f\left(\sum_{s=0}^{t-1} a_s \right) \le B \cdot f(a_0) + \int_{a_0}^{\sum_{t=0}^{T} a_t} f(u)\mathrm{d}u.$$

引理 24（引理 3.5，见 Auer 等人 [2002]）。令 a_1，……，a_T和 δ均为非负实数。则

$$\sum_{t=1}^{T} \frac{a_t}{\sqrt{\delta + \sum_{s=1}^{t} a_s}} \le 2 \left(\sqrt{\delta + \sum_{t=1}^{T} a_t} - \sqrt{\delta} \right).$$

引理 25 (Jacobsen 和 Cutkosky [2024]中的引理 G.2)。令 $\beta \in (0, 1]$, $\lambda > 0$, $\mathbf{z}_t \in \mathbb{R}^d$，并定义$A_0 = \lambda I$ 和

$$A_t = \mathbf{z}_t \mathbf{z}_t^\top + \beta A_{t-1} \qquad \textit{for each } t > 0.$$

那么，对于任意序列 $c_1, c_2, \ldots \in \mathbb{R}$,

$$\sum_{t=1} c_t^2 \mathbf{z}_t^\top A_t^{-1} \mathbf{z}_t \le d \ln\left(\frac{1}{\beta}\right) \sum_{t=1}^{T} c_t^2 + \left\{ \max_{t\in[T]} c_t^2 \right\} \cdot d \ln\left(1 + \frac{\sum_{t=1}^{T} \beta^{T-t} \|\mathbf{z}_t\|_2^2}{\lambda d} \right).$$