arXiv:2301.03125v1 [stat.ML] 9 Jan 2023

小批量随机近端点方法的深入分析：稳定性、光滑性与偏差

Xiao-Tong Yuan 和 Ping Li

百度研究院认知计算实验室，中国北京市西北旺东路10号，邮编100193；美国华盛顿州贝尔维尤市东北第8街10900号，邮编98004 电子邮箱：{xtyuan1980，pingli98}@gmail.com

摘要

随机近端点（SPP）方法近年来在随机优化中备受关注。与经典随机梯度下降（SGD）方法相比，这些方法不仅具有强有力的收敛保证，而且鲁棒性更强，计算开销却几乎没有增加。本文研究用于求解凸复合风险最小化问题的 SPP 小批量变体，即 M-SPP。本文的核心贡献是从算法稳定性理论的视角推导出一组关于 M-SPP 的新颖超额风险界。特别是在满足光滑性和二次增长条件时，我们证明，采用小批量大小为 n 、迭代次数为 T 的 M-SPP，其期望收敛速率为快速收敛率，由一个 $\mathcal{O}\left(\frac{1}{T^2}\right)$ 偏差衰减项和一个 $\mathcal{O}\left(\frac{1}{nT}\right)$ 方差衰减项组成。在小-n、大-T 的情形下，这一结果显著改进了 SPP类方法已知的最佳结果，并揭示了模型噪声水平对收敛速率的影响。在互补的小-T、大-n 情形下，我们提出 M-SPP 的两阶段扩展，以实现可比的收敛速率。此外，我们还针对 M-SPP 的无放回抽样变体，推导出一个近乎紧的参数估计误差高概率界（概率取自数据随机性）。我们还提供了数值证据，以支持将理论预测应用于 Lasso 和逻辑回归模型时所得的结果。

1 引言

我们考虑在闭凸子集$\mathcal{W} \subseteq \mathbb{R}^p$上进行正则化风险最小化的如下问题：

$$\min_{w \in \mathcal{W}} R(w) := R^{\ell}(w) + r(w), \quad \text{where} \quad R^{\ell}(w) := \mathbb{E}_{z \sim \mathcal{D}}[\ell(w; z)], \tag{1}$$

其中，$\ell : \mathcal{W} \times \mathcal{Z} \mapsto R^+$是一个非负凸损失函数，其值$\ell(w; z)$用于衡量一个假设在数据样本$z \in \mathcal{Z}$上的损失，该假设由$w \in \mathcal{W}$参数化；$\mathcal{D}$表示$\mathcal{Z}$上的一个分布；$r : \mathcal{W} \mapsto \mathbb{R}^+$是一个与数据无关的非负凸函数，其值$r(w)$用于衡量该假设的某种复杂度。我们尤其关注复合总体风险R在其极小化器附近强凸的情形，不过在这种情况下，并不一定要求R^{ℓ}和r同时具有强凸性。例如，ℓ_1范数正则项$r(w) = \mu\|w\|_1$或其分组变体常用于采用二次或逻辑损失函数进行稀疏广义线性模型学习（Van de Geer，2008年；Ravikumar 等，2009年；Negahban 等，2012）。

在统计机器学习中，通常假设估计量只能通过批量训练集或在线/增量方式，获取一组从$\mathcal{D}$中抽取的独立同分布随机数据实例$S = \{z_i\}_{i=1}^N$。我们的目标是利用对S的了解计算一个随机估计量$\hat{w}_S$，并希望它能够良好泛化，成为总体风险的近似极小化器。更确切地说，我们旨在推导一个合适的大数定律，即样本量的消失速率δ_N，使得在期望意义下或以高概率，对S而言，$\hat{w}_S$处的超额风险$R(\hat{w}_S) - R^* \leq \delta_N$满足相应条件，其中$R^* := \min_{w \in \mathcal{W}} R(w)$表示复合风险的最小值。

在本研究中，受到随机近端点（SPP）算法近期取得显著成功的启发（Patrascu 和 Necoara，2017年；Asi 和 Duchi，2019a，b；Davis 和 Drusvyatskiy，2019年）及其小批量扩展（Wang 等人，2017b；Zhou 等人，2019年；Asi 等人，2020年），我们针对一类用于求解随机复合风险最小化问题（1）的小批量 SPP 方法，给出了更精细的泛化性能分析。

1.1 算法与研究动机

小批量随机近端点算法。设$S_t = \{z_{i,t}\}_{i=1}^n$为从分布$\mathcal{D}$中抽取的$n$个独立同分布样本，在时间点$t \geq 1$，并记

$$R_{S_t}(w) := \frac{1}{n} \sum_{i=1}^{n} \ell(w; z_{i,t}) + r(w)$$

作为关于S_t的正则化经验风险。我们考虑算法 1 中所概述的用于复合风险最小化的小批量随机近端点（M-SPP）算法，该算法基于数据小批量序列$S = \{S_t\}_{t=1}^T$。算法中的精度值ϵ_t量化了w_t在小批量S_t上求解内循环正则化经验风险最小化问题时的次优性。M-SPP 算法具有通用性，涵盖了若干现有 SPP 方法作为特例。例如，在

算法 1：小批量随机近端点（M
SPP）输入：正则化模量 $\{\gamma_t\}_{t\geq 1}$。输出：将 $\{w_t\}_{1\leq t\leq T}$ 加权平均得到 $\bar{w}_T$。初始化：指定 w_0的值。通常 $w_0 = 0$。对于 $t = 1$、2、.....、T，执行：抽取一个小批量 S_t：$= \{z_{i,t}\}_{i=1}^{n}$ 独立同分布$\sim$ $\mathcal{D}^n$，并估计满足以下条件的 w_t：其中$R_{S_t}(w) := \frac{1}{n}\sum_{i=1}^{n}\ell(w;z_i,\ t)$ $+ r(w)$，且 $\epsilon_t \geq 0$ 衡量估计的次优性。结束

$$F_t(w_t) \leq \min_{w\in\mathcal{W}}\left\{F_t(w) := R_{S_t}(w) + \frac{\gamma_t}{2}\|w - w_{t-1}\|^2\right\} + \epsilon_t, \tag{2}$$

当 $n = 1$ 且 $\epsilon_t \equiv 0$ 处于极端情形时，M-SPP 会简化为标准 SPP 方法的一种复合变体（Bertsekas，2011），其形式如式（5）所示。一般而言，递归更新形式（2）可视为现有用于统计估计的小批量随机近端点方法的一种自然复合扩展（Wang 等人，2017b；Asi 等人，2020年）。

已有结果及其局限性。本文着重分析 M-SPP 在凸复合风险优化中的泛化性能。近期，Asi 等人（2020，定理 2）证明，若瞬时损失函数关于参数是强凸的，则 M-SPP 算法以 $\mathcal{O}\left(\frac{\log(nT)}{nT}\right)$ 的速率收敛。此前，Wang 等人（2017b，定理 5）证明，当个体损失函数是利普希茨连续且强凸时，M-SPP 的收敛速率为 $\mathcal{O}(\frac{1}{nT})$。这些结果以及 SPP 的相关结果（Patrascu 和 Necoara，2017；Davis 和 Drusvyatskiy，2019）通常要求每个瞬时损失函数都具有强凸性，而这一条件过于严格，在高维空间或无限维空间中难以满足。例如，对于特征-标签对 $z = (x, y)$，二次损失 $\ell(w; z) = \frac{1}{2}(w^\top x - y)^2$ 是凸的，但通常并非强凸；不过，只要随机特征 x 的协方差矩阵非退化，总体风险 $R^\ell(w) = \frac{1}{2}\mathbb{E}(y - w^\top x)^2$ 便是强凸的。与此同时，分析中采用的利普希茨损失假设（Wang 等人，2017b，定理 5）使其适用范围局限于二次损失等光滑损失；更不用说利普希茨连续性与强凸性之间的相互制约（Agarwal 等人，2012；Asi 和 Duchi，2019b）。

上述已有研究的不足促使我们探究复合风险最小化中 M-SPP 的收敛行为，并将研究范围拓展到单个损失函数不一定强凸且利普希茨连续的情形。从优化的角度来看，光滑性对于在（6）中求解内循环强凸风险最小化子问题并获得强收敛保证至关重要，例如，采用方差缩减随机算法（Johnson 和 Zhang，2013；Xiao 和 Zhang，2014）或通信高效分布式优化算法（Shamir 等人，2014；Zhang 和 Lin，2015；Yuan 和 Li，2020）。为涵盖这一重要但尚未得到充分理解的问题情形，我们.....

尚未得到充分理解的问题情形，我们着重分析凸损失函数光滑且风险函数具有二次增长性质时 M-SPP 的收敛行为（正式定义见假设2）。

1.2 我们的贡献与主要结果

本工作的主要贡献是通过算法稳定性理论（Bousquet 和 Elisseeff，2002年；Feldman，2018）对 M-SPP 算法进行更精细的非渐近收敛分析。令 $W^* := \{w \in \mathcal{W} : R(w) = R^*\}$ 为复合总体风险 R的极小化解集合。我们特别关注这样一种情形：损失函数ℓ 是凸且光滑的，但不一定满足利普希茨条件（例如二次损失），而总体风险 R 满足二次增长条件，即 $R(w) - R^* \geq \frac{\lambda}{2}\min_{w^* \in W^*} \|w - w^*\|^2, \forall w \in \mathcal{W}$，其中 $\lambda > 0$为某个常数；强凸目标函数以及其他各种统计估计问题都满足这一条件（例如参见 Karimi 等人，2016；Drusvyatskiy 和 Lewis，2018）。对于一族 L-光滑损失函数，若 $\gamma_t = \mathcal{O}(\lambda\rho t)$，其中标量 $\rho \in$ 任意取值于$(0, 0.5]$ 且 $\epsilon_t \equiv 0$，我们在定理 1 中证明，加权平均输出 $\bar{w}_T = {}_T{}^2(_{T+1}) \sum_{t=1}^{T} t w_t$ 的超额风险在期望意义下有如下上界：

$$R(\bar{w}_T) - R^* \lesssim \frac{\rho\,[R(w_0) - R^*]}{T^2} + \frac{LR^*}{\rho\lambda nT}. \tag{3}$$

在这个复合界中，与初始间隙 $R(w_0) - R^*$ 相关的第一个偏差分量以 $\mathcal{O}\left(\frac{1}{T^2}\right)$ 的速率衰减，而与 R^* 相关的第二个方差分量则以 $\mathcal{O}\left(\frac{1}{\lambda nT}\right)$的速率收敛。方差衰减率实际上与强凸优化中 SGD 类方法相应的最优速率一致（Rakhlin 等人，2012；Dieuleveut 等人，2017年；Woodworth 和 Srebro，2021年）。此外，这种 $\mathcal{O}\left(\frac{1}{T^2} + \frac{1}{\lambda nT}\right)$ 界与 SPP（Davis 和 Drusvyatskiy，2019）或 M-SPP（Wang 等人，2017b）的界相当，而后两者是在显著更强的假设下得到的，即每个单个损失函数也都必须是强凸且利普希茨的。在可实现或近似可实现的机器学习情形中，若 R^* 等于或近似为零，则式 (3) 中的方差项会比 Wang 等人（2017b）和 Davis 和 Drusvyatskiy（2019）的界更紧。据我们所知，对于光滑凸损失函数，式 (3) 中的界是 SPP 类方法中的新结果。更一般地，对于任意凸风险函数，我们在定理 3 中给出了精确 M-SPP 的一个 $\mathcal{O}(\frac{1}{\sqrt{nT}})$ 超额风险界。此外，如定理 4 和定理5 所示，只要内循环次优性足够小，类似结果也可推广至非精确 M-SPP。

在 $T \ll n$ 这一情形下，对于采用大型数据批次的离线增量学习而言尤其重要，若在超额风险界（ $\rho = \sqrt{\frac{T}{n\lambda}}$3）中取近乎最优的值，则可得到 $\mathcal{O}\left(\frac{1}{T\sqrt{\lambda nT}}\right)$收敛速率。就 n而言，该速率远慢于此前小 $\mathcal{O}(\frac{1}{\lambda nT})$-大$n$-$T$ 设置下可达到的速率。为解决这一不足，我们提出 M-SSP 的两阶段变体（见算法2），以提升其在小T-大n 情形下的性能：

第一阶段，在 S_1 上调用小批量大小足够小的 M-SPP，以获得 w_1，然后以 w_1 作为初始化。第二阶段则对其余小批次应用 M-SPP。随后，在定理2中，我们证明，第二阶段输出的期望超额风险可加速至如下量级

$$R(\bar{w}_T) - R^* \lesssim \frac{L^2(R(w_0) - R^*)}{\lambda^2 n^2 T^2} + \frac{LR^*}{\lambda n T}, \tag{4}$$

其中成立 ，无论小批量大小 n 与迭代次数 T之间的相对关系如何，均成立。

除了上述期望风险界之外，我们还基于算法稳定性理论推导出了 M-SPP 的高概率模型估计误差界。我们的偏差分析针对 M-SPP 的一种无放回抽样变体进行（见算法 3）。对于具有二次增长性质的总体风险，除内循环次优性ϵ_t带来的加性项外，我们在定理 6 中建立了以下估计误差偏差界 $D(\bar{w}_T, W^*)$，该界以至少 $1-\delta$ 的概率成立， S 同时对抽样随机性取期望：

$$D(\bar{w}_T, W^*) \lesssim \frac{\sqrt{L\log(1/\delta)}\log(T)}{\lambda\sqrt{nT}} + \sqrt{\frac{[R(w_0) - R^*]}{\lambda T^2} + \frac{LR^*}{\rho\lambda^2 nT}}.$$

当 $T =\Omega(n$时，忽略对数因子后，上述界在总样本量 $N = nT$方面与已知的统计估计极小极大下界相匹配，即使不受计算限制亦然（Tsybakov，2008年）。

为凸优化中的 SPP 类方法凸显本工作的核心贡献，以下关于 M-SPP 的三项新见解使我们的结果有别于目前已知的最佳结果：

1. 首先，式（3）和（4）中的快速收敛率揭示了噪声水平对收敛速率的影响，该噪声水平由 R^* 量化；这一点此前尚未在 SPP 类方法中得到认识。这些界适用于光滑损失函数，因此补充了此前针对 Lipschitz 损失的结果（Patrascu 和 Necoara，2017年；Wang 等人，2017b；Davis 和 Drusvyatskiy，2019年）。

2. 其次，式（3）和（4）中的风险界是在总体风险满足二次增长条件的前提下建立的。这一条件远弱于以往分析中通常采用的逐点损失强凸性假设；后者用于使 SPP 类方法达到相当的收敛速率（Toulis 和 Airoldi，2017年；Wang 等人，2017b；Asi 等人，2020年）。

3. 第三，我们从一致算法稳定性的角度对 M-SPP 进行偏差分析，据我们所知，此前关于 SPP 类方法的研究尚未涉及这一问题。

需要强调的是，尽管我们通过实证研究对 M-SPP 的数值特性提供了一些见解，但本研究主要是一项理论贡献。

1.3 相关工作

我们的工作处于机器学习研究中两个方向的交汇处：随机优化和算法稳定性理论。这两个方向都得到了广泛研究，文献中已建立了大量出色且富有洞见的理论成果。接下来，我们将不完全地回顾一些与本研究密切相关的代表性工作。

随机优化。自 Robbins 和 Monro（1951）的开创性工作以来，随机梯度下降（SGD）方法已得到广泛研究，用于近似求解问题 (1) 的简化版本，其中 $r \equiv 0$（Zhang，2004；Nemirovski 等人，2009年；Rakhlin 等人，2012；Bottou 等人，2018）。对于复合形式，人们开发了大量近端随机梯度下降方法，以便在存在潜在非光滑正则项的情况下高效优化（Hu 等人，2009年；Duchi 等人，2010；Ghadimi 和 Lan，2012；Lan，2012；Kulunchakov 和 Mairal，2019年）。为应对随机梯度下降在步长选择和数值不稳定性方面的挑战（Nemirovski 等人，2009年；Bach 和 Moulines，2011），近年来，人们研究了多种更复杂的方法，包括隐式随机/在线学习（Crammer 等人，2006；Kulis 和 Bartlett，2010；Toulis 等人，2016；Toulis 和 Airoldi，2017年）以及随机近端点（SPP）方法（Bertsekas，2011；Patrascu 和 Necoara，2017年；Asi 和 Duchi，2019a,b；Davis 和 Drusvyatskiy，2019年），以提高随机（复合）优化的稳定性和自适应性。例如，在本文考虑的复合优化情形下，标准 SPP 的迭代过程可表示为以下递推形式，其中 $i \geq 1$:

$$\hat{w}_i^{\text{spp}} := \arg\min_{w \in \mathcal{W}} \ell(w; z_i) + r(w) + \frac{\gamma_i}{2}\|w - \hat{w}_{i-1}^{\text{spp}}\|^2, \tag{5}$$

其中，$z_i \sim \mathcal{D}$ 是一个随机数据样本，γ_i 是正则化模量，而 $\|\cdot\|$ 表示欧几里得范数。与标准随机梯度下降方法相比，后者每次迭代的建模简单，但容易受到步长选择的影响；SPP 方法在目标函数近似方面更加准确，因此对算法超参数选择的稳定性显著提升，同时还能保证最优收敛性质（Asi 和 Duchi，2019a，b）。

上述（近端）随机优化方法的一项突出优势是，其收敛保证可直接应用于总体风险；对于强凸问题，经过$\mathcal{O}(\frac{1}{T})$轮 T 迭代后，即可达到极小极大最优速率（Nemirovski 等人，2009；Agarwal 等人，2012；Rakhlin 等人，2012）。对于大规模机器学习而言，内存效率的提升是采用随机优化方法而非批量优化方法的另一项实际优势。然而，由于按顺序处理的特性，随机优化方法在并行化方面往往效率较低，尤其是在分布式计算环境中，模型更新需要节点间进行大量通信（Bottou 等人，2018）。

经验风险最小化。与随机梯度下降类方法和在线学习处于相反的一端，下面定义的（复合）经验风险最小化（ERM，又称M估计）是另一种广受欢迎的

统计学习中受到广泛研究的形式Lehmann 和 Casella，2006）：

$$\hat{w}_S^{\text{erm}} := \arg\min_{w \in \mathcal{W}} \left\{ R_S(w) := \frac{1}{N} \sum_{i=1}^{N} \ell(w; z_i) + r(w) \right\}.$$

得益于有限和结构，研究者已为经验风险最小化建立了大量具有线性收敛速率的随机增量算法，包括随机方差缩减梯度（Johnson and Zhang，2013；Xiao and Zhang，2014）、SAGA（Defazio 等人，2014）和 Katyusha（Allen-Zhu，2017）等。从分布式计算的角度看，经验风险最小化相较于 SGD 类方法的一项固有优势在于，它能够更好地利用数据样本之间的统计相关性，设计通信高效分布式优化算法（Jaggi 等人，2014；Shamir 等人，2014；Zhang and Lin，2015；Lee 等人，2017）。与随机优化方法不同，批量算法或增量算法的泛化性能本质上由经验风险最小化的泛化性能决定（Bottou and Bousquet，2007）；经验风险最小化长期以来一直受到研究，并已有许多富有启发性的成果（Vapnik，1999年；Bartlett 等人，2005；Srebro 等人，2010；Mei 等人，2018）。特别是对于强凸风险函数，经验风险最小化有可能达到 $\mathcal{O}(\frac{1}{N})$ 收敛速率（Bartlett 等人，2005；Koltchinskii，2006；Zhang 等人，2017），但对于参数化学习模型，这些快速收敛率通常依赖于维数。

人们已经认识到，就泛化、运行时间、存储和并行化效率而言，随机梯度下降类方法与经验风险最小化类方法无法相互占优。这促使近期研究者尝试提出所谓的随机模型方法，以兼得两者之长。其中，一种广受欢迎的组合范式是小批量近端更新，即在每次迭代中，通过（近似）求解随机小批量上的局部经验风险最小化问题来更新模型（Li 等人，2014；Wang 等人，2017b；Asi 等人，2020年；邓和高，2021年）。这种策略可视为 SPP 算法的小批量扩展，研究表明，它能够在计算、通信和内存效率之间实现显著更优的权衡，适用于大规模分布式机器学习（Li 等人，2014；Wang 等人，2017年a）。另一种做法是提出多种增量有限和算法的在线扩展，例如流式随机方差缩减梯度（Frostig 等人，2015）和流式SAGA（Jothimurugesan 等人，2018），用于随机优化；这些方法在保证方面可与经验风险最小化相媲美，但计算成本更低。

算法稳定性与泛化。自 Bousquet 和 Elisseeff（2002年）的开创性工作以来，算法稳定性得到了广泛研究，并在为强凸经验风险最小化估计量建立泛化界方面取得了显著成果（Zhang，2003；Mukherjee 等人，2006；Shalev-Shwartz 等人，2010）。特别是，基于一致稳定性概念的方法给出了强凸经验风险最小化的最新风险界（Feldman 和 Vondr´ak，2018，2019；Bousquet 等人，2020；Klochkov 和 Zhivotovskiy，2021）。Hardt 等人（2016）证明，通过（随机）梯度下降得到的解对于光滑凸或非凸损失函数具有稳定性。对于非光滑凸损失，随机梯度下降所诱导的稳定性泛化界已在期望意义下（Lei 和 Ying，2020）或偏差意义下建立，

（Bassily 等人，2020）。在稀疏学习中，算法稳定性理论已被用于推导广泛使用的迭代硬阈值算法（IHT）的泛化界（Yuan 和 Li，2022年）。从一致算法稳定性的视角出发，研究者已针对凸（Wang 等人，2017b）和弱凸（邓和高，2021年）利普希茨损失研究了 M-SPP 的收敛速率。尽管我们的研究与 Wang 等人（2017b）；邓和高（2021）的研究思路相似，但针对光滑凸损失函数所定制的分析截然不同，所得快速收敛率在低噪声统计情形下尤为重要（Srebro 等人，2010）。

1.4 符号说明与论文结构

符号说明。我们分析中频繁使用的关键量和符号汇总于表1。

符号说明	定义
n	小批量大小
T	迭代轮次
N	访问的样本总数，即 $N = nT$
f	假设
ℓ	损失函数
r	正则化项
R^ℓ	总体风险： $R^\ell(w) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f_w(x)\ y)]$
R	复合总体风险： $R(w) := R^\ell(w) + r(w)$
R^*	复合风险的最优值，即 $R^* := \min_{w\in\mathcal{W}} R(w)$
W^*	复合风险的最优解集，即 $W^* := \arg\min_{w\in\mathcal{W}} R(w)$
S_t	时间点 t 的数据小批量
S_I	I 上数据小批量的并集，即 $S_I := \{S_t\}_{t\in I}$
R_S^ℓ	S 上的经验风险，即 $R_S^\ell(w) := \frac{1}{\lvert S\rvert}\sum_{(x,\ y)\ \in S} \ell\ (f_w\ (x,y)$
R_S	S 上的复合经验风险，即 $R_S(w) := R_S^\ell(w) + r(w)$
ϵ_t	时间点 t 的小批量风险最小化精度
$\lVert w\rVert_1$	向量 w 的 ℓ_1-范数，即 $\lVert w\rVert_1 := \sum_i \lvert[w]_i\rvert$
$\lVert w\rVert$	向量 w 的欧几里得范数
$D(w, W^*)$	从 w 到 W^* 的距离，即 $D(w, W^*) = \min_{w^*\in W^*} \lVert w - w^*\rVert$
$[T]$	$[T] := \{1, ..., T\}$
$\mathbf{1}_{\{C\}}$	条件 C 的指示函数

表1：符号说明表。

文章结构。本文内容安排如下：第2节分析具有凸且光滑损失函数的精确 M-SPP 的风险界，并提出一种两阶段变体以进一步提升收敛性能。第3节将分析扩展至允许非精确 M-SPP 迭代这一更贴近实际的情形。第4节研究 M-SPP 的高概率估计误差界。第5节对一些密切相关的结果进行了全面比较。第6节介绍用于验证理论和评估算法的数值研究。第7节给出结论。所有主要结果的证明以及关于 M-SPP 迭代稳定性的一些补充结果均列于附录。

2 光滑损失下 M-SPP 的精细分析

本节将运用算法稳定性理论中的工具，分析 M-SPP 在光滑凸损失函数下的收敛速率。为简化符号并清晰呈现核心思想，以下暂且假设 M-SPP 迭代过程中的内循环复合经验风险最小化（2）已通过 $\epsilon_t \equiv 0$ 精确求解，即

$$w_t = \arg\min_{w\in\mathcal{W}} \left\{ F_t(w) := R_{S_t}(w) + \frac{\gamma_t}{2}\|w - w_{t-1}\|^2 \right\}. \tag{6}$$

非精确变体（即 $\epsilon_t > 0$）的完整收敛分析将在第 3 节中通过略为复杂的扰动分析给出。

2.1 基本假设

我们首先介绍后续分析中将用到的一些基本假设。若可微函数 $g: \mathcal{W} \mapsto \mathbb{R}$ 满足 $\forall s,t \in \mathbb{R}$,则称其为 L-光滑。

$$\left|g(w) - g(w') - \langle \nabla g(w), w - w'\rangle\right| \le \frac{L}{2}|w - w'|^2.$$

如下述假设所正式表述的，我们假设个体损失函数是凸且 L-光滑的，例如，二次损失（用于回归）和逻辑损失（用于预测）均满足这一条件。

假设 1。损失函数 ℓ 关于其第一个参数是凸且 L光滑的。此外，我们假设正则化项 r 在 $\mathcal{W}$上是凸的。

定义 $D(w, W^*) := \min_{w^*\in W^*} \|w - w^*\|$ 为从 w到极小化解集合 W^*的距离。下一个假设要求总体风险具有如下特征：偏离极小化解集合时呈现二次增长（Anitescu，2000；Karimi 等人，2016）。

假设2。总体风险函数 R满足 $R(w) \ge R^* + \frac{\lambda}{2}D^2(w, W^*)$ $\forall w \in \mathcal{W}$，其中某个 $\lambda > 0$。

显然，二次增长性质可由传统的强凸性条件（在极小化器附近）推出；许多流行的学习模型都满足该条件，包括线性回归、逻辑回归、广义线性模型、平滑 Huber 损失以及其他各种统计估计问题。特别地，当 2 满足 R^ℓ 为强凸且r 为凸时，假设 2 成立。请注意，对于具有二次增长性质的风险函数，针对 Lipschitz 损失函数的 M-SPP 先前分析（Wang 等人，2017b）通常不适用，因为如果允许无穷大的距离 $\min_{w^* \in W^*} \|w - w^*\| \to \infty$，假设 2 意味着损失函数的 Lipschitz 常数可能任意大。

2.2 主要结果

下面的定理给出了我们的主要结果，刻画了在损失函数光滑且总体风险函数满足二次增长条件时，精确 M-SPP 的期望收敛速率。回顾一下，$N = nT$ 表示截至迭代计数器 T 所访问的数据点总数。

定理 1。假设 *1* 和 *2* 成立。考虑 $\epsilon_t \equiv 0$ 以及加权平均输出 $\bar{w}_T = {}^{2}{}_{T\,(T+1)} \sum_{t=1}^{T} t w_t$，其中算法为 *1*。令 $\rho \in$ ~~$(0, 0.5]$~~ 为任意标量。

(a) 假设 $n \geq \frac{64L}{\lambda\rho}$。令 $\gamma_t = \frac{\lambda\rho t}{4}$，其中 $t \geq 1$。那么对于任意 $T \geq 1$,

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \leq \frac{4\rho\left[R(w_0) - R^*\right]}{T^2} + \frac{2^9 L}{\lambda\rho n T} R^*.$$

(b) 对于 $t \geq 1$，令 $\gamma_t = \frac{\lambda\rho t}{4} + \frac{16L}{n}$。那么对于任意 $T \geq 1$,

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \leq \left(\frac{4\rho}{T^2} + \frac{2^8 L}{\lambda n T}\right)\left[R(w_0) - R^*\right] + \left(\frac{2^{16} L^2}{\lambda^2\rho^2 n^2 T} + \frac{2^9 L}{\lambda\rho n T}\right) R^*,$$

证明。证明方法受到 Wang 等人(2017b) 针对利普希茨损失和逐样本强凸损失提出的一致稳定性论证的启发，并加入了若干新要素，以处理光滑损失和风险函数的二次增长。作为一个非平凡的关键步骤，我们证明，鉴于 Srebro 等人 (2010，引理 2.1) 的一项经典结果（该结果允许用光滑损失的函数值来界定其导数），可以将这些稳定性论证推广到光滑损失。关于该结果的完整证明，请参见附录 A.1 。

□

关于定理 1 有几点说明。

注释 1. 在部分（a）中，小批量大小必须足够大。在这种情况下，超额风险界由两个部分组成：第一项是与初始间隙相关的偏差部分 $R(w_0) - R^*$，其衰减速率为 $\mathcal{O}(\frac{1}{T^2})$；第二项是与 R^*相关的方差部分，其以 $\mathcal{O}(\frac{1}{\lambda n T})$的主导速率趋于零。方差项表明，在低噪声设置下，当因子 R^* 相对较小时，收敛速率可以得到提升。特别是在满足 $R^* = 0$的可分情形下，定理 *1* 的超额风险界可达到 $\mathcal{O}(\frac{1}{T^2})$的速率。

注释 2. 部分（a）的结果有一个缺点：小批量大小必须远大于总体风险的条件数 R。相比之下，部分（b）的超额风险界适用于任意小批量大小。然而，其代价是偏差衰减项 $\mathcal{O}(\frac{1}{T^2}+\frac{1}{\lambda nT})$ 相对较慢；当 $T \gg n$ 时，该项由 $\mathcal{O}(\frac{1}{\lambda nT})$ 主导。

注 3. 设 $N = nT$ 为访问的数据点总数。当 $T \gg n$时，定理 *1* 中的 $\mathcal{O}(\frac{1}{N})$主导收敛速率与 SPP 类方法已有的结果相符（Wang 等人，*2017b*；Davis 和 Drusvyatskiy，*2019*），不过这些结果是在假设每个单独的损失函数都应为利普希茨连续且强凸的条件下得到的。与具有光滑损失函数的随机梯度下降（SGD）的$\mathcal{O}(\frac{1}{N})$ 收敛速率相比（Lei 和 Ying，*2020*，定理 12），定理 *1* 中的结果更强，且条件更宽松，体现在以下两方面：1）我们的界明确展示了 R^* 的影响，而它通常表示模型的噪声水平；2）我们只要求总体风险具有二次增长性质，而 Lei 和 Ying 的界*(2020*，定理 12）不仅要求损失函数为利普希茨连续，还假设经验风险是强凸的。

我们进一步探讨定理 ρ 中的标量选择1。我们重点讨论部分（a），类似的观察也适用于部分（b）。根据小批量大小 n 与迭代轮次 T相对大小的不同，下面分两种互补情形进行讨论：

- 情形I：小-n-大-T。假设 $n = \mathcal{O}(1)$，且允许 $T \to \infty$。在这种情形下，只需设置 $\rho = 0.5$，即可得到部分（a）中阶为 $\mathcal{O}\left(\frac{1}{T^2}+\frac{1}{\lambda nT}\right)$ 的收敛速率。

- 情形II：小-T-大-n。假设 $T = \mathcal{O}(1)$，且允许 $n \to \infty$。在此设置下，已知 $n \geq \frac{4T}{\lambda}$，那么通过近似最优的选择 $\rho = \sqrt{\frac{T}{n\lambda}}$，定理 1(a) 中的超额风险界将为 $\mathcal{O}\left(\frac{1}{T\sqrt{\lambda nT}}\right)$ 量级，这明显慢于情形 I 中先前的快速收敛率。这是符合直觉的，因为采用大批次的 M-SPP 表现得更像正则化经验风险最小化；众所周知，即使对于强凸问题，它也会呈现较慢的收敛速率（Shalev-Shwartz 等人，2010；Srebro 等人，2010）。尽管如此，这种小-T-大-n设置对于采用大批次的离线增量学习和分布式统计学习具有特殊意义（Li 等，2014；Wang 等人，2017b；You 等人，2020）。我们将在下一小节讨论这一关键情形。

2.3 两阶段 M-SPP 方法

算法 2：两阶段 M-SPP（M-SPP-TP）

输入：数据集 $S = \{S_t\}_{t=}^{T}{}_{1}$ 其中 $S_t := \{z_{i,t}\}_{i=}^{n}{}_{1}$ 独立同分布。$\sim \mathcal{D}^n$，正则化模量 $\{\gamma_t > 0\}_{t\in[T]}$。

输出：$\bar{w}_T$ 作为 $\{w_t\}_{2\le t\le T}$的加权平均。

初始化：指定 w_0的一个值。通常 $w_0 = 0$。

/* 第一阶段 */

将样本 S_1 划分为大小相等的互不相交的小批量 m；

在这些小批量上运行 M-SPP，以获得输出 w_1；

/* 第二阶段 */

以 w_1 初始化，在数据小批次 $\{S_t\}_{2\le t\le T}$ 上使用 $\{\gamma_t\}_{2\le t\le T}$ 运行 M-SPP，得到序列 $\{w_t\}_{2\le t\le T}$。

为弥补前文讨论中提到的不足，我们提出 M-SSP 的两阶段变体，如算法 2 所示，以提升其在小-T-大-n 情形下的性能。该过程可视为 M-SPP 的一种重启论证（Nemirovskii 和 Nesterov，1985；Renegar 和 Grimmer，2022；Zhou 等人，2022）。具体而言，第一阶段作为初始化步骤，调用 M-SPP 将 S_1 均匀划分为小批次，且小批量大小为 m，从而得到 w_1。随后从 w_1出发，第二阶段仅调用 M-SPP 处理后续的大批次 $\{S_t\}_{t\ge 2}$；若适用，这适合大规模并行化。下面的定理是将定理 1应用于这一两阶段 M-SPP 过程所得的推论。

定理 2. 假设*1* 和*2* 成立。考虑在算法*2*的第一阶段和第二阶段中实施 M-SPP 的 $\epsilon_t \equiv 0$。考虑第二阶段中的加权平均输出 $\bar{w}_T = \frac{2}{(T-1)(T+2)}\sum_{t=2}^{T} t w_t$。

(a) 假设 $n \ge \frac{128L}{\lambda}$。在第一阶段设置 $m = \frac{128L}{\lambda}$，并设置 $\gamma_t = \frac{\lambda t}{8}$ 以在第一阶段和第二阶段中实施 M-SPP。则对于任意 $T \ge 2$，$\bar{w}_T$ 满足

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \frac{L^2\left[R(w_0) - R^*\right]}{\lambda^2 n^2 T^2} + \frac{L}{\lambda n T} R^*.$$

(b) 在第一阶段设置 $m = \mathcal{O}(1)$，并设置 $\gamma_t = \frac{\lambda t}{8} + \frac{16L}{n}$ 以在第一阶段和第二阶段中实施 M-SPP。则对于任意 $T \ge 2$，$\bar{w}_T$ 满足

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \frac{L^2\left[R(w_0) - R^*\right]}{\lambda^2 n T} + \frac{L^3}{\lambda^3 n T} R^*.$$

证明。 有关该结果的证明，请参见附录 A.2 。 □

注 4. 定理 2的 (a) 部分表明，当小批量大小足够大时，两阶段 M-SPP 的超额风险界包含一个尺度为 $\mathcal{O}\left(\frac{1}{n^2T^2}\right)$ 的偏差衰减项，以及一个以 $\mathcal{O}(\frac{1}{nT})$的速率衰减的方差项。即使 T 的尺度相对较小，该速率仍然成立，因此优于定理 $\mathcal{O}\left(\frac{1}{T\sqrt{nT}}\right)$ 所隐含的 1 在小-T-大-n 情形下针对原始 M-SPP 的速率。值得一提的是，我们为 M-SPP 得出的界中的偏差分量和方差分量，均比强凸经验风险最小化的相应结果更快（Srebro 等人，2010）。

备注 5. 定理 2 部分（b）中的超额风险界适用于任意小批量大小，但代价是 $\mathcal{O}(\frac{1}{nT})$ 偏差衰减速率相对较慢。

2.4 任意凸风险的结果

我们进一步分析所提出的 M-SPP 算法在损失函数 ℓ 凸且光滑、但不要求复合风险 R 具有二次增长性质时的表现。以下是在这种一般设定下的主要结果。

定理 3. 假设 1 成立。令 $\gamma_t \equiv \gamma \geq \frac{16L}{n}$。设 $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$ 为算法1的平均输出。则

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \frac{\gamma}{T} D^2(w_0, W^*) + \frac{L}{\gamma n} R^*.$$

特别地，对于 $\gamma = \sqrt{\frac{T}{n}} + \frac{16L}{n}$，有

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \left(\frac{1}{\sqrt{nT}} + \frac{L}{nT}\right) D^2(w_0, W^*) + \frac{L}{\sqrt{nT}} R^*.$$

证明。 该结果的证明见附录A.3 。 □

注 6. 定理 3 的第一个界表明，对于任意 $\epsilon \in (0, 1)$，通过设定 $\gamma = \mathcal{O}\left(\frac{L}{\epsilon n}\right)$,$R(\bar{w}_t)$ 以 $\mathcal{O}(\frac{1}{nT\epsilon})$ 的速率收敛到 $(1+\epsilon)R^*$。该界与 Lei 和 Ying（2020，定理 4）针对光滑随机梯度下降方法所得的结果相符。定理 3 的第二个界进一步表明，通过设定 $\gamma = \mathcal{O}(\sqrt{\frac{T}{n}} + \frac{L}{n})$，$\bar{w}_T$ 的超额风险中偏差项和方差项均以 $\mathcal{O}(\frac{1}{\sqrt{nT}})$ 的速率衰减；就阶而言，这与针对利普希茨损失所得的相应界相符（Wang 等人，2017b，定理 4）。据我们所知，这种偏差—方差复合收敛速率对于采用凸且光滑损失函数的 SPP 类方法而言是新的。

类似于对 SPP 的鲁棒性分析（Asi 和 Duchi，2019年a，b），我们也分析了凸损失下 M-SPP 的迭代稳定性，以及其对正则化模量 γ_t 的选择的敏感程度。相关结果见附录 A.4，这些结果证实，选择 γ_t 对于生成不发散的估计误差序列而言，不受损失函数梯度尺度的影响。

3 非精确 M-SPP 的扰动分析

在上一节中，我们分析了 M-SPP 的收敛速率，假设其迭代过程中构造的内循环近端经验风险最小化子问题（2）均得到精确求解，即 $\epsilon_t \equiv 0$。为使分析更贴近实际，本节进一步对内循环近端经验风险最小化子问题仅需以一定精度 $\epsilon_t > 0$ 近似求解时的 M-SPP 进行扰动分析。首先，我们需要对正则化项 r 作出以下利普希茨连续性假设。

假设 3. 正则化项 r 在 $\mathcal{W}$ 上是利普希茨连续的，即 $|r(w) - r(w')| \leq G\|w - w'\|, \forall w, w' \in \mathcal{W}$。

例如，ℓ_1-范数正则项 $r(w) = \mu\|w\|_1$ 关于欧几里得范数满足此假设，因为 $|r(w) - r(w')| = \mu|\|w\|_1 - \|w'\|_1| \leq \mu\|w - w'\|_1 \leq \mu\sqrt{p}\|w - w'\|$。

以下定理给出了我们关于具有光滑损失函数的复合随机凸优化中非精确 M-SPP 收敛速率的主要结果。

定理 4. 假设 *1*、*2* 和 *3* 均成立。令 $\rho \in (0, 1/4]$ 为任意标量，并令 $\gamma_t = \frac{\lambda\rho t}{4}$。假设 $n \geq \frac{76L}{\lambda\rho}$。设 $\epsilon_t \leq \frac{\epsilon}{nt^4}$ 对某个 $\epsilon \in [0, 1]$成立。那么对于任意$T \geq 1$，算法 $\bar{w}_t = {}_{(T+1)}^{2} \sum_{t=1}^{T} t w_t$ 的加权平均输出满足*1*

$$\mathbb{E}\left[R(\bar{w}_t) - R^*\right] \lesssim \frac{\rho}{T^2}(R(w_0) - R^*) + \frac{L}{\lambda\rho n T}R^* + \frac{\sqrt{\epsilon}}{T^2}\left(\frac{L}{\lambda\rho} + G\sqrt{\frac{1}{\lambda\rho}}\right).$$

证明。关于该结果的证明，请参见附录 B.1。我们想强调的是，我们针对光滑损失的扰动分析与 Wang 等人（2017b）针对利普希茨损失函数所做的分析有很大不同。这主要是因为在光滑损失的情况下，损失函数的变化不再能由预测的变化作为上界，因此我们需要更仔细地处理正则化小批量经验风险近似最小化所引起的扰动。

□

我们依次对定理4作几点评论。

注 7. 定理 *4* 表明，定理 *1* 部分（a）中精确 M-SPP 的超额风险界可以推广到其非精确版本，前提是内循环小批量经验风险最小化问题 (2)求解得足够精确，例如 $\epsilon_t \leq \mathcal{O}\left(\frac{1}{nt^4}\right)$。类似地，定理 *1* 部分（b）中关于任意小批量大小的结果也可以推广到非精确 M-SPP；为避免冗余，此处略去。由于内循环小批量经验风险最小化问题是强凸的，且损失函数是光滑的，通常可以通过方差缩减随机梯度下降方法在对数时间内达到所需精度 $\mathcal{O}\left(\log\left(\frac{1}{\epsilon_t}\right)\right)$（Xiao 和 Zhang，*2014*）。

注 8. 与第 2.2 节末尾的讨论类似，通过指定ρ的取值，我们可以直接得到定理 4 的推论，从而更明确地展示关于 $N = nT$的收敛速率。对于两阶段 M-SPP，根据定理 4 ，我们还可以证明，只要小批量优化足够精确，定理 2 中的界就可以推广到非精确情形。这些推广基本上都很直接，因此略去。

在下面的定理中，我们给出非精确 M-SPP 的超额风险界，其中复合风险 R 是凸的，但不一定具有二次增长性质。

定理5。 假设 1 和 3成立。令 $\gamma_t \equiv \gamma \geq \frac{19L}{n}$ 。假设 ϵ_t ≤最小值 $\left\{\frac{\epsilon}{n^2t^5},\ \frac{2G^2}{9n^2\gamma}\right\}$ 其中 $\epsilon \in [0,1]$。则算法 $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$1 的平均输出 满足

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \frac{\gamma}{T}D^2(w_0, W^*) + \frac{L}{\gamma n}R^* + \left(\frac{L}{\gamma n} + \frac{\gamma}{LnT} + \frac{G}{\sqrt{\gamma}nT}\right)\sqrt{\epsilon}.$$

特别地，对于 $\gamma = \sqrt{\frac{T}{n}} + \frac{19L}{n}$，有

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \left(\frac{1}{\sqrt{nT}} + \frac{L}{nT}\right)D^2(w_0, W^*) + \frac{L}{\sqrt{nT}}R^* + \left(\frac{L+G}{\sqrt{nT}} + \frac{1}{nT}\right)\sqrt{\epsilon}.$$

证明。 该结果的证明见附录B.2 。 □

备注9。 定理 5 证实，定理 3 中针对精确 M-SPP 得到的超额风险界，对小批量近端经验风险最小化子问题中足够小的次优性 $\epsilon_t \leq \mathcal{O}(\frac{1}{n^2t^5})$ 具有容忍性。

4 高概率性能保证

在前两节中，我们分析了 M-SPP 的期望超额风险界。本节将继续研究 M-SPP 关于训练数据随机性的高概率保证，仍采用算法稳定性的概念。为此，我们首先介绍 M-SPP 的一种变体，该变体通过对给定的数据小批次进行无放回抽样来执行近端点更新。随后，我们证明，所提出算法的输出关于抽样随机性在期望意义下是一致稳定的。本节的主要结果是，对于强凸总体风险，我们建立了一个近乎最优的高概率（关于数据）估计误差界 $\|\bar{w}_t - w^*\|$，该界在内部数据抽样的随机性下成立于期望意义。此外，我们还为任意凸损失函数提供了一个高概率泛化界。

4.1 无放回抽样 M-SPP

算法 3：无放回抽样 M-SPP (M-SPP-SWoR)

输入：数据集 $S = \{S_t\}_{t=}^{T}{}_{1}$ 其中 $S_t := \{z_{i,t}\}_{i=}^{n}{}_{1} \overset{\text{独立同分布}}{\sim} \mathcal{D}^n$，正则化模量 $\{\gamma_t > 0\}_{t\in[T]}$。

输出：$\bar{w}_T$ 作为 $\{w_t\}_{1\le t\le T}$的加权平均值。

初始化：指定一个 w_0值。通常 $w_0 = 0$。

对于 $t = 1, 2, ..., T$ **do**

均匀随机地无放回抽取一个索引 $\xi_t \in [T]$ 。

估计满足以下条件的 w_t

$$F_t(w_t) \le \text{最小化}_{w\in\mathcal{W}} \left\{ F_t(w) := R_{S_{\xi_t}}(w) + \frac{\gamma_t}{2}\|w - w_{t-1}\|^2 \right\} + \epsilon_t, \tag{7}$$

其中，$\epsilon_t \ge 0$ 衡量次优性。

end

我们来考察算法 3 中所述的 M-SPP-SWoR（通过无放回抽样实现的 M-SPP）过程。给定一组 S 个 T 数据小批次后，该算法在每次迭代时都会从 S 中均匀随机地抽取一个小批次，且不放回，用于近端更新。经过T 轮迭代后，所有小批次均已用于更新模型。由于该过程仅仅是算法 1 中所述 M-SPP 的一种随机打乱变体，因此可以看出，前几节中针对 M-SPP 建立的所有期望界，在任何打乱实现方式下都可直接推广至 M-SPP-SWoR。正如我们将在下一小节中很快证明的，这种随机打乱方案有助于提升 M-SPP 的平均算法稳定性，进而为 M-SPP-SWoR 带来强有力的高概率保证。

4.2 一致稳定性分析

设 $S = \{S_t\}_{t\in[T]}$ 和 $S' = \{S'_t\}_{t\in[T]}$ 是两组数据小批次。记 $S_t \doteq S'_t$当 S_t 和 S'_t 仅在一个数据点上不同时，记为 $S \doteq S'$ 当 S 和 S' 仅在一个小批次以及该小批次中的一个数据点上不同时，记为。我们引入 M-SPP 的一致稳定性这一概念，以具体阐明一致算法稳定性的概念；一致算法稳定性是分析统计估计量及其学习算法的泛化界的有力工具（Bousquet 和 Elisseeff，2002年；Hardt 等人，2016；Feldman 和 Vondr´ak，2019年）。

定义 1（M-SPP 的一致稳定性）。若 M-SPP 算法相对于映射 h 是 ϱ-一致稳定的：$\mathcal{W} \mapsto \mathbb{R}^q$ 当且仅当 $\|h(\bar{w}_T) - h(\bar{w}'_T)\| \le \varrho$ 对任意一对数据集 $S \doteq S'$ 均成立。

下面的结果给出了原始 M-SPP（算法 1）的一致稳定性（相对于相同映射）界，该界确定性成立；同时还给出了 M-SPP- SWoR（算法 3）的相应界，该界在数据小批次抽样的随机性下以期望形式成立。

命题 1。 假设假设 *1* 成立，且损失函数有界，即满足 $0 \leq \ell(y', y) \leq M$对所有 y, y'均成立。令 $S = \{S_t\}_{t\in[T]}$和 $S' = \{S'_t\}_{t\in[T]}$为满足 $S \doteq S'$的两组数据小批次。那么

(a) 由 M-SPP（算法 *1*）分别在S 和 S' 上生成的加权平均输出 $\bar{w}_T$ 和$\bar{w}'_T$ 满足

$$\sup_{S,S'} \|\bar{w}_T - \bar{w}'_T\| \leq \frac{4\sqrt{2LM}}{n \min_{t\in[T]} \gamma_t} + \sum_{t=1}^{T} 2\sqrt{\frac{2\epsilon_t}{\gamma_t}}.$$

(b) 由 M-SPP-SWoR（算法 *3*）分别在 S 和 S' 上生成的加权平均输出 $\bar{w}_T$ 和$\bar{w}'_T$ 满足

$$\sup_{S,S'} \mathbb{E}_{\xi_{[T]}} \left[\|\bar{w}_T - \bar{w}'_T\|\right] \leq \sum_{t=1}^{T} \left\{ \frac{4\sqrt{2LM}}{nT\gamma_t} + 2\sqrt{\frac{2\epsilon_t}{\gamma_t}} \right\}.$$

证明。 证明见附录C.1 。 □

注释 10. 假设次优性 $\{\epsilon_t\}_{t\in[T]}$ 足够小。若将 $\gamma_t = \mathcal{O}(t)$ 设置为适用于具有二次增长性质的总体风险的取值，则命题 *1* 表明 M-SPP 是$\mathcal{O}(\frac{1}{n})$-一致稳定的；而对无放回抽样的随机性取期望时，M-SPP-SWoR 的一致稳定性参数显著改善，按 $\mathcal{O}\left(\frac{\log(T)}{nT}\right)$ 的速率缩放。若将$\gamma_t \equiv \sqrt{\frac{T}{n}}$设置为适用于一般凸损失的取值，则 M-SPP 将是 $\mathcal{O}(\frac{1}{\sqrt{nT}})$-一致稳定的，而 M-SPP-SWoR 在对抽样取期望时具有相同的一致稳定性参数。

在下面的定理中，基于命题 1中的一致稳定性界，我们推导出 M-SPP-SWoR 的估计误差上界 $D(\bar{w}_T, W^*)$；该上界对于数据分布以高概率成立，同时对于用于更新的小批次随机抽样以期望成立。

定理 6。 假设假设条件 *1*、*2*、*3* 成立，且损失函数 ℓ 在区间 $(0, M]$上有界。令 $\rho \in (0, 1/4]$ 为任意标量，并设 $\gamma_t = \frac{\lambda\rho t}{4}$。假设 $n \geq \frac{76L}{\lambda\rho}$。设 $\epsilon_t \leq \min\left\{\frac{\epsilon}{nt^4}, \frac{LM}{\lambda\rho n^2 T^2 t}\right\}$，对于某个 $\epsilon \in [0, 1]$成立。那么，以至少 $1-\delta$的概率，关于S，M-SPP-SWoR（算法 *3*）的加权平均输出满足

$$\begin{aligned}
&\mathbb{E}_{\xi_{[T]}} \left[D(\bar{w}_T, W^*)\right] \\
\lesssim &\frac{\sqrt{LM\log(1/\delta)}\log(T)}{\lambda\rho\sqrt{nT}} + \sqrt{\frac{\rho\left[R(w_0) - R^*\right]}{\lambda T^2} + \frac{L}{\lambda^2\rho nT}R^* + \frac{\sqrt{\epsilon}}{\lambda T^2}\left(\frac{L}{\lambda\rho} + G\sqrt{\frac{1}{\lambda\rho}}\right)}.
\end{aligned}$$

证明。该结果的证明见附录C.2 。 □

注释 11。我们讨论定理6中界的最优性。考虑 $\rho = \mathcal{O}(1)$。尺度为 $\mathcal{O}\big(\sqrt{\log(1/\delta)}\log(T)\sqrt{nT}\big)$ 的第一项表示在数据上以高概率获得泛化保证所需的额外开销。第二项与定理4中相应的期望估计误差界一致，也与已知的强凸随机梯度下降最优速率一致（Rakhlin 等人，*2012*；Dieuleveut 等人，2017年）。根据统计估计的极小极大下界（*Tsybakov*，2008年），定理6 中建立的估计误差界对于强凸风险最小化而言近乎最优。

最后，我们为任意凸总体风险函数给出了 M-SPP 的高概率泛化界。

定理 7. 假设 *1*和 *3* 成立，且损失函数 ℓ 在区间 $[0, M]$内有界。令 $\gamma_t \equiv \sqrt{\frac{T}{n}}$。假设 $\epsilon_t \leq \frac{LM}{4nT^2\sqrt{nT}}$。那么，以至少 $1-\delta$的概率，在S上，M-SPP（算法 *1*）的平均输出 $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$ 满足

$$|R(\bar{w}_T) - R_S(\bar{w}_T)| \lesssim \frac{(LM + G\sqrt{LM})\log(N)\log(1/\delta)}{\sqrt{nT}} + M\sqrt{\frac{\log(1/\delta)}{nT}}.$$

证明。 该结果的证明见附录 C.3 。 □

顺便指出，采用类似的一致稳定性论证，可以证明定理 7 中的高概率泛化界对凸非光滑损失函数同样成立。由于本文聚焦于光滑损失函数，详细分析超出了本文的范围，故此处略去。

5 与先前方法的比较

与 M-SPP 和 SPP 方法的比较。本文所考虑的 M-SPP 算法是 SPP 方法的小批量扩展。近年来，随机优化领域广泛关注 SPP 的收敛分析。特别是，对于包含N 个数据点的有限和优化，(Bertsekas, 2011) 提出并分析了一种增量 SPP 方法。对于采用线性预测模型和强凸 Lipschitz 损失的学习问题，(Toulis 等人, 2016) 针对 SPP 建立了一组在适当 $\gamma \in (0.5, 1]$条件下的 $\mathcal{O}(\frac{1}{N^\gamma})$ 收敛速率，其中 N为迭代计数器。对于任意凸损失函数，研究人员分析了 SPP 的非渐近收敛性能：对于 Lipschitz 损失，得到 $\mathcal{O}(\frac{1}{\sqrt{N}})$ 速率 (Patrascu 和 Necoara, 2017; Davis 和 Drusvyatskiy, 2019)；对于强凸且 Lipschitz 的损失函数 (Davis 和 Drusvyatskiy, 2019) 或光滑损失函数 (Patrascu 和 Necoara, 2017)，得到 $\mathcal{O}(\frac{1}{N})$ 速率；对于强凸非光滑损失函数 (Asi 和 Duchi, 2019b)，则得到 $\mathcal{O}\left(\frac{\log(N)}{N}\right)$ 速率。近期研究表明，对于强凸损失函数，$\mathcal{O}\left(\frac{\log(N)}{N}\right)$ 速率也适用于 M-SPP (Asi 等，2020)。针对弱凸损失函数（例如，凸损失与光滑映射的复合），研究人员已在有 (Duchi and Ruan, 2018) 或无 (Davis and Drusvyats

kiy,

2019) 复合结构的随机优化中，研究了 SPP 的渐近和非渐近行为。在相关研究中，与我们工作最密切的是为通信高效分布式优化开发的小批量近端更新方法 (Wang 等人，2017b)。类似地，从算法稳定性的角度出发，研究者为该方法建立了适用于凸性任意的 Lipschitz 损失函数（ $\mathcal{O}(\frac{1}{N^\gamma})$ 或强凸损失函数（$\gamma = 0.5$) 的 速率。与这些已有结果相比，我们关于 M-SPP 的收敛结果在以下方面具有新意：

- 针对光滑损失函数推导了收敛速率，并明确展示了统计模型的噪声水平（如 R^* 所编码）对收敛性能的影响；此前对于 SPP 类方法尚未发现这种影响。

- 本文获得的 $\mathcal{O}(N^{-1})$ 快速收敛率适用于具有二次增长性质的总体风险，而无需每个瞬时损失函数都是强凸的。

- 我们给出了 M-SPP 的一种无放回抽样变体的近最优模型估计误差界：对于数据的随机性，该界以高概率成立；对于抽样的随机性，则在期望意义下成立。

与随机梯度下降和经验风险最小化的比较。与定理 1 和定理 3 中的结果类似，针对最小二乘回归的加速 SGD（Dieuleveut 等人，2017年）以及适用于一般凸光滑学习问题的小批量 SGD（M-SGD）（Woodworth 和 Srebro，2021年），偏差—方差复合速率早已为人所知。虽然这些结果的风格相似，但我们采用了截然不同的算法框架，并运用了相当不同的证明技术。具体而言，与 Woodworth 和 Srebro（2021年）不同，我们的分析既不需要了解模型尺度（这在实际问题中通常无法获取），也不依赖针对强凸问题的重启论证。此外，对于采用光滑损失函数的随机梯度下降，在最优总体风险为零这一理想的干净情形下，近期已通过稳定性理论证明了快速收敛率 $\mathcal{O}(\frac{1}{N})$（Lei 和 Ying，2020年，定理 4）。对于 $\gamma = \mathcal{O}(\frac{1}{n})$, 我们定理 3 的第一个界与 M-SPP 中的相应界相匹配。对于强凸问题，我们在定理 1 中的结果强于（Lei 和 Ying，2020年，定理 12）：前者（我们的结果）只要求总体风险具有二次增长性质，而后者要求损失函数是利普希茨的，且经验风险是强凸的。最后，对于凸经验风险最小化，Srebro 等人（2010）和 Zhang 等人（2017）在一些更严格的条件下证明了类似的复合风险界，例如研究定义域有界且样本量巨大，并满足 $N \gg p$。

表2 总结了本文所得风险界与此前针对（M-）SPP、（M-）SGD 和经验风险最小化的若干风险界之间的比较。

方法	文献	风险界	条件		
			Loss	R	R_S
M-SPP	Asi 等人 (2020年)	$\mathcal{O}\left(\frac{\log(N)}{N}\right)$	光滑凸	—	—
	Wang 等人(2017b)	$\mathcal{O}\left(\frac{1}{N}\right)$	Lipschitz & 光滑凸	—	—
	定理1 (我们的工作)	$\mathcal{O}\left(\frac{1}{T^2}+\frac{R^*}{N}\right)$ 或 $\mathcal{O}\left(\frac{1}{T^2}+\frac{1+R^*}{N}\right)$	光滑且凸	qg	—
	定理3 (我们的工作)	$\mathcal{O}\left(\frac{1}{N}+R^*\right)$ 或 $\mathcal{O}\left(\frac{1+R^*}{\sqrt{N}}\right)$	光滑且凸	—	—
SPP	Asi 和 Duchi (2019年b)	$\mathcal{O}\left(\frac{\log(N)}{N}\right)$	光滑凸	—	—
	Patrascu 和 Necoara (2017年)	$\mathcal{O}\left(\frac{1}{N}\right)$	光滑强凸且光滑凸	—	
	Davis 和 Drusvyatskiy(2019年)	$\mathcal{O}\left(\frac{1}{N^2}+\frac{1}{N}\right)$	Lipschitz 且光滑凸	—	
M-SGD	Woodworth 和 Srebro (2021年)	$\mathcal{O}\left(\frac{1}{T^2}+\frac{1}{N}+\sqrt{\frac{R^*}{N}}\right)$	光滑强凸且凸	—	—
		$\mathcal{O}\left(e^{-T}+\frac{R^*}{N}\right)$	光滑且凸	qg	—
SGD	Dieuleveut 等人 (2017年)	$\mathcal{O}\left(\frac{1}{N^2}+\frac{R^*}{N}\right)$	二次	光滑凸	—
	Lei 和 Ying (2020年)	$\mathcal{O}\left(\frac{1}{N}+R^*\right)$ 或 $\mathcal{O}\left(\frac{1+R^*}{\sqrt{N}}\right)$	光滑且凸	—	光滑凸
	Rakhlin 等人(2012)	$\mathcal{O}\left(\frac{1}{N}\right)$	Lipschitz & 光滑且凸	光滑凸	—
ERM	Zhang 等人（2017年）	$\mathcal{O}\left(\frac{p}{N}+\frac{R^*}{N}\right)$ 或 $\mathcal{O}\left(\frac{1}{N^2}+\frac{R^*}{N}\right)$ 对于 $N \gtrsim p$	光滑且凸	Lip 以及 光滑凸	—
	Srebro 等人（2010）	$\mathcal{O}\left(\frac{1}{N}+\sqrt{\frac{R^*}{N}}\right)$	光滑且凸	—	—

表 2：我们的方法与此前一些关于 M-SPP 和 SPP 以及随机梯度下降和经验风险最小化的风险界的比较。回顾一下， T 是迭代次数， N 是访问的样本总数。此处列出的所有界均在期望意义下成立。这里使用了以下缩写：凸（凸），s.cvx（强凸），Lipschitz（利普希茨连续），sm（光滑），qg（二次增长）。

6 实验

我们开展了一组数值研究，以展示小批量随机近端点方法在（复合）统计学习问题中的收敛性能，并回答与本文确立的关键理论和算法相关的以下3个问题：

- 问题1： 小批量大小和统计学习模型的噪声水平如何影响光滑损失函数下 M-SPP 的收敛速度? 这个问题主要是为了验证定理 1 和定理 5，并通过第 6.1 节中关于 Lasso 估计的仿真实验予以回答。

- 问题2： 在小-T-大-n设定下，M-SPP 的两阶段变体能否优于 M-SPP? 第 6.1 节中给出的仿真结果也回答了这个与定理2 的验证相关的问题。

- 问题3： M-SPP（-TP）方法与 M-SGD 的收敛性能相比如何? 第 6.2 节中逻辑回归任务的真实数据实验结果回答了这个算法比较问题。

6.1 仿真实验

我们首先通过仿真实验，验证当应用于广泛使用的 Lasso 回归模型（Wainwright，2009年）时，我们关于光滑损失的理论结果。该模型采用二次损失函数$\ell(f_w(x)\ y) = \frac{1}{2}(y - w^\top x)^2$和 $r(f_w) = \mu\|w\|_1$，其中 μ是 ℓ_1-惩罚模量。给定模型参数 $\bar{w} \in \mathbb{R}^p$以及从标准高斯分布 $\mathcal{N}(0, I_{p\times p})$中抽取的特征点 $x \in \mathbb{R}^p$，响应变量 y根据线性模型 $y = \bar{w}^\top x + \varepsilon$生成，并加入随机高斯噪声 $\varepsilon \sim \mathcal{N}(0, \sigma^2)$。在这种情况下，总体风险函数可以用闭式表达为

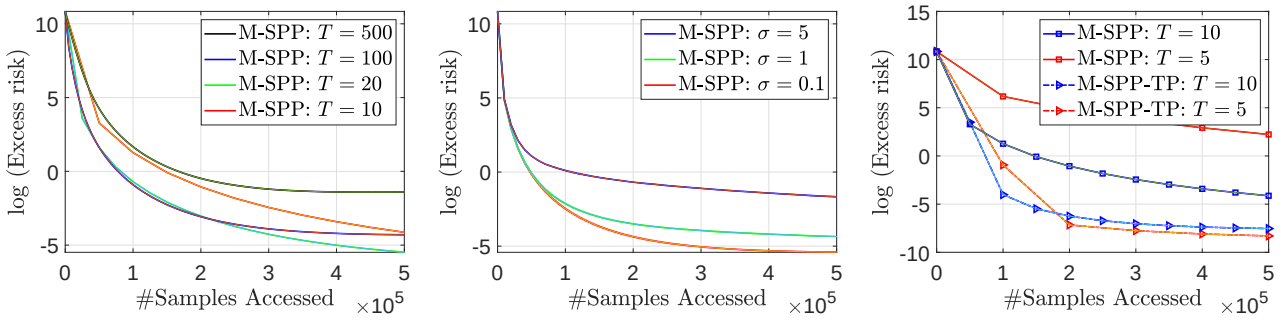
$$R(w) = \frac{1}{2}\|w - \bar{w}\|^2 + \frac{\sigma^2}{2} + \mu\|w\|_1.$$

给定一组从上述数据分布中抽取的 T个随机 n-小批量$\{S_t = \{x_{i,t}, y_{i,t}\}_{i\in[n]}\}_{t\in[T]}$，我们旨在评估 M-SPP 向 R 的极小化解收敛的性能，该解可表示为

$$w^* = (\bar{w} - \mu)_+ - (-\bar{w} - \mu)_+,$$

where $(\cdot)_+$ 是一个逐元素函数，用于保留向量的正值部分 r.

我们使用 $p = 5000$ 和 $N = nT = 100p$进行测试，并考虑一个设定正确的稀疏回归模型，其中真实参数向量 $\bar{w}$ 是 k-稀疏的，满足 $k = 0.2p$，其非零项从均值为零的高斯分布中采样。我们设置 $\mu = 10^{-3}$ 并初始化 $w^{(0)} = 0$。内循环的小批量近端 Lasso 子问题通过标准近端梯度下降法进行优化，并采用以下两种终止条件之一：1）相邻两次迭代的目标值之差小于 10^{-3}；2）迭代步数达到 1000。



(a) 不同 T 下的结果。 (b) 不同 σ 下的结果。 (c) M-SPP *versus* M-SPP-TP

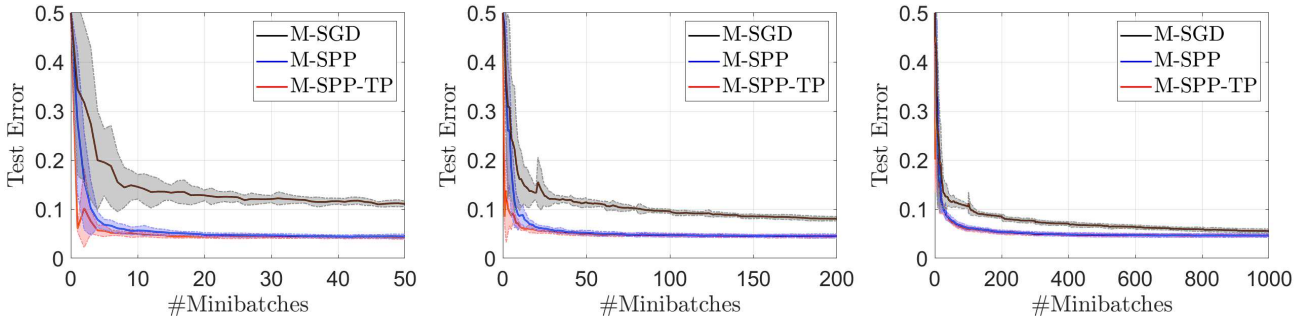
图1：Lasso 回归的仿真实验：M-SPP 和 M-SPP-TP 的收敛表现。y 轴表示超额风险的对数刻度。

为验证理论，我们考虑以下两种实验设置：

- 我们固定噪声水平 $\sigma = 0.1$ 并研究不同 $T \in \{10, 20, 100, 500\}$ 对 M-SPP 收敛性的影响。图 1(a) 展示了超额风险随样本量变化的曲线，采用半对数布局，y 轴表示超额风险的对数刻度。从这组曲线中可以观察到一个明显趋势：在初期，当小批次总数相对较大时（例如，$T \in \{20, 100\}$），M-SPP 收敛得更快。这与定理 1 关于 T和 n 对收敛速率影响的预测一致。而在后期，当 T 相对较大时（例如，$T \in \{100, 500\}$），则表现出相对较慢的收敛行为。定理 4 中的非精确分析可以解释这一现象：该定理表明，为了保证所需的收敛速率，当 T 相对较大时，内循环近端经验风险最小化更新需要达到极高的精度。因此，这组结果回答了问题 1 中关于小批量大小对收敛速率影响的问题。

  此外，在这一设置下，我们还比较了 M-SPP 与其两阶段变体 M-SPP-TP 在$T \in \{5, 10\}$方面的表现。相关结果如图1(c)所示，结果表明，M-SPP-TP 显著提升了 M-SPP 在小T-大n 情形下的收敛速度。这一观察结果支持定理2 的结论，并对问题 2 作出了肯定回答。

- 我们固定 $T = 50$ 并研究噪声水平变化 $\sigma \in \{0.1, 1, 5\}$ 对收敛性能的影响。结果如图 1(b)所示。从这组结果可以看出，在噪声水平 σ相对较小时，收敛速度更快；而当 σ 足够小时（例如 $\sigma \leq 1$），速度对噪声水平的变化不再敏感。这与定理 1的预测一致，同时要记住 $R^* = \frac{1}{2}\|w^* - \bar{w}\|^2 + \frac{\sigma^2}{2} + \mu\|w^*\|_1 \leq \|\bar{w}\|^2 + \frac{1}{2}\sigma^2$这一事实。通过这组结果，问题 1 中关于噪声水平对收敛性能影响的疑问得到了解答。



(a) $n = N/5$ (b) $n = N/20$ (c) $n = N/100$

图 2：逻辑回归的真实数据结果：不同小批量大小下 gisette 的测试误差收敛对比。

6.2 真实数据实验

我们进一步将所提出的方法与用于二元预测问题的 M-SGD 进行比较，并采用逻辑损失函数$\ell(w^\top x, y) = \log(1+\exp(-yw^\top x))$。此处，M-SGD 方法由 SGDLibrary（Kasai，2017年）中的随机梯度下降求解器实现。对于 M-SPP 和 M-SPP-TP，内循环中的小批量近端经验风险最小化问题由同一个随机梯度下降求解器求解，随机梯度下降批量大小固定为 10，数据处理仅进行一个训练轮次。对于所有考察的方法，我们均将 $w^{(0)} = 0$ 初始化为相同值。

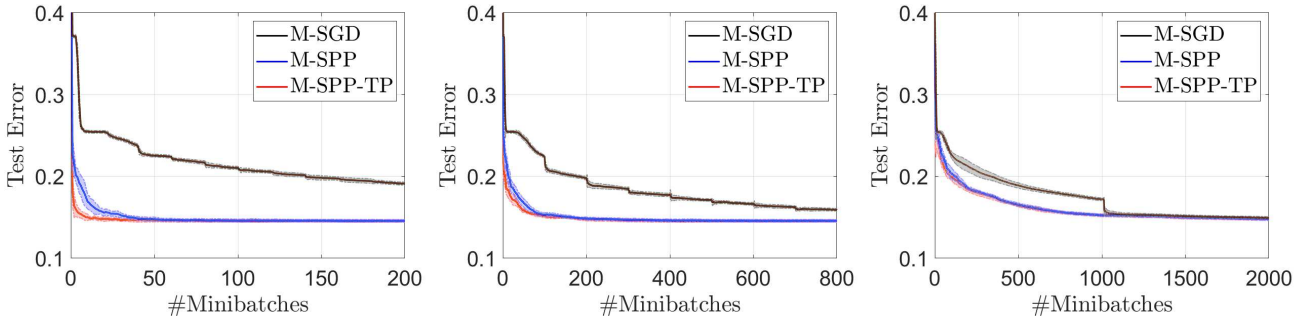
我们使用两个公开数据集进行评估：`gisette` 数据集（Guyon 等人，2004），包含 $p = 5000$，$N = 6000$ 以及 `covtype.binary` 数据集（Collobert 等人，2001年），包含 $p = 54$，$N = 581,012$[1]。对于每个数据集，我们将一半样本用作训练集，其余样本用作测试集。我们关注小批量大小 n 对模型预测性能的影响，并以测试误差衡量预测性能。所有纳入考察的随机算法均执行 10 个训练轮次的数据处理，因此小批次总数为 $T = N/n \times 10$。我们对每个实验进行 10 次重复，每次都随机划分数据，并报告均值及误差条。

在图2中，我们展示了在不同小批量大小 $n \in \{\frac{N}{5}, \frac{N}{20}, \frac{N}{100}\}$下，gisette上测试误差随访问的小批次数量变化的曲线（误差条以颜色阴影表示）。从这些曲线中我们可以观察到：

- 在相同的小批量大小下，M-SPP 和 M-SPP-TP 比 M-SGD 收敛得更快、更稳定，尤其是在小批量大小相对较大时（见图2(a)）。这是符合预期的，因为当小批量大小变大时，M-SGD 趋近于梯度下降法，而 M-SPP 趋近于经验风险最小化问题。这回答了实验部分开头提出的问题3。

- 在迭代初期，M-SPP-TP 的收敛表现比 M-SPP 更为出色，尤其是在小批量大小相对较大的情况下。这与我们在定理 1和定理 2中的理论结果一致。

[1]两个数据集均可在https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/获取。



(a) $n = N/20$ (b) $n = N/100$ (c) $n = N/1000$

图3：逻辑回归的真实数据结果：不同小批量大小下 covtype.binary 上测试误差收敛情况的比较。

图 3展示了 `covtype`数据集上的相应结果$n \in \{\frac{N}{20}, \frac{N}{100}, \frac{N}{1000}\}$。从这组结果中，我们再次看到，在相同的小批量大小下，M-SPP 和 M-SPP-TP 的表现始终优于 M-SGD；而在小批量大小相对较大时（例如 $n = \frac{N}{20}$），M-SPP-TP 的收敛速度快于 M-SPP。

7 结论与未来展望

本文针对具有光滑凸损失的小批量随机近端点方法，提出了改进的收敛分析。在总体风险满足二次增长条件时，我们证明，小批量大小为 n、迭代次数为 T 的 M-SPP 以复合速率收敛，该速率由一个 $\mathcal{O}(\frac{1}{T^2})$ 偏差衰减项和一个 $\mathcal{O}(\frac{1}{N})$ 方差衰减项构成。在小-n大-T 情形下，这一结果显著改进了此前相关的 SPP 类方法结果；这些方法通常要求每个瞬时损失函数均为 Lipschitz 且强凸。另一方面，在小-T大-n 情形下，我们提出了 M-SPP 的两阶段加速方法，将 $\mathcal{O}(\frac{1}{T^2})$ 偏差衰减速率提升至 $\mathcal{O}\left(\frac{\log(N)}{N^2}\right)$。或许最有趣的理论发现是，（占主导地位的）方差衰减项依赖于总体风险最小值的一个因子，这解释了 M-SPP 在低噪声统计情形下更快的收敛表现，我们的数值证据也支持了这一点。除了期望风险界之外，我们还为 M-SPP 的一种随机打乱变体推导出了近乎最优的参数估计误差界；该界对数据分布具有高概率保证，对随机打乱取期望时也成立。总之，我们的理论为理解凸 M-SPP 类算法奠定了新颖且更坚实的基础。这类算法近年来在大规模机器学习的理论研究和实践应用中均受到广泛关注（Li等，2014；Wang等，2017a；Asi等，2020）。

未来对我们的理论进行研究，有几个值得关注的方向：

- 对于训练数据上的（后缀）平均迭代序列或最后一个迭代序列，推导适用于 M-SPP 的近最优指数超额风险界仍然是一个悬而未决的问题。
- 受近期关于带动量加速的 M-SPP 理解方面所取得进展的启发（邓和高，2021年；Chadha 等人，2022年），为光滑损失函数的理论提供动量和弱凸扩展是很有意义的。
- 最后但同样重要的是，我们预计，本文发展的理论可以扩展到采用小批量随机近端点方法的非参数学习设置。

致谢

作者衷心感谢匿名审稿人提出的建设性意见。Xiao-Tong Yuan 的工作还部分获得了中国国家重点研发计划（项目编号：2018AAA0100400）以及中国国家自然科学基金（NSFC）（项目编号：U21B2049、61876090 和 61936005）的资助。

A 第 2 节结果的证明

本节将给出第2节中主要结果的技术性证明。

A.1 定理 1 的证明

这里证明定理 1 ，为方便起见，下面重述该定理。

定理 1. 假设条件 *1* 和 *2* 成立。考虑 $\epsilon_t \equiv 0$ 以及加权平均输出 $\bar{w}_T = {}_T{}^{2}{}_{(T+1)\sum_{t=1}^{T} t w_t}$ 算法 *1*。令 $\rho \in (0, 0.5]$ ~~为任~~意标量。

(a) 假设 $n \geq \frac{64L}{\lambda\rho}$。令 $\gamma_t = \frac{\lambda\rho t}{4}$，其中 $t \geq 1$。则对于任意 $T \geq 1$,

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \leq \frac{4\rho\left[R(w_0) - R^*\right]}{T^2} + \frac{2^9 L}{\lambda\rho n T} R^*.$$

(b) 令 $\gamma_t = \frac{\lambda\rho t}{4} + \frac{16L}{n}$，其中 $t \geq 1$。则对于任意 $T \geq 1$,

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \leq \left(\frac{4\rho}{T^2} + \frac{2^8 L}{\lambda n T}\right)\left[R(w_0) - R^*\right] + \left(\frac{2^{16} L^2}{\lambda^2\rho^2 n^2 T} + \frac{2^9 L}{\lambda\rho n T}\right) R^*,$$

我们首先给出以下引理，它将在证明中用到。可以将其视为先前结果（Wang 等人，2017b，引理 1）在复合最小化情形下的直接推广。为完整起见，这里给出证明。

引理 1. 假设损失函数 ℓ 关于其第一个参数是凸的，且正则化函数 r 是凸的。那么对于任意 $w \in \mathcal{W}$，有

$$R_{S_t}(w_t) - R_{S_t}(w) \leq \frac{\gamma_t}{2}\left(\|w - w_{t-1}\|^2 - \|w - w_t\|^2 - \|w_t - w_{t-1}\|^2\right).$$

证明. 由于 ℓ 和 r 均为凸函数，R_{S_t} 在 $\mathcal{W}$上是凸的。w_t 的最优性意味着，对于任意 $w \in \mathcal{W}$ 和 $\eta \in (0, 1)$

$$\begin{aligned} &R_{S_t}(w_t) + \frac{\gamma_t}{2}\|w_t - w_{t-1}\|^2 \leq R_{S_t}((1-\eta)w_t + \eta w) + \frac{\gamma_t}{2}\|(1-\eta)w_t + \eta w - w_{t-1}\|^2 \\ \leq &(1-\eta)R_{S_t}(w_t) + \eta R_{S_t}(w) + \frac{\gamma_t}{2}\left[(1-\eta)\|w_t - w_{t-1}\|^2 + \eta\|w - w_{t-1}\|^2 - \eta(1-\eta)\|w - w_t\|^2\right], \end{aligned}$$

其中，在最后一个不等式中，我们用到了范数 $\|\cdot\|$的定义。对上述不等式两边重新整理可得

$$\eta(R_{S_t}(w_t) - R_{S_t}(w)) \leq \frac{\eta\gamma_t}{2}\left[\|w - w_{t-1}\|^2 - (1-\eta)\|w - w_t\|^2 - \|w_t - w_{t-1}\|^2\right],$$

这意味着（请记住，$\eta > 0$）

$$R_{S_t}(w_t) - R_{S_t}(w) \leq \frac{\gamma_t}{2}\left[\|w - w_{t-1}\|^2 - (1-\eta)\|w - w_t\|^2 - \|w_t - w_{t-1}\|^2\right].$$

对上述不等式取极限 $\eta \to 0^+$，即可得到所需的界。 □

以下关于光滑函数有界性的结果源自Srebro 等人（2010，引理 3.1）。

引理 2. 若 g 非负且 L-光滑，则 $\|\nabla g(w)\| \leq \sqrt{2Lg(w)}$。

令 $\{\mathcal{F}_t\}_{t\geq 1}$ 为由迭代序列生成的滤过 $\{w_t\}_{t\geq 1}$，其中 $\mathcal{F}_t = \sigma\ (w_1,\ w_2,\ \ldots\ldots,\ w_t)$。由引理1 和引理2，我们可以进一步建立以下关键引理，它在证明定理1中起着基础性作用。

引理 3。假设条件*1* 成立。令$\gamma_t \geq \frac{16L}{n}$。则有

$$\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \leq \gamma_t \left(D^2(w_{t-1}, W^*) - \mathbb{E}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right]\right) + \frac{16L}{\gamma_t n} R^*.$$

证明。考虑样本集 $S_t^{(i)}$，它与 S_t 完全相同，只是其中一个 $z_{i,\ t}$ 被另一个随机样本 $z'_{i,\ t}$替换。记

$$w_t^{(i)} = \arg\min_{w\in\mathcal{W}} \left\{F_t^{(i)}(w) := R_{S_t^{(i)}}(w) + \frac{\gamma_t}{2}\|w - w_{t-1}\|^2\right\},$$

其中 $R_{S_t^{(i)}}(w) := \frac{1}{n}\left(\sum_{j\neq i} \ell(w; z_{j,\ t}) + \ell(w; z'_{i,\ t})\right) + r(w)$。然后我们可以证明

$$\begin{aligned}
&F_t(w_t^{(i)}) - F_t(w_t)\\
=&\frac{1}{n}\sum_{j\neq i}\left(\ell(w_t^{(i)}; z_{j,t}) - \ell(w_t; z_{j,t})\right) + \frac{1}{n}\left(\ell(w_t^{(i)}; z_{i,t}) - \ell(w_t; z_{i,t})\right)\\
&+ r(w_t^{(i)}) - r(w_t) + \frac{\gamma_t}{2}\|w_t^{(i)} - w_{t-1}\|^2 - \frac{\gamma_t}{2}\|w_t - w_{t-1}\|^2\\
=&F_t^{(i)}(w_t^{(i)}) - F_t^{(i)}(w_t) + \frac{1}{n}\left(\ell(w_t^{(i)}; z_{i,t}) - \ell(w_t; z_{i,t})\right) - \frac{1}{n}\left(\ell(w_t^{(i)}; z'_{i,t}) - \ell(w_t; z'_{i,t})\right)\\
\leq&\frac{1}{n}\left|\ell(w_t^{(i)}; z_{i,t}) - \ell(w_t; z_{i,t})\right| + \frac{1}{n}\left|\ell(w_t^{(i)}; z'_{i,t}) - \ell(w_t; z'_{i,t})\right|\\
\overset{\zeta_1}{\leq}&\frac{\|\nabla\ell(w_t^{(i)}; z_{i,t})\| + \|\nabla\ell(w_t; z'_{i,t})\|}{n}\|w_t^{(i)} - w_t\|\\
\overset{\zeta_2}{\leq}&\frac{\sqrt{2L\ell(w_t^{(i)}; z_{i,t})} + \sqrt{2L\ell(w_t; z'_{i,t})}}{n}\|w_t^{(i)} - w_t\|,
\end{aligned}$$

其中“ζ_1”是由损失函数的凸性得出的，而在“ζ_2”中我们使用了引理2。引理1 中的界表明

$$F_t(w_t^{(i)}) - F_t(w_t) \geq \frac{\gamma_t}{2}\|w_t^{(i)} - w_t\|^2.$$

结合前面的两个不等式可得

$$\frac{\gamma_t}{2}\|w_t^{(i)} - w_t\| \leq \frac{\sqrt{2L\ell(w_t^{(i)}; z_{i,t})} + \sqrt{2L\ell(w_t; z'_{i,t})}}{n},$$

由此立即得到

$$\|w_t^{(i)} - w_t\| \leq \frac{2\left(\sqrt{2L\ell(w_t^{(i)}; z_{i,t})} + \sqrt{2L\ell(w_t; z'_{i,t})}\right)}{\gamma_t n}. \tag{8}$$

现在考虑关于损失函数 ℓ、定义在 S_t 上的以下总体风险和经验风险:

$$R^\ell(w) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(w; z)], \quad R^\ell_{S_t}(w) := \frac{1}{n}\sum_{i=1}^n \ell(w; z_{i,t}).$$

由于 S_t 和 $S_t^{(i)}$ 都是来自数据分布的独立同分布样本。因此可得 t

$$\begin{aligned}&\mathbb{E}_{S_t}\left[R^\ell(w_t) \mid \mathcal{F}_{t-1}\right] = \mathbb{E}_{S_t\cup\{z'_{i,t}\}}\left[\ell(w_t; z'_{i,t}) \mid \mathcal{F}_{t-1}\right] \\ =&\mathbb{E}_{S_t^{(i)}}\left[R^\ell(w_t^{(i)}) \mid \mathcal{F}_{t-1}\right] = \mathbb{E}_{S_t^{(i)}\cup\{z_{i,t}\}}\left[\ell(w_t^{(i)}; z_{i,t}) \mid \mathcal{F}_{t-1}\right]\end{aligned}$$

.

由于上述结论对所有 $i = 1, ..., n$均成立，我们可以进一步证明

$$\begin{aligned}&\mathbb{E}_{S_t}\left[R^\ell(w_t) \mid \mathcal{F}_{t-1}\right] \\ =&\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S_t^{(i)}\cup\{z_{i,t}\}}\left[\ell(w_t^{(i)}; z_{i,t}) \mid \mathcal{F}_{t-1}\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S_t\cup\{z'_{i,t}\}}\left[\ell(w_t^{(i)}; z_{i,t}) \mid \mathcal{F}_{t-1}\right] \\ =&\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S_t\cup\{z'_{i,t}\}}\left[\ell(w_t; z'_{i,t}) \mid \mathcal{F}_{t-1}\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S_t^{(i)}\cup\{z_{i,t}\}}\left[\ell(w_t; z'_{i,t}) \mid \mathcal{F}_{t-1}\right].\end{aligned} \tag{9}$$

关于经验情形，我们发现

$$\begin{aligned}&\mathbb{E}_{S_t}\left[R^\ell_{S_t}(w_t) \mid \mathcal{F}_{t-1}\right] \\ =&\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S_t}\left[\ell(w_t; z_{i,t}) \mid \mathcal{F}_{t-1}\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{S_t\cup\{z'_{i,t}\}}\left[\ell(w_t; z_{i,t}) \mid \mathcal{F}_{t-1}\right].\end{aligned}$$

结合前两个等式可得

$$
\begin{aligned}
&\left|\mathbb{E}_{S_t}\left[R(w_t)-R_{S_t}(w_t)\mid\mathcal{F}_{t-1}\right]\right| \\
=&\left|\mathbb{E}_{S_t}\left[R^{\ell}(w_t)-R^{\ell}_{S_t}(w_t)\mid\mathcal{F}_{t-1}\right]\right| \\
=&\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S_t\cup\{z'_{i,t}\}}\left[\ell(w_t^{(i)};z_{i,t})-\ell(w_t;z_{i,t})\mid\mathcal{F}_{t-1}\right]\right| \\
\leq&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S_t\cup\{z'_{i,t}\}}\left[\left|\ell(w_t^{(i)};z_{i,t})-\ell(w_t;z_{i,t})\right|\mid\mathcal{F}_{t-1}\right] \\
\leq&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S_t\cup\{z'_{i,t}\}}\left[\sqrt{2L\ell(w_t^{(i)};z_{i,t})}\|w_t^{(i)}-w_t\|\mid\mathcal{F}_{t-1}\right] \\
\overset{(8)}{\leq}&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S_t^{(i)}\cup\{z_{i,t}\}}\left[\frac{4L\ell(w_t^{(i)};z_{i,t})}{\gamma_t n}+\frac{4L\sqrt{\ell(w_t^{(i)};z_{i,t})\ell(w_t;z'_{i,t})}}{\gamma_t n}\mid\mathcal{F}_{t-1}\right] \\
\overset{\zeta_1}{\leq}&\left(\frac{L}{\gamma_t n}\right)\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{S_t\cup\{z'_{i,t}\}}\left[6\ell(w_t^{(i)};z_{i,t})+2\ell(w_t;z'_{i,t})\mid\mathcal{F}_{t-1}\right] \\
\overset{(9)}{=}&\frac{8L}{\gamma_t n}\mathbb{E}_{S_t}\left[R^{\ell}(w_t)\mid\mathcal{F}_{t-1}\right]\leq\frac{8L}{\gamma_t n}\mathbb{E}_{S_t}\left[R(w_t)\mid\mathcal{F}_{t-1}\right],
\end{aligned}
$$

其中在“ζ_1”中，我们使用了这一事实： $a^2+b^2\geq 2ab$ ，最后一个不等式则由这一事实得出： $r\geq 0$.

现在记 $w_t^*=\arg\min_{w\in W^*}\|w-w_t\|$。在给定 $\mathcal{F}_{t-1}$的条件下，对引理 1 中关于 $w=w_{t-1}^*$ 的界两边取期望，得到

$$
\begin{aligned}
&\mathbb{E}_{S_t}\left[R_{S_t}(w_t)-R^*\mid\mathcal{F}_{t-1}\right] \\
\leq&\frac{\gamma_t}{2}\mathbb{E}_{S_t}\left[\|w_{t-1}^*-w_{t-1}\|^2-\|w_{t-1}^*-w_t\|^2-\|w_t-w_{t-1}\|^2\mid\mathcal{F}_{t-1}\right] \\
\leq&\frac{\gamma_t}{2}\left(\|w_{t-1}^*-w_{t-1}\|^2-\mathbb{E}_{S_t}\left[\|w_t^*-w_t\|^2\mid\mathcal{F}_{t-1}\right]\right).
\end{aligned}
$$

结合前两个不等式可得

$$
\begin{aligned}
&\mathbb{E}_{S_t}\left[R(w_t)-R^*\mid\mathcal{F}_{t-1}\right] \\
=&\mathbb{E}_{S_t}\left[R(w_t)-R_{S_t}(w_t)+R_{S_t}(w_t)-R^*\mid\mathcal{F}_{t-1}\right] \\
\leq&\left|\mathbb{E}_{S_t}\left[R(w_t)-R_{S_t}(w_t)\mid\mathcal{F}_{t-1}\right]\right|+\mathbb{E}_{S_t}\left[R_{S_t}(w_t)-R^*\mid\mathcal{F}_{t-1}\right] \\
\leq&\frac{\gamma_t}{2}\left(\|w_{t-1}^*-w_{t-1}\|^2-\mathbb{E}_{S_t}\left[\|w_t^*-w_t\|^2\mid\mathcal{F}_{t-1}\right]\right)+\frac{8L}{\gamma_t n}\mathbb{E}_{S_t}\left[R(w_t)\mid\mathcal{F}_{t-1}\right] \\
=&\frac{\gamma_t}{2}\left(\|w_{t-1}^*-w_{t-1}\|^2-\mathbb{E}_{S_t}\left[\|w_t^*-w_t\|^2\mid\mathcal{F}_{t-1}\right]\right)+\frac{8L}{\gamma_t n}\mathbb{E}_{S_t}\left[R(w_t)-R^*\mid\mathcal{F}_{t-1}\right]+\frac{8L}{\gamma_t n}\mathbb{E}_{S_t}\left[R^*\right] \\
\leq&\frac{\gamma_t}{2}\left(\|w_{t-1}^*-w_{t-1}\|^2-\mathbb{E}_{S_t}\left[\|w_t^*-w_t\|^2\mid\mathcal{F}_{t-1}\right]\right)+\frac{1}{2}\mathbb{E}_{S_t}\left[R(w_t)-R^*\mid\mathcal{F}_{t-1}\right]+\frac{8L}{\gamma_t n}R^*,
\end{aligned}
$$

其中，在最后一个不等式中，我们使用了条件 $\gamma_t\geq\frac{52L}{n}$ 。重新整理

上述不等式中的各项后，我们得到

$$\begin{aligned}\mathbb{E}_{S_t}\left[R(w_t)-R^* \mid \mathcal{F}_{t-1}\right] \leq & \gamma_t\left(\|w_{t-1}^*-w_{t-1}\|^2-\mathbb{E}_{S_t}\left[\|w_t^*-w_t\|^2 \mid \mathcal{F}_{t-1}\right]\right)+\frac{16L}{\gamma_t n}R^* \\ = & \gamma_t\left(D^2(w_{t-1},W^*)-\mathbb{E}_{S_t}\left[D^2(w_t,W^*) \mid \mathcal{F}_{t-1}\right]\right)+\frac{16L}{\gamma_t n}R^*.\end{aligned}$$

这就得到了所需的界。 □

下面的引理是引理3的直接推论。

引理4。 假设*1* 成立。令$\gamma_t \geq \frac{16L}{n}$。则对所有$t \geq 1$，以下结论成立:

$$\mathbb{E}\left[D^2(w_t,W^*)\right] \leq D^2(w_0,W^*)+\sum_{\tau=1}^{t}\frac{16L}{\gamma_\tau^2 n}R^*.$$

证明。 由于 $R(w_t) \geq R^*$ 且$\gamma_t \geq \frac{52L}{n}$，引理3 中的界立即推出

$$\mathbb{E}_{S_t}\left[D^2(w_t,W^*) \mid \mathcal{F}_{t-1}\right] \leq D^2(w_{t-1},W^*)+\frac{16L}{\gamma_t^2 n}R^*. \tag{10}$$

将上述递推关系从时间点 t 展开至零，可得对所有 $t \geq 1$,

$$\mathbb{E}\left[D^2(w_t,W^*)\right] \leq D^2(w_0,W^*)+\sum_{\tau=1}^{t}\frac{16L}{\gamma_\tau^2 n}R^*.$$

这就证明了所需的界。 □

有了这些引理，我们现在可以证明定理 1中的主要结果。

定理 *1* 的证明。（a）部分：注意，关于 n 的条件 $\gamma_t = \frac{\lambda\rho t}{4} \geq \frac{\lambda\rho}{4} \geq \frac{16L}{n}$。结合条件 $R(w_t) - R^* \geq \frac{\lambda}{2}D^2(w_t,W^*$，应用引理 3 可得

$$\begin{aligned}&(1-\rho)\mathbb{E}\left[R(w_t)-R^* \mid \mathcal{F}_{t-1}\right] \\ \leq & \gamma_t D^2(w_{t-1},W^*)-\left(\gamma_t+\frac{\lambda\rho}{2}\right)\mathbb{E}\left[D^2(w_t,W^*) \mid \mathcal{F}_{t-1}\right]+\frac{2^4 L}{\gamma_t n}R^* \\ \leq & \frac{\lambda\rho t}{4}D^2(w_{t-1},W^*)-\frac{\lambda\rho(t+2)}{4}\mathbb{E}\left[D^2(w_t,W^*) \mid \mathcal{F}_{t-1}\right]+\frac{2^6 L}{\lambda\rho t n}R^* \\ \leq & \frac{\lambda\rho t}{4}D^2(w_{t-1},W^*)-\frac{\lambda\rho(t+2)}{4}\mathbb{E}\left[D^2(w_t,W^*) \mid \mathcal{F}_{t-1}\right]+\frac{2^7 L}{\lambda\rho(t+1) n}R^*,\end{aligned}$$

其中在 t 最后一个不等式中，我们已对 $t \geq 1$ 使用了$\frac{1}{t} \leq \frac{2}{t+1}$。上述不等式意味着

$$\begin{aligned}&t\mathbb{E}\left[R(w_t)-R^* \mid \mathcal{F}_{t-1}\right] \\ \leq & (t+1)\mathbb{E}\left[R(w_t)-R^* \mid \mathcal{F}_{t-1}\right] \\ \leq & \frac{\lambda\rho t(t+1)}{4(1-\rho)}D^2(w_{t-1},W^*)-\frac{\lambda\rho(t+1)(t+2)}{4(1-\rho)}\mathbb{E}\left[D^2(w_t,W^*) \mid \mathcal{F}_{t-1}\right]+\frac{2^7 L}{\lambda n\rho(1-\rho)}R^*.\end{aligned}$$

然后，根据全期望定律并经过适当整理，我们得到

$$\begin{aligned}&t\mathbb{E}\left[R(w_t)-R^*\right]\\ \leq&\frac{\lambda\rho t(t+1)}{4(1-\rho)}\mathbb{E}\left[D^2(w_{t-1},W^*)\right]-\frac{\lambda\rho(t+1)(t+2)}{4(1-\rho)}\mathbb{E}\left[D^2(w_t,W^*)\right]+\frac{2^7L}{\lambda n\rho(1-\rho)}R^*.\end{aligned}\tag{11}$$

将上述不等式从 $t=1$ 到 T 求和并归一化后，我们得到

$$\begin{aligned}\frac{2}{T(T+1)}\sum_{t=1}^{T}t\mathbb{E}\left[R(w_t)-R^*\right]\leq&\frac{\lambda\rho}{T(T+1)(1-\rho)}D^2(w_0,W^*)+\frac{2^8L}{\lambda\rho(1-\rho)(T+1)n}R^*\\ \leq&\frac{2\lambda\rho}{T(T+1)}D^2(w_0,W^*)+\frac{2^9L}{\lambda\rho(T+1)n}R^*,\end{aligned}$$

其中，在最后一个不等式中我们使用了 $\rho\leq 0.5$。考虑加权输出 $\bar{w}_T=\frac{2}{T(T+1)}\sum_{t=1}^{T}tw_t$，根据上述不等式，以及风险函数$R$ 的凸性和二次增长性质，我们有

$$\mathbb{E}\left[R(\bar{w}_T)-R^*\right]\leq\frac{4\rho\left[R(w_0)-R^*\right]}{T(T+1)}+\frac{2^9L}{\lambda\rho n(T+1)}R^*,$$

由此可得部分（a）中所需的界。

部分（b）：注意，$\gamma_t=\frac{\lambda\rho t}{4}+\frac{16L}{n}\geq\frac{16L}{n}$ 对所有 $t\geq 1$成立。根据引理 4 ，对所有 $t\geq 1$, 有:

$$\begin{aligned}&\mathbb{E}\left[D^2(w_t,W^*)\right]\\ \leq&D^2(w_0,W^*)+\sum_{\tau=1}^{t}\frac{16L}{\gamma_\tau^2n}R^*\leq D^2(w_0,W^*)+\frac{2^8L}{\lambda^2\rho^2n}R^*\sum_{\tau=1}^{t}\frac{1}{\tau^2}\leq D^2(w_0,W^*)+\frac{2^9L}{\lambda^2\rho^2n}R^*.\end{aligned}\tag{12}$$

与部分（a）的论证类似，应用引理 3 以及二次增长条件$R(w_t)-R^*\geq\frac{\lambda}{2}D^2(w_t,W^*)$ 和 $\rho\leq 0.5$，得到

$$\begin{aligned}&\frac{1}{2}\mathbb{E}\left[R(w_t)-R^*\mid\mathcal{F}_{t-1}\right]\\ \leq&(1-\rho)\mathbb{E}\left[R(w_t)-R^*\mid\mathcal{F}_{t-1}\right]\\ \leq&\gamma_tD^2(w_{t-1},W^*)-\left(\gamma_t+\frac{\lambda\rho}{2}\right)\mathbb{E}\left[D^2(w_t,W^*)\mid\mathcal{F}_{t-1}\right]+\frac{2^4L}{\gamma_tn}R^*\\ \leq&\frac{\lambda\rho t}{4}D^2(w_{t-1},W^*)-\frac{\lambda\rho(t+2)}{4}\mathbb{E}\left[D^2(w_t,W^*)\mid\mathcal{F}_{t-1}\right]\\ &+\frac{16L}{n}\left(D^2(w_{t-1},W^*)-\mathbb{E}\left[D^2(w_t,W^*)\mid\mathcal{F}_{t-1}\right]\right)+\frac{2^6L}{\lambda\rho tn}R^*,\end{aligned}$$

其中，在第二个不等式中我们使用了 $\gamma_t\geq\frac{52L}{n}$ ，在最后一个不等式中我们使用了

$\gamma_t \geq \frac{\lambda\rho t}{4}$。然后，根据全期望定律并经过适当整理，我们得到

$$\begin{aligned}
&\mathbb{E}\left[R(w_t) - R^*\right] \\
\leq&\frac{\lambda\rho t}{2}\mathbb{E}\left[D^2(w_{t-1}, W^*)\right] - \frac{\lambda\rho(t+2)}{2}\mathbb{E}\left[D^2(w_t, W^*)\right] \\
&+ \frac{2^5 L}{n}\left(\mathbb{E}\left[D^2(w_{t-1}, W^*)\right] - \mathbb{E}\left[D^2(w_t, W^*)\right]\right) + \frac{2^7 L}{\lambda t n\rho}R^*,
\end{aligned}$$

这意味着

$$\begin{aligned}
&t\mathbb{E}\left[R(w_t) - R^*\right] \\
\leq&(t+1)\mathbb{E}\left[R(w_t) - R^*\right] \\
\leq&\frac{\lambda\rho t(t+1)}{2}\mathbb{E}\left[D^2(w_{t-1}, W^*)\right] - \frac{\lambda\rho(t+1)(t+2)}{2}\mathbb{E}\left[D^2(w_t, W^*)\right] \\
&+ \frac{2^5 L(t+1)}{n}\left(\mathbb{E}\left[D^2(w_{t-1}, W^*)\right] - \mathbb{E}\left[D^2(w_t, W^*)\right]\right) + \frac{2^7 L(t+1)}{\lambda t n\rho}R^* \\
\leq&\frac{\lambda\rho t(t+1)}{2}\mathbb{E}\left[D^2(w_{t-1}, W^*)\right] - \frac{\lambda\rho(t+1)(t+2)}{2}\mathbb{E}\left[D^2(w_t, W^*)\right] \\
&+ \frac{2^6 L t}{n}\left(\mathbb{E}\left[D^2(w_{t-1}, W^*)\right] - \mathbb{E}\left[D^2(w_t, W^*)\right]\right) + \frac{2^8 L}{\lambda n\rho}R^*,
\end{aligned}$$

其中，在最后一个不等式中，我们用到了 $t+1 \leq 2t$ as $t \geq 1$ 的事实。将上述不等式从 $t=1$ 到 T 求和并归一化后，我们得到

$$\begin{aligned}
&\frac{2}{T(T+1)}\sum_{t=1}^{T} t\mathbb{E}\left[R(w_t) - R^*\right] \\
\leq&\frac{2\lambda\rho}{T(T+1)}D^2(w_0, W^*) + \frac{2^7 L}{nT(T+1)}\sum_{t=1}^{T} D^2(w_{t-1}, W^*) + \frac{2^9 L}{\lambda\rho(T+1)n}R^* \\
\leq&\frac{2\lambda\rho}{T(T+1)}D^2(w_0, W^*) + \frac{2^7 L}{nT(T+1)}\sum_{t=1}^{T}\left(D^2(w_0, W^*) + \frac{2^9 L}{\lambda^2\rho^2 n}R^*\right) + \frac{2^9 L}{\lambda\rho(T+1)n}R^* \\
=&\left(\frac{2\lambda\rho}{T(T+1)} + \frac{2^7 L}{n(T+1)}\right)D^2(w_0, W^*) + \left(\frac{2^{16}L^2}{\lambda^2\rho^2 n^2(T+1)} + \frac{2^9 L}{\lambda\rho n(T+1)}\right)R^*,
\end{aligned}$$

其中，在最后一个不等式中，我们用到了（12）。利用上述不等式中的凸性和二次增长性质可得

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \leq \left(\frac{4\rho}{T(T+1)} + \frac{2^8 L}{\lambda n(T+1)}\right)\left[R(w_0) - R^*\right] + \left(\frac{2^{16}L^2}{\lambda^2\rho^2 n^2(T+1)} + \frac{2^9 L}{\lambda\rho n(T+1)}\right)R^*,$$

由此可得部分（b）所需的界。证明至此结束。 □

A.2 定理 2 的证明

在本小节中，我们证明定理2，其内容重述如下。

定理 2. 假设条件 *1* 和*2* 成立。考虑 $\epsilon_t \equiv 0$，用于在算法 *2* 的第一阶段和第二阶段中实现 M-SPP。考虑第二阶段中的加权平均输出 $\bar{w}_T = \frac{2}{(T-1)(T+2)} \sum_{t=2}^{T} t w_t$。

(a) 假设 $n \geq \frac{128L}{\lambda}$。在第一阶段设定 $m = \frac{128L}{\lambda}$，并在第一阶段和第二阶段中设定 $\gamma_t = \frac{\lambda t}{8}$ 以实现 M-SPP。则对任意 $T \geq 2$，$\bar{w}_T$ 满足

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \frac{L^2\left[R(w_0) - R^*\right]}{\lambda^2 n^2 T^2} + \frac{L}{\lambda n T} R^*.$$

(b) 在第一阶段设定 $m = \mathcal{O}(1)$，并在第一阶段和第二阶段中设定 $\gamma_t = \frac{\lambda t}{8} + \frac{16L}{n}$ 以实现 M-SPP。则对任意 $T \geq 2$，$\bar{w}_T$ 满足

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \frac{L^2\left[R(w_0) - R^*\right]}{\lambda^2 n T} + \frac{L^3}{\lambda^3 n T} R^*.$$

证明。 部分 (a)： 在第一阶段，通过应用定理 1 的第一部分 $\rho = 1/2$ 和 $T = n/m \geq 1$（稍微滥用一下符号说明）我们立即得到

$$\mathbb{E}_{S_1}\left[R(w_1) - R^*\right] \leq \frac{2m^2\left[R(w_0) - R^*\right]}{n^2} + \frac{2^{10} L}{\lambda n} R^*. \tag{13}$$

在第二阶段，给定 $\mathcal{F}_1$，对递推形式（11）从 $t = 2, \ldots, T$ 求和，并结合 $\rho = 1/2$进行适当归一化，可得

$$\begin{aligned}
&\frac{2}{(T-1)(T+2)} \sum_{t=2}^{T} t \mathbb{E}_{S_{2:t}}\left[R(w_t) - R^* \mid \mathcal{F}_1\right] \\
\leq &\frac{6\lambda D^2(w_1, W^*)}{(T-1)(T+2)} + \frac{2^{10} L}{\lambda n (T+2)} R^* \leq \frac{3\left(R(w_1) - R^*\right)}{(T-1)(T+2)} + \frac{2^{10} L}{\lambda n (T+2)} R^*,
\end{aligned}$$

在最后一个不等式中，我们用到了二次增长性质。考虑加权平均输出 $\bar{w}_T = \frac{2}{(T-1)(T+2)} \sum_{t=2}^{T} t w_t$。根据上述~~不等式和~~全期望定律，我们必有

$$\begin{aligned}
\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \leq &\frac{6\mathbb{E}_{S_1}\left[R(w_1) - R^*\right]}{(T-1)(T+2)} + \frac{2^{10} L}{\lambda n (T+2)} R^* \\
\leq &\frac{6\mathbb{E}_{S_1}\left[R(w_1) - R^*\right]}{T^2} + \frac{2^{102} L}{\lambda n T} R^* \\
\leq &\frac{12 m^2\left[R(w_0) - R^*\right]}{n^2 T^2} + \frac{2^{13} L}{\lambda n T} R^* \\
\leq &\frac{2^{22} L^2\left[R(w_0) - R^*\right]}{\lambda^2 n^2 T^2} + \frac{2^{13} L}{\lambda n T} R^*,
\end{aligned}$$

其中我们在多处用到了这一事实 $T \geq 2$，并且在倒数第二步中用到了（13）。这立即推出了部分（a）中所需的界。

部分（b）：在第一阶段，通过应用定理 1 的第二部分（取 $\rho = 1/2$ 和 $T = n/m \geq 1$），并保留主导项，我们得到

$$\begin{aligned}\mathbb{E}_{S_1}\left[R(w_1) - R^*\right] \lesssim & \left(\frac{m^2}{n^2} + \frac{L}{\lambda n}\right)\left[R(w_0) - R^*\right] + \left(\frac{L^2}{\lambda^2 mn} + \frac{L}{\lambda n}\right)R^* \\ \lesssim & \frac{L}{\lambda n}[R(w_0) - R^*] + \frac{L^2}{\lambda^2 n}R^*.\end{aligned} \tag{14}$$

在第二阶段，根据定理 1 部分（b）的证明思路，我们可以证明加权平均输出 $\bar{w}_T = {}^{2\,(T-1)(}_{T+2)} \sum_{t=2}^{T} t w_t$ 满足

$$\begin{aligned}\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim & \left(\frac{1}{T^2} + \frac{L}{\lambda n T}\right)\mathbb{E}_{S_1}\left[R(w_1) - R^*\right] + \left(\frac{L^2}{\lambda^2 n^2 T} + \frac{L}{\lambda n T}\right)R^* \\ \lesssim & \left(\frac{L}{\lambda n T^2} + \frac{L^2}{\lambda^2 n^2 T}\right)\left[R(w_0) - R^*\right] + \left(\frac{L^3}{\lambda^3 n^2 T} + \frac{L^2}{\lambda^2 n T}\right)R^* \\ \lesssim & \frac{L^2}{\lambda^2 n T}[R(w_0) - R^*] + \frac{L^3}{\lambda^3 n T}R^*,\end{aligned}$$

其中在 第二步中，我们使用了（14）。这证明了部分（b）中所需的界 . □

A.3 定理 3 的证明

本小节证明如下重述的定理 3。

定理 3. 假设假设 *1* 成立。设 $\gamma_t \equiv \gamma \geq \frac{16L}{n}$。令 $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$ 为算法*1*的平均输出。则

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \frac{\gamma}{T}D^2(w_0, W^*) + \frac{L}{\gamma n}R^*.$$

特别对于 $\gamma = \sqrt{\frac{T}{n}} + \frac{16L}{n}$，有

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \left(\frac{1}{\sqrt{nT}} + \frac{L}{nT}\right)D^2(w_0, W^*) + \frac{L}{\sqrt{nT}}R^*.$$

证明。 由于 $\gamma_t \equiv \gamma \geq \frac{16L}{n}$，引理 3 中的界成立。根据全期望定律，并将该不等式从 $t = 1$，....., T 求和并进行适当归一化，我们得到

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[R(w_t) - R^*\right] \leq \frac{\gamma}{T}D^2(w_0, W^*) + \frac{16L}{\gamma n}R^*.$$

考虑 $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$。根据上述不等式以及 R 的凸性，我们有

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \leq \frac{\gamma}{T}D^2(w_0, W^*) + \frac{16L}{\gamma n}R^*.$$

这就证明了第一个所需的界。将 $\gamma = \sqrt{\frac{T}{n}} + \frac{16L}{n} > \frac{16L}{n}$ 代入上述界即可立即得到第二个界。证明至此完成。 □

A.4 关于 M-SPP 的（迭代）稳定性

在本附录小节中，我们进一步分析 M-SPP 对正则化模量选择的敏感性 $\{\gamma_t\}_{t\geq 1}$，依据以下迭代稳定性概念，该概念本质上由Asi 和 Duchi (2019a,b)。

定义2. 对于生成迭代序列 $\{w_t\}_{t\geq 1}$ 以最小化总体风险 $R(w)$ 的随机优化算法，如果满足以下条件，则称其是稳定的：

$$\sup_{t\geq 1} D(w_t, W^*) < \infty, \quad \textit{with probability 1.}$$

在介绍关于 M-SPP 迭代稳定性的主要结果之前，我们首先回顾 Robbins-Siegmund 非负几乎超鞅收敛引理，该引理通常用于证明包括 SPP 在内的随机优化方法的稳定性和收敛性（Asi 和 Duchi，2019年b）。

引理5 （Robbins 和 Siegmund （1971年））。 考虑四个非负随机变量序列 $\{U_t\}$、$\{V_t\}$、$\{\alpha_t\}$、$\{\beta_t\}$，它们关于滤过 $\{\mathcal{F}_t\}_{t\geq 0}$均可测。假设 $\sum_t \alpha_t < \infty$、$\sum_t \beta_t < \infty$，并且

$$\mathbb{E}[U_{t+1} \mid \mathcal{F}_t] \leq (1+\alpha_t)U_t + \beta_t - V_t.$$

则存在 U_∞，使得 $U_t \xrightarrow{a.s.} U_\infty$ 和 $\sum_t V_t < \infty$ 以概率1成立

.

下面的命题表明，估计误差序列 $\{\|w_t - w^*\|\}$ 在期望意义下不会发散，而是收敛到某个有限值，并且以概率 1 有界。

命题 2。 假设条件 *1* 成立。设 $\gamma_t \geq \frac{16L}{n}$ 且 $\sum_{t\geq 1} L\gamma_t^{-2} < \infty$。则有以下结论：

(a) $\mathbb{E}\left[D(w_t, W^*)\right] < \infty$;

部分（b） $D(w_t, W^*)$收敛到某个有限值，且 $\sup_{t\geq 1} D(w_t, W^*) < \infty$ 以概率 1 成立。

证明。 应用引理 4 可得，对所有 $t \geq 1$

$$\mathbb{E}\left[D^2(w_t, W^*)\right] \lesssim D^2(w_0, W^*) + \sum_{\tau=1}^{t} \frac{L}{\gamma_\tau^2 n} R^* < \infty,$$

其中用到了给定的关于 γ_t的条件。这就证明了部分（a）。为证明部分（b），将引理 5 与 $\alpha_t = V_t \equiv 0$和 $\beta_t = \frac{16L}{\gamma_t^2 n} R^*$ 应用于（10）可得 D（w_t， W^*）收敛到某个有限值，因此 $\sup_{t\geq 1} D$（w_t， W^*）$< \infty$ 几乎必然成立。 □

注释 12. 命题 *2* 表明，与小批量 SGD 不同，为了生成估计误差不发散的序列，M-SPP 中 γ_t 的选择不受损失函数梯度尺度的影响。

第 B 节 第3节结果的证明

本节给出第 3 节所述主要结果的技术证明。

B.1 定理 4的证明

在本小节中，我们证明定理 4，其重述如下。

定理 4。 假设 1、2 和 3 均成立。令 $\rho \in (0,\ 1/4]$ 为任意标量，并令 $\gamma_t = \frac{\lambda\rho t}{4}$ 。假设 $n \geq \frac{76L}{\lambda\rho}$ 。设 $\epsilon_t \leq \frac{\epsilon}{nt^4}$ 对于某个 $\epsilon \in [0,1]$成立。则对于任意$T \geq 1$，算法 1 的加权平均输出 $\bar{w}_t = \sum_{t=1}^{T} tw_t / \frac{T}{2}_{(T+1)}$ 满足

$$\mathbb{E}\left[R(\bar{w}_t) - R^*\right] \lesssim \frac{\rho}{T^2}(R(w_0) - R^*) + \frac{L}{\lambda\rho nT}R^* + \frac{\sqrt{\epsilon}}{T^2}\left(\frac{L}{\lambda\rho} + G\sqrt{\frac{1}{\lambda\rho}}\right).$$

预备知识。 下文中，我们用 $\tilde{w}_t := \operatorname{argmin}_{w\in\mathcal{W}} F_t(w$表示内循环小批量经验风险最小化优化的精确解，其作用与第 w_t 2节中的 相同。我们首先给出以下引理，用于界定非精确最小化器 w_t 与精确最小化器 $\tilde{w}_t$之间的差异。

引理6。假设损失函数 ℓ 关于其第一个参数是凸的，且 r 是凸的。那么对于任意 $w \in \mathcal{W}$，我们有

$$\|w_t - \tilde{w}_t\| \leq \sqrt{\frac{2\epsilon_t}{\gamma_t}}.$$

证明。使用与引理相同的论证1 可得，对于所有 $w \in \mathcal{W}$，

$$R_{S_t}(\tilde{w}_t) - R_{S_t}(w) \leq \frac{\gamma_t}{2}\left(\|w - w_{t-1}\|^2 - \|w - \tilde{w}_t\|^2 - \|\tilde{w}_t - w_{t-1}\|^2\right). \tag{15}$$

将 $w = w_t$ 代入上式可得

$$\frac{\gamma_t}{2}\|w_t - \tilde{w}_t\|^2 \leq F_t(w_t) - F_t(\tilde{w}_t) \leq \epsilon_t,$$

wh这直接推出 $\|w_t - \tilde{w}_t\| \leq \sqrt{2\epsilon_t/\gamma_t}$。 这就证明了所需的第二个界 d. □

下面的引理将引理3推广到了非精确情形。

引理 7。 假设条件 1、2 和 3 均成立。设 $\gamma_t \geq \frac{19L}{n}$ 。那么，对于任意 $\rho \in (0,\ 1)$，以下界成立：

$$\begin{aligned}\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \leq & \gamma_t\left(D^2(w_{t-1}, W^*) - \mathbb{E}\left[\left(1 - \frac{\rho\lambda}{2\gamma_t}\right)D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right]\right) \\ & + \frac{19L}{\gamma_t n}R^* + \left(3n + \frac{4\gamma_t}{\rho\lambda}\right)\epsilon_t + 3G\sqrt{\frac{2\epsilon_t}{\gamma_t}}.\end{aligned}$$

或者，对于任意 $w^* \in W^*$，在条件 *1* 和 *3* 成立时，有

$$\begin{aligned}\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \leq & \gamma_t \left(\|w_{t-1} - w^*\|^2 - \mathbb{E}\left[\|w_t - w^*\|^2 \mid \mathcal{F}_{t-1}\right]\right) + \frac{19L}{\gamma_t n} R^* \\ & + 3n\epsilon_t + \left(2\sqrt{2\gamma_t}\mathbb{E}\left[\|w_t - w^*\| \mid \mathcal{F}_{t-1}\right] + 3G\sqrt{\frac{2}{\gamma_t}}\right)\sqrt{\epsilon_t}.\end{aligned}$$

证明。我们将 $\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right]$ 分解为以下三项 s:

$$\begin{aligned}&\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \\ =&\underbrace{\mathbb{E}\left[R(w_t) - R(\tilde{w}_t) \mid \mathcal{F}_{t-1}\right]}_{A} + \underbrace{\mathbb{E}\left[R(\tilde{w}_t) - R_{S_t}(\tilde{w}_t) \mid \mathcal{F}_{t-1}\right]}_{B} + \underbrace{\mathbb{E}\left[R_{S_t}(\tilde{w}_t) - R^* \mid \mathcal{F}_{t-1}\right]}_{C}.\end{aligned}$$

接下来，我们分别界定这三项。为界定项 A，我们可以证明

$$\begin{aligned}&|A| := \left|\mathbb{E}\left[R(w_t) - R(\tilde{w}_t) \mid \mathcal{F}_{t-1}\right]\right| \\ =&\left|\mathbb{E}\left[R^\ell(w_t) - R^\ell(\tilde{w}_t) \mid \mathcal{F}_{t-1}\right] + \mathbb{E}\left[r(w_t) - r(\tilde{w}_t)\right] \mid \mathcal{F}_{t-1}\right| \\ \leq&\mathbb{E}\left[\mathbb{E}_z|\ell(w_t; z) - \ell(\tilde{w}_t; z)| \mid \mathcal{F}_{t-1}\right] + \mathbb{E}\left[|r(w_t) - r(\tilde{w}_t)| \mid \mathcal{F}_{t-1}\right] \\ \overset{\zeta_1}{\leq}&\mathbb{E}\left[\mathbb{E}_z\left[\sqrt{2L\ell(w_t; z)}\|w_t - \tilde{w}_t\|\right] \mid \mathcal{F}_{t-1}\right] + \mathbb{E}\left[G\|w_t - \tilde{w}_t\| \mid \mathcal{F}_{t-1}\right] \\ \leq&\mathbb{E}\left[\mathbb{E}_z\left[\frac{L}{\gamma_t n}\ell(w_t; z) + \frac{\gamma_t n}{2}\|w_t - \tilde{w}_t\|^2\right] \mid \mathcal{F}_{t-1}\right] + \mathbb{E}\left[G\|w_t - \tilde{w}_t\| \mid \mathcal{F}_{t-1}\right] \\ =&\mathbb{E}\left[\frac{L}{\gamma_t n}R^\ell(w_t) \mid \mathcal{F}_{t-1}\right] + \mathbb{E}_{S_t}\left[\frac{\gamma_t n}{2}\|w_t - \tilde{w}_t\|^2 + G\|w_t - \tilde{w}_t\| \mid \mathcal{F}_{t-1}\right] \\ \leq&\mathbb{E}\left[\frac{L}{\gamma_t n}R(w_t) \mid \mathcal{F}_{t-1}\right] + n\epsilon_t + G\sqrt{\frac{2\epsilon_t}{\gamma_t}},\end{aligned}$$

其中在“ζ_1”中，我们使用了损失函数的凸性、引理 2 以及假设 3 ，并且在最后一个不等式中使用了 $r > 0$ 以及引理 6的扰动界。

为界定第 项 B，采用与引理 3 大致相同的证明论证，我们可以证明

$$\begin{aligned}B :=&\mathbb{E}\left[R(\tilde{w}_t) - R_{S_t}(\tilde{w}_t) \mid \mathcal{F}_{t-1}\right] \\ \leq&\frac{8L}{\gamma_t n}\mathbb{E}\left[R(\tilde{w}_t) \mid \mathcal{F}_{t-1}\right] \\ =&\frac{8L}{\gamma_t n}\mathbb{E}\left[R(\tilde{w}_t) - R(w_t)\right] + \frac{8L}{\gamma_t n}\mathbb{E}\left[R(w_t) \mid \mathcal{F}_{t-1}\right] \\ \leq&\frac{1}{2}|A| + \frac{8L}{\gamma_t n}\mathbb{E}\left[R(w_t) \mid \mathcal{F}_{t-1}\right],\end{aligned}$$

其中我们使用了关于小批量大小 γ_t 的条件。

为了界定该项 C，根据 $\tilde{w}_t$ 的定义，并应用引理 1 以及 $w = w_{t-1}^*$

我们可以验证

$$
\begin{aligned}
C :=& \mathbb{E}\left[R_{S_t}(\tilde{w}_t) - R^* \mid \mathcal{F}_{t-1}\right] \\
\leq& \frac{\gamma_t}{2}\mathbb{E}\left[\|w_{t-1}^* - w_{t-1}\|^2 - \|w_{t-1}^* - \tilde{w}_t\|^2 - \|\tilde{w}_t - w_{t-1}\|^2 \mid \mathcal{F}_{t-1}\right] \\
\leq& \frac{\gamma_t}{2}\mathbb{E}\left[\|w_{t-1}^* - w_{t-1}\|^2 - \|w_{t-1}^* - \tilde{w}_t\|^2 \mid \mathcal{F}_{t-1}\right] \\
=& \frac{\gamma_t}{2}\left(\|w_{t-1}^* - w_{t-1}\|^2 - \mathbb{E}\left[\|w_{t-1}^* - w_t + w_t - \tilde{w}_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) \\
=& \frac{\gamma_t}{2}\left(\|w_{t-1}^* - w_{t-1}\|^2 - \mathbb{E}\left[\|w_{t-1}^* - w_t\|^2 + 2\langle w_{t-1}^* - w_t, w_t - \tilde{w}_t\rangle + \|w_t - \tilde{w}_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) \\
\leq& \frac{\gamma_t}{2}\left(\|w_{t-1}^* - w_{t-1}\|^2 - \mathbb{E}\left[\left(1 - \frac{\rho\lambda}{2\gamma_t}\right)\|w_{t-1}^* - w_t\|^2 - \frac{2\gamma_t}{\rho\lambda}\|w_t - \tilde{w}_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) \\
\leq& \frac{\gamma_t}{2}\left(\|w_{t-1}^* - w_{t-1}\|^2 - \mathbb{E}\left[\left(1 - \frac{\rho\lambda}{2\gamma_t}\right)\|w_t^* - w_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) + \frac{2\gamma_t\epsilon_t}{\rho\lambda} \\
=& \frac{\gamma_t}{2}\left(D^2(w_{t-1}, W^*) - \mathbb{E}\left[\left(1 - \frac{\rho\lambda}{2\gamma_t}\right)D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right]\right) + \frac{2\gamma_t\epsilon_t}{\rho\lambda}.
\end{aligned}
$$

结合上述三个界可得

$$
\begin{aligned}
&\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] = A + B + C \\
\leq& \frac{3}{2}|A| + \frac{8L}{\gamma_t n}\mathbb{E}\left[R(w_t) \mid \mathcal{F}_{t-1}\right] \\
&+ \frac{\gamma_t}{2}\left(D^2(w_{t-1}, W^*) - \mathbb{E}\left[\left(1 - \frac{\rho\lambda}{2\gamma_t}\right)D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right]\right) + \frac{2\gamma_t\epsilon_t}{\rho\lambda} \\
\leq& \mathbb{E}\left[\frac{3L}{2\gamma_t n}R(w_t) \mid \mathcal{F}_{t-1}\right] + \frac{3n}{2}\epsilon_t + \frac{3G}{2}\sqrt{\frac{2\epsilon_t}{\gamma_t}} + \frac{8L}{\gamma_t n}\mathbb{E}\left[R(w_t) \mid \mathcal{F}_{t-1}\right] \\
&+ \frac{\gamma_t}{2}\left(D^2(w_{t-1}, W^*) - \mathbb{E}\left[\left(1 - \frac{\rho\lambda}{2\gamma_t}\right)D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right]\right) + \frac{2\gamma_t\epsilon_t}{\rho\lambda} \\
\leq& \frac{\gamma_t}{2}\left(D^2(w_{t-1}, W^*) - \mathbb{E}\left[\left(1 - \frac{\rho\lambda}{2\gamma_t}\right)D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right]\right) + \frac{9.5L}{\gamma_t n}\mathbb{E}\left[R(w_t) \mid \mathcal{F}_{t-1}\right] \\
&+ \left(\frac{3n}{2} + \frac{2\gamma_t}{\rho\lambda}\right)\epsilon_t + \frac{3G}{2}\sqrt{\frac{2\epsilon_t}{\gamma_t}} \\
=& \frac{\gamma_t}{2}\left(D^2(w_{t-1}, W^*) - \mathbb{E}\left[\left(1 - \frac{\rho\lambda}{2\gamma_t}\right)D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right]\right) + \frac{9.5L}{\gamma_t n}\mathbb{E}\left[R^*\right] + \frac{9.5L}{\gamma_t n}\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \\
&+ \left(\frac{3n}{2} + \frac{2\gamma_t}{\rho\lambda}\right)\epsilon_t + \frac{3G}{2}\sqrt{\frac{2\epsilon_t}{\gamma_t}} \\
\leq& \frac{\gamma_t}{2}\left(D^2(w_{t-1}, W^*) - \mathbb{E}\left[\left(1 - \frac{\rho\lambda}{2\gamma_t}\right)D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right]\right) + \frac{9.5L}{\gamma_t n}\mathbb{E}\left[R^*\right] + \frac{1}{2}\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \\
&+ \left(\frac{3n}{2} + \frac{2\gamma_t}{\rho\lambda}\right)\epsilon_t + \frac{3G}{2}\sqrt{\frac{2\epsilon_t}{\gamma_t}},
\end{aligned}
$$

其中，在最后一个不等式中，我们使用了条件 $\gamma_t \geq \frac{19L}{n}$。重新整理上述不等式中的各项后，我们得到所需的第一个界。

为推导第二个界，对于任意固定的 $w^* \in W^*$，我们注意到，项 C 也可以按如下方式进行界定：

$$\begin{aligned}
C \leq & \frac{\gamma_t}{2}\left(\|w^* - w_{t-1}\|^2 - \mathbb{E}\left[\|w^* - w_t\|^2 + 2\langle w^* - w_t, w_t - \tilde{w}_t\rangle + \|w_t - \tilde{w}_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) \\
\leq & \frac{\gamma_t}{2}\left(\|w^* - w_{t-1}\|^2 - \mathbb{E}\left[\|w^* - w_t\|^2 - 2\|w_t - w^*\|\|w_t - \tilde{w}_t\| \mid \mathcal{F}_{t-1}\right]\right) \\
\leq & \frac{\gamma_t}{2}\left(\|w^* - w_{t-1}\|^2 - \mathbb{E}\left[\|w^* - w_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) + \sqrt{2\gamma_t\epsilon_t}\mathbb{E}\left[\|w^* - w_t\| \mid \mathcal{F}_{t-1}\right].
\end{aligned}$$

与第一个界的证明类似，我们可以推导出

$$\begin{aligned}
& \mathbb{E}_{S_t}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] = A + B + C \\
\leq & \frac{3}{2}|A| + \frac{8L}{\gamma_t n}\mathbb{E}\left[R(w_t) \mid \mathcal{F}_{t-1}\right] + \frac{\gamma_t}{2}\left(\|w^* - w_{t-1}\|^2 - \mathbb{E}\left[\|w^* - w_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) + \sqrt{2\gamma_t\epsilon_t}\mathbb{E}\left[\|w^* - w_t\| \mid \mathcal{F}_{t-1}\right] \\
\leq & \frac{\gamma_t}{2}\left(\|w^* - w_{t-1}\|^2 - \mathbb{E}\left[\|w^* - w_t\|^2 \mid \mathcal{F}_{t-1}\right]\right) + \frac{9.5L}{\gamma_t n}R^* \\
& + \frac{1}{2}\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] + \frac{3n}{2}\epsilon_t + \sqrt{2\gamma_t\epsilon_t}\mathbb{E}\left[\|w^* - w_t\| \mid \mathcal{F}_{t-1}\right] + \frac{3G}{2}\sqrt{\frac{2\epsilon_t}{\gamma_t}}.
\end{aligned}$$

对上述不等式中的各项重新整理后，我们得到所需的第二个界。 □

有了上述初步结果，我们现在可以证明定理4。

定理 4 的证明。 根据假设 $R(w_t) - R^* \geq \frac{\lambda}{2}D^2(w_t, W^*)$ 且 $\gamma_t = \frac{\lambda\rho t}{4} \geq \frac{\lambda\rho}{4} \geq \frac{19L}{n}$，根据引理 7 中的第一个界，我们可以证明

$$\begin{aligned}
& (1-2\rho)\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \\
\leq & \gamma_t D^2(w_{t-1}, W^*) - \left(\gamma_t + \frac{\rho\lambda}{2}\right)\mathbb{E}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right] + \frac{19L}{\gamma_t n}R^* + \left(3n + \frac{4\gamma_t}{\rho\lambda}\right)\epsilon_t + 3G\sqrt{\frac{2\epsilon_t}{\gamma_t}} \\
\leq & \frac{\lambda\rho t}{4}D^2(w_{t-1}, W^*) - \frac{\rho\lambda(t+2)}{4}\mathbb{E}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right] + \frac{76L}{\lambda\rho n t}R^* + (3n+t)\,\epsilon_t + 6G\sqrt{\frac{2\epsilon_t}{\lambda\rho t}}.
\end{aligned}$$

现在设 $\epsilon_t \leq \frac{\epsilon}{nt^4}$ 对某个 $\epsilon \in [0,1]$ 成立。由于 $\rho \leq 1/4$，上述结果意味着

$$\begin{aligned}
& \mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \\
\leq & \frac{\lambda\rho t}{2}D^2(w_{t-1}, W^*) - \frac{\rho\lambda(t+2)}{2}\mathbb{E}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right] + \frac{152L}{\lambda\rho n t}R^* + \left(\frac{6}{t^4} + \frac{2}{t^3} + 12G\sqrt{\frac{2}{\lambda\rho t^5}}\right)\sqrt{\epsilon}.
\end{aligned}$$

于是，上述不等式蕴含

$$\begin{aligned}
t\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \leq & (t+1)\mathbb{E}\left[R(w_t) - R^* \mid \mathcal{F}_{t-1}\right] \\
\leq & \frac{\lambda\rho t(t+1)}{2}D^2(w_{t-1}, W^*) - \frac{\lambda\rho(t+1)(t+2)}{2}\mathbb{E}\left[D^2(w_t, W^*) \mid \mathcal{F}_{t-1}\right] \\
& + \frac{304L}{\lambda\rho n}R^* + \left(\frac{12}{t^3} + \frac{4}{t^2} + \frac{24G}{t}\sqrt{\frac{2}{\lambda\rho t}}\right)\sqrt{\epsilon},
\end{aligned}$$

其中我们用到了如下事实：$\frac{t+1}{t} \leq 2$ 对 $t \geq 1$。根据全期望定律，将上述不等式从 $t = 1,...,T$ 求和并进行自然归一化，可得

$$\begin{aligned}
&\frac{2}{T(T+1)} \sum_{t=1}^{T} t\mathbb{E}\left[R(w_t) - R^*\right] \\
\leq &\frac{2\lambda\rho}{T(T+1)} D^2(w_0, W^*) + \frac{608L}{\lambda\rho(T+1)n} R^* + \frac{\sqrt{\epsilon}}{T(T+1)} \left(64 + 192G\sqrt{\frac{2}{\lambda\rho}}\right) \\
\leq &\frac{4\rho}{T(T+1)} (R(w_0) - R^*) + \frac{608L}{\lambda\rho(T+1)n} R^* + \frac{\sqrt{\epsilon}}{T(T+1)} \left(64 + 192G\sqrt{\frac{2}{\lambda\rho}}\right),
\end{aligned}$$

由此立即得到所需的界。证明至此完成。 □

B.2 定理5的证明

本小节将证明定理5，其表述如下。

定理5。假设*1* 和*3*成立。令 $\gamma_t \equiv \gamma \geq \frac{19L}{n}$ 。假设 $\epsilon_t \leq \min\left\{\frac{\epsilon}{n^2 t^5}, \frac{2G^2}{9n^2\gamma}\right\}$ 其中 $\epsilon \in [0, 1]$。则算法 $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$ 1的平均输出满足

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \frac{\gamma}{T} D^2(w_0, W^*) + \frac{L}{\gamma n} R^* + \left(\frac{L}{\gamma n} + \frac{\gamma}{LnT} + \frac{G}{\sqrt{\gamma} nT}\right)\sqrt{\epsilon}.$$

特别地，对于 $\gamma = \sqrt{\frac{T}{n}} + \frac{19L}{n}$，有

$$\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \left(\frac{1}{\sqrt{nT}} + \frac{L}{nT}\right) D^2(w_0, W^*) + \frac{L}{\sqrt{nT}} R^* + \left(\frac{L+G}{\sqrt{nT}} + \frac{1}{nT}\right)\sqrt{\epsilon}.$$

以下引理可通过数学归纳法证明（例如参见Schmidt等人，2011），将用于证明主要结果。

引理 8. 假设非负序列 $\{u_\tau\}_{\tau \geq 1}$ 对所有 $t \geq 1$ 都满足以下递推关系：

$$u_t^2 \leq S_t + \sum_{\tau=1}^{t} \alpha_\tau u_\tau,$$

其中 $\{S_\tau\}_{\tau \geq 1}$为递增序列，$S_0 \geq u_0^2$且 $\alpha_\tau \geq 0$对所有 τ均成立。于是，对所有 $t \geq 1$，以下界成立：

$$u_t \leq \sqrt{S_t} + \sum_{\tau=1}^{t} \alpha_\tau.$$

以下引理给出了期望估计误差 $\mathbb{E}\left[\|w_0^* - w_t\|\right]$ 的上界。

引理9. 在定理 5的条件下，对所有 $t \geq 1$，以下界成立：

$$\mathbb{E}\left[\|w_t - w_0^*\|\right] \leq \|w_0 - w_0^*\| + \sqrt{\frac{t}{\gamma}R^*} + \frac{6tG}{\gamma}.$$

证明. 回顾 $w_0^* = \arg\min_{w \in W^*} \|w_0 - w\|$。由于 $\gamma_t \equiv \gamma \geq \frac{19L}{n}$，引理 7中的第二个界成立。对于任意 $t \in [T]$，将该不等式与 $w^* = w_0^*$ 从 $\tau = 1$，……，t 相加，可得

$$\begin{aligned}
&\sum_{\tau=1}^{t} \mathbb{E}\left[R(w_\tau) - R^*\right] + \gamma \mathbb{E}\left[\|w_t - w_0^*\|^2\right] \\
\leq &\gamma \|w_0 - w_0^*\|^2 + \frac{19L}{\gamma n} t R^* + 3n \sum_{\tau=1}^{t} \epsilon_\tau + \sum_{\tau=1}^{t} \left(2\sqrt{2\gamma}\mathbb{E}\left[\|w_0^* - w_\tau\|\right] + 3G\sqrt{\frac{2}{\gamma}}\right)\sqrt{\epsilon_\tau}.
\end{aligned} \tag{16}$$

从上述不等式中去掉非负项 $\sum_{\tau=1}^{t} \mathbb{E}_{S_{[\tau]}}\left[R(w_\tau) - R^*\right]$，可得

$$\begin{aligned}
&\underbrace{\mathbb{E}\left[\|w_t - w_0^*\|^2\right]}_{u_t^2} \\
\leq &\|w_0 - w_0^*\|^2 + \frac{19L}{\gamma^2 n} t R^* + \frac{3n}{\gamma} \sum_{\tau=1}^{t} \epsilon_\tau + \sum_{\tau=1}^{t} \left(2\sqrt{\frac{2}{\gamma}}\mathbb{E}\left[\|w_0^* - w_\tau\|\right] + 3G\frac{\sqrt{2}}{\gamma\sqrt{\gamma}}\right)\sqrt{\epsilon_\tau} \\
\overset{\zeta_1}{\leq} &\|w_0 - w_0^*\|^2 + \frac{t}{\gamma} R^* + \sum_{\tau=1}^{t} \left(\frac{3n}{\gamma}\epsilon_\tau + \frac{3G\sqrt{2}}{\gamma\sqrt{\gamma}}\sqrt{\epsilon_\tau}\right) + \sum_{\tau=1}^{t} \left(2\sqrt{\frac{2\epsilon_\tau}{\gamma}}\sqrt{\mathbb{E}\left[\|w_0^* - w_\tau\|^2\right]}\right) \\
\leq &\underbrace{\|w_0 - w_0^*\|^2 + \frac{t}{\gamma} R^* + \sum_{\tau=1}^{t} \frac{4G\sqrt{2\epsilon_\tau}}{\gamma\sqrt{\gamma}}}_{S_t} + \sum_{\tau=1}^{t} \left(\underbrace{2\sqrt{\frac{2\epsilon_\tau}{\gamma}}}_{\alpha_\tau} \underbrace{\sqrt{\mathbb{E}\left[\|w_0^* - w_\tau\|^2\right]}}_{u_\tau}\right),
\end{aligned}$$

其中，在“ζ_1”中我们使用了 $\gamma \geq \frac{19L}{n}$ 和基本不等式 $\mathbb{E}^2[X] \leq \mathbb{E}[X^2]$；在最后一个不等式中，我们对所有 $\tau \geq 1$使用了条件 $\epsilon_\tau \leq \frac{2G^2}{9n^2\gamma}$。对上述递推形式应用引理 8，可得对所有 $t \geq 1$,

$$\begin{aligned}
\sqrt{\mathbb{E}\left[\|w_t - w_0^*\|^2\right]} \leq &\sqrt{\|w_0 - w_0^*\|^2 + \frac{t}{\gamma} R^* + \sum_{\tau=1}^{t} \frac{4G\sqrt{2\epsilon_\tau}}{\gamma}} + \sum_{\tau=1}^{t} 2\sqrt{\frac{2\epsilon_\tau}{\gamma}} \\
\leq &\|w_0 - w_0^*\| + \sqrt{\frac{t}{\gamma} R^*} + \sum_{\tau=1}^{t} \sqrt{\frac{4G\sqrt{2\epsilon_\tau}}{\gamma\sqrt{\gamma}}} + \sum_{\tau=1}^{t} 2\sqrt{\frac{2\epsilon_\tau}{\gamma}} \\
\leq &\|w_0 - w_0^*\| + \sqrt{\frac{t}{\gamma} R^*} + \frac{6Gt}{\gamma},
\end{aligned}$$

其中，最后一个不等式是由条件 $\epsilon_\tau \leq \frac{2G^2}{9\gamma}$ 对所有 $\tau \geq 1$ 成立得到的。因此，上述不等式直接推出对所有 $t \in [T]$ 成立的所需界。 □

现在我们可以证明定理5了。

定理5的证明。去掉非负项 $\gamma\mathbb{E}\left[\|w_t - w^*\|^2\right]$，在（16）中进行自然归一化后可得

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[R(w_t) - R^*\right] \\
\leq&\frac{\gamma}{T}\|w_0 - w_0^*\|^2 + \frac{19L}{\gamma n}R^* + \frac{3n}{T}\sum_{t=1}^{T}\epsilon_t + \frac{1}{T}\sum_{t=1}^{T}\left(2\sqrt{2\gamma}\mathbb{E}\left[\|w_t - w_0^*\|\right] + 3G\sqrt{\frac{2}{\gamma}}\right)\sqrt{\epsilon_t} \\
\overset{\zeta_1}{\leq}&\frac{\gamma}{T}\|w_0 - w_0^*\|^2 + \frac{19L}{\gamma n}R^* + \frac{3n}{T}\sum_{t=1}^{T}\epsilon_t \\
&+ \frac{1}{T}\sum_{t=1}^{T}\left(2\sqrt{2}\left(\sqrt{\gamma}\|w_0 - w_0^*\| + \sqrt{tR^*} + \frac{6Gt}{\sqrt{\gamma}}\right) + 3G\sqrt{\frac{2}{\gamma}}\right)\sqrt{\epsilon_t} \\
\overset{\zeta_2}{\leq}&\frac{\gamma}{T}\|w_0 - w_0^*\|^2 + \frac{19L}{\gamma n}R^* + \frac{1}{T}\sum_{t=1}^{T}\left(3n\epsilon_t + 2\sqrt{2\gamma\epsilon_t}\|w_0 - w_0^*\| + 2\sqrt{2tR^*\epsilon_t} + \frac{15\sqrt{2\epsilon_t}Gt}{\sqrt{\gamma}}\right) \\
\overset{\zeta_3}{\leq}&\frac{\gamma}{T}\|w_0 - w_0^*\|^2 + \frac{19L}{\gamma n}R^* \\
&+ \frac{1}{T}\sum_{t=1}^{T}\left(3n\epsilon_t + \frac{\gamma\|w_0 - w_0^*\|^2}{t^2} + 2t^2\epsilon_t + \frac{2LR^*}{\gamma n} + \frac{\gamma nt\epsilon_t}{L} + \frac{15\sqrt{2\epsilon_t}Gt}{\sqrt{\gamma}}\right) \\
\leq&\frac{3\gamma}{T}\|w_0 - w_0^*\|^2 + \frac{21L}{\gamma n}R^* + \frac{1}{T}\sum_{t=1}^{T}\left(3n\epsilon_t + 2t^2\epsilon_t + \frac{\gamma nt\epsilon_t}{L} + \frac{15\sqrt{2\epsilon_t}Gt}{\sqrt{\gamma}}\right),
\end{aligned}
$$

其中“ζ_1”由引理9得出，“ζ_2”源于 $t \geq 1$，“ζ_3”源于 $ab \leq (a^2 + b^2)/2$。现在考虑$\epsilon_t \leq \frac{\epsilon}{n^2t^5}$，其中 $\epsilon \in [0, 1]$。那么由前述不等式可得

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[R(w_t) - R^*\right] \\
\leq&\frac{3\gamma}{T}\|w_0 - w_0^*\|^2 + \frac{21L}{\gamma n}R^* + \frac{1}{T}\sum_{t=1}^{T}\left(\frac{3}{nt^5} + \frac{2}{nt^3} + \frac{\gamma}{nLt^4} + \frac{15\sqrt{2}G}{nt^{1.5}\sqrt{\gamma}}\right)\sqrt{\epsilon} \\
\leq&\frac{3\gamma}{T}\|w_0 - w_0^*\|^2 + \frac{21L}{\gamma n}R^* + \frac{1}{T}\left(\frac{6}{n} + \frac{4}{n} + \frac{2\gamma}{nL} + \frac{45\sqrt{2}G}{n\sqrt{\gamma}}\right)\sqrt{\epsilon}.
\end{aligned}
$$

令 $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$。结合 R 的凸性，上述不等式意味着

$$
\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \frac{\gamma}{T}D^2(w_0, W^*) + \frac{L}{\gamma n}R^* + \left(\frac{1}{nT} + \frac{\gamma}{LnT} + \frac{G}{nT\sqrt{\gamma}}\right)\sqrt{\epsilon}.
$$

这就证明了第一个界。将 $\gamma = \sqrt{\frac{T}{n}} + \frac{19L}{n} > \frac{19L}{n}$ 代入上述界，并保留主导项，可得所需的第二个界：

$$
\mathbb{E}\left[R(\bar{w}_T) - R^*\right] \lesssim \left(\frac{1}{\sqrt{nT}} + \frac{L}{nT}\right)D^2(w_0, W^*) + \frac{L}{\sqrt{nT}}R^* + \left(\frac{L+G}{\sqrt{nT}} + \frac{1}{nT}\right)\sqrt{\epsilon}.
$$

证明结束。 □

C 第 4 节结果的证明4

本节给出第4节中所述高概率估计误差界的证明。

C.1 命题 1 的证明1

本小节将证明如下重述的命题 1 。

命题 1。 假设 *1* 成立，且损失函数有界，即 $0 \leq \ell(y', y) \leq M$ 对所有 y, y'。设 $S = \{S_t\}_{t\in[T]}$ 和 $S' = \{S'_t\}_{t\in[T]}$ 是满足 $S \doteq S'$ 的两组数据小批次。那么

(a) M-SPP（算法 *1*）分别在 S 和 S' 上生成的加权平均输出 $\bar{w}_T$ 和 $\bar{w}'_T$ 满足

$$\sup_{S,S'} \|\bar{w}_T - \bar{w}'_T\| \leq \frac{4\sqrt{2LM}}{n \min_{t\in[T]} \gamma_t} + \sum_{t=1}^{T} 2\sqrt{\frac{2\epsilon_t}{\gamma_t}}.$$

(b) 加权平均输出 $\bar{w}_T$ 和 $\bar{w}'_T$ 分别由 M-SPP-SWoR（算法 *3*）在 S 和 S' 上生成，满足

$$\sup_{S,S'} \mathbb{E}_{\xi_{[T]}} \left[\|\bar{w}_T - \bar{w}'_T\|\right] \leq \sum_{t=1}^{T} \left\{ \frac{4\sqrt{2LM}}{nT\gamma_t} + 2\sqrt{\frac{2\epsilon_t}{\gamma_t}} \right\}.$$

我们首先需要证明以下初步结果，该结果涉及 M-SPP 更新在相同或不同小批次上执行时的扩张性质。

引理 10。 假设假设条件 *1* 成立，且损失函数 ℓ 在区间 $[0, M]$ 内有界 $w_0 = w'_0$。令序列 $\{w_t\}_{t\in[T]}$ 和 $\{w'_t\}_{t\in[T]}$ 分别根据以下规则在 $\{S_t\}_{t\in[T]}$ 和 $\{S'_t\}_{t\in[T]}$ 上生成:

$$F_t(w_t) \leq \min_{w\in\mathcal{W}} \left\{ F_t(w) := R_{S_t}(w) + \frac{\gamma_t}{2}\|w - w_{t-1}\|^2 \right\} + \epsilon_t,$$
$$F'_t(w'_t) \leq \min_{w\in\mathcal{W}} \left\{ F'_t(w) := R_{S'_t}(w) + \frac{\gamma_t}{2}\|w - w'_{t-1}\|^2 \right\} + \epsilon_t.$$

假设对于所有 $t \in [T]$, $S_t = S'_t$ 或 $S_t \doteq S'_t$ 成立。令 $\beta_t = \mathbf{1}_{\{S_t \neq S'_t\}}$。则对于所有 $t \in [T]$，以下界成立，

$$\|w_t - w'_t\| \leq \sum_{\tau=1}^{t} \left\{ \beta_\tau \frac{4\sqrt{2LM}}{n\gamma_\tau} + 2\sqrt{\frac{2\epsilon_\tau}{\gamma_\tau}} \right\}.$$

证明。 令 $w_t^* = \arg\min_w F_t(w)$ 和 $w_t'^* = \arg\min_w F_t'(w)$。由引理 1 可得

$$\begin{aligned}
R_{S_t}(w_t^*) - R_{S_t}(w_t'^*) \leq & \frac{\gamma_t}{2}\left(\|w_t'^* - w_{t-1}\|^2 - \|w_t'^* - w_t^*\|^2 - \|w_t^* - w_{t-1}\|^2\right) \\
R_{S_t'}(w_t'^*) - R_{S_t'}(w_t^*) \leq & \frac{\gamma_t}{2}\left(\|w_t^* - w_{t-1}'\|^2 - \|w_t'^* - w_t^*\|^2 - \|w_t'^* - w_{t-1}'\|^2\right).
\end{aligned}$$

将上述两个不等式两边相加可得

$$\begin{aligned}
& R_{S_t}(w_t^*) - R_{S_t}(w_t'^*) + R_{S_t'}(w_t'^*) - R_{S_t'}(w_t^*) \\
\leq & \frac{\gamma_t}{2}\left(\|w_t'^* - w_{t-1}\|^2 - \|w_t^* - w_{t-1}\|^2 + \|w_t^* - w_{t-1}'\|^2 - \|w_t'^* - w_{t-1}'\|^2 - 2\|w_t'^* - w_t^*\|^2\right) \\
= & \frac{\gamma_t}{2}\left(2\langle w_t^* - w_t'^*, w_{t-1} - w_{t-1}'\rangle - 2\|w_t'^* - w_t^*\|^2\right) \\
\leq & \frac{\gamma_t}{2}\left(\|w_{t-1} - w_{t-1}'\|^2 - \|w_t^* - w_t'^*\|^2\right)
\end{aligned}.$$

我们需要区分以下两种互补情形。

情形I：$S_t = S_t'$。在这种情形下，前一个不等式可直接推出

$$\|w_t^* - w_t'^*\| \leq \|w_{t-1} - w_{t-1}'\|.$$

利用三角不等式和引理6 可得

$$\|w_t - w_t'\| \leq \|w_t - w_t^*\| + \|w_t^* - w_t'^*\| + \|w_t' - w_t'^*\| \leq \|w_{t-1} - w_{t-1}'\| + 2\sqrt{\frac{2\epsilon_t}{\gamma_t}}. \tag{17}$$

情形II： S_t 和 S_t' 仅有一个元素不同。在这种情形下，我们有

$$\begin{aligned}
& \|w_t^* - w_t'^*\|^2 \\
\leq & \|w_{t-1} - w_{t-1}'\|^2 + \frac{2}{\gamma_t}\left(R_{S_t}(w_t'^*) - R_{S_t}(w_t^*) + R_{S_t'}(w_t^*) - R_{S_t'}(w_t'^*)\right) \\
= & \|w_{t-1} - w_{t-1}'\|^2 + \frac{2}{\gamma_t}\left(R_{S_t}^{\ell}(w_t'^*) - R_{S_t}^{\ell}(w_t^*) + R_{S_t'}^{\ell}(w_t^*) - R_{S_t'}^{\ell}(w_t'^*)\right) \\
= & \|w_{t-1} - w_{t-1}'\|^2 + \frac{2}{\gamma_t}\left(\frac{1}{|S_t|}\sum_{z\in S_t}(\ell(w_t'^*; z) - \ell(w_t^*; z) + \frac{1}{|S_t'|}\sum_{z\in S_t'}(\ell(w_t^*; z) - \ell(w_t'^*; z))\right) \\
\leq & \|w_{t-1} - w_{t-1}'\|^2 + \frac{4\sqrt{2LM}}{n\gamma_t}\|w_t^* - w_t'^*\|.
\end{aligned}$$

其中，在最后一个不等式中，我们使用了 $\ell(\cdot;\cdot)$关于其第一个参数是 $\sqrt{2LM}$-Lipschitz的，这由引理2推出，并且 S_t 和 S_t' 也仅有一个元素不同。由于 $x^2 \leq y^2 + ax$蕴含 $x \leq y + a$ 对所有 x， y，$a > 0$，由上述结果可得

$$\|w_t^* - w_t'^*\| \leq \|w_{t-1} - w_{t-1}'\| + \frac{4\sqrt{2LM}}{n\gamma_t}.$$

然后，根据三角不等式和引理6 ，我们有

$$\|w_t - w_t'\| \leq \|w_t - w_t^*\| + \|w_t^* - w_t'^*\| + \|w_t' - w_t'^*\| \leq \|w_{t-1} - w_{t-1}'\| + \frac{4\sqrt{2LM}}{n\gamma_t} + 2\sqrt{\frac{2\epsilon_t}{\gamma_t}}. \tag{18}$$

令 $\beta_t = \mathbf{1}_{\{S_t \neq S'_t\}}$ 其中 $\mathbf{1}_{\{C\}}$ 是条件 C的指示函数。根据递推形式（17）和（18），以及条件 $w_0 = w'_0$，我们可以证明对所有 $t \in [T]$

$$\|w_t - w'_t\| \leq \sum_{\tau=1}^{t} \left\{ \frac{4\beta_\tau \sqrt{2LM}}{n\gamma_\tau} + 2\sqrt{\frac{2\epsilon_\tau}{\gamma_\tau}} \right\},$$

这就是所需的界。 □

现在我们可以证明命题 1 的主要结果了。

命题 *1* 的证明。 考虑一对固定的小批次集合 $S \doteq S'$。

部分 (a)：设 $\{w_t\}_{t\in[T]}$和 $\{w'_t\}_{t\in[T]}$是分别由算法 1 在 $\{S_t\}_{t\in[T]}$ 和 $\{S'_t\}_{t\in[T]}$ 上生成的两个解序列。在每个时间点 t，定义随机变量$\beta_t := \mathbf{1}_{\{S_t \neq S'_t\}}$。由于根据假设， S 和 S' 仅在一个小批次上不同，因此必定存在唯一的一个 $t \in [T]$，使得 $\beta_t = 1$，且对于所有 $j \in [T], j \neq t$， $\beta_j = 0$。那么，在 $\beta_\tau = 1$ 关于 $\tau = \arg\min^{i\in[t]} \gamma_i$ 的最坏情况下，由引理 10 可知，对于所有 $t \in [T]$,

$$\|w_t - w'_t\| \leq \frac{4\sqrt{2LM}}{n \min_{i\in[t]} \gamma_i} + \sum_{i=1}^{t} 2\sqrt{\frac{2\epsilon_i}{\gamma_i}} \leq \frac{4\sqrt{2LM}}{n \min_{i\in[T]} \gamma_i} + \sum_{i=1}^{T} 2\sqrt{\frac{2\epsilon_i}{\gamma_i}}.$$

那么， $\bar{w}_T$ 和 $\bar{w}'_T$ 的凸组合性质意味着

$$\|\bar{w}_T - \bar{w}'_T\| \leq \frac{\sum_t \gamma_t \|w_t - w'_t\|}{\sum_t \gamma_t} \leq \frac{4\sqrt{2LM}}{n \min_{t\in[T]} \gamma_t} + \sum_{t=1}^{T} 2\sqrt{\frac{2\epsilon_t}{\gamma_t}}.$$

由于上述界对任意一对 $\{S, S'\}$都成立，因此所需结果立即得出。

部分（b）：回顾一下， $\{\xi_t\}_{t\in[T]}$ 是用于从 S 和 S'中迭代选择数据小批量的均匀随机索引。设 $\{w_t\}_{t\in[T]}$ 和$\{w'_t\}_{t\in[T]}$ 是由算法 3分别在 $\{S_{\xi_t}\}_{t\in[T]}$和 $\{S'_{\xi_t}\}_{t\in[T]}$上生成的两个解序列。定义随机变量$\beta_t := \mathbf{1}_{\left\{S_{\xi_t} \neq S'_{\xi_t}\right\}}$。由于根据假设， S 和 S' 仅在一个小批量上不同，在无放回抽样方案下，必定存在唯一一个 $t \in [T]$，使得 $\beta_t = 1$，并且对所有 $j \in [T]$、$j \neq t$均有$\beta_j = 0$。定义事件 $\mathcal{E}_t$:= $\{\beta_t = 1$ 且 $\beta_{j\neq t,\ j\in[T]} = 0\}$ 对所有 $t \in [T]$均成立。于是， ξ_t 的均匀随机性意味着

$$R(\mathcal{E}_t) = \frac{1}{T}, \quad t \in [T].$$

给定 $t \in [T]$，假设 $\mathcal{E}_\tau$ 在某个 $\tau \in [t]$处发生。那么由引理 10 可得

$$\|w_t - w'_t\| \leq \frac{4\sqrt{2LM}}{n\gamma_\tau} + \sum_{i=1}^{t} 2\sqrt{\frac{2\epsilon_i}{\gamma_i}}.$$

假设 $\mathcal{E}_\tau$在某个 $\tau \in \{t+1, t+2, ..., T\}$处发生，再次由引理 10 可得

$$\|w_t - w_t'\| \leq \sum_{i=1}^{t} 2\sqrt{\frac{2\epsilon_i}{\gamma_i}}.$$

那么我们有

$$\begin{aligned}
\mathbb{E}_{\xi_{[t]}}\left[\|w_t - w_t'\|\right] &= \sum_{\tau=1}^{T} R\left(\mathcal{E}_\tau\right)\left[\|w_t - w_t'\| \mid \mathcal{E}_\tau\right] \\
&\leq \sum_{\tau=1}^{t}\left\{\frac{4\sqrt{2LM}}{nT\gamma_t} + \sum_{i=1}^{t}\frac{2}{T}\sqrt{\frac{2\epsilon_i}{\gamma_i}}\right\} + \sum_{\tau=t+1}^{T}\left\{\sum_{i=1}^{t}\frac{2}{T}\sqrt{\frac{2\epsilon_t}{\gamma_t}}\right\} \\
&= \sum_{\tau=1}^{t}\left\{\frac{4\sqrt{2LM}}{nT\gamma_\tau} + 2\sqrt{\frac{2\epsilon_\tau}{\gamma_\tau}}\right\} \leq \sum_{t=1}^{T}\left\{\frac{4\sqrt{2LM}}{nT\gamma_t} + 2\sqrt{\frac{2\epsilon_t}{\gamma_t}}\right\}.
\end{aligned}$$

由此可得

$$\mathbb{E}_{\xi_{[T]}}\left[\|\bar{w}_T - \bar{w}_T'\|\right] \leq \frac{\sum_t \gamma_t \mathbb{E}_{\xi_{[t]}}\left[\|w_t - w_t'\|\right]}{\sum_t \gamma_t} \leq \sum_{t=1}^{T}\left\{\frac{4\sqrt{2LM}}{nT\gamma_t} + 2\sqrt{\frac{2\epsilon_t}{\gamma_t}}\right\}.$$

由于上述界对任意一对 $\{S, S'\}$ 均成立，因此所需结果立即得出。 □

C.2 定理 6 的证明

在本小节中，我们证明定理 6，其内容重述如下。

定理 6。 假设 *1*、*2*、*3* 成立，且损失函数 ℓ 在区间 $(0, M]$内有界。令 $\rho \in (0, 1/4]$ 为任意标量，并设 $\gamma_t = \frac{\lambda\rho t}{4}$。假设 $n \geq \frac{76L}{\lambda\rho}$。设 $\epsilon_t \leq$最小值 $\left\{\frac{\epsilon}{nt^4}, \frac{LM}{\lambda\rho n^2 T^2 t}\right\}$对于某个 $\epsilon \in [0, 1]$成立。那么，以至少 $1-\delta$的概率，*S*M-SPP-SWoR（算法 *3*）的加权平均输出满足

$$\begin{aligned}
&\mathbb{E}_{\xi_{[T]}}\left[D(\bar{w}_T, W^*)\right] \\
\lesssim &\frac{\sqrt{LM\log(1/\delta)}\log(T)}{\lambda\rho\sqrt{nT}} + \sqrt{\frac{\rho\left[R(w_0) - R^*\right]}{\lambda T^2} + \frac{L}{\lambda^2\rho nT}R^* + \frac{\sqrt{\epsilon}}{\lambda T^2}\left(\frac{L}{\lambda\rho} + G\sqrt{\frac{1}{\lambda\rho}}\right)}.
\end{aligned}$$

为证明这一结果，我们需要使用以下重新表述的 McDiarmid 不等式（McDiarmid，1989），该不等式也称为有界差分不等式。

引理 11（McDiarmid/有界差分不等式）。设 X_1、X_2、...、X_N 是取值于 $\mathcal{X}$ 的相互独立随机变量。假设函数 h： $\mathcal{X}^N \mapsto \mathbb{R}$ 满足有界差分性质，即对于任意 $i \in [N]$ 以及任意 x_1、...、x_N、x_i'，以下不等式成立：

$$|h(x_1, ..., x_{i-1}, x_i, x_{i+1}, ..., x_N) - h(x_1, ..., x_{i-1}, x_i', x_{i+1}, ..., x_N)| \leq c_i.$$

那么，对于任意 $\varepsilon > 0$，

$$\mathbb{P}\left(h(X_1, ..., X_N) - \mathbb{E}\left[h(X_1, ..., X_N)\right] \geq \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{N} c_i^2}\right).$$

现在我们可以证明定理 6。

定理 *6* 的证明。设 $S = \{S_t\}_{t\in[T]}$和 $S' = \{S'_t\}_{t\in[T]}$是两组满足 $S \doteq S'$ 的数据小批次。根据命题 1，由算法 $\bar{w}_T$和 $\bar{w}'_T$ 分别在 S 和 S' 上生成的加权平均输出满足

$$\sup_{S,S'} \mathbb{E}_{\xi_{[T]}} \left[\|\bar{w}_T - \bar{w}'_T\|\right] \leq \sum_{t=1}^{T} \left\{ \frac{4\sqrt{2LM}}{nT\gamma_t} + 2\sqrt{\frac{2\epsilon_t}{\gamma_t}} \right\} \leq \sum_{t=1}^{T} \left\{ \frac{5\sqrt{2LM}}{nT\gamma_t} \right\} \leq \frac{20\sqrt{2LM}(1+\log(T))}{\lambda\rho nT},$$

其中，在倒数第二个不等式中，我们使用了条件 $\epsilon_t \leq \frac{LM}{4n^2T^2\gamma_t} = \frac{LM}{\lambda\rho N^2 t}$。由三角不等式和上述界可得

$$\sup_{S,S'} \mathbb{E}_{\xi_{[T]}} \left[\left|D(\bar{w}_T, W^*) - D(\bar{w}'_T, W^*)\right|\right] \leq \sup_{S,S'} \mathbb{E}_{\xi_{[T]}} \left[\|\bar{w}_T - \bar{w}'_T\|\right] \leq \frac{20\sqrt{2LM}(1+\log(T))}{\lambda\rho nT}.$$

由于 $\xi_{[T]}$ 在 S上相互独立，直接对 $h(S) := D(\bar{w}_T, W^*)$应用 McDiarmid 不等式（取 $c_i \equiv c = \frac{20\sqrt{2LM}(1+\log(T))}{\lambda\rho nT}$）可知，以至少 $1-\delta$的概率（关于 S的随机性），

$$\mathbb{E}_{\xi_{[T]}} \left[D(\bar{w}_T, W^*) - \mathbb{E}_S \left[\mathbb{E}_{\xi_{[T]}} \left[D(\bar{w}_T, W^*)\right]\right]\right] \leq c\sqrt{\frac{nT\log(1/\delta)}{2}} = \frac{20\sqrt{LM\log(1/\delta)}(1+\log(T))}{\lambda\rho\sqrt{nT}}.$$

接下来，我们为 $\mathbb{E}_S\left[D(\bar{w}_T, W^*)\right]$推导一个界。根据 Jensen 不等式以及 F 的二次增长性质，我们有

$$\begin{aligned}
\mathbb{E}_S \left[\mathbb{E}_{\xi_{[T]}} \left[D(\bar{w}_T, W^*)\right]\right] =& \mathbb{E}_{\xi_{[T]}} \left[\mathbb{E}_S \left[D(\bar{w}_T, W^*)\right]\right] \\
\leq& \mathbb{E}_{\xi_{[T]}} \left[\sqrt{\mathbb{E}_S \left[D^2(\bar{w}_T, W^*)\right]}\right] \\
\leq& \mathbb{E}_{\xi_{[T]}} \left[\sqrt{\frac{2}{\lambda}\mathbb{E}_S \left[R(\bar{w}_T) - R^*\right]}\right] \\
\lesssim& \mathbb{E}_{\xi_{[T]}} \left[\sqrt{\frac{\rho\left[R(w_0) - R^*\right]}{\lambda T^2} + \frac{L}{\lambda^2\rho nT}R^* + \frac{\sqrt{\epsilon}}{\lambda T^2}\left(\frac{L}{\lambda\rho} + G\sqrt{\frac{1}{\lambda\rho}}\right)}\right] \\
=& \sqrt{\frac{\rho\left[R(w_0) - R^*\right]}{\lambda T^2} + \frac{L}{\lambda^2\rho nT}R^* + \frac{\sqrt{\epsilon}}{\lambda T^2}\left(\frac{L}{\lambda\rho} + G\sqrt{\frac{1}{\lambda\rho}}\right)},
\end{aligned}$$

其中，在最后一个不等式中，我们应用了定理 4。因此，根据前两个不等式可得，以至少 $1-\delta$ 关于 S的概率，

$$\begin{aligned}
&\mathbb{E}_{\xi_{[T]}} \left[D(\bar{w}_T, W^*)\right] \\
\lesssim& \frac{\sqrt{LM\log(1/\delta)}\log(T)}{\lambda\rho\sqrt{nT}} + \sqrt{\frac{\rho\left[R(w_0) - R^*\right]}{\lambda T^2} + \frac{L}{\lambda^2\rho nT}R^* + \frac{\sqrt{\epsilon}}{\lambda T^2}\left(\frac{L}{\lambda\rho} + G\sqrt{\frac{1}{\lambda\rho}}\right)},
\end{aligned}$$

由此得到所需的界。 □

C.3 定理 7 的证明

下面我们证明以下重述的定理7。

定理 7。 假设 *1*和 *3* 成立，且损失函数 ℓ 在区间 $[0, M]$内有界。令 $\gamma_t \equiv \sqrt{\frac{T}{n}}$。假设 $\epsilon_t \leq \frac{LM}{4nT^2\sqrt{nT}}$。则以至少 $1-\delta$的概率S，M-SPP（算法 *1*）的平均输出 $\bar{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$ 满足

$$|R(\bar{w}_T) - R_S(\bar{w}_T)| \lesssim \frac{(LM + G\sqrt{LM})\log(N)\log(1/\delta)}{\sqrt{nT}} + M\sqrt{\frac{\log(1/\delta)}{nT}}.$$

我们需要以下引理，该引理本质上源自 Bousquet 等人（2020年，推论 8），它为关于损失函数具有一致稳定性的学习算法给出了近乎紧的泛化界。

引理 12（Bousquet 等人（2020年））。设学习算法 $\mathcal{A}_w$ 由 w 参数化，并且对于任意（x， y）$\in \mathcal{X} \times \mathcal{Y}$ 和 $S \doteq S'$，满足 $|\ell(\mathcal{A}_{w_S}(x)\ y) - \ell(\mathcal{A}_{w_{S'}}(x)\ y)| \leq \varrho$。假设损失函数对于所有 y、y' 满足 $0 \leq \ell(y',\ y) \leq M$。则对于任意 $\delta \in (0, 1)$，以至少 $1-\delta$ 的概率，在大小为 N 的独立同分布数据集 S 上，

$$|R(\mathcal{A}_{w_S}) - R_S(\mathcal{A}_{w_S})| \lesssim \varrho\log(N)\log\left(\frac{1}{\delta}\right) + M\sqrt{\frac{\log(1/\delta)}{N}}.$$

有了这个引理，我们就可以证明定理7中的主要结果

定理7的证明。设 $S = \{S_t\}_{t\in[T]}$和 $S' = \{S'_t\}_{t\in[T]}$是满足$S \doteq S'$的两组数据小批次。注意到 $\gamma_t \equiv \gamma = \sqrt{\frac{T}{n}}$。根据命题1，由算法1 分别在 S 和 S' 上生成的平均输出 $\bar{w}_T$和 $\bar{w}'_T$满足

$$\sup_{S,S'}\|\bar{w}_T - \bar{w}'_T\| \leq \frac{4\sqrt{2LM}}{n\gamma} + \sum_{t=1}^{T} 2\sqrt{\frac{2\epsilon_t}{\gamma}} \leq \frac{5\sqrt{2LM}}{n\gamma} = \frac{5\sqrt{2LM}}{\sqrt{nT}}.$$

其中，在倒数第一个不等式中，我们使用了条件 $\epsilon_t \leq \frac{LM}{4nT^2\sqrt{N}}$。由此可得

$$|\ell(\bar{w}_T; z) - \ell(\bar{w}'_T; z)| \leq \sqrt{2ML}\|\bar{w}_T - \bar{w}'_T\| \leq \frac{10LM}{\sqrt{nT}},$$

其中我们用到了 $\ell(\cdot;\ \cdot)$ 是 $\sqrt{2LM}$关于其第一个参数的 -Lipschitz 函数（这可由引理2推出）。根据假设3，我们有

$$|r(\bar{w}_T) - r(\bar{w}'_T)| \leq G\|\bar{w}_T - \bar{w}'_T\| \leq \frac{5G\sqrt{2LM}}{\sqrt{nT}}.$$

上述两个不等式表明，M-SPP 关于复合损失函数 $\frac{10LM+5G\sqrt{2LM}}{\sqrt{nT}}$是一致稳定的，稳定参数为 $\ell + r$。将引理12 应用于 M-SPP，可得

$$|R(w_S) - R_S(w_S)| \lesssim \frac{(LM + G\sqrt{LM})\log(nT)}{\sqrt{nT}}\log\left(\frac{1}{\delta}\right) + M\sqrt{\frac{\log(1/\delta)}{nT}}.$$

证明结束。 □

参考文献

Alekh Agarwal、Peter L. Bartlett、Pradeep Ravikumar 和 Martin J. Wainwright。随机凸优化预言机复杂度的信息论下界。IEEE 信息论汇刊，58(5):3235–3249，2012。3，6

Zeyuan Allen-Zhu。Katyusha：首个直接加速随机梯度方法的算法。机器学习研究杂志，18:221:1–221:51，2017年。7

Mihai Anitescu。具有二次增长条件的退化非线性规划。SIAM 优化杂志，10(4):1116–1135，2000年。9

Hilal Asi 和 John C Duchi。更好的模型在随机优化中的重要性。美国国家科学院院刊，116(46):22924–22930，2019年a。2，6，13，35

Hilal Asi 和 John C. Duchi。随机（近似）近端点方法：收敛、最优性与自适应性。SIAM 优化杂志，29(3):2257–2290，2019年b。2，3，6，13，18，20，35

Hilal Asi、Karan N. Chadha、Gary Cheng 和 John C. Duchi。小批量随机近似近端点方法。载于神经信息处理系统进展（NeurIPS），线上会议，2020年。2，3，5，7，18，20，24

Francis R. Bach 和Eric Moulines。机器学习随机逼近算法的非渐近分析。载于神经信息处理系统进展（NIPS），第 451– 459 页，西班牙格拉纳达，2011年。6

Peter L Bartlett、Olivier Bousquet 和 Shahar Mendelson。局部 Rademacher 复杂度。 统计年刊，33(4):1497–1537，2005年。7

Raef Bassily、Vitaly Feldman、Crist´obal Guzm´an 和 KunalTalwar。随机梯度下降在非光滑凸损失上的稳定性。载于神经信息处理系统进展（NeurIPS），线上会议，2020年。8

Dimitri P. Bertsekas. 大规模凸优化的增量近端方法。 数学规划，129(2):163–195，2011。3，6，18

L´eon Bottou 和 Olivier Bousquet。大规模学习的权衡。在 神经信息处理系统进展（NIPS），第 161–168 页，温哥华，加拿大，2007 年。7

L´eon Bottou、Frank E. Curtis 和 Jorge Nocedal。大规模机器学习的优化方法。SIAM 综述，60(2)：223–311，2018。6Olivier Bousquet 和 Andr´e Elisseeff。稳定性与泛化。机器学习研究杂志，2：499–526，2002年。4，7，16

Olivier Bousquet、Yegor Klochkov 和 Nikita Zhivotovskiy。一致稳定算法的更紧界。在学习理论会议（COLT）论文集，第610–626页，线上会议 [奥地利格拉茨], 2020。7，48

Karan N. Chadha、Gary Cheng 和 John C. Duchi。加速、最优且并行：基于模型的随机优化的一些结果。载于 国际机器学习会议 (ICML) 论文集，第 2811–2827 页，巴尔的摩，马里兰州，2022年。 25

Ronan Collobert、Samy Bengio 和 Yoshua Bengio。面向超大规模问题的并行支持向量机混合。载于神经信息处理系统进展 (NIPS)，第 633–640 页，温哥华，加拿大], 2001。 23

Koby Crammer、Ofer Dekel、Joseph Keshet、Shai Shalev-Shwartz 和 Yoram Singer。在线被动攻击算法。 机器学习研究杂志，7:551–585，2006。 6

Damek Davis 和 Dmitriy Drusvyatskiy。弱凸函数的随机基于模型的最小化。 SIAM 优化杂志，29(1):207–239，2019年。 2，3，4，5，6，11，18，20

Aaron Defazio、Francis R. Bach 和 Simon Lacoste-Julien。SAGA：一种快速增量梯度方法，支持非强凸复合目标函数。载于 神经信息处理系统进展（NIPS），第 1646–1654 页，加拿大蒙特利尔，2014。 7

Qi Deng 和 Wenzhi Gao。用于随机弱凸优化的小批量和动量模型方法。载于 神经信息处理系统进展（NeurIPS），第 23115–23127 页，线上，2021年。 7，8，25

Aymeric Dieuleveut、Nicolas Flammarion 和 Francis R. Bach。更难、更好、更快、更强：最小二乘回归的收敛速率。 机器学习研究杂志，18:101:1–101:51，2017年。 4，18，19， 20

Dmitriy Drusvyatskiy 和 Adrian S. Lewis。误差界、二次增长与近端方法的线性收敛。 数学运筹学研究，43(3):919–948，2018。 4

John C. Duchi 和 Feng Ruan。复合优化与弱凸优化问题的随机方法。SIAM 优化杂志，28(4):3229–3259，2018。18

John C. Duchi、ShaiShalev-Shwartz、Yoram Singer 和 Ambuj Tewari。复合目标镜像下降。载于第 23 届学习理论会议（COLT）论文集，第 14–26 页，海法，以色列，2010。6

Vitaly Feldman 和 Jan Vondr´ak。一致稳定算法的泛化界。载于神经信息处理系统进展（NeurIPS），第 9770–9780 页，蒙特´利，加拿大，2018年。 4， 7

Vitaly Feldman 和 Jan Vondr´ak。具有近乎最优速率的一致稳定算法的高概率泛化界。载于学习理论会议论文集（COLT），第 1270–1279 页，美国亚利桑那州凤凰城，2019年。 7，16

Roy Frostig、Rong Ge、Sham M. Kakade 和 Aaron Sidford。单次遍历中与经验风险最小化器竞争。在第 28 届学习理论会议（COLT）论文集，第 728–763 页，法国巴黎，2015 年。7

Saeed Ghadimi 和 Guanghui Lan。强凸随机复合优化的最优随机逼近算法 I：一个通用算法框架。SIAM 优化杂志，22(4):1469–1492，2012 年。6

Isabelle Guyon、Steve R. Gunn、Asa Ben-Hur 和 Gideon Dror。NIPS 2003 特征选择挑战赛的结果分析。载于神经信息处理系统进展（NIPS），第 545–552 页，加拿大温哥华], 2004。23

Moritz Hardt、Ben Recht 和 Yoram Singer。训练更快，泛化更好：随机梯度下降的稳定性。载于第 33 届国际机器学习会议（ICML）论文集，第 1225–1234 页，纽约市，纽约州，2016 年。7，16

Chonghai Hu、James T. Kwok 和 Weike Pan。用于随机优化和在线学习的加速梯度方法。载于神经信息处理系统进展（NIPS），第 781–789 页，加拿大温哥华，2009年。6

Martin Jaggi、Virginia Smith、Martin Tak´ac、Jonathan Terhorst、Sanjay Krishnan、Thomas Hofmann 和 Michael I. Jordan。通信高效的分布式对偶坐标上升。载于神经信息处理系统进展（NIPS），第 3068–3076 页，加拿大蒙特利尔，2014。7

Rie Johnson 和 Tong Zhang。使用预测方差缩减加速随机梯度下降。载于神经信息处理系统进展（NIPS），第 315–323 页，美国内华达州太浩湖，2013。3，7

Ellango Jothimurugesan、AshrafTahmasbi、Phillip B. Gibbons 和 Srikanta Tirthapura。流式数据上的方差缩减随机梯度下降。载于神经信息处理系统进展（NeurIPS），第 9928–9937 页，蒙特´利尔，加拿大，2018。7

Hamed Karimi、Julie Nutini 和 Mark Schmidt。满足 Polyak–Łojasiewicz 条件时梯度方法和近端梯度方法的线性收敛。载于 欧洲机器学习与数据库知识发现会议（ECML/PKDD）论文集，第一部分，第 795–811 页，意大利加尔达湖畔里瓦，2016 年。4，9

笠井裕之。SGDLibrary：一个用于随机优化算法的 MATLAB 库。机器学习研究杂志，18:215:1–215:5，2017年。23

Yegor Klochkov 和 Nikita Zhivotovskiy。具有收敛速率的稳定性与偏差最优风险界 $o(1/n)$。载于神经信息处理系统进展（NeurIPS），第 5065–5076 页，线上举行，2021年。7

Vladimir Koltchinskii. 风险最小化中的局部 Rademacher 复杂度与预言机不等式。统计年刊，34(6):2593–2656，2006。7

Brian Kulis 和 Peter L. Bartlett。隐式在线学习。载于 第 27 届国际机器学习会议（ICML）论文集，第 575–582 页，海法，以色列，2010。6

Andrei Kulunchakov 和 Julien Mairal。随机复合优化的通用加速框架。载于神经信息处理系统进展（NeurIPS），第 12556– 12567 页，温哥华，加拿大，2019年。6

Guanghui Lan。随机复合优化的一种最优方法。 数学规划，133 (1-2):365–397，2012。6

Jason D. Lee、Qihang Lin、Tengyu Ma 和 Tianbao Yang。通过有放回地额外采样数据实现的分布式随机方差缩减梯度方法。 机器学习研究杂志，18:122:1– 122:43，2017年。7

Erich L Lehmann 和 George Casella。点估计理论。施普林格科学与商业媒体，2006年。7

Yunwen Lei 和 Yiming Ying。随机梯度下降稳定性与泛化的细粒度分析。载于第37届国际机器学习会议（ICML）论文集，第5809–5819页，线上会议，2020年。7，11，13，19，20

Mu Li、Tong Zhang、Yuqiang Chen 和 Alexander J. Smola。面向随机优化的高效小批量训练。载于第20届ACM SIGKDD知识发现与数据挖掘国际会议（KDD）论文集，第661–670页，纽约，纽约州，2014。7，11，24

C. McDiarmid。组合学综述，1989：有界差分法。1989。46Song Mei、Yu Bai 和 Andrea Montanari。非凸损失的经验风险景观。统计年刊，46(6A):2747–2774，2018年。7

Sayan Mukherjee、Partha Niyogi、Tomaso A. Poggio 和 Ryan M. Rifkin。学习理论：稳定性足以保证泛化，并且是经验风险最小化一致性的充要条件。*Adv. Comput. Math.*，25(1-3):161–193，2006。7

Sahand N Negahban、Pradeep Ravikumar、Martin J Wainwright、Bin Yu 等。带有可分解正则项的 m-估计量高维分析统一框架。*Statistical Science*，27(4)：538–557，2012。2

Arkadi Nemirovski、Anatoli B. Juditsky、Guanghui Lan 和 Alexander Shapiro。随机规划的稳健随机逼近方法。SIAM 优化杂志，19(4)：1574–1609，2009年。6

Arkaddii S Nemirovskii 和 Yu E Nesterov。光滑凸最小化的最优方法。苏联计算数学与数学物理，25(2):21–30，1985年。 12

AndreiPatrascu 和 Ion Necoara。约束凸优化中随机近端点方法的非渐近收敛。 机器学习研究杂志，18:198:1–198:42，2017年。 2，3，5，6，18，20

Alexander Rakhlin、Ohad Shamir 和 Karthik Sridharan。使梯度下降在强凸随机优化中达到最优。在第29届国际机器学习会议（ICML）论文集，爱丁堡，苏格兰，英国，2012。 4，6，18，20

Pradeep Ravikumar、John Lafferty、Han Liu 和 Larry Wasserman。稀疏加性模型。英国皇家统计学会期刊：B辑（统计方法学），71(5):1009–1030，2009年。 2

James Renegar 和 Benjamin Grimmer。一种简单且近乎最优的重启方案，用于加速一阶方法。计算数学基础，22(1):211–256，2022 年。 12

Herbert Robbins 和 Sutton Monro。一种随机逼近方法。 数学统计年刊，第 400–407 页，1951 年。 6

Herbert Robbins 和 David Siegmund。一个关于非负几乎超鞅的收敛定理及其若干应用。载于统计学中的优化方法，第 233–257 页。1971年。35

Mark Schmidt、Nicolas Le Roux 和 Francis R. Bach。凸优化中非精确近端梯度方法的收敛速率。载于 神经信息处理系统进展（NIPS），第 1458–1466 页，西班牙格拉纳达，2011 年。 40

Shai Shalev-Shwartz、Ohad Shamir、Nathan Srebro 和 Karthik Sridharan。可学习性、稳定性与一致收敛。 机器学习研究杂志，11:2635–2670，2010。 7， 11

Ohad Shamir、Nathan Srebro 和 Tong Zhang。使用近似牛顿型方法进行通信高效分布式优化。载于 第 31 届国际机器学习会议 (ICML) 论文集，第 1000–1008 页，中国北京，2014 年。 3,7

Nathan Srebro、Karthik Sridharan 和 Ambuj Tewari。光滑性、低噪声与快速收敛率。载于神经信息处理系统进展（NIPS），第 2199–2207 页，加拿大温哥华，2010。 7，8，10，11，13，19，20，27

Panos Toulis 和 Edoardo M Airoldi。基于随机梯度的估计量的渐近性质与有限样本性质。统计年刊，45(4):1694–1727，2017年。 5，6

Panos Toulis、Dustin Tran 和 Edoardo M. Airoldi。论随机梯度下降的稳定性与最优性。载于第19届国际人工智能与统计会议（AISTATS）论文集，第1290–1298页，西班牙加的斯，2016。 6,18

Alexandre B Tsybakov。 非参数估计导论。施普林格科学与商业媒体，2008年。 5,18

Sara A Van de Geer。高维广义线性模型与 Lasso。 统计年刊，36(2):614–645，2008年。 2

Vladimir Vapnik。统计学习理论综述。 IEEE 神经网络汇刊，10(5): 988–999，1999年。 7

Martin J. Wainwright。利用 l1约束二次规划（Lasso）实现高维含噪稀疏恢复的尖锐阈值。IEEE 信息论汇刊，55(5):2183–2202，2009年。21

Jialei Wang、Mladen Kolar、Nathan Srebro 和 Tong Zhang。利用稀疏性实现高效的分布式学习。载于第34届国际机器学习会议（ICML）论文集，第3636–3645页，澳大利亚悉尼，2017年a。 7, 24

Jialei Wang、Weiran Wang 和 Nathan Srebro。采用小批量近端法实现节省内存和通信的分布式随机优化。载于第30届学习理论会议（COLT）论文集，第1882–1919页，荷兰阿姆斯特丹，2017b。 2, 3, 4, 5, 7, 8, 10, 11,13, 14, 19,20, 26

Blake E. Woodworth 和 Nathan Srebro。一种更优的随机优化算法：小批量处理与插值学习。载于神经信息处理系统进展（NeurIPS），第7333–7345页，线上举行，2021年。 4,19, 20

Lin Xiao 和 Tong Zhang。采用渐进式方差缩减的近端随机梯度法。SIAM 优化杂志，24(4):2057–2075，2014。3，7，14

Yang You、Jing Li、Sashank J. Reddi、Jonathan Hseu、Sanjiv Kumar、Srinadh Bhojanapalli、Xiao-dan Song、James Demmel、Kurt Keutzer 和 Cho-Jui Hsieh。深度学习的大批量优化：76 分钟训练 BERT。载于第八届国际学习表征会议（ICLR）论文集，埃塞俄比亚亚的斯亚贝巴，2020年。11

Xiao-Tong Yuan 和 Ping Li. 分布式近似牛顿方法的收敛：全局化、更精确的界及其他。 机器学习研究杂志，21:206:1–206:51，2020年。 3

Xiao-Tong Yuan 和 Ping Li. 迭代硬阈值的稳定性与风险界。 IEEE 信息论汇刊，68(10):6663–6681，2022年。 8

Lijun Zhang、Tianbao Yang 和 Rong Jin。随机凸优化的经验风险最小化： $o(1/n)$-型和 $o(1/n^2)$-型风险界。载于 学习理论会议，第 1954–1979 页，2017 年。7，19，20

Tong Zhang。核方法的留一界。神经计算，15(6)：1397–1437，2003 年。7

Tong Zhang。使用随机梯度下降算法解决大规模线性预测问题。载于第二十一届国际机器学习会议（ICML）论文集，加拿大艾伯塔省班夫，2004 年。6

Yuchen Zhang 和 Xiao Lin。DiSCO：自协调经验损失的分布式优化。载于第 32 届国际机器学习会议（ICML）论文集，第 362–370 页，法国里尔，2015 年。3，7

Kaiwen Zhou、Lai Tian、Anthony Man-Cho So 和 James Cheng。寻找凸有限和近似驻点的实用方案。载于国际人工智能与统计会议（AISTATS）论文集，第 3684–3708 页，线上会议，2022 年。12

Pan Zhou、Xiaotong Yuan、Huan Xu、Shuicheng Yan 和 Jiashi Feng。通过小批量近端更新实现高效元学习。载于神经信息处理系统进展（2019年），第 1532–1542 页，温哥华，加拿大，2019年。 2