arXiv:2302.02607v2 [cs.LG] 8 Jun 2023

基于目标的随机优化代理函数

Jonathan Wilder La vington [*1] **Sharan Vaswani** [*2] **Reza Babanezhad** [3] **Mark Schmidt** [1 4] **Nicolas Le Roux** [5 6]

摘要

我们考虑最小化那些计算（可能是随机）梯度代价高昂的函数。这类函数在强化学习、模仿学习和对抗训练中十分常见。我们的目标优化框架利用（代价高昂的）梯度计算，在目标空间（例如，线性模型用于分类时输出的逻辑值）中构造可高效最小化的代理函数。这样便能对模型进行多次参数更新，从而摊销梯度计算的成本。在全批量设置下，我们证明代理函数是损失的全局上界，并且可以使用黑箱优化算法对其进行（局部）最小化。我们证明，由此得到的上界最小化算法能够确保收敛到损失的驻点。接下来，我们将框架扩展到随机设置，并提出 SSO 算法；该算法可视为在目标空间中进行的投影随机梯度下降。这一联系使我们能够证明 SSO 在最小化凸函数时的理论保证。我们的框架允许使用标准随机优化算法来构造代理函数，再通过任意确定性优化方法将其最小化。为评估该框架，我们考察了一系列监督学习和模仿学习问题。实验结果表明，目标优化具有诸多优势，且 SSO 效果显著。

[*]共同贡献 [1]英属哥伦比亚大学 [2]西蒙菲莎大学 [3]三星SAIT人工智能实验室，蒙特利尔 [4] 加拿大CIFAR人工智能讲席（Amii） [5] 微软研究院 [6] 加拿大CIFAR人工智能讲席（MILA）。通讯作者：Jonathan WilderLavington <wilderlavington@gmail.com>、Sharan Vaswani<vaswani.sharan@gmail.com>。

论文集收录于 40^{th} 国际机器学习会议，美国夏威夷檀香山。PMLR 202, 2023年。版权所有 © 2023年，作者所有。

1. 引言

随机梯度下降（SGD）（Robbins 和 Monro，1951年）及其变体（Duchi等人，2011年；Kingma 和 Ba，2015年）是机器学习（ML）中无处不在的优化方法。对于监督学习，迭代一阶方法需要计算各个小批量样本上的梯度，而梯度计算的成本往往占这些算法总计算成本的很大部分。例如，在强化学习（RL）（Williams，1992；Sutton 等人，2000）或在线模仿学习（IL）（Ross等人，2011）中，策略优化需要通过与真实或模拟环境进行可能代价高昂的交互来收集数据。

我们关注这样一类算法：它们通过昂贵的梯度预言机构造一系列代理函数。通常，这些代理函数被选为底层函数的全局上界，因此，最小化代理函数便可迭代地最小化原始函数。从算法角度看，这些代理函数可以高效地最小化，而无需额外调用梯度预言机，这使得该技术在所关注的应用中具有优势。逐步构造并最小化代理函数的技术通常称为主化最小化，期望最大化（EM）算法（Dempster 等人，1977年）就是一个例子。在强化学习中，常见算法（Schulman 等人，2015年；2017年）也依赖于最小化代理函数。

通常，代理函数是利用底层函数的凸性和/或光滑性构造的。这类代理函数已应用于随机设定（Mairal，2013年；2015年）。例如，近端线性算法（Drusvyatskiy，2017年）利用损失函数的复合结构，在参数空间中构造代理函数。与这些已有工作不同，我们在精心选择的目标空间而非参数空间上构造代理函数，利用复合结构来构建更好的代理函数，这种结构在机器学习（ML）中十分常见。例如，在监督学习中，典型的损失函数形式为 $h(\theta) = \ell(f(\theta))$，其中 ℓ 是（通常为）凸损失（例如，回归中的平方损失或分类中的逻辑损失），而 f 对应于输入的一种变换（例如线性变换，或如神经网络中的高维非凸变换）

（例如神经网络的情形）。类似地，在模仿学习（IL）中，ℓ 衡量正在学习的策略与真实专家策略之间的差异，而 f 则对应于正在学习的策略的一种特定参数化。更正式地说，若 Θ 是参数的可行集，$f\colon \Theta \to \mathcal{Z}$ 是从参数空间 $\Theta \subseteq \mathbb{R}^d$ 映射到目标空间 $\mathcal{Z} \subseteq \mathbb{R}^p$ 的一个可能非凸的映射，而$\ell\colon \mathcal{Z} \to \mathbb{R}$ 是一个凸损失函数。例如，对于线性回归，$h(\theta) = \frac{1}{2}\|X\theta - y\|_2^2$，$z = f(\theta) = X\theta$，且$\ell(z) = \frac{1}{2}\|z - y\|_2^2$。在我们关注的应用中，计算 $\nabla_z \ell(z)$需要调用成本高昂的梯度预言机，但 $\nabla_\theta f(\theta)$可以高效计算。Nguyen 等人（2022）利用这种复合结构证明全局收敛，而我们将利用它在目标空间中构造代理函数。Johnson 和 Zhang（2020）也使用目标空间构造代理函数，但每次模型更新都需要调用（随机）梯度预言机，因此其提出的算法在我们的设定中效率低下。此外，与我们的工作不同，Johnson 和 Zhang（2020）没有针对随机设定的理论保证。与我们同期，Woodworth 等人（2023）也考虑最小化计算成本高昂的函数，但他们通过设计与原函数相似的“代理”损失函数来实现。我们的贡献如下。

目标平滑性代理函数：在第 3 节中，我们利用 ℓ 相对于 z 的光滑性来定义目标平滑性代理函数，并证明它是底层函数 h 的全局上界。具体而言，这些代理函数利用原函数更紧的上界构造而成。这确保了在重新计算代价高昂的梯度之前，可以通过多次模型更新进一步向极小化点推进。

确定性情形下的目标优化：在第 3 节中，我们设计了一种主化最小化算法：迭代地构造目标平滑性代理函数，然后使用任意黑盒算法对其进行（局部）最小化。尽管构造目标平滑性代理函数需要访问代价高昂的梯度预言机，但对其进行最小化无需额外的预言机调用，因此可以对模型进行多次计算高效的更新。我们将这一框架称为目标优化。近年来，在为强化学习设计高效的离策略算法的背景下，已有研究探索了在目标空间中构造代理函数这一思想（Vaswani 等人，2021）。然而，与我们的工作不同，Vaswani 等人 (2021) 未考虑随机设定，也未提供理论收敛保证。在算法 1 中，我们具体化了目标优化框架，并证明其以次线性速率收敛到驻点（附录 C.1 中的引理 C）。

随机目标平滑性代理：在第4节中，我们考虑这样一种情形：我们可以访问一个计算开销较大的随机梯度预言机，它返回真实梯度的含噪但无偏的估计。与确定性情形类似，我们访问梯度预言机，以构造随机目标平滑性代理函数。尽管该代理函数是通过目标空间中的随机梯度构造的，但相对于参数而言，它是一个确定性函数，因此可以使用任何标准优化算法对其进行最小化。通过这种方式，我们的框架将 $\nabla_z \ell(z)$ 中的随机性（位于目标空间）与 f 中潜在的非凸性（位于参数空间）区分开来。

随机设定下的目标优化：与确定性情形类似，我们使用 m 步梯度下降来最小化随机目标平滑性代理，并将所得算法称为随机代理优化（SSO）。随后，我们将 SSO 解释为目标空间中的非精确投影随机梯度下降。这一对SSO 的解释使我们能够使用标准随机优化算法来构造代理函数，而这些代理函数可以通过任意确定性优化方法进行最小化。在目标空间中最小化代理函数也具有优势，因为这使我们能够选择约束更新幅度的空间。具体而言，对于深度神经网络等过参数化模型，参数空间中的更新与目标空间中的更新之间仅存在松散的联系。为了直接约束目标空间中的更新，诸如自然梯度（Amari，1998；Kakade，2001）等方法需要执行计算开销较大的操作。相比之下，SSO 能够直接控制目标空间中的更新，同时也可以高效实现。

SSO的理论结果：假设 $h(\theta)$ 是光滑且强凸的，我们在第4.1节证明，SSO（在目标空间中采用恒定步长）会线性收敛到真实极小化点的一个邻域（定理 4.2）。在命题 4.3 中，我们证明，该邻域的大小取决于目标空间中随机梯度的噪声（σ_z^2），以及误差项 ζ^2；该误差项取决于最优解处随机梯度的不相似性。在命题 4.4 中，我们给出一个二次示例，说明一般情况下这一误差项不可或缺。然而，随着小批量大小的增加，或者模型过度参数化到足以插值数据（Schmidt 和 Le Roux，2013年；Vaswani 等人，2019a），ζ_t^2 会变小。在恰好满足插值条件的特殊情况下，我们证明，采用 O 次迭代（$\log(1/\epsilon)$）的 SSO 足以保证收敛到真实极小化点的一个 ϵ 邻域。最后，我们指出，对于代价高昂的梯度预言机和常见损失函数而言，SSO 可能比采用参数化随机梯度下降更高效（第4.2节）。

实验评估：为评估我们的目标优化框架，我们将在线模仿学习（OIL）作为主要示例。在 OIL 中进行策略优化时，

计算 $\nabla_z \ell$ 需要通过与计算成本高昂的模拟环境进行交互来收集数据。使用 Mujoco 基准套件（Todorov 等人，2012年），我们证明 SSO 能够带来更优的实证表现（第 5 节）。随后，我们考察标准监督学习问题，将采用不同目标替代项的SSO 与标准优化方法进行比较。这些实证结果表明，目标优化和SSO具有实际优势。

2. 问题表述

我们关注的是最小化具有复合结构且梯度计算成本高昂的函数。形式上，我们的目标是求解以下问题：$\min_{\theta\in\Theta} h(\theta) := \ell(f(\theta))$，其中 $\Theta \subseteq \mathbb{R}^d$, $\mathcal{Z} \subseteq \mathbb{R}^p$, $f{:}\Theta \to \mathcal{Z}$，且 $\ell: \mathcal{Z} \to \mathbb{R}$。在本文中，我们假设 h 关于参数 θ 是 L_θ-光滑的，并且$\ell(z)$关于目标 z是 L-光滑的。对于所有广义线性模型，包括线性回归和逻辑回归，$f = X^\top\theta$是关于 θ 的线性映射，而 ℓ关于 z是凸的。

示例：对于特征为$X \in \mathbb{R}^{n\times d}$ 、标签为 $y \in \{-1,+1\}^n$的逻辑回归，$h(\theta) = \sum_{i=1}^n \log(1+\exp(-y_i\langle X_i,\theta\rangle))$。如果我们将逻辑值[1]，作为“目标”，那么 $\mathcal{Z} = \{z|z = X\theta\} \subseteq \mathbb{R}^n$ 和 $\ell(z) = \sum_{i=1}^n \log(1+\exp(-y_i z_i))$。在这种情况下，$L_\theta$ 是 $X^\top X$的最大特征值，而 $L = \frac{1}{4}$。线性回归也有类似的例子。

在使用神经网络的场景中，函数 f（将 X 映射到 y）通常是非凸的，而损失 ℓ 则是凸的。例如，在 OIL 中，目标空间是参数化策略的空间，其中 ℓ 是使用策略分布 $\pi := f(\theta)$ 时的累积损失，该分布的密度由 θ参数化。尽管我们的算法框架能够处理非凸的 ℓ 和 f，但根据具体场景，我们的理论结果会假设 ℓ（或 h）关于 θ 是（强）凸的，且 f 是仿射映射。

对于我们关注的应用，计算 $\nabla_z \ell(z)$的计算成本很高，而 $f(\theta)$（及其梯度）则可以高效地计算。例如，在 OIL 中，计算策略的累积损失 ℓ（以及相应的梯度 $\nabla_z \ell(z)$）需要在环境中对其进行评估。由于这一操作涉及与环境或模拟器的交互，因此计算成本很高。另一方面，计算 $\nabla_\theta f(\theta)$的成本仅取决于策略参数化，不需要与环境进行额外交互。在某些情况下，更自然的做法是使用随机梯度预言机，该预言机返回一个带噪声、无偏的梯度 $\nabla\ell(z)$

[1]目标空间并非唯一。例如，我们可以直接以逻辑回归的分类概率为目标，从而得到不同的 f 和 ℓ。

~使得 $\mathbb{E}[\nabla\ell(z)] = \nabla\ell(z)$。我们在第4节考察随机性的影响。

如果不利用复合结构，也不显式考虑梯度预言机的成本，就可以直接使用梯度下降或随机梯度下降等迭代一阶方法来最小化 $h(\theta)$。在第 $t \in [T]$次迭代中，参数化梯度下降的更新为：$\theta_{t+1} = \theta_t - \eta\nabla h_t(\theta_t)$，其中 η 是根据 h的性质选取或调整的步长。由于 h 是 L_θ-光滑的，参数化梯度下降的每次迭代都可视为针对参数，精确地最小化由光滑性条件导出的二次代理函数。具体而言，$\theta_{t+1} :=\arg\min g_t^p(\theta)$，其中 g_t^p 是参数化光滑性代理函数：$g_t^p(\theta) := h(\theta_t) + \langle\nabla h(\theta_t)\ \theta-\theta_t\rangle + \frac{1}{2\eta}\|\theta-\theta_t\|_2^2$。

二次代理函数在 θ_t 处是紧的，即 $h(\theta_t) = g_t^p(\theta_t)$，并且随着我们远离 θ_t 而变得更松。对于 $\eta \le \frac{1}{L_\theta}$，该代理函数是 h 的全局（对所有 θ）上界，即 $g_t^p(\theta) \ge h(\theta)$。最小化全局上界会使 h 下降，因为 $h(\theta_{t+1}) \le g_t^p(\theta_{t+1}) \le g_t^p(\theta_t) = h_t(\theta_t)$。类似地，参数化随机梯度下降更新包括访问随机梯度预言机，以获得 $(\tilde{h}(\theta)\ \nabla\tilde{h}(\theta))$，使得 $\mathbb{E}[\tilde{h}(\theta)] = h(\theta)$~ 且 $\mathbb{E}[\nabla h(\theta)] = \nabla h(\theta)$，并迭代构造随机参数化光滑性代理函数 ~$g^p{}_t(\theta)$。具体而言，$\theta_{t+1} = \arg\min$ ~$g^p{}_t(\theta)$，其中 ~$g^p{}_t(\theta) := \tilde{h}(\theta_t) + \langle\nabla\tilde{h}(\theta_t)\ \theta-\theta_t\rangle + \frac{1}{2\eta_t}\|\theta-\theta_t\|_2^2$。此处，$\eta_t$ 是依赖迭代的步长，其衰减方式取决于 h 的性质（Robbins 和 Monro，1951年）。与这些方法不同，在下一节中，我们将利用损失关于目标空间的光滑性，并提出一种确定性情形下的主化-最小化算法。

3. 确定性情形

我们考虑在确定性情形下最小化 $\ell(f)$，此时我们可以精确计算梯度 $\nabla_z\ell(z)$。与第2节中的参数化情形类似，我们利用 $\ell(z)$ 关于目标空间的光滑性，并定义围绕 z_t 的目标光滑代理函数如下：$\ell(z_t) + \langle\nabla_z\ell(z_t)\ z-z_t\rangle + \frac{1}{2\eta}\|z-z_t\|_2^2$，其中 η 是目标空间中的步长，其值将通过理论分析确定。由于 $\overline{z} = f(\theta)$，该代理函数可以表示为关于 θ 的函数：$g_t(\theta) := \left[\ell(z_t) + \langle\nabla_z\ell(z_t)\ f(\theta) - z_t\rangle + \frac{1}{2\eta}\|f(\theta) - z_t\|_2^2\right]$，一般而言，它并非关于 θ 的二次函数。

Example: For linear regression, $f(\theta) = X^\top\theta$ and $g_t(\theta) = \frac{1}{2}\|X\theta_t - y\|_2^2 + \langle[X\theta_t - y], X(\theta-\theta_t)\rangle + \frac{1}{2\eta}\|X(\theta-\theta_t)\|_2^2$.

与参数化光滑性代理函数类似，我们看到$h(\theta_t) = \ell(f(\theta_t)) = g_t(\theta_t)$。如果 ℓ 在目标空间 $\mathcal{Z}$中是 L-光滑的，那么对于 $\eta \le \frac{1}{L}$，我们有 $g_t(\theta) \ge \ell(f(\theta)) = h(\theta)$，对所有 θ均成立，也就是说，该代理函数是 h上的全局上界。由于 g_t 是 h的全局上界，与梯度下降类似，我们可以

通过在每次迭代中最小化代理函数来最小化 h，即$\theta_{t+1} = \operatorname{argmin}_\theta g_t(\theta)$。然而，与梯度下降不同的是，一般而言，$g_t$的极小化点不存在闭式解，因此我们将考虑对其进行近似最小化。

算法 1（随机）替代函数优化

输入：θ_0（初始化），T（迭代次数），m_t（内循环次数），η（目标空间的步长），α（参数空间的步长）

对于 $t = 0$ 至 $T-1$，**执行**

调用（随机）梯度预言机以构造 $\tilde{g}_t(\theta)$

初始化内循环：$\omega_0 = \theta_t$

对 $k \leftarrow 0$ 至 m_{t-1} **do**

$\omega_{k+1} = \omega_k - \alpha \nabla_\omega \tilde{g}_t(\omega_k)$

结束循环

$\theta_{t+1} = \omega_m \quad ; \quad z_{t+1} = f(\theta_{t+1})$

结束循环

返回 θ_T

由此得到以下元算法：对于每个 $t \in [T]$，在迭代点 θ_t处，构造代理函数 g_t，并通过（近似）最小化 $g_t(\theta)$来计算 θ_{t+1}。该元算法允许在每次迭代中使用任意黑盒算法来最小化代理函数。在算法 1中，我们通过使用 $g_t(\theta)$ 步梯度下降来最小化 $m \geq 1$，从而实例化该元算法。对于 $m = 1$，算法1 得到以下更新：$\theta_{t+1} = \theta_t - \alpha \nabla g_t(\theta_t) = \theta_t - \alpha \nabla h(\theta_t)$，因此等价于步长为 α的参数化梯度下降。

示例：对于线性回归，使用 1号算法并令 $m = 1$，可以得到最小二乘目标上的参数化梯度下降。另一方面，精确最小化代理函数（对应于 $m = \infty$）来计算 θ_{t+1}，得到以下更新：$\theta_{t+1} = \theta_t - \eta (X^\intercal X)^{-1} [X^\intercal (X\theta_t - y)]$，并得到参数空间中的牛顿更新。在这种情况下，使用 g_t 步梯度下降来近似最小化 $m \in (1, \infty)$可在参数空间中实现一阶方法与二阶方法之间的插值。在这种情况下，我们的框架类似于拟牛顿方法（Nocedal 和 Wright, 1999），这类方法试图在不显式建模海森矩阵的情况下对损失的曲率进行建模。与这些方法不同，我们的框架不会带来额外的内存开销。

在附录C.1 中的附录C中，我们证明了算法 1对于任意取值的 $m \geq 1$ 以及对 α 和η 作出的适当选择，都会使 $O(1/T)$ 收敛到 h 的一个驻点。重要的是，该结果仅依赖于 ℓ和 g_t^z的光滑性，无需 $\ell(z)$ 或 $f(\theta)$ 为凸。因此，该结果适用于使用非凸模型（如深度神经网络）的情形，也适用于具有非凸损失函数的问题，例如强化学习中的问题。在下一节中，我们将此框架扩展到随机设定，此时我们只能获得带噪声（但无偏）的梯度估计。

4. 随机设定

在随机设定下，我们使用来自梯度预言机的带噪但无偏的$\tilde{\ell}(z)$ $\tilde{\nabla}\ell(z)$估计值（$\ell(z)$ $\nabla\ell(z)$），来构造随机目标代理。为简化理论分析，我们将聚焦于一种特殊情况：ℓ 是有限损失和，即 $\ell(z) = \frac{1}{n}\sum_{i=1}^n \ell_i(z)$。在这种情况下，在迭代 t时查询随机梯度预言机，会返回与损失索引 i_t 对应的单个损失及其梯度，即 $\left(\tilde{\ell}(z)\ \tilde{\nabla}\ell(z)\right) = (\ell_{i_t}(z)\ \nabla\ell_{i_t}(z))$。这种结构存在于我们关注的应用场景中，例如，使用包含 n 个训练样本的数据集进行监督学习，或在在线模仿学习中使用迭代 t时的策略收集多条轨迹。在此设定下，确定性代理函数为 $g_t(\theta) := \frac{1}{n}\left[\sum [\ell_i(z) + \langle \nabla\ell_i(z_t),\ f(\theta) - z_t\rangle] + \frac{1}{2\eta_t}\|f(\theta) - z_t\|_2^2\right]$。

为了实现随机替代函数及由此得到的算法，我们仅考虑关于目标空间可分离的损失函数，即对于 $z \in \mathcal{Z}$，如果 $z^i \in \mathbb{R}$ 表示 z 的坐标 i，那么 $\ell(z) = \frac{1}{n}\sum_i \ell_i(z^i)$。例如，在所有监督学习问题的损失函数中都存在这种结构，其中 $\mathcal{Z} \subseteq \mathbb{R}^n$ 和 $z^i = f_i(\theta) := f(X_i, \theta)$。在这种情况下，对于所有 i 和 $j \neq i$，$\frac{\partial \ell_i}{\partial z^j} = 0$，并将随机目标代理定义为 $\tilde{g}_t(\theta) := \ell_{i_t}(z_t) + \frac{\partial \ell_{i_t}}{\partial z^{i_t}}\Big|_{z_t}\left[f_{i_t}(\theta) - z_t^{i_t}\right] + \frac{1}{2\eta_t}\left[f_{i_t}(\theta) - z_t^{i_t}\right]^2$，其中 η_t 是迭代 t 时目标空间中的步长。注意，$\tilde{g}_t(\theta)$ 仅取决于 i_t，并且只需要访问 $\partial\ell_{i_t}(z_t)$（这意味着可以高效地构造它）。关于随机替代函数，我们有两点观察：(i) 不同于使用 i_t 构造随机梯度的参数化随机平滑代理 $\tilde{g}^p$，$\tilde{g}_t(\theta)$ 对随机梯度 $\partial\ell_{i_t}(z_t)$ 和正则化项$\left[f_{i_t}(\theta) - z_t^{i_t}\right]^2$ 都使用 i_t（同一个随机样本）；(ii) 虽然 $\mathbb{E}_{i_t}[\tilde{g}_t(\theta)] = g_t(\theta)$，$\mathbb{E}_{i_t}[\arg\min \tilde{g}_t(\theta)] \neq \arg\min g_t(\theta)$，但在参数化情形下，$\mathbb{E}[\arg\min \tilde{g}_t^p(\theta)] = \arg\min g_t^p(\theta)$。

示例：对于线性回归，$z^i = f_i(\theta) = X_i\theta$ 和$\ell_i(z^i) = \frac{1}{2}(z^i - y_i)^2$。在这种情况下，随机目标代理等于 $\tilde{g}_t(\theta) = \frac{1}{2}(X_{i_t}\theta_t - y_{i_t})^2 + [X_{i_t}\theta_t - y_{i_t}] \cdot X_{i_t}(\theta - \theta_t) + \frac{1}{2\eta_t}\|X_{i_t}(\theta - \theta_t)\|_2^2$。

与确定性情形类似，下一个迭代点可以通过（近似）最小化 $\tilde{g}_t(\theta)$来获得。在算法上，我们可以在迭代 t 时构造代理函数 $\tilde{g}_t$，并使用任意黑盒算法对其进行近似最小化。我们将由此得到的框架称为随机代理优化(SSO)。例如，我们可以使用 m 步梯度下降来最小化 $\tilde{g}_t$。由此得到的算法与算法 1 相同，但使用 $\tilde{g}_t$（算法中的改动以绿色标出）。注意，代理函数取决于

随机抽取的 i_t，因此是随机的。不过，一旦构造出代理函数，就可以使用任意确定性算法将其最小化；也就是说，算法1的内循环中不会再引入额外的随机性。此外，对于 $m = 1$这一特殊情形，算法1的更新与参数化随机梯度下降相同。Newton 等人（Newtonetal. (2021）之前的工作，例如回顾式优化框架，也考虑在同一批样本上进行多次更新。然而，与算法1要求目标空间中相邻迭代点彼此接近不同，Newton 等人(2021）在每次迭代中使用特定的停止准则，并采用逐渐增大的批量大小。

为了证明 SSO的理论保证，我们将其解释为目标空间中的投影随机梯度下降。具体而言，我们在附录中证明以下等价关系D.1。

引理 4.1。 以下更新等价：

$$
\begin{aligned}
&(1)\quad \tilde{\theta}_{t+1} = \arg\min_{\theta} \tilde{g}_t(\theta); && \text{(SSO)}\\
&\qquad \tilde{z}^{(1)}_{t+1} = f(\tilde{\theta}_{t+1})\\
&(2)\quad z_{t+1/2} = z_t - \eta_t \nabla_z \ell_{i_t}(z_t); && \text{(Target-space SGD)}\\
&\qquad \tilde{z}^{(2)}_{t+1} = \arg\min_{z\in\mathcal{Z}} \frac{1}{2}\left\|z_{t+1/2} - z\right\|^2_{\mathcal{P}_t}
\end{aligned}
$$

其中， $\mathcal{P}_t \in \mathbb{R}^{p\times p}$ 是一个随机对角矩阵，使得$\mathcal{P}_t(i_t, i_t) = 1$ 且 $\mathcal{P}_t(j, j) = 0$对所有 $j \neq i_t$均成立。也就是说，*SSO (1)* 和目标空间随机梯度下降 *(2)* 在每一步得到相同的迭代点；即如果 $z_t = f(\theta_t)$，那么 $\tilde{z}_{t+1} := \tilde{z}^{(1)}_{t+1} = \tilde{z}^{(2)}_{t+1}$。

目标空间随机梯度下降的第二步对应于将 i_t（使用随机采样的索引）投影到 $\mathcal{Z}$[2]上。

利用这一等价关系，我们可以将对 $\tilde{g}_t$ 的非精确最小化（例如，在算法 1 中使用 m 步梯度下降）理解为向 $\mathcal{Z}$ 进行非精确投影，这将有助于证明 SSO 的收敛保证。上述解释还使我们能够借鉴现有的随机梯度下降文献（Robbins 和 Monro，1951年；Li 等人，2021年；Vaswani 等，2019b），在目标空间中指定步长序列 $\{\eta_t\}_{t=1}^T$，从而完成随机替代函数的实例化。此外，在目标空间中采用其他随机优化算法，例如正则化领导者跟随算法（Abernethy 等人，2009年），以及自适应梯度方法，如 Adagrad（Duchi等人，2011年）、在线牛顿法（Hazan 等人，2007年）和随机镜像下降（Bubeck 等人，2015年，第 6 章），会得到不同的随机替代函数（参见附录 B），随后可使用黑盒确定性算法对其进行优化。接下来，我们证明当 ℓ（z）是光滑强凸函数时，SSO 的收敛保证。

[2]对于线性参数化， $z^{i_t} = \langle X_{i_t}, \theta\rangle$ 且集合 $\mathcal{Z}$ 是凸的。对于非凸的 f，集合 $\mathcal{Z}$ 可能是非凸的，因此投影未必定义良好。不过，仍然可以最小化 $\tilde{g}_t$，但无法保证其收敛。

4.1. 理论结果

对于 $\ell(z)$ 是光滑且强凸函数的情形，我们首先分析目标空间中非精确投影随机梯度下降的收敛性（第4.1.1节），然后在第4.1.2节中界定投影误差。

4.1.1. 收敛性分析

在进行理论分析时，我们假设选择 f 能确保投影定义良好（例如，对于满足 $f = X^\intercal\theta$ 的线性参数化）。我们定义 $\bar{z}_{t+1} := f(\theta_{t+1})$，其中 $\theta_{t+1} := \arg\min \tilde{q}_t(\theta)$，且 $\tilde{q}_t := \ell_{i_t}(z_t) + \frac{\partial \ell_{i_t}}{\partial z^{i_t}_t}\left[f_{i_t}(\theta) - z^{i_t}_t\right] + \frac{1}{2\eta'_t}\|f(\theta) - z_t\|^2_2$。为了确保 $\mathbb{E}[\tilde{q}_t(\theta)] = g_t(\theta)$，我们将设定 $\eta'_t = \eta_t n$。注意， $\tilde{q}_t$ 与 $\tilde{g}_t$ 相似，但正则化项中没有随机性。与 $\tilde{z}_{t+1}$ 类似， $\bar{z}_{t+1}$ 可解释为目标空间中投影后的（投影相对于 ℓ_2-范数）随机梯度下降的结果（引理 D.1）。注意， $\tilde{q}_t$ 仅在 SSO 的分析中定义。

对于理论分析，定义$\epsilon_{t+1} := \|z_{t+1} - \bar{z}_{t+1}\|_2$ 为第 t 次迭代中的投影误差。其中， $z_{t+1} = f(\theta_{t+1})$ 其中， θ_{t+1} 通过（近似）最小化 $\tilde{g}_t$得到。请注意，投影误差同时包含随机投影（取决于 i_t）以及对$\tilde{g}_t$进行非精确最小化所产生的影响，并将在第 4.1.2 节中对其进行界定。在以下引理中（证明见附录 D.2），我们利用引理 4.1中的等价关系，推导出针对 η_t的两种选择的以下保证：(a) 常数步长；(b) 指数步长（Li 等人，2021；Vaswani 等人，2022年）。

定理 4.2。 假设 *(i)* ℓ_{i_t} 是 L-光滑且凸的，*(ii)* ℓ 是 μ-强凸的，*(iii)* $z^* = \operatorname{argmin}_{z\in\mathcal{Z}} \ell(z)$ *(iv)* 并且对于所有 t、$\epsilon_t \le \epsilon$、 T 次迭代，*SSO* 均可对 $z_T = f(\theta_T)$得到以下界，

$$
\begin{aligned}
\mathbb{E}\left\|z_{T+1} - z^*\right\| \le \Bigg(&\left(\prod_{i=1}^T \rho_i\right)\|z_1 - z^*\|_2^2\\
&+ 2\sigma^2 \sum_{t=1}^T \prod_{i=t+1}^T \rho_i \eta'^2_t\Bigg)^{1/2}\\
&+ 2\epsilon \sum_{t=1}^T \prod_{i=t+1}^T \rho_i\,,
\end{aligned}
$$

其中 $\rho_t = (1 - \mu\eta'_t)$，$\sigma^2 := \mathbb{E}\left[\|\nabla\ell(z^*) - \nabla\ell_t(z^*)\|_2^2\right]$，

(a) 恒定步长：当 $\eta_t = \frac{1}{2Ln}$ 且对于 $\rho = 1 - \frac{\mu}{2L}$，

$$
\begin{aligned}
\mathbb{E}\left\|z_{T+1} - z^*\right\| \le \|z_1 - z^*\|\left(1 - \frac{1}{2\kappa}\right)^{\frac{T}{2}}\\
+ \frac{\sigma}{\sqrt{\mu L}} + \frac{2\epsilon}{1 - \sqrt{1 - \frac{1}{2\kappa}}}\,.
\end{aligned}
$$

(b) 指数步长：当 $\eta_t = \frac{1}{2Ln}\alpha^t$ 对于 $\alpha =$

$(\frac{\beta}{T})^{\frac{1}{4}}$,

$$\mathbb{E}\,\|z_{T+1}-z^*\| \le c_1\,\exp\left(-\frac{T}{4\kappa}\frac{\alpha}{\ln(T/\beta)}\right)\|z_1-z^*\| + \frac{4\kappa c_1(\ln(T/\beta))}{Le\alpha\sqrt{T}}\sigma + 2\epsilon\,c_2\,.$$

where $c_1=\exp\left(\frac{1}{4\kappa}\,\frac{2\beta}{\ln(T/\beta)}\right)$, *and* $c_2=\exp\left(\frac{\beta\ln(T)}{2\kappa\ln(T/\beta)}\right)$

在上述结果中，σ^2 是无约束情形下噪声的自然对应物。在无约束情形下，$\nabla\ell(z^*)=0$，因此我们恢复了随机梯度下降分析中使用的标准噪声概念（Bottou 等人，2018；Gower 等人，2019）。与参数化随机梯度下降不同，σ^2 和 κ 都不依赖于特定的模型参数化，而只取决于目标空间中 ℓ 的性质。对于恒定步长和指数步长，上述引理将 Bottou 等人（2018）和 Li 等人（2021）、Vaswani 等人（2022）的随机梯度下降证明推广到投影随机梯度下降，并且能够处理类似于 Schmidt 等人（2011）的近似投影误差。对于恒定步长，SSO 收敛到极小化点的一个邻域，该邻域取决于 σ^2 和投影误差 ϵ_t^2。对于指数步长，SSO 会产生噪声自适应（Vaswani 等人，2022）的 $O\big(\exp\left(\frac{-T}{\kappa}\right)+\frac{\sigma^2}{T}\big)$ 收敛到解的一个邻域，而该邻域仅取决于投影误差。下一节中，我们将界定投影误差。

4.1.2. 控制投影误差

为了完成SSO的理论分析，我们需要控制投影误差 $\epsilon_{t+1}=\|\bar{z}_{t+1}-z_{t+1}\|_2$，该误差可分解为以下两个部分：$\epsilon_{t+1}\le\|\bar{z}_{t+1}-\tilde{z}_{t+1}\|_2+\|\tilde{z}_{t+1}-z_{t+1}\|_2$，其中 $\tilde{z}_{t+1}=\tilde{f}(\theta_{t+1})$ 且 θ_{t+1} 是 $\tilde{g}_t$的极小化点。该分解的第一部分源于使用了依赖于 i_t 的随机投影（在 $\tilde{z}_{t+1}$ 的定义中），而非确定性投影（在 $\bar{z}_{t+1}$的定义中）。第二部分则源于对随机代理函数进行非精确最小化，可通过将 $\tilde{g}_t$ 最小化至所需容差来控制。为了界定 ϵ_{t+1}，我们将作出以下额外假设：f 关于 L_f是 θ-利普希茨连续的。[3] 以下命题（证明见附录D.3）给出了 ϵ_{t+1}的上界。

命题 4.3. 假设 *(i)* f 是 L_f-利普希茨连续，*(ii)* $\tilde{g}_t$ 是 μ_g 强凸且 L_g 光滑，满足 $\kappa_g:=L_g/\mu_g$[4]，*(iii)* $\tilde{q}_t$ 是 μ_q 强凸，*(iv)*$\zeta_t^2:=\frac{8}{\text{最小值}\{\mu_g,\mu_q\}}([\text{最小值}\{\mathbb{E}_{i_t}[\tilde{g}_t]\}-\mathbb{E}_{i_t}[\text{最小值}\{\tilde{g}_t\}]])$
$+\left(\frac{8}{\min\{\mu_g,\mu_q\}}\left[\min\{\mathbb{E}_{i_t}[\tilde{q}_t]\}-\mathbb{E}_{i_t}[\min\{\tilde{q}_t\}]\right]\right)$，*(v)* $\sigma_z^2:=\text{最小值}_{z\in\mathcal{Z}}\{\mathbb{E}_{i_t}[\ell_{i_t}(z)]\}-\mathbb{E}_{i_t}[\text{最小值}_{z\in\mathcal{Z}}\{\ell_{i_t}(z)\}]$，如果 $\tilde{g}_t$ 通过适当步长的梯度下降内循环进行最小化，那么，$\mathbb{E}[\epsilon_{t+1}^2]\le L_f^2\,\zeta_t^2+\frac{4L_f^2}{\mu_g}\left[\text{指数}\,(-m_t/\kappa_g)\left[\mathbb{E}[\ell(z_t)-\ell(z^*)]+\sigma_z^2\right]\right]$。

[3]例如，在线性参数化中，$L_f=\|X\|$ 某些神经网络也满足这一性质。[4]为简单起见，我们考虑强凸函数，并且这些结果应该可以推广到这样的情形：$\tilde{g}$ 是非凸的，但满足 PL 不等式，或者是弱凸的。

注意，ζ_t^2 和 σ_z^2 的定义都与 Vaswani 等人（2022）在随机梯度下降分析中使用的定义相似，可以将其解释为方差项。具体而言，利用强凸性，$\min\{\mathbb{E}_{i_t}[\tilde{g}_t]\}-\mathbb{E}_{i_t}[\min\{\tilde{g}_t\}]\le\frac{2}{\mu_g}\mathbb{E}[\|\nabla\tilde{g}_t(\theta'_{t+1})-\mathbb{E}[\nabla\tilde{g}_t(\theta'_{t+1})]\|_2^2]$，这就是 $\overline{\nabla}\tilde{g}_t(\theta'_{t+1})$中的方差。类似地，我们可以将 ζ_t^2定义中的另一项解释为方差。第一项 $L_f^2\,\zeta_t^2$是由随机投影与确定性投影之间的差异导致的，而第二项$4L_f^2/\mu_g\exp(-m_t/\kappa_g)\left[\mathbb{E}[\ell(z_t)-\ell(z^*)]+\sigma_z^2\right]$则是由于随机替代函数的非精确最小化而产生的。令 $m_t=O(\log(1/\epsilon))$可以将第二项降至 $O(\epsilon)$。在 SSO 的每次迭代中对示例进行小批量采样时，随着批量大小增加，ζ_t^2和 σ_z^2 都会减小（这是标准有放回采样界的结果（Lohr，2019）），并且在全批量时变为零。另一种值得关注的情形是使用能够插值数据的过参数化模型（Schmidt 和 Le Roux，2013；Ma 等人，2018；Vaswani 等人，2019a）。插值条件意味着随机梯度在最优解处变为零，且非参数回归（Liang 和 Rakhlin，2018；Belkin 等人，2019）和过参数化深度神经网络（Zhang 等人，2017）等模型满足这一条件。在此条件下，$\zeta_t^2=\sigma_z^2=0$（Vaswani 等人，2020；Loizou 等人，2021）。从算法角度来看，上述结果表明，在噪声占主导的情况下，对 SSO 使用较大的 m 可能不会带来显著改善。不过，在使用较大批量大小或过参数化模型时，使用较大的 m 可以获得更优的性能。

根据上述结果，自然会产生一个问题：在一般（非插值）情形下，对 ζ_t^2 的依赖是否不可避免。为说明这一点，我们构造了一个例子（详见附录D.4），并证明，即使目标是两个一维二次函数之和，且精确地最小化代理函数，即 $m=\infty$对于任意收敛的步长序列，SSO也会收敛到解的某个邻域。

命题*4.4.* 考虑使用 *SSO* 最小化两个一维二次函数之和 ~~$h(\theta):=h_1(\theta)+h_2(\theta)$~~ *2*，其中 $h_1(\theta):=1\,2(\theta-1)^2$ 和 $h_2(\theta):=\frac{1}{2}(2\theta+1/2)^2$,，采用 $m_t=\infty$，对于任意的 α_t 序列以及任意常数$c\in(0,1]$。*SSO* 将收敛到解的某个邻域。具体而言，若 θ^* 是 h 的极小化点，且$\theta_1>0$，则 $\mathbb{E}(\theta_T-\theta^*)\ge min\,(\theta_1,\frac{3}{8})$。

为了证明上述结果，我们利用这样一个事实：对于二次函数，SSO（使用 $m=\infty$）等价于子采样牛顿法；而在一维情况下，

我们可以还原 Vaswani 等人 (2022年中的例子。由于上述例子适用于 $m = \infty$，我们得出结论：这种偏差并非源于代理函数最小化不精确。此外，由于该例子适用于所有步长（包括 任意 递减步长），我们可以得出结论：优化误差并非随机性的副作用。为了避免这种偏差项，子采样牛顿方法会使用不同的批次来计算子采样梯度和海森矩阵，并采用递增的批量大小（Bollapragada 等人，2019年），或考虑使用过参数化模型（Meng 等人，2020年）。

Agarwal 等人(2020) 也证明了在同一批次上执行多步随机梯度下降时的理论保证。与我们的工作不同，他们的动机是分析数据回响的性能（Choi 等人，2019）。从技术角度来看，他们考虑了 (i) 参数空间中的更新，以及 (ii) 随着 m 增加，其内循环步长会减小。最后，我们指出，我们的框架及后续理论保证也适用于这种情形。

4.2. 目标优化的优势

为了直观了解目标优化可能带来的优势，我们考虑一个简单情形：每个h_i 都是 μ_θ-强凸、 L_θ-光滑的，并且 $\kappa_\theta = {}^{L_\theta}/_{\mu_\theta}$。为方便起见，我们定义 $\zeta^2 := \max_{t\in[T]} \zeta_t^2$。下文将说明，在某些条件下，对于病态最小二乘问题（其中 $\kappa_\theta >> 1$），目标优化相较于参数化随机梯度下降具有可证明的优势。

示例：对于最小二乘情形， κ_θ 是 $X^\intercal X$ 矩阵的条件数。为了达到 ϵ 次优性，在完全知晓 σ_θ^2 的情况下：$= \mathbb{E}\,\|\nabla h_i(\theta^*)\|_2^2$ 以及所有问题相关常数时，参数化随机梯度下降需要 $T_{\text{param}} = O\big(\max\big\{\kappa_\theta \log\big({}^1/_\epsilon\big), {}^{\sigma_\theta^2}/_{\mu^2\epsilon}\big\}\big)$ 次迭代（Gower 等人，2019，定理 3.1），其中第一项是偏差项，第二项是噪声的影响。为了使 SSO 达到 ϵ 次优性，我们要求$\zeta^2 \le \epsilon$ （例如，通过使用足够大的批次大小），并进行 $O(\kappa_\theta \log(1/\epsilon))$次内迭代。与参数化情形类似，SSO 所需的外迭代次数为 $T_{\text{target}} = O\big(\max\big\{\log\big({}^1/_\epsilon\big), {}^{\sigma^2}/_\epsilon\big\}\big)$，因为相对于目标的条件数等于 1。

为了对计算成本高昂的梯度预言机的影响建模，设计算 ℓ 的梯度的成本为 τ，计算代理函数的梯度的成本为 1。当梯度噪声较小，且偏差主导参数化优化和目标优化的迭代次数时，参数化随机梯度下降的成本主要由 $\tau T_{\text{param}} = O(\kappa_\theta\tau)$ 决定，而目标优化的成本为 $T_{\text{target}} \times [\tau + \kappa_\theta \log(1/\epsilon)] = O(\tau + \kappa_\theta)$。反之，当噪声占主导时，我们需要将参数化情形下的 $O\big(\tau\sigma_\theta^2/\mu^2\epsilon\big)$ 成本与 $O\big({}^{\sigma^2}/_\epsilon\,[\tau + \kappa_\theta \log(1/\epsilon)]\big)$ 目标优化成本进行比较。根据噪声的定义， $\sigma_\theta^2 = \mathbb{E}_i\,\|X_i^\intercal(X_i\theta^* - y_i)\|_2^2 \le L\sigma^2$。将 σ_θ^2 替换为 $L\sigma^2$ 后，我们再次看到，参数化随机梯度下降的复杂度取决于 $O(\kappa_\theta\tau)$，而 SSO 的复杂度取决于 $O(\kappa_\theta + \tau)$。

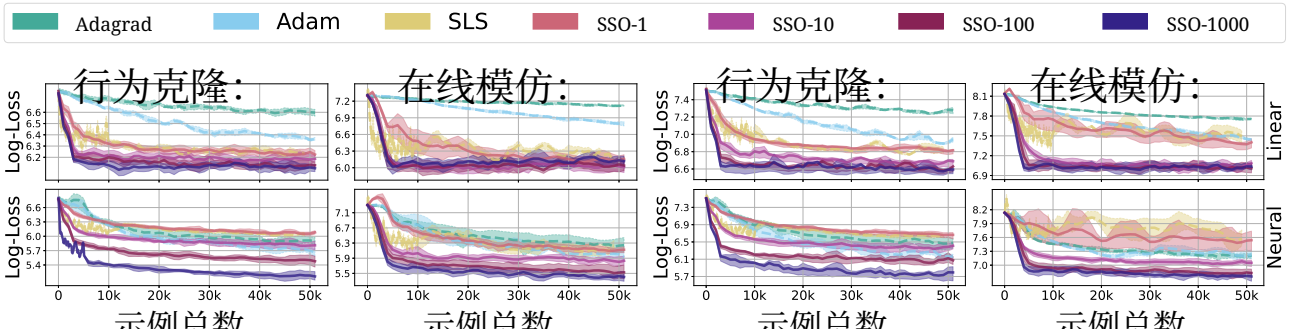
逻辑回归也可以证明类似的性质。同样，我们也可以使用其他随机优化算法，为目标优化构造代理函数。另请参阅附录 B 和附录 E 中的进一步讨论。

5. 实验评估

我们评估目标优化框架在在线模仿学习和监督学习中的表现[5]。在后续实验中，在理论上可确定步长时，我们使用理论选择的步长；否则使用 Paszke 等人（2019年）提供的默认步长。正文中的实验均不采用衰减调度，但在附录E中，我们同时考察标准 ${}^1/_{\sqrt{t}}$ 调度（Bubeck 等人，2015年）以及指数步长调度（Vaswani 等人，2022年；Orabona，2019）。对于 SSO，由于代理函数的优化是一个确定性问题，我们在所有实验中使用具有相同超参数的标准回溯 Armijo 线搜索（Armijo，1966）。对于每个实验，我们绘制平均损失随（随机）梯度预言机调用次数的变化曲线。均值和相关分位数均基于三个随机种子进行报告。

在线模仿学习：我们考虑这样一种设置，其中损失是通过与模拟器交互生成的。在这种设置下，行为策略通过观察模拟环境的状态，并在该状态下采取动作来收集样本。对于通过交互收集到的每个状态，专家策略都会提供其在该状态下会采取的动作。模仿学习的目标是生成一个模仿专家的策略。损失衡量学习策略 π 与专家策略之间的差异。在这种情况下， $z = \pi$ 其中 π 是给定状态时动作上的分布，$\ell_{i_t}(z) = \mathbb{E}_{s_t}\,[\text{KL}(\pi(\cdot|s_t)||\pi_{\text{expert}}(\cdot|s_t))]$其中，期望是针对行为策略访问的状态计算的，而 $f(\theta)$ 是策略参数化。由于计算 ℓ_{i_t} 需要行为策略与环境交互，因此代价高昂。此外，KL 散度在 ℓ_1-范数下是 1-强凸的，因此 OIL 满足我们的所有假设。当行为策略本身就是专家时，我们将该问题称为行为克隆。当学习策略用于与环境交互时（Florence 等人，2022年），我们将该问题称为在线模仿学习（OIL）（Lavington 等人，2022年；Ross等人，2011年）。

[5]代码可在此处获取 http://github.com/WilderLavington/Target-Based-Surrogates-For-Stochastic-Optimization。



(a) Hopper-v2 环境　　(b) Walker2d-v2 环境

图 1：比较 随机梯度下降、SLS、Adam、Adagrad 和 SSO 所产生的对数策略损失（专家标签与策略模型生成的平均动作之间的均方误差），该损失是总交互次数（等于 t，见算法 1）的函数。图例中的 SSO-m 表示代理函数经过 m 次梯度下降步数的最小化。底行展示了策略由神经网络参数化的实验，顶行则展示了策略由线性模型参数化的示例。在所有环境、行为策略和模型类型中，SSO 均优于其他在线优化算法。此外，随着 m 增大，SSO 的性能也会提升。

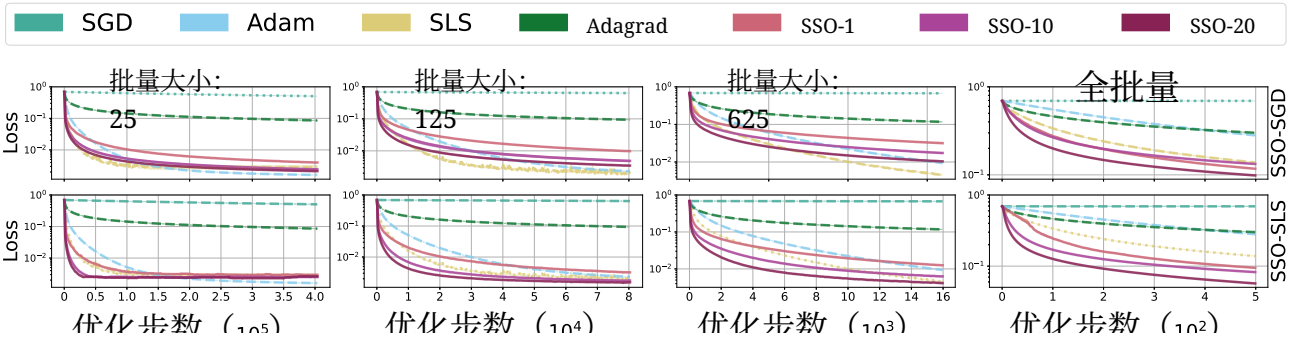


图 2：在 rcv1 数据集（随机梯度下降 及其 SSO 变体（顶行）、SLS 及其 SSO 变体（底行）的比较（Chang 和 Lin，2011年）中使用逻辑损失。 Adam 和 Adagrad 作为基线方法纳入比较。所有图均采用对数空间，横轴表示优化步数（等于 t，见算法 1）。我们注意到，随机梯度下降 采用理论步长时，表现不如 SLS 或 Adam 等更复杂的算法。相比之下，SSO 采用理论步长时，其表现可与使用默认超参数的 SLS 和 Adam 相媲美。值得注意的是，在不同 m 和批量大小下，SLS 和随机梯度下降的 SSO 变体均优于各自的参数化对应方法。

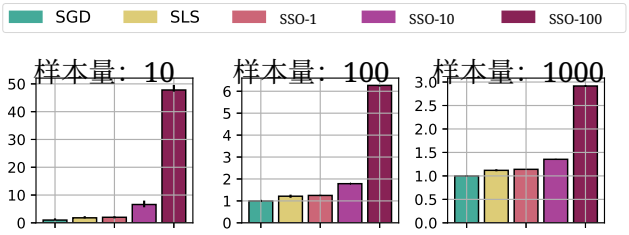


图 3：SSO 与相关基线方法之间的运行时间比较（以单次随机梯度下降更新的成本为基准进行归一化）。这些图表表明，当数据收集成本高昂时，即使 m 相对较大，SSO 的速度也能与随机梯度下降相当。对于图 1，$\tau \approx 1000$，计算 ℓ 与 f 所需时间之比约为 0.001。

在图 1 中，我们考察 Mujoco 基准套件中的连续控制环境（Todorov 等人，2012年）。策略对应于一个标准正态分布，其均值由线性函数或神经网络参数化。为收集状态，我们从随机（多元正态分布）策略中采样，以采取动作。每轮环境交互中会收集 1000 个状态，并用于更新策略。专家策略由正态分布定义，并通过双层多层感知机参数化，使用软演员-评论家算法（Haarnoja 等人，2018）进行训练。图1 显示，SSO 采用理论步长时，在环境交互次数（调用梯度预言机的次数）对应的对数损失方面，显著优于标准优化算法。此外，我们发现，无论采用线性还是神经网络参数化，

随着 m增加，学习策略的性能持续提升。

运行时间比较：在图3中，我们展示了各算法之间的相对运行时间。每一列表示完成单个优化步骤（采样状态并更新模型参数）所需的平均运行时间，并以随机梯度下降的运行时间进行归一化。我们改变每一步采集的状态数量（称为样本量），并考察样本量为10、100 和1000的情况。比较在使用双层多层感知机的`Hopper-v2`环境中进行。我们发现，对于较小的样本量，采集状态所需的时间并非模型更新时间的主要因素（例如，在第 1 列中，`SSO-100`所需时间几乎是随机梯度下降的 50 倍）。另一方面，对于较大的样本量，`SSO` 执行多次模型更新不再是主要耗时因素（例如，最右侧一列显示，`SSO-100` 所需时间仅约为随机梯度下降的三倍，但经验性能要好得多）。这项实验表明，在数据访问是计算随机梯度的主要瓶颈时，目标优化可能带来益处，这与第4.2节提出的理论直觉一致。对于模拟器开销较大的应用，例如自动驾驶车辆相关应用（Dosovitskiy 等人，2017年），`SSO` 还可以带来进一步的改进。

监督学习：为了探索在目标优化框架中使用其他优化方法，我们考虑一个简单的监督学习设定。具体而言，我们使用来自 LIBSVM（`rcv1` 数据集（Chang 和 Lin，2011年），在逻辑损失下采用四种不同的批量大小。附录E中还包含了针对其他数据集和优化算法的补充实验。[6]

随机替代函数的扩展：我们考虑在目标空间中采用不同的优化算法——随机线搜索（Vaswani 等，2019b）（SLS），以构造不同的代理函数。我们将由此得到的算法称为 SSO-SLS。对于 SSO-SLS，在每次迭代中，我们都会在目标空间中执行回溯线搜索，以设定满足 Armijo 条件的 η_t：$\ell_{i_t}(z_t - \eta_t \nabla_z \ell_{i_t}(z_t)) \leq \ell_{i_t}(z_t) - \frac{\eta_t}{2} \|\nabla_z \ell_{i_t}(z_t)\|_2^2$。我们使用选定的 η_t 来实例化代理函数 $\tilde{g}_t$，并遵循算法 1。在图 2 中，我们将 SSO-SGD（上排）和 SSO-SLS（下排）分别与其参数化变体进行比较。我们观察到，SGD 和 SLS 的 SSO 变体（i）在所有 m 下都不逊于其参数化对应方法；（ii）当 m 足够大时，表现优于后者。

[6]我们还将`SSO` 与随机方差缩减方法 SVRG（Johnson 和 Zhang，2013年）进行了比较，发现`SSO` 在我们考察的各种批量大小和数据集上始终表现更优。有关这一现象的示例，请参见附录15 中的图E。

6. 讨论

未来，我们计划将理论结果扩展到更广泛的函数类别，并通过实证评估目标优化在更复杂模型上的表现。由于我们的框架允许在目标空间中使用任意优化器，我们将探索其他优化算法，以构造更好的代理函数。另一个未来方向是构造能够利用模型中额外结构的更优代理函数。例如，Taylor 等人(2016); Amid 等人(2022年) 利用深度神经网络模型中的复合结构，构造局部的逐层代理函数，从而实现大规模并行化。最后，我们还计划将框架扩展到强化学习等应用，其中 ℓ是非凸的。

7. 致谢

我们感谢 Frederik Kunstner 和 Reza Asad 指出了本文早期版本中的错误。本研究部分得到了加拿大 CIFAR人工智能计划以及加拿大自然科学与工程研究委员会（NSERC）探索资助 RGPIN-2022-03669 和 RGPIN-2022-04816 的支持。

参考文献

Abernethy, J. D., Hazan, E., 和 Rakhlin, A.（2009）。在黑暗中竞争：一种用于多臂老虎机线性优化的高效算法。*COLT*。（引文见5）

Agarwal, N., Anil, R., Koren, T., Talwar, K., 和 Zhang, C.（2020）。采用滞后数据管道的随机优化。神经信息处理系统进展，33:10282–10293。（引文见7）

Amari, S.（1998）。自然梯度在学习中高效发挥作用。神经计算。（引文见2）

Amid, E., Anil, R., 和 Warmuth, M.（2022）。Locoprop：通过局部损失优化增强反向传播。载于 Camps-Valls, G., Ruiz, F. J. R., 和 Valera, I. 编，第*25*届国际人工智能与统计会议论文集，第151卷，机器学习研究论文集，第9626–9642页。PMLR。（引文见9）

Armijo, L.（1966）。具有 Lipschitz 连续一阶偏导数的函数的最小化。太平洋数学杂志，16(1):1–3。（引自7）

Belkin, M.、Rakhlin, A. 和 Tsybakov, A. B.（2019）。数据插值是否与统计最优性相矛盾？载于*AISTATS*。（引用于6）

Bollapragada, R.、Byrd, R. H. 和 Nocedal, J.（2019）。用于优化的精确与非精确子采样牛顿方法。*IMA* 数值分析期刊，39(2):545–578。（引用于7）

Bottou, L.、Curtis, F. E. 和 Nocedal, J.（2018）。大规模机器学习的优化方法。*SIAM* 综述，60(2):223–311。（引用见6）

Bubeck, S. 等（2015）。凸优化：算法与复杂性。*Machine Learning* 基础与趋势®，8(3-4):231–357。（引用见5，7）

Chang, C.-C. 和 Lin, C.-J.（2011年）。LIBSVM：一个支持向量机库。ACM 智能系统与技术汇刊（TIST），2(3):1–27。（引自第 8、9、30、36 页）Choi, D.、Passos, A.、Shallue, C. J. 和 Dahl, G. E.（2019年）。通过数据回响加速神经网络训练。arXiv 预印本 arXiv:1907.05550。（引自第 7 页）Dempster, A. P.、Laird, N. M. 和 Rubin, D. B.（1977年）。通过 EM 算法从不完整数据中进行最大似然估计。英国皇家统计学会会刊：B辑（方法论），39(1):1–22。（引自第 1 页）Dosovitskiy, A.、Ros, G.、Codevilla, F.、Lopez, A. 和 Koltun, V.（2017年）。CARLA：一个开放的城市驾驶模拟器。载于《第 1 届机器人学习年会论文集》，第 1–16 页。（引自第 9 页）Drusvyatskiy, D.（2017年）。重新审视近端点方法。arXiv 预印本 arXiv:1712.06038。（引自第 1 页）Duchi, J.、Hazan, E. 和 Singer, Y.（2011年）。用于在线学习和随机优化的自适应次梯度方法。机器学习研究期刊。（引自第 1、5 页）Florence, P.、Lynch, C.、Zeng, A.、Ramirez, O. A.、Wahid, A.、Downs, L.、Wong, A.、Lee, J.、Mordatch, I. 和 Tompson, J.（2022年）。隐式行为克隆。载于 Faust, A.、Hsu, D. 和 Neumann, G.（编），《第 5 届机器人学习会议论文集》，机器学习研究论文集第 164 卷，第 158–168 页。PMLR。（引自第 7 页）Gower, R. M.、Loizou, N.、Qian, X.、Sailanbayev, A.、Shulgin, E. 和 Richtárik, P.（2019年）。随机梯度下降：一般分析与改进的收敛速率。载于《国际机器学习会议论文集》，第 5200–5209 页。PMLR。（引自第 6、7 页）Haarnoja, T.、Zhou, A.、Abbeel, P. 和 Levine, S.（2018年）。软演员-评论家：采用随机演员的离策略最大熵深度强化学习。载于《国际机器学习会议论文集》，第 1861–1870 页。PMLR。（引自第 8 页）

Hazan, E.、Agarwal 和 Kale（2007年）。用于在线凸优化的对数遗憾算法。*MachineLearning*，69(2):169–192。（引自第 5）

Johnson, R. 和 Zhang, T.（2013）。利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，*NeurIPS*。（引用见9，36）

Johnson, R. 和 Zhang, T.（2020）。通过连续函数梯度优化引导非凸模型的学习。载于国际机器学习会议，第4921–4930页。PMLR。（引用见2）

Kakade, S. M.（2001）。自然策略梯度。载于《神经信息处理系统进展》第14卷*[*神经信息处理系统：自然与合成，NIPS 2001，2001年12月3日至8日，温哥华，不列颠哥伦比亚省，加拿大*]*。（引用见第2页）Kingma, D. 和 Ba, J.（2015）。Adam：一种随机优化方法。载于 ICLR。（引用见第1、30页）Lavington, J. W., Vaswani, S. 和 Schmidt, M.（2022）。在线模仿学习的改进策略优化。arXiv 预印本 arXiv:2208.00088。（引用见第7、37页）Li, X., Zhuang, Z. 和 Orabona, F.（2021）。重新审视指数步长和余弦步长：简洁性、自适应性与性能。载于《国际机器学习会议》，第6553–6564页。PMLR。（引用见第5、6页）Liang, T. 和 Rakhlin, A.（2018）。只需插值：核无岭回归也能泛化。arXiv 预印本 arXiv:1808.00387。（引用见第6页）Lohr, S. L.（2019）。《抽样：设计与分析：设计与分析》。Chapman and Hall/CRC。（引用见第6页）Loizou, N., Vaswani, S., Laradji, I. H. 和 Lacoste-Julien, S.（2021）。随机 Polyak 步长用于随机梯度下降：实现快速收敛的自适应学习率。载于《国际人工智能与统计会议》，第1306–1314页。PMLR。（引用见第6页）Ma, S., Bassily, R. 和 Belkin, M.（2018）。插值的力量：理解随机梯度下降在现代过参数化学习中的有效性。载于 ICML。（引用见第6页）Mairal, J.（2013）。用于大规模优化的随机主化-最小化算法。《神经信息处理系统进展》，第26卷。（引用见第1页）Mairal, J.（2015）。增量主化-最小化优化及其在大规模机器学习中的应用。《SIAM 优化期刊》，25(2):829–855。（引用见第1页）

Meng, S. Y., Vaswani, S., Laradji, I. H., Schmidt, M. 和 Lacoste-Julien, S. （ 2020 ）。快速而迅猛的收敛：插值条件下的随机二阶方法。载于国际人工智能与统计会议，第1375–1386页。PMLR。（引用见第7）

Nesterov, Y.（2003）。凸优化导论：基础课程，第87卷。Springer Science & Business Media。（引自22）

Newton, D.、Bollapragada, R.、Pasupathy, R. 和 Yip, N. K.（2021）。光滑随机优化的回顾性近似。*arXiv* 预印本 *arXiv:2103.04392*。（引自5）

Nguyen, L. M., Tran, T. H. 和 van Dijk, M.（2022）。有限和优化：关于收敛到全局解的新视角。arXiv 预印本 arXiv:2202.03524。（引用于 2）Nocedal, J. 和 Wright, S. J. （ 1999 ）。数值优化。Springer。（引用于 4）Orabona, F.（2019）。在线学习现代导论。arXiv 预印本 arXiv:1912.13213。（引用于 7、30）Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Rai-son, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. 和 Chintala, S.（2019）。PyTorch：一种命令式风格、高性能的深度学习库。载于《神经信息处理系统进展第32卷》，第8024–8035页。Curran Associates, Inc.（引用于 7、33、34、35、38）Robbins, H. 和 Monro, S.（1951年）。一种随机逼近方法。《数学统计年刊》，第400–407页。（引用于 1、3、5）Ross, S., Gordon, G. 和 Bagnell, D.（2011年）。将模仿学习和结构化预测化归为无遗憾在线学习。《机器学习研究杂志》，第627–635页。（引用于 1、7）Schmidt, M. 和 Le Roux, N.（2013年）。强增长条件下随机梯度下降的快速收敛。arXiv 预印本 arXiv:1308.6370。（引用于 2、6）Schmidt, M., Roux, N. 和 Bach, F.（2011年）。凸优化中非精确近端梯度方法的收敛率。神经信息处理系统进展，24。（引用于 6、28）Schulman, J., Levine, S., Abbeel, P., Jordan, M. 和 Moritz, P.（2015年）。信赖域策略优化。载于国际机器学习大会（ICML），第1889–1897页。（引用于 1）

Schulman, J., Wolski, F., Dhariwal, P., Radford, A. 和 Klimov, O.（2017年）。近端策略优化算法。 *CoRR*, abs/1707.06347。（引用于 1）

Sutton, R. S.、McAllester, D. A.、Singh, S. P. 和 Mansour, Y.（2000）。采用函数逼近的强化学习策略梯度方法。载于神经信息处理系统进展（*NeurIPS*），第1057–1063页。（引自1，38）

Taylor, G.、Burmeister, R.、Xu, Z.、Singh, B.、Patel, A. 和 Goldstein, T.（2016）。无需梯度训练神经网络：一种可扩展的交替方向乘子法方法。载于 Balcan, M. F. 和 Weinberger, K. Q. 编，第*33*届国际机器学习会议论文集，第48卷，机器学习研究论文集，第2722–2731页，纽约，纽约州，美国。PMLR。（引自9）

Todorov, E., Erez, T. 和 Tassa, Y.（2012年）。Mujoco：一种用于基于模型的控制的物理引擎。载于2012年IEEE/RSJ国际智能机器人与系统会议。（引用于第3、8、37页）Vaswani, S., Bach, F. 和 Schmidt, M.（2019a）。过参数化模型的随机梯度下降快速乃至更快收敛，以及一种加速感知机。载于第22届国际人工智能与统计会议，第1195–1204页。PMLR。（引用于第2、6页）Vaswani, S., Bache m, O., Totaro, S., Müller, R., Garg, S., Geist, M., Machado, M. C., Castro, P. S. 和 Roux, N. L.（2021）。一类用于稳定高效强化学习的通用代理函数。arXiv预印本 arXiv:2108.05828。（引用于第2页）Vaswani, S., Dubois-Taine, B. 和 Babanezhad, R.（2022年）。迈向噪声自适应、问题自适应的（加速）随机梯度下降。载于国际机器学习会议，第22015–22059页。PMLR。（引用于第5、6、7、20、25、30页）Vaswani, S., Laradji, I., Kunstner, F., Meng, S. Y., Schmidt, M. 和 Lacoste-Julien, S.（2020）。采用过参数化时，自适应梯度方法收敛更快（但应进行线搜索）。arXiv预印本 arXiv:2006.06835。（引用于第6页）Vaswani, S., Mishkin, A., Laradji, I., Schmidt, M., Gidel, G. 和 Lacoste-Julien, S.（2019b）。轻松实现随机梯度：插值、线搜索与收敛率。载于神经信息处理系统进展，第3727–3740页。（引用于第5、9、30、33、34、35、38页）Ward, R., Wu, X. 和 Bottou, L.（2020）。Adagrad步长：非凸景观上的尖锐收敛结果。《机器学习研究杂志》，21(1):9047–9076。（引用于第34页）

Williams, R. J.（1992）。简单统计梯度跟随算法：用于连接主义强化学习。 机器学习，8(3-4):229–256。（引用于1）

Woodworth, B., Mishchenko, K., 和 Bach, F.（2023年）。两个损失胜过一个：利用更廉价的代理实现更快的优化。*arXiv*预印本 *arXiv:2302.03542*。（引用于2）

Zhang, C., Bengio, S., Hardt, M., Recht, B. 和 Vinyals, O.（2017年）。理解深度学习需要重新思考泛化问题。载于*ICLR*。（引自6）

附录结构

A 定义 B 算法 C 确定性情形下的
证明 D 随机设定下的证明 E 其他
实验结果

A. 定义

我们的主要假设是，每个单独的函数 f_i 都是可微的，具有有限最小值 f_i^*，并且是 L_i-光滑的，这意味着对于所有 v 和 w，

$$f_i(v) \leq f_i(w) + \langle \nabla f_i(w), v - w \rangle + \frac{L_i}{2} \|v - w\|_2^2 , \qquad \text{(Individual Smoothness)}$$

这也意味着 f 是 L-光滑的，其中 L 是各个函数的最大光滑常数。光滑性的一个推论是随机梯度范数满足以下界，

$$\|\nabla f_i(w) - \nabla f_i^*\|^2 \leq 2L(f_i(w) - f_i^* - \langle \nabla f_i^*, w - w_i^* \rangle). \tag{1}$$

我们还假设每个 f_i 都是凸的，这意味着对于所有 v 和 w，

$$f_i(v) \geq f_i(w) + \langle \nabla f_i(w), v - w \rangle, \qquad \text{(Convexity)}$$

根据具体情形，我们还将假设 f 是 μ 强凸的，这意味着对于所有 v 和 w，

$$f(v) \geq f(w) + \langle \nabla f(w), v - w \rangle + \frac{\mu}{2} \|v - w\|_2^2 , \qquad \text{(Strong Convexity)}$$

B. 算法

本节将介绍目标空间中超越标准随机梯度下降更新的算法。我们将通过两种方式来实现这一点——(i) 将随机梯度下降扩展为利用二阶信息的在线牛顿步，详见附录B.1；以及 (ii) 将随机梯度下降扩展为更一般的随机镜像下降算法，详见附录 B.2。对于 (i) 和 (ii)，我们将分别针对平方损失和逻辑损失实例化相应算法。

B.1. 在线牛顿步

下面考虑针对目标的在线牛顿步。相应的更新为：

$$z_{t+1/2} = z_t - \eta_t [\nabla_z^2 \ell_t(z_t)]^{-1} \nabla_z \ell_t(z_t) \quad ; \quad \bar{z}_{t+1} = \arg\min_{z \in \mathcal{Z}} \frac{1}{2} \left\| z - z_{t+1/2} \right\|_{\mathcal{P}_t}^2 \tag{2}$$

$$z_{t+1} = f(\theta_{t+1}) \quad ; \quad \theta_{t+1} = \arg\min_{\theta} \left[\langle \nabla_z \ell_t(z_t), f(\theta) - z_t \rangle + \frac{1}{2\eta_t} \|f(\theta) - z_t\|_{\nabla^2 \ell_t(z_t)}^2 \right] \tag{3}$$

其中，$\nabla_z^2 \ell_t(z_t)$ 是关于 z 的、与样本 i_t 对应的损失海森矩阵。下面针对平方损失实例化此更新。在这种情况下，$\ell_t(z) = \frac{1}{2} \|z - y_t\|_2^2$，因此，$\nabla \ell_t(z) = z - y_t$，$[\nabla^2 \ell_t(z)]_{i_t, i_t} = 1$，且 $[\nabla^2 \ell_t(z)]_{j,j} = 0$ 对所有 $j \neq i_t$ 均成立。因此，对于平方损失，式 (2) 等同于目标空间中的梯度下降。

对于逻辑损失，$\ell_t(z) = \log(1 + \exp(-y_t z))$。如果 i_t 是在迭代 t 时抽取的损失索引，那么，对于所有$j \neq i_t$，$[\nabla \ell_t(z)]_j = 0$。同样，除了 $[i_t, i_t]$ 之外，$\nabla^2 \ell_t(z)$ 的所有元素均为零。

$$[\nabla \ell_t(z)]_{i_t} = \frac{-y_t}{1 + \exp(y_t z_t)} \quad ; \quad [\nabla^2 \ell(z)]_{i_t, i_t,} = \frac{1}{1 + \exp(y_t z_t)} \frac{1}{1 + \exp(-y_t z_t)} = (1 - p_t) p_t ,$$

其中，$p_t = \frac{1}{1+\exp(-y_t z_t)}$ 是将示例 i_t 分类为具有 $+1$ 标签的概率。在这种情况下，代理函数可以写为：

$$\tilde{g}_t^z(\theta) = \frac{-y_t}{1+\exp(y_t\, z_t)} \left(f_{i_t}(\theta) - z_t^{i_t}\right) + \frac{(1-p_t)\, p_t}{2\eta_t} \left(f_{i_t}(\theta) - z_t^{i_t}\right)^2 \tag{4}$$

与之前一样，上述代理函数可以高效地实现。

B.2. 随机镜像下降

如果 ϕ 是一个可微的严格凸镜像映射，它会在 x 和 y 之间诱导出布雷格曼散度：$D_\phi(y,\ x) := \phi(y) - \phi(x) - \langle \nabla\phi(x)\ y - x\rangle$。为了高效实现随机镜像下降，我们要求布雷格曼散度是可分离的，即 $D_\phi(y,\ x) = \sum_{j=1}^p D_{\phi_j}(y^j,\ x^j) = \sum_{j=1}^p \phi_j(y^j) - \phi_j(x^j) - \frac{\partial \phi_j}{(\partial x^j}^x [y^j - x^j])$。当 ϕ 为欧几里得范数或负熵时，就满足这种可分离结构。我们将目标空间中的随机镜像下降更新定义如下，

$$\nabla\phi(z_{t+1/2}) = \nabla\phi(z_t) - \eta_t \nabla_z \ell_t(z_t) \quad ; \quad \bar{z}_{t+1} = \arg\min_{z\in\mathcal{Z}} \sum_{j=1}^p \mathbb{I}(j = i_t) D_{\phi_j}(z^j, z^j_{t+1/2})) \tag{5}$$

$$\implies \bar{z}_{t+1} = \arg\min_{z\in\mathcal{Z}} \left[\langle \nabla_z \ell_t(z_t), z - z_t\rangle + \frac{1}{\eta_t} \sum_{j=1}^p \mathbb{I}(j = i_t) D_{\phi_j}(z^j, z^j_{t+1/2}) \right] \tag{6}$$

其中 $\mathbb{I}$ 是示性函数，i_t 对应于在迭代 t中选取的样本索引。对于欧几里得镜像映射，$\phi(z) = \frac{1}{2}\|z\|_2^2$, $D_\phi(z,\ z_t) = \frac{1}{2}\|z - z_t\|_2^2$，由此可恢复随机梯度下降更新。

镜像映射的另一种常见选择是负熵函数：$\phi(x) = \sum_{i=1}^K x^i \log(x^i)$，其中 x^i 是 $x \in \mathbb{R}^K$的坐标i 。这会导出作为布雷格曼散度的（广义）KL 散度，

$$D_\phi(x, y) = \sum_{k=1}^K x^k \log\left(\frac{x^k}{y^k}\right) + \sum_{k=1}^K x^k - \sum_{k=1}^K y^k\,.$$

如果 x 和 y 都对应概率分布，即 $\sum_{k=1}^K x^k = \sum_{k=1}^K y^k = 1$，那么诱导出的布雷格曼散度就是这两个分布之间的标准 KL 散度。对于多分类，$\mathcal{Z} \subseteq \mathbb{R}^{p\times K}$，且每个 $z^i \in \Delta_K$ 都满足 Δ_K 是 K维单纯形。我们将 z^i 的坐标 j 称为 $[z^i]_j$。由于 $z^i \in \Delta_K$，$[z^i]_k \geq 0$且$\sum_{k=1}^K [z^i]_k = 1$。

当使用负熵镜像映射时，让我们具体写出公式（5）中的一般 SMD 更新。在此情况下，对于 $z \in \Delta_K$，$[\nabla\phi(z)]_k = 1 + \log([z]_k)$。记 $\nabla_t := \nabla_z \ell_t(z_t)$ 并使用 $[\nabla_t]_k$ 表示 k 维向量 K的坐标∇_t。因此，公式（5）可以写成：

$$[z^{i_t}_{t+1/2}]_k = [{z_t}^{i_t}]_k \exp\left(-\eta_t [\nabla_t]_k\right) \tag{7}$$

对于多分类，$y^i \in \{0,1\}^K$ 是独热向量。如果 $[y^i]_k$ 表示向量 y^i的坐标 k，那么 n个观测值对应的对数似然可以写成：

$$\ell(z) = \sum_{i=1}^n \sum_{k=1}^K [y^i]_k \log([z^i]_k)\,.$$

在我们的目标优化框架中，目标对应于将样本分类为某一类别的概率。我们使用一种参数化来建模向量值函数 $f_i(\theta)$：$\mathbb{R}^d \to \mathbb{R}^K$。这确保对于所有 i，$\sum_{k=1}^K [f_i(\theta)]_k = 1$。因此，公式（5）中的投影步骤可以改写为：

$$\min_{z\in\mathcal{Z}} D_\phi(z, z_{t+1/2}) = \sum_{k=1}^K [f_{i_t}(\theta)]_k \log\left(\frac{[f_{i_t}(\theta)]_k}{[z^{i_t}_{t+1/2}]_k}\right)$$

其中 $[z^{i_t}_{t+1/2}]_k$ 根据公式 (7) 计算。由于 $[z^{i_t}_{t+1/2}]_k$ 的计算以及由此产生的投影仅取决于样本 i_t，因此可以高效地实现。

C. 确定性情形中的证明

引理 C.1。 假设 $g_t^z(\theta)$ 关于欧几里得范数是 β-光滑的，且 $\eta \leq \frac{1}{L}$，那么对于 $\alpha = 1/\beta$，算法 *t1* 的迭代 *1* 可保证 $h(\theta_{t+1}) \geq h(\theta_t)$。对于任意数量 $m \geq 1$ 的代理步骤，在此设定下，若再假设 h 以下界 h^*，则算法 *1* 可得到如下保证：

$$\min_{t\in\{0,\dots,T-1\}} \|\nabla h(\theta_t)\|_2^2 \leq \frac{2\beta\left[h(\theta_0) - h^*\right]}{T}.$$

证明。使 使用算法 1 中的更新以及 $\alpha = \frac{1}{\beta}$ 和 β-光滑性，可得 $g_t^z(\theta)$对所有 $k \in [m-1]$,

$$g_t^z(\omega_{k+1}) \leq g_t^z(\omega_k) - \frac{1}{2\beta}\|\nabla g_t^z(\omega_k)\|_2^2$$

m 步之后,

$$g_t^z(\omega_m) \leq g_t^z(\omega_0) - \frac{1}{2\beta}\sum_{k=0}^{m-1}\|\nabla g_t^z(\omega_k)\|_2^2$$

由于 $\theta_{t+1} = \omega_m$ 和 $\omega_0 = \theta_t$ 在算法 1中,

$$\implies g_t^z(\theta_{t+1}) \leq g_t^z(\theta_t) - \frac{1}{2\beta}\|\nabla g_t^z(\theta_t)\|_2^2 - \sum_{k=1}^{m-1}\|\nabla g_t^z(\omega_k)\|_2^2$$

注意， $h(\theta_t) = g_t^z(\theta_t)$，并且如果 $\eta \leq \frac{1}{L}$，那么 $h(\theta_{t+1}) \leq g_t^z(\theta_{t+1})$。利用这些关系,

$$h(\theta_{t+1}) \leq h(\theta_t) - \left[\underbrace{\frac{1}{2\beta}\|\nabla g_t^z(\theta_t)\|_2^2 + \sum_{k=1}^{m-1}\|\nabla g_t^z(\omega_k)\|_2^2}_{\geq 0}\right] \implies h(\theta_{t+1}) \leq h(\theta_t).$$

这证明了引理的第一部分。由于 $\sum_{k=1}^{m-1}\|\nabla g_t^z(\omega_k)\|_2^2 \geq 0$,

$$h(\theta_{t+1}) \leq h(\theta_t) - \frac{1}{2\beta}\|\nabla g_t^z(\theta_t)\|_2^2 \implies \|\nabla h(\theta_t)\|_2^2 \leq 2\beta\left[h(\theta_t) - h(\theta_{t+1})\right] \qquad \text{(Since } \nabla h(\theta_t) = \nabla g_t^z(\theta_t)\text{)}$$

从 $k = 0$ 到 $T-1$ 求和，并除以 T,

$$\frac{\|\nabla h(\theta_t)\|_2^2}{T} \leq \frac{2\beta\left[h(\theta_t) - h^*\right]}{T} \implies \min_{t\in\{0,\dots,T-1\}}\|\nabla h(\theta_t)\|_2^2 \leq \frac{2\beta\left[h(\theta_t) - h^*\right]}{T}$$

□

D. 随机设定下的证明

D.1. 目标空间中 SSO 与随机梯度下降的等价性

引理 4.1. 以下更新是等价的：

$$
\begin{aligned}
&(1)\quad \tilde{\theta}_{t+1} = \arg\min_{\theta} \tilde{g}_t(\theta); && \text{(SSO)}\\
&\qquad \tilde{z}_{t+1}^{(1)} = f(\tilde{\theta}_{t+1})\\
&(2)\quad z_{t+1/2} = z_t - \eta_t \nabla_z \ell_{i_t}(z_t); && \text{(Target-space SGD)}\\
&\qquad \tilde{z}_{t+1}^{(2)} = \arg\min_{z\in\mathcal{Z}} \frac{1}{2}\left\|z_{t+1/2} - z\right\|_{\mathcal{P}_t}^2
\end{aligned}
$$

其中，$\mathcal{P}_t \in \mathbb{R}^{p\times p}$ 是一个随机对角矩阵，使得 $\mathcal{P}_t(i_t, i_t) = 1$ 且 $\mathcal{P}_t(j, j) = 0$ 对所有 $j \neq i_t$ 均成立。也就是说，*SSO (1)* 和目标空间中的随机梯度下降 *(2)* 在每一步都会得到相同的迭代点，即若 $z_t = f(\theta_t)$，则 $\tilde{z}_{t+1} := \tilde{z}_{t+1}^{(1)} = \tilde{z}_{t+1}^{(2)}$。

证明。 由于 ℓ 是可分离的，若 i_t 是从 z 中采样得到的坐标，我们可以将目标空间中的更新重写如下：

$$
\begin{aligned}
{z_{t+1/2}}^{i_t} &= {z_t}^{i_t} - \eta_t \frac{\partial \ell_{i_t}(z_t)}{\partial z^{i_t}}\\
{z_{t+1/2}}^{j} &= {z_t}^{j} \quad \text{when } j \neq i_t\ .
\end{aligned}
$$

将上述更新代入投影步骤，可得：

$$
\begin{aligned}
\tilde{z}_{t+1} &= \arg\min_{z\in\mathcal{Z}} \frac{1}{2}\{\left\|{z_{t+1/2}}^{i_t} - z^{i_t}\right\|_2^2\}\\
&= \arg\min_{z\in\mathcal{Z}} \frac{1}{2}\{\left\|{z_t}^{i_t} - \eta_t \frac{\partial \ell_{i_t}(z_t)}{\partial z^{i_t}} - z^{i_t}\right\|_2^2\}\\
&= \arg\min_{z\in\mathcal{Z}} \left\{\left[\frac{\partial \ell_{i_t}(z_t)}{\partial z^{i_t}}\left[z^{i_t} - z_t^{i_t}\right] + \frac{1}{2\eta_t}\left\|z^{i_t} - z_t^{i_t}\right\|_2^2\right]\right\} && \text{(Due to separability of } \ell)
\end{aligned}
$$

由于对于所有 $z \in \mathcal{Z}, z = f(\theta)$ 和 $z^i = f_i(\theta)$，对于所有 i 均成立。因此存在 $\tilde{z}_{t+1} = f(\tilde{\theta}_{t+1})$，使得，

$$
\tilde{\theta}_{t+1} = \arg\min_{\theta\in\Theta} \left\{\frac{\partial \ell_{i_t}(z_t)}{\partial z^{i_t}}\left[f_{i_t}(\theta) - f_{i_t}(\theta_t)\right] + \frac{1}{2\eta_t}\left\|f_{i_t}(\theta) - f_{i_t}(\theta_t)\right\|_2^2\right\} = \arg\min_{\theta\in\Theta} \tilde{g}_t^z(\theta)
$$

□

引理 D.1. 考虑以下更新：

$$
\begin{aligned}
&\bar{\theta}_{t+1} = \arg\min_{\theta} \tilde{q}_t(\theta) \quad ; \quad \bar{z}_{t+1}^{(1)} = f(\bar{\theta}_{t+1}) && \text{(SSO)}\\
&z_{t+1/2} = z_t - \eta_t' \nabla_z \ell_{i_t}(z_t); && \text{(Target-space SGD)}\\
&\bar{z}_{t+1}^{(2)} = \arg\min_{z\in\mathcal{Z}} \frac{1}{2}\left\|z_{t+1/2} - z\right\|_2^2
\end{aligned}
$$

SSO 与目标 空间随机梯度下降在每一步都得到相同的迭代点，即如果 $z_t = f(\theta_t)$，那么 $\bar{z}_{t+1} := \bar{z}_{t+1}^{(1)} = \bar{z}^{(2)}_{t+1}$.

证明。

$$
\begin{aligned}
\bar{z}_{t+1} &= \arg\min_{z\in\mathcal{Z}} \frac{1}{2}\{\left\|z_{t+1/2} - z\right\|_2^2\}\\
&= \arg\min_{z\in\mathcal{Z}} \frac{1}{2}\{\left\|z_t - \eta_t' \nabla_z \ell_{i_t}(z_t) - z\right\|_2^2\}\\
&= \arg\min_{z\in\mathcal{Z}} \left\{\left[\frac{\partial \ell_{i_t}(z_t)}{\partial z^{i_t}}\left[z^{i_t} - z_t^{i_t}\right] + \frac{1}{2\eta_t'}\left\|z - z_t\right\|_2^2\right]\right\} && \text{(Due to separability of } \ell)
\end{aligned}
$$

¯由于对所有 $z \in \mathcal{Z}, z = f(\theta)$。因此存在 $\bar{z}_{t+1} = f(\theta_{t+1})$，使得，

$$\begin{aligned}\bar{\theta}_{t+1} &= \arg\min_{\theta\in\Theta}\left\{\frac{\partial \ell_{i_t}(z_t)}{\partial z^{i_t}}\left[f_{i_t}(\theta) - f_{i_t}(\theta_t)\right] + \frac{1}{2\eta_t'}\left\|f(\theta) - f(\theta_t)\right\|_2^2\right\} \\ &= \arg\min_{\theta\in\Theta} \tilde{q}_t(\theta)\end{aligned}$$

□

D.2 强凸函数的证明

我们考虑 $\ell(z)$ 为强凸函数且集合 $\mathcal{Z}$ 为凸集的情况。我们将重点讨论目标空间中的随机梯度下降，并考虑以下更新:

$$\begin{aligned} z_{t+1/2} &= z_t - \eta_t' \nabla \ell_t(z_t) \\ \bar{z}_{t+1} &= \Pi_{\mathcal{Z}}[z_{t+1/2}] := \arg\min_{z\in\mathcal{Z}} \frac{1}{2}\left\|z - z_{t+1/2}\right\|_2^2 \\ \left\|z_{t+1} - \bar{z}_{t+1}\right\| &\le \epsilon_{t+1}\end{aligned}$$

引理 D.2。 根据 $\bar{z}_{t+1}$ 和 ϵ_{t+1} 的次优性，界定 z_{t+1} 相对于 z^* 的次优性，可得

$$\left\|z_{t+1} - z^*\right\|_2^2 \le \left\|\bar{z}_{t+1} - z^*\right\|_2^2 + 2\epsilon_{t+1}\left\|z_{t+1} - z^*\right\| . \tag{8}$$

证明。

$$\begin{aligned}\left\|z_{t+1} - z^*\right\|_2^2 &= \left\|z_{t+1} - \bar{z}_{t+1} + \bar{z}_{t+1} - z^*\right\|_2^2 \\ &= \left\|z_{t+1} - \bar{z}_{t+1}\right\|_2^2 + \left\|\bar{z}_{t+1} - z^*\right\|_2^2 + 2\langle z_{t+1} - \bar{z}_{t+1}, \bar{z}_{t+1} - z^*\rangle \\ &= \left\|z_{t+1} - \bar{z}_{t+1}\right\|_2^2 + \left\|\bar{z}_{t+1} - z^*\right\|_2^2 + 2\langle z_{t+1} - \bar{z}_{t+1}, \bar{z}_{t+1} - z_{t+1} + z_{t+1} - z^*\rangle \\ &= \left\|z_{t+1} - \bar{z}_{t+1}\right\|_2^2 + \left\|\bar{z}_{t+1} - z^*\right\|_2^2 + 2\langle z_{t+1} - \bar{z}_{t+1}, z_{t+1} - z^*\rangle - 2\left\|z_{t+1} - \bar{z}_{t+1}\right\|_2^2 \\ &\le \left\|\bar{z}_{t+1} - z^*\right\|_2^2 + 2\left\|z_{t+1} - \bar{z}_{t+1}\right\|\left\|z_{t+1} - z^*\right\| \\ &\le \left\|\bar{z}_{t+1} - z^*\right\|_2^2 + 2\epsilon_{t+1}\left\|z_{t+1} - z^*\right\|\end{aligned}$$

□

现在，我们用第 t 次迭代时的不精确次优性来界定第 $t+1$ 次迭代时的精确次优性，从而得到一个递推。

引理 D.3。 假设 ***(i)*** 每个 ℓ_t 都是 L-光滑的，且 ***(ii)*** ℓ 是 μ-强凸的，并且 ***(iii)*** $\eta_t' \le \frac{1}{2L}$，则有

$$\mathbb{E}\left\|\bar{z}_{t+1} - z^*\right\|_2^2 \le (1 - \mu\eta_t')\mathbb{E}\left\|z_t - z^*\right\|_2^2 + 2{\eta_t'}^2\sigma^2 \tag{9}$$

其中 $\sigma^2 = \mathbb{E}\left\|\nabla\ell(z^*) - \nabla\ell_t(z^*)\right\|_2^2$。

证明。由于 $z^* \in \mathcal{Z}$ 且最优， $z^* = \Pi_{\mathcal{Z}}[z^* - \eta_t'\nabla\ell(z^*)]$。

$$\begin{aligned}\left\|\bar{z}_{t+1} - z^*\right\|_2^2 &= \left\|\Pi_{\mathcal{Z}}[z_t - \eta_t'\nabla\ell_t(z_t)] - \Pi_{\mathcal{Z}}[z^* - \eta_t'\nabla\ell(z^*)]\right\|_2^2 \\ &\le \left\|[z_t - \eta_t'\nabla\ell_t(z_t)] - [z^* - \eta_t'\nabla\ell(z^*)]\right\|_2^2 && \text{(Since projections are non-expansive)} \\ &= \left\|[z_t - \eta_t'\nabla\ell_t(z_t)] - [z^* - \eta_t'\nabla\ell_t(z^*)] + \eta_t'\left[\nabla\ell(z^*) - \nabla\ell_t(z^*)\right]\right\|_2^2 \\ &= \left\|z_t - z^*\right\|_2^2 + {\eta_t'}^2\left\|\nabla\ell(z^*) - \nabla\ell_t(z^*)\right\|_2^2 + {\eta_t'}^2\left\|\nabla\ell_t(z_t) - \nabla\ell_t(z^*)\right\|_2^2 + \underbrace{2\eta_t'\langle z_t - z^*, \nabla\ell(z^*) - \nabla\ell_t(z^*)\rangle}_{A_t} \\ &\quad - 2\eta_t'\langle z_t - z^*, \nabla\ell_t(z_t) - \nabla\ell_t(z^*)\rangle + 2{\eta_t'}^2\underbrace{\langle\nabla\ell_t(z^*) - \nabla\ell(z^*), \nabla\ell_t(z_t) - \nabla\ell_t(z^*)\rangle}_{B_t} \\ &\le \left\|z_t - z^*\right\|_2^2 + 2{\eta_t'}^2\left\|\nabla\ell(z^*) - \nabla\ell_t(z^*)\right\|_2^2 + 2{\eta_t'}^2\left\|\nabla\ell_t(z_t) - \nabla\ell_t(z^*)\right\|_2^2 \\ &\quad + A_t - 2\eta_t'\langle z_t - z^*, \nabla\ell_t(z_t) - \nabla\ell_t(z^*)\rangle && \text{(Young inequality on } B_t\text{)}\end{aligned}$$

T对 i_t 取期望，已知 $\mathbb{E}A_t = 0$，并利用 $\sigma^2 = \mathbb{E}\|\nabla\ell(z^*) - \nabla\ell_t(z^*)\|_2^2$.

$$
\begin{aligned}
\mathbb{E}\|\bar{z}_{t+1} - z^*\|_2^2 &\le \mathbb{E}\|z_t - z^*\|_2^2 + 2\eta_t'^2\mathbb{E}\|\nabla\ell_t(z_t) - \nabla\ell_t(z^*)\|_2^2 - 2\eta_t'\langle z_t - z^*, \nabla\ell(z_t) - \nabla\ell(z^*)\rangle + 2\eta_t'^2\sigma^2 \\
&\le \mathbb{E}\|z_t - z^*\|_2^2 + 4\eta_t'^2 L\,\mathbb{E}\{\ell_t(z_t) - \ell_t(z^*) - \langle\nabla\ell_t(z^*), z_t - z^*\rangle\} - 2\eta_t'\langle z_t \\
&\qquad - z^*, \nabla\ell(z_t) - \nabla\ell(z^*)\rangle + 2\eta_t'^2\sigma^2 \qquad (10) \\
&\le \mathbb{E}\|z_t - z^*\|_2^2 + 2\eta_t'\{\ell(z_t) - \ell(z^*) - \langle\nabla\ell(z^*), z_t - z^*\rangle\} - 2\eta_t'\langle z_t - z^*, \nabla\ell(z_t) - \nabla\ell(z^*)\rangle + 2\eta_t'^2\sigma^2 \\
&\qquad\qquad (\eta_t' < \tfrac{1}{2L}) \\
&\le \mathbb{E}\|z_t - z^*\|_2^2 + 2\eta_t'\{\ell(z_t) - \ell(z^*) - \langle\nabla\ell(z_t), z_t - z^*\rangle\} + 2\eta_t'^2\sigma^2 \\
&\le \mathbb{E}\|z_t - z^*\|_2^2 - \mu\eta_t'\mathbb{E}\|z_t - z^*\|_2^2 + 2\eta_t'^2\sigma^2 \qquad \text{(strong convexity of } \ell) \\
&\le (1 - \mu\eta_t')\mathbb{E}\|z_t - z^*\|_2^2 + 2\eta_t'^2\sigma^2
\end{aligned}
$$

其中，在公式（10）中，我们利用了 ℓ_t的光滑性。 □

定理 4.2。 假设 *(i)* ℓ_{i_t} 是 L-光滑且凸的，*(ii)* ℓ 是 μ-强凸的，*(iii)* $z^* = \operatorname{argmin}_{z\in\mathcal{Z}}\ell(z)$ *(iv)* 并且对于所有 t，$\epsilon_t \le \epsilon$， T次 *SSO* 迭代，对于 $z_T = f(\theta_T)$均满足以下界，

$$
\begin{aligned}
\mathbb{E}\|z_{T+1} - z^*\| \le \Bigg(&\left(\prod_{i=1}^T \rho_i\right)\|z_1 - z^*\|_2^2 \\
&+ 2\sigma^2\sum_{t=1}^T\prod_{i=t+1}^T\rho_i\eta_t'^2\Bigg)^{1/2} \\
&+ 2\epsilon\sum_{t=1}^T\prod_{i=t+1}^T\rho_i\,,
\end{aligned}
$$

其中 $\rho_t = (1 - \mu\eta_t')$， $\sigma^2 := \mathbb{E}\left[\|\nabla\ell(z^*) - \nabla\ell_t(z^*)\|_2^2\right]$.

(a) 恒定步长：当 $\eta_t = \frac{1}{2Ln}$且对于 $\rho = 1 - \frac{\mu}{2L}$时，

$$
\begin{aligned}
\mathbb{E}\|z_{T+1} - z^*\| \le \|z_1 - z^*\|&\left(1 - \frac{1}{2\kappa}\right)^{\frac{T}{2}} \\
&+ \frac{\sigma}{\sqrt{\mu L}} + \frac{2\epsilon}{1 - \sqrt{1 - \frac{1}{2\kappa}}}\,.
\end{aligned}
$$

(b) 指数步长：当$\eta_t = \frac{1}{2Ln}\alpha^t$ 对于 $\alpha = (\frac{\beta}{T})^{\frac{1}{T}}$，

$$
\begin{aligned}
\mathbb{E}\|z_{T+1} - z^*\| \le c_1\,&\exp\left(-\frac{T}{4\kappa}\frac{\alpha}{\ln(T/\beta)}\right)\|z_1 - z^*\| \\
&+ \frac{4\kappa c_1(\ln(T/\beta))}{Le\alpha\sqrt{T}}\sigma + 2\epsilon\,c_2\,.
\end{aligned}
$$

其中 $c_1 = \exp\left(\frac{1}{4\kappa}\right)$，且 $c_2 = \exp^{\beta}\left(\frac{2\beta\ln(T/\beta)}{2\kappa\ln(T/\beta)}\right)$

证明。 根据引理D.2和引理D.3可得

$$
\begin{aligned}
\mathbb{E}\|z_{t+1} - z^*\|_2^2 &\le \mathbb{E}\|\bar{z}_{t+1} - z^*\|_2^2 + 2\mathbb{E}[\epsilon_{t+1}\|z_{t+1} - z^*\|] \\
&\le \underbrace{(1 - \mu\eta_t')}_{\rho_t}\mathbb{E}\|z_t - z^*\|_2^2 + 2\eta_t'^2\sigma^2 + 2\mathbb{E}[\epsilon_{t+1}\|z_{t+1} - z^*\|]
\end{aligned}
$$

从 $t = 1$ 递归到 T,

$$
\mathbb{E}[\|z_{T+1} - z^*\|_2^2] \le \left(\prod_{t=1}^T\rho_t\right)\|z_1 - z^*\|_2^2 + 2\sigma^2\sum_{t=1}^T\prod_{i=t+1}^T\rho_i\eta_t'^2 + 2\sum_{t=1}^T\prod_{i=t+1}^T\rho_i\mathbb{E}[\epsilon_{t+1}\|z_{t+1} - z^*\|]
$$

D记 $u_t := \mathbb{E}\|z_t - z^*\|$。根据詹森不等式，我们知道 $u_T^2 \le \mathbb{E}[\|z_T - z^*\|_2^2]$.

$$\implies u_{T+1}^2 \le \left(\prod_{t=1}^{T}\rho_t\right)u_1^2 + 2\sigma^2\sum_{t=1}^{T}\prod_{i=t+1}^{T}\rho_i{\eta'_t}^2 + 2\epsilon\sum_{t=1}^{T}\prod_{i=t+1}^{T}\rho_i\, u_{t+1}$$

将前一个不等式的两边除以 $\prod_{i=1}^{T}\rho_i$，得到:

$$\left(\prod_{i=1}^{T}\rho_i\right)^{-1}u_{T+1}^2 \le u_1^2 + 2\sigma^2\left(\prod_{i=1}^{T}\rho_i\right)^{-1}\sum_{t=1}^{T}{\eta'_t}^2\left(\prod_{i=t+1}^{T}\rho_i\right) + 2\epsilon\left(\prod_{i=1}^{T}\rho_i\right)^{-1}\sum_{t=1}^{T}u_{t+1}\left(\prod_{i=t+1}^{T}\rho_i\right)$$

将上述不等式化简为适用于任意 $\tau \le T$的形式,

$$\left[\left(\prod_{i=1}^{\tau}\rho_i\right)^{-\frac{1}{2}}u_{\tau+1}\right]^2 \le u_1^2 + 2\sigma^2\sum_{t=1}^{\tau}{\eta'_t}^2\left(\prod_{i=1}^{t}\rho_i\right)^{-1} + \sum_{t=1}^{\tau}2\epsilon\left(\prod_{i=1}^{t}\rho_i\right)^{-\frac{1}{2}}\left[\left(\prod_{i=1}^{t}\rho_i\right)^{-\frac{1}{2}}u_{t+1}\right]$$

令 $v_\tau := (\prod_{i=1}^{\tau}\rho_i)^{-\frac{1}{2}}u_{\tau+1}$且 $S_\tau := u_1^2 + 2\sigma^2\sum_{t=1}^{\tau}{\eta'_t}^2\left(\prod_{i=1}^{t}\rho_i\right)^{-1}$。再记 $\lambda_t := 2\epsilon\left(\prod_{i=1}^{t}\rho_i\right)^{-\frac{1}{2}}$。注意到 - $S_0 = u_1^2 = v_0^2$且 $S_{\tau+1} = S_\tau + 2\sigma^2{\eta'_{t\tau+1}}^2\left(\prod_{i=1}^{\tau+1}\rho_i\right)^{-1}$。因此， S_τ是一个递增数列。使用新变量重写前一个不等式，得到以下不等式:

$$v_\tau^2 \le S_\tau + \sum_{t=1}^{\tau}\lambda_t v_t$$

利用引理的结果D.10可得:

$$\begin{aligned} v_\tau &\le \frac{1}{2}\sum_{t=1}^{\tau}\lambda_t + \left(S_\tau + \left(\frac{1}{2}\sum_{t=1}^{\tau}\lambda_t\right)^2\right)^{\frac{1}{2}} \\ &\le \sum_{t=1}^{\tau}\lambda_t + \sqrt{S_\tau} \qquad (\text{ using } \sqrt{a+b} \le \sqrt{a} + \sqrt{b} \text{ for } a, b \ge 0) \end{aligned}$$

将上述不等式用原始变量表示，可得:

$$\left(\prod_{i=1}^{\tau}\rho_i\right)^{-\frac{1}{2}}u_{\tau+1} \le \sum_{t=1}^{\tau}2\epsilon\left(\prod_{i=1}^{t}\rho_i\right)^{-\frac{1}{2}} + \left(u_1^2 + 2\sigma^2\sum_{t=1}^{\tau}{\eta'_t}^2\left(\prod_{i=1}^{t}\rho_i\right)^{-1}\right)^{\frac{1}{2}}$$

$$u_{\tau+1} \le 2\epsilon\sum_{t=1}^{\tau}\left(\prod_{i=1}^{t}\rho_i\right)^{-\frac{1}{2}}\left(\prod_{i=1}^{\tau}\rho_i\right)^{\frac{1}{2}} + u_1\left(\prod_{i=1}^{\tau}\rho_i\right)^{\frac{1}{2}} + \sqrt{2}\sigma\left(\sum_{t=1}^{\tau}{\eta'_t}^2\left(\prod_{i=1}^{t}\rho_i\right)^{-1}\right)^{\frac{1}{2}}\left(\prod_{i=1}^{\tau}\rho_i\right)^{\frac{1}{2}}$$

(a) 恒定步长：选择恒定步长 $\eta'_t = \eta = \frac{1}{2L}$ 意味着 $\rho_i = \rho = 1 - \mu\eta = 1 - \frac{\mu}{2L}$。将其代入前一个不等式可得:

$$\begin{aligned} u_{\tau+1} &\le 2\epsilon\rho^{\frac{\tau}{2}}\sum_{t=1}^{\tau}\rho^{-\frac{t}{2}} + u_1\rho^{\frac{\tau}{2}} + \sqrt{2}\sigma\eta\,\rho^{\frac{\tau}{2}}\sqrt{\sum_{t=1}^{\tau}\rho^{-t}} \\ &\le \frac{2\epsilon}{1-\sqrt{\rho}} + u_1\rho^{\frac{\tau}{2}} + \frac{\sqrt{2}\sigma\eta}{\sqrt{1-\rho}} \qquad (\text{applying the formula for finite geometric series}) \end{aligned}$$

$$\mathbb{E}\|z_{T+1} - z^*\| \le \|z_1 - z^*\|\left(1 - \frac{1}{2\kappa}\right)^{\frac{T}{2}} + \frac{\sigma}{\sqrt{\mu L}} + \frac{2\epsilon}{1-\sqrt{1-\frac{1}{2\kappa}}}.$$

（b）指数步长。

从式（10）出发，并使用 Vaswani 等人（2022年）的证明，我们有

$$\begin{aligned}
\mathbb{E}\left\|\bar{z}_{t+1}-z^*\right\|_2^2 &\le \mathbb{E}\left\|z_t-z^*\right\|_2^2+4\eta_t'^2L\mathbb{E}\left\{\ell_t(z_t)-\ell_t(z^*)-\langle\nabla\ell_t(z^*),z_t-z^*\rangle\right\}-2\eta_t'\langle z_t-z^*,\nabla\ell(z_t)\\
&\quad-\nabla\ell(z^*)\rangle+2\eta_t'^2\sigma^2\\
&\le \mathbb{E}\left\|z_t-z^*\right\|_2^2+\frac{\alpha^{2t}}{L}\mathbb{E}\left\{\ell_t(z_t)-\ell_t(z^*)-\langle\nabla\ell_t(z^*),z_t-z^*\rangle\right\}-\frac{\alpha_t}{L}\langle z_t-z^*,\nabla\ell(z_t)\\
&\quad-\nabla\ell(z^*)\rangle+2\eta_t'^2\sigma^2 \qquad (\text{using } \eta_t'=\tfrac{\alpha_t}{2L})\\
&\le \mathbb{E}\left\|z_t-z^*\right\|_2^2+\frac{\alpha_t}{L}\left\{\ell(z_t)-\ell(z^*)-\langle\nabla\ell(z^*),z_t-z^*\rangle\right\}-\frac{\alpha_t}{L}\langle z_t-z^*,\nabla\ell(z_t)-\nabla\ell(z^*)\rangle+2\eta_t'^2\sigma^2\\
&\qquad (\text{using } \alpha_t\le 1)\\
&= \mathbb{E}\left\|z_t-z^*\right\|_2^2+\frac{\alpha_t}{L}\left\{\ell(z_t)-\ell(z^*)-\langle\nabla\ell(z_t),z_t-z^*\rangle\right\}+2\eta_t'^2\sigma^2\\
&\le \mathbb{E}\left\|z_t-z^*\right\|_2^2-\frac{\mu\alpha_t}{2L}\mathbb{E}\left\|z_t-z^*\right\|_2^2+2\eta_t'^2\sigma^2=\left(1-\frac{1}{2\kappa}\alpha_t\right)\mathbb{E}\left\|z_t-z^*\right\|_2^2+2\eta_t'^2\sigma^2\\
&\qquad (\text{using strong convexity of } \ell)
\end{aligned}$$

结合上述结果与引理D.2，可得：

$$\begin{aligned}
\mathbb{E}\left\|z_{t+1}-z^*\right\|_2^2 &\le \left(1-\frac{1}{2\kappa}\alpha_t\right)\mathbb{E}\left\|z_t-z^*\right\|_2^2+2\eta_t'^2\sigma^2+2\mathbb{E}[\epsilon_{t+1}\left\|z_{t+1}-z^*\right\|]\\
&\le \exp\left(-\frac{1}{2\kappa}\alpha_t\right)\mathbb{E}\left\|z_t-z^*\right\|_2^2+\frac{\sigma^2}{2L^2}\alpha^{2t}+2\mathbb{E}[\epsilon_{t+1}\left\|z_{t+1}-z^*\right\|] \qquad (1-x\le\exp(-x))\\
&\le \exp\left(-\frac{1}{2\kappa}\alpha_t\right)\mathbb{E}\left\|z_t-z^*\right\|_2^2+\frac{\sigma^2}{2L^2}\alpha^{2t}+2\epsilon\mathbb{E}[\left\|z_{t+1}-z^*\right\|] \qquad (\epsilon_{t+1}\le\epsilon)
\end{aligned}$$

从 $t=1$ 到 T展开递推，记 $u_t:=\mathbb{E}\left\|z_t-z^*\right\|$，并应用詹森不等式推导出 $u_{T+1}^2\le\mathbb{E}\left\|z_{T+1}-z^*\right\|_2^2$，可得：

$$u_{T+1}^2\le u_1^2\exp\left(-\frac{1}{2\kappa}\sum_{t=1}^{T}\alpha_t\right)+\frac{\sigma^2}{2L^2}\sum_{t=1}^{T}\alpha^{2t}\exp\left(-\frac{1}{2\kappa}\sum_{i=t+1}^{T}\alpha_i\right)+2\epsilon\sum_{t=1}^{T}\exp\left(-\frac{1}{2\kappa}\sum_{i=t+1}^{T}\alpha_i\right)u_{t+1}$$

将两边同时乘以 $\exp\left(\frac{1}{2\kappa}\sum_{t=1}^{T}\alpha_t\right)$，可得：

$$\left(\exp\left(\frac{1}{4\kappa}\sum_{t=1}^{T}\alpha_t\right)u_{T+1}\right)^2\le u_1^2+\frac{\sigma^2}{2L^2}\sum_{t=1}^{T}\alpha^{2t}\exp\left(\frac{1}{2\kappa}\sum_{i=1}^{t}\alpha_i\right)+2\epsilon\sum_{t=1}^{T}\exp\left(\frac{1}{2\kappa}\sum_{i=1}^{t}\alpha_i\right)u_{t+1}.$$

现在定义 $v_\tau:=\exp\left(\frac{1}{4\kappa}\sum_{t=1}^{\tau}\alpha_t\right)u_{\tau+1}$，$S_\tau:=u_1^2+\frac{\sigma^2}{2L^2}\sum_{t=1}^{\tau}\alpha^{2t}\exp\left(\frac{1}{2\kappa}\sum_{i=1}^{t}\alpha_i\right)$ 以及 $\lambda_t:=2\epsilon\exp\left(\frac{1}{4\kappa}\sum_{i=1}^{t}\alpha_i\right)$。注意，$S_\tau$是递增的，且 $S_0=u_1^2=v_0^2\ge 0$和 $\lambda_t>0$，$v_\tau>0$。应用引理 D.10，与固定步长情形类似，对于 $\tau=T$，我们得到：

$$\exp\left(\frac{1}{4\kappa}\sum_{t=1}^{T}\alpha_t\right)u_{T+1}\le\left(u_1^2+\frac{\sigma^2}{2L^2}\sum_{t=1}^{T}\alpha^{2t}\exp\left(\frac{1}{2\kappa}\sum_{i=1}^{t}\alpha_i\right)\right)^{1/2}+2\epsilon\sum_{t=1}^{T}\exp\left(\frac{1}{4\kappa}\sum_{i=1}^{t}\alpha_i\right)$$

两边同时乘以 $\exp\left(-\frac{1}{4\kappa}\sum_{t=1}^{T}\alpha_t\right)$，得到

$$u_{T+1} \le \left(\exp\left(-\frac{1}{2\kappa}\underbrace{\sum_{t=1}^{T}\alpha_t}_{:=A}\right)u_1^2 + \frac{\sigma^2}{2L^2}\underbrace{\sum_{t=1}^{T}\alpha^{2t}\exp\left(\frac{1}{2\kappa}\sum_{i=t+1}^{T}\alpha_i\right)}_{:=B_T}\right)^{1/2} + 2\epsilon\underbrace{\sum_{t=1}^{T}\exp\left(-\frac{1}{4\kappa}\sum_{i=t+1}^{T}\alpha_i\right)}_{C_T}$$
$$= \left(u_1^2\exp\left(-\frac{1}{2\kappa}A\right) + \frac{\sigma^2}{2L^2}B_T\right)^{1/2} + 2\epsilon\, C_T$$

为了界定 A，我们使用引理 D.6并得到

$$u_1^2\,\exp\left(-\frac{1}{2\kappa}A\right) \le \|z_1 - z^*\|_2^2\underbrace{\exp\left(\frac{1}{2\kappa}\frac{2\beta}{\ln(T/\beta)}\right)}_{:=c_1^2}\exp\left(-\frac{T}{2\kappa}\frac{\alpha}{\ln(T/\beta)}\right)$$

为了界定 B_T，我们使用引理 D.7

$$B_T \le \frac{16\kappa^2 c_1^2(\ln(T/\beta))^2}{e^2\alpha^2 T}$$

最后，利用引理 D.8对 C_T进行界定，我们得到

$$\mathbb{E}\,\|z_{T+1} - z^*\| \le \left(c_1^2\exp\left(-\frac{T}{2\kappa}\frac{\alpha}{\ln(T/\beta)}\right)\|z_1 - z^*\|_2^2 + \frac{16\kappa^2 c_1^2(\ln(T/\beta))^2}{2L^2e^2\alpha^2T}\sigma^2\right)^{1/2} + 2\epsilon\underbrace{\exp\left(\frac{\beta\ln(T)}{2\kappa\ln(T/\beta)}\right)}_{c_2}$$

$$\implies \mathbb{E}\,\|z_{T+1} - z^*\| \le c_1\,\exp\left(-\frac{T}{4\kappa}\frac{\alpha}{\ln(T/\beta)}\right)\|z_1 - z^*\| + \frac{4\kappa c_1(\ln(T/\beta))}{Le\alpha\sqrt{T}}\sigma + 2\epsilon\, c_2\,.$$

□

D.3. 控制投影误差

回顾一下理论分析所需的以下定义：

$$\tilde{\theta}_{t+1} := \arg\min_{\theta} \tilde{g}_t(\theta); \quad \tilde{g}_t(\theta) := \ell_{i_t}(z_t) + \frac{\partial \ell_{i_t}(z_t)}{\partial z^{i_t}} \left[f_{i_t}(\theta) - z_t^{i_t}\right] + \frac{1}{2\eta_t}\left[f_{i_t}(\theta) - z_t^{i_t}\right]^2 \quad ; \quad \tilde{z}_{t+1} = f(\theta_{t+1})$$

$$\bar{\theta}_{t+1} := \arg\min_{\theta} \tilde{q}_t(\theta) \quad ; \quad \tilde{q}_t(\theta) := \ell_{i_t}(z_t) + \frac{\partial \ell_{i_t}(z_t)}{\partial z^{i_t}} \left[f_{i_t}(\theta) - z_t^{i_t}\right] + \frac{1}{2\eta_t'}\left\|f(\theta) - z_t\right\|_2^2 \quad ; \quad \bar{z}_{t+1} = f(\bar{\theta}_{t+1})$$

$$\theta'_{t+1} := \arg\min_{\theta} g_t(\theta) \quad ; \quad g_t(\theta) := \frac{1}{n}\left[\left[\sum_i \ell_i(z_t) + \langle \nabla \ell_i(z_t), f(\theta) - z_t\rangle\right] + \frac{1}{2\eta_t}\left\|f(\theta) - z_t\right\|_2^2\right]$$

$$z_{t+1} = f(\theta_{t+1})\,,$$

其中， θ_{t+1} 是通过对 $\tilde{g}_t(\theta)$运行 m_t 次梯度下降而得到的。我们将使用这些定义来证明以下命题，以控制每次迭代中的投影误差。

命题 *4.3.* 假设 *(i)* f 是 L_f-利普希茨连续的，*(ii)* $\tilde{g}_t$ 是 μ_g 强凸且 L_g 光滑，其中 $\kappa_g := L_g/\mu_g$[7]，*(iii)* $\tilde{q}_t$ 是 μ_q 强凸的，*(iv)* $\zeta_t^2 := {}^{8}\min_{\{\mu_g,\mu_q\}}\left(\left[\min\left\{\mathbb{E}_{i_t}\left[\tilde{g}_t\right]\right\} - \mathbb{E}_{i_t}\left[\min\left\{\tilde{g}_t\right\}\right]\right] + \left[\min\left\{\mathbb{E}_{i_t}\left[\tilde{q}_t\right]\right\} - \mathbb{E}_{i_t}\left[\min\left\{\tilde{q}_t\right\}\right]\right]\right)$，*(v)* $\sigma_z^2 := \min_{z\in\mathcal{Z}}\left\{\mathbb{E}_{i_t}\left[\ell_{i_t}(z)\right]\right\} - \mathbb{E}_{i_t}\left[\min_{z\in\mathcal{Z}}\left\{\ell_{i_t}(z)\right\}\right]$，若使用适当的步长，通过 m_t 次梯度下降内循环来最小化$\tilde{g}_t$ ，则 $\mathbb{E}[\epsilon_{t+1}^2] \leq L_f^2\,\zeta_t^2 + \frac{4L_f^2}{\mu_g}\left[\exp\left(-m_t/\kappa_g\right)\left[\mathbb{E}[\ell(z_t) - \ell(z^*)] + \sigma_z^2\right]\right]$。

证明。 由于我们通过 θ_{t+1} 最小化 $\tilde{g}_t(\theta)$ ，使用 m_t 次梯度下降迭代，并从 θ_t开始，根据梯度下降的收敛保证（Nesterov，2003），

$$\begin{aligned}
\left\|\tilde{\theta}_{t+1} - \theta_{t+1}\right\|_2^2 &\leq \exp\left(-m_t/\kappa_g\right)\left\|\tilde{\theta}_{t+1} - \theta_t\right\|_2^2 \\
\left\|\theta_{t+1} - \bar{\theta}_{t+1}\right\|_2^2 &= \left\|\theta_{t+1} - \tilde{\theta}_{t+1} + \tilde{\theta}_{t+1} - \bar{\theta}_{t+1}\right\|_2^2 \leq 2\left\|\theta_{t+1} - \tilde{\theta}_{t+1}\right\|_2^2 + 2\left\|\tilde{\theta}_{t+1} - \bar{\theta}_{t+1}\right\|_2^2 \\
&\qquad\qquad (\|a+b\|_2^2 \leq 2\|a\|_2^2 + 2\|b\|_2^2) \\
&\leq 2\exp\left(-m_t/\kappa_g\right)\left\|\tilde{\theta}_{t+1} - \theta_t\right\|_2^2 + 2\left\|\tilde{\theta}_{t+1} - \bar{\theta}_{t+1}\right\|_2^2 \\
&\leq \frac{4}{\mu_g}\exp\left(-m_t/\kappa_g\right)\left[\tilde{g}_t(\theta_t) - \tilde{g}_t(\tilde{\theta}_{t+1})\right] + 2\left\|\tilde{\theta}_{t+1} - \bar{\theta}_{t+1}\right\|_2^2
\end{aligned}$$

对 i_t取期望，

$$\mathbb{E}_{i_t}\left\|\theta_{t+1} - \bar{\theta}_{t+1}\right\|_2^2 \leq \frac{4}{\mu_g}\left[\exp\left(-m_t/\kappa_g\right)\,\mathbb{E}_{i_t}[\tilde{g}_t(\theta_t) - \tilde{g}_t(\tilde{\theta}_{t+1})]\right] + 2\mathbb{E}\left\|\tilde{\theta}_{t+1} - \bar{\theta}_{t+1}\right\|_2^2$$

我们先简化 $\mathbb{E}\left\|\tilde{\theta}_{t+1} - \bar{\theta}_{t+1}\right\|_2^2$。

$$\begin{aligned}
\mathbb{E}\left\|\tilde{\theta}_{t+1} - \bar{\theta}_{t+1}\right\|_2^2 &= \mathbb{E}\left\|\tilde{\theta}_{t+1} - \theta'_{t+1} + \theta'_{t+1} - \bar{\theta}_{t+1}\right\|_2^2 \\
&\leq 2\mathbb{E}\left\|\tilde{\theta}_{t+1} - \theta'_{t+1}\right\|_2^2 + 2\mathbb{E}\left\|\theta'_{t+1} - \bar{\theta}_{t+1}\right\|_2^2 \\
&\leq \frac{4}{\mu_g}\mathbb{E}[\tilde{g}_t(\theta'_{t+1}) - \tilde{g}_t(\tilde{\theta}_{t+1})] + \frac{4}{\mu_q}\mathbb{E}[\tilde{q}_t(\theta'_{t+1}) - \tilde{q}_t(\bar{\theta}_{t+1})] \\
&= \frac{4}{\mu_g}\left[\min\left\{\mathbb{E}_{i_t}\left[\tilde{g}_t\right]\right\} - \mathbb{E}_{i_t}\left[\min\left\{\tilde{g}_t\right\}\right]\right] + \frac{4}{\mu_q}\left[\min\left\{\mathbb{E}_{i_t}\left[\tilde{q}_t\right]\right\} - \mathbb{E}_{i_t}\left[\min\left\{\tilde{q}_t\right\}\right]\right] \\
&\qquad\qquad (\text{Since } \mathbb{E}[\tilde{g}] = \mathbb{E}[\tilde{q}] = g)
\end{aligned}$$

[7]为简单起见，我们考虑强凸函数；这些结果也应当可以推广到 $\tilde{g}$ 为非凸但满足 PL 不等式，或为弱凸的情形。

$$2\mathbb{E}\left\|\tilde{\theta}_{t+1}-\bar{\theta}_{t+1}\right\|_2^2 \leq \underbrace{\frac{8}{\min\{\mu_g,\mu_q\}}\left(\left[\min\left\{\mathbb{E}_{i_t}\left[\tilde{g}_t\right]\right\}-\mathbb{E}_{i_t}\left[\min\left\{\tilde{g}_t\right\}\right]\right]+\left[\min\left\{\mathbb{E}_{i_t}\left[\tilde{q}_t\right]\right\}-\mathbb{E}_{i_t}\left[\min\left\{\tilde{q}_t\right\}\right]\right]\right)}_{:=\zeta_t^2}$$

$$\implies 2\mathbb{E}\left\|\tilde{\theta}_{t+1}-\bar{\theta}_{t+1}\right\|_2^2 \leq \zeta_t^2$$

利用上述关系,

$$\begin{aligned}
\mathbb{E}_{i_t}\left\|\theta_{t+1}-\bar{\theta}_{t+1}\right\|_2^2 &\leq \frac{4}{\mu_g}\left[\exp\left(-m_t/\kappa_g\right)\,\mathbb{E}_{i_t}[\tilde{g}_t(\theta_t)-\tilde{g}_t(\tilde{\theta}_{t+1})]\right]+\zeta_t^2 \\
&\leq \frac{4}{\mu_g}\left[\exp\left(-m_t/\kappa_g\right)\,\mathbb{E}_{i_t}[h_t(\theta_t)-h_t(\tilde{\theta}_{t+1})]\right]+\zeta_t^2 \\
&\qquad\qquad \text{(Since } \tilde{g}_t(\theta_t)=h_t(\theta_t) \text{ and } \tilde{g}_t(\theta)\geq h_t(\theta) \text{ for all } \theta\text{)} \\
&= \frac{4}{\mu_g}\left[\exp\left(-m_t/\kappa_g\right)\,\mathbb{E}_{i_t}[h_t(\theta_t)-h_t^*+h_t^*-h_t(\tilde{\theta}_{t+1})]\right]+\zeta_t^2 \qquad (h_t^*:=\min_\theta h_t(\theta)) \\
&\leq \frac{4}{\mu_g}\left[\exp\left(-m_t/\kappa_g\right)\,\mathbb{E}_{i_t}[h_t(\theta_t)-h_t^*]+\zeta_t^2\right] \qquad \text{(Since } h_t^*\leq h_t(\theta) \text{ for all } \theta\text{)} \\
&= \frac{4}{\mu_g}\left[\exp\left(-m_t/\kappa_g\right)\,[\mathbb{E}_{i_t}[h_t(\theta_t)-h_t(\theta^*)]+\mathbb{E}_{i_t}[h_t(\theta^*)-h_t^*]]\right]+\zeta_t^2 \\
&= \frac{4}{\mu_g}\left[\exp\left(-m_t/\kappa_g\right)\,[\mathbb{E}_{i_t}[\ell_t(z_t)-\ell_t(z^*)]+\mathbb{E}_{i_t}[\ell_t(z^*)-\ell_t^*]]\right]+\zeta_t^2 \\
&\qquad\qquad \text{(Si度 } h(\theta) \quad \ell(\theta) \quad \ell(\ ))
\end{aligned}$$

$$\mathbb{E}_{i_t}[\left\|\theta_{t+1}-\bar{\theta}_{t+1}\right\|_2^2] \leq \frac{4}{\mu_g}\left[\exp\left(-m_t/\kappa_g\right)\,[\mathbb{E}_{i_t}[\ell_t(z_t)-\ell_t(z^*)]+\mathbb{E}_{i_t}[\ell_t(z^*)-\ell_t^*]]\right]+\zeta_t^2$$

$$\leq \frac{4}{\mu_g}\left[\exp\left(-m_t/\kappa_g\right)\,\left[[\ell(z_t)-\ell(z^*)]+\underbrace{\mathbb{E}_{i_t}[\ell_t(z^*)-\ell_t^*]}_{:=\sigma_z^2}\right]\right]+\zeta_t^2$$

(由于 z_t 和 z^* 均独立于 ℓ_t 和 $\mathbb{E}_{i_t}[\ell_t]=\ell$中的随机性)

$$\implies \mathbb{E}[\left\|\theta_{t+1}-\bar{\theta}_{t+1}\right\|_2^2] \leq \frac{4}{\mu_g}\left[\exp\left(-m_t/\kappa_g\right)\,\left[\ell(z_t)-\ell(z^*)+\sigma_z^2\right]\right]+\zeta_t^2$$

现在，我们将利用上述不等式和 f的利普希茨连续性来界定 $\mathbb{E}[\epsilon_{t+1}^2]$。

$$\begin{aligned}
&\mathbb{E}[\epsilon_{t+1}^2]=\|z_{t+1}-\bar{z}_{t+1}\|_2^2=\left\|f(\theta_{t+1})-f(\bar{\theta}_{t+1})\right\|_2^2\leq L_f^2\left\|\theta_{t+1}-\bar{\theta}_{t+1}\right\|_2^2 \qquad \text{(Since } f \text{ is } L_f\text{-Lipschitz)} \\
&\implies \mathbb{E}[\epsilon_{t+1}^2]\leq\frac{4L_f^2}{\mu_g}\left[\exp\left(-m_t/\kappa_g\right)\,\left[\ell(z_t)-\ell(z^*)+\sigma_z^2\right]\right]+L_f^2\,\zeta_t^2
\end{aligned}$$

对从迭代 $k=0$到 t 中的随机性取期望,

$$\mathbb{E}[\epsilon_{t+1}^2]\leq\frac{4L_f^2}{\mu_g}\left[\exp\left(-m_t/\kappa_g\right)\,\left[\mathbb{E}[\ell(z_t)-\ell(z^*)]+\sigma_z^2\right]\right]+L_f^2\,\zeta_t^2$$

□

D.4. 展示 ζ^2 项必要性的示例

命题*4.4.* 考虑使用 *SSO* 最小化两个一维二次函数之和 $h(\theta) := h_1(\theta)+h_2(\theta)$ *2*，其中 $h_1(\theta) := \frac{1}{2}(\theta-1)^2$和 $h_2(\theta) = \frac{1}{2}(2\theta+1/2)^2$,，其中 $m_t = \infty$ 和 $\eta_t = c\,\alpha_t$ 为任意迭代序列，且 $c \in (0,1]$为任意常数。*SSO* 收敛到解的一个邻域，具体而言，如果 θ^* 是 h 和 $\theta_1 > 0$的极小化点，则，$\mathbb{E}(\theta_T - \theta^*) \geq min\left(\theta_1, \frac{3}{8}\right)$。

证明。 我们先计算 $\theta^* := \arg\min h(\theta)$。

$$h(\theta) = \frac{1}{4}(\theta-1)^2 + \frac{1}{4}\left(2\theta+\frac{1}{2}\right)^2 = \frac{5}{4}\theta^2 + \frac{1}{4} + \frac{1}{16} \Rightarrow \theta^* = 0$$

对于 h_1,

$$\ell_1(z) = \frac{1}{2}(z-1)^2 \quad \text{where } z = \theta$$
$$\tilde{g}_t(\theta) := \frac{1}{2}(\theta_t-1)^2 + (\theta_t-1)(\theta_t-\theta) + \frac{1}{2\eta_t}(\theta-\theta_t)^2$$

如果 $m_t = \infty$，SSO 将精确地最小化 $\tilde{g}_t$。由于 $\nabla\tilde{g}_t(\theta_{t+1}) = 0$,

$$\implies \frac{1}{\eta_t}(\theta_{t+1}-\theta_t) = -(\theta_t-1) \implies \theta_{t+1} = \theta_t - \eta_t(\theta_t-1) \implies \theta_{t+1} = \theta_t - \eta_t\nabla h_1(\theta_t)$$

同理，对于 h_2,

$$\ell_2(z) = \frac{1}{2}(z+1/2)^2 \quad \text{where } z = 2\theta$$
$$\tilde{g}_t(\theta) := \frac{1}{2}(2\theta_t+1/2)^2 + (2\theta_t+1/2)(2\theta_t-2\theta) + \frac{1}{2\eta_t}(2\theta-2\theta_t)^2$$

如果 $m_t = \infty$，SSO 将精确地最小化 $\tilde{g}_t$。由于 $\nabla\tilde{g}_t(\theta_{t+1}) = 0$,

$$\implies \frac{4}{\eta_t}(\theta_{t+1}-\theta_t) = -(4\theta_t+1) \implies \theta_{t+1} = \theta_t - \eta_t(\theta_t+1/4) \implies \theta_{t+1} = \theta_t - \frac{\eta_t}{4}\nabla h_2(\theta_t)$$

如果 $i_t = 1$ 和 $\eta = c\,\alpha_t$

$$\theta_{t+1} = \theta_t - c\,\alpha_t(\theta_t-1) = c\,\alpha_t + (1-c\alpha_t)\theta_t$$

如果 $i_k = 2$,

$$\theta_{t+1} = \theta_t - c\,\alpha_t\frac{2}{4}(2\theta_t+\frac{1}{2}) = (1-c\,\alpha_t)\theta_t - \frac{1}{4}c\,\alpha_t$$

Then

$$\mathbb{E}\theta_{t+1} = (1-c\,\alpha_t)\theta_t + \frac{1}{2}c\,\alpha_t - \frac{1}{8}c\,\alpha_t = (1-c\,\alpha_t)\theta_t + \frac{3}{8}c\,\alpha_t$$

and

$$\mathbb{E}\theta_T = \mathbb{E}(\theta_T-\theta^*) = (\theta_1-\theta^*)\prod_{t=1}^{T}(1-c\,\alpha_t) + \frac{3}{8}\sum_{t=1}^{T}2(1-c)\alpha_t\prod_{i=t+1}^{T}(1-c\,\alpha_i)$$

利用引理 D.9 以及以下事实 $c\,\alpha_t \leq 1$ 对所有 t，可得若 $\theta_1 - \theta^* = \theta_1 > 0$，则:

$$\mathbb{E}(\theta_T-\theta^*) \geq \min\left(\theta_1, \frac{3}{8}\right).$$

□

D.5. 辅助引理

引理D.4、引理D.6和引理 D.7 的证明可见于（Vaswani 等人，2022年）。**引理 D.4**。
对所有 $x > 1$,

$$\frac{1}{x-1} \leq \frac{2}{\ln(x)}$$

证明。 对于 $x > 1$，我们有

$$\frac{1}{x-1} \leq \frac{2}{\ln(x)} \iff \ln(x) < 2x - 2$$

定义 $f(x) = 2x - 2 - \ln(x)$。我们有 $f'(x) = 2 - \frac{1}{x}$。因此，对于 $x \geq 1$，我们有 $f'(x) > 0$，所以 f在 $[1, \infty)$ 上递增。此外，我们有 $f(1) = 2 - 2 - \ln(1) = 0$，这表明对于所有 $x > 1$， $f(x) \geq 0$，从而证明完毕。 □

引理 D.5。对于所有 x， $\gamma > 0$,

$$\exp(-x) \leq \left(\frac{\gamma}{ex}\right)^{\gamma}$$

证明。 设 $x > 0$。定义 $f(\gamma) = \left(\frac{\gamma}{ex}\right)^{\gamma} - \exp(-x)$。我们有

$$f(\gamma) = \exp\left(\gamma \ln(\gamma) - \gamma \ln(ex)\right) - \exp(-x)$$

and

$$f'(\gamma) = \left(\gamma \cdot \frac{1}{\gamma} + \ln(\gamma) - \ln(ex)\right) \exp\left(\gamma \ln(\gamma) - \gamma \ln(ex)\right)$$

Thus

$$f'(\gamma) \geq 0 \iff 1 + \ln(\gamma) - \ln(ex) \geq 0 \iff \gamma \geq \exp\left(\ln(ex) - 1\right) = x$$

因此， f在 $(0, x]$ 上递减，并在 $[x, \infty)$ 上递增。此外,

$$f(x) = \left(\frac{x}{ex}\right)^{x} - \exp(-x) = \left(\frac{1}{e}\right)^{x} - \exp(-x) = 0$$

因此，对于所有 $\gamma > 0$， $f(\gamma) \geq 0$，这就证明了该引理。 □

引理 D.6. 假设 $\alpha < 1$， 我们有

$$A := \sum_{t=1}^{T} \alpha^t \geq \frac{\alpha T}{\ln(T/\beta)} - \frac{2\beta}{\ln(T/\beta)}$$

证明。

$$\sum_{t=1}^{T} \alpha^t = \frac{\alpha - \alpha^{T+1}}{1-\alpha} = \frac{\alpha}{1-\alpha} - \frac{\alpha^{T+1}}{1-\alpha}$$

我们有

$$\frac{\alpha^{T+1}}{1-\alpha} = \frac{\alpha\beta}{T(1-\alpha)} = \frac{\beta}{T} \cdot \frac{1}{1/\alpha - 1} \leq \frac{\beta}{T} \cdot \frac{2}{\ln(1/\alpha)} = \frac{\beta}{T} \cdot \frac{2}{\frac{1}{T}\ln(T/\beta)} = \frac{2\beta}{\ln(T/\beta)} \tag{11}$$

wher在不等式中，我们使用了引理 D.4 以及以下事实 $1/\alpha > 1$。代回 A 后得到 t,

$$\begin{aligned} A &\geq \frac{\alpha}{1-\alpha} - \frac{2\beta}{\ln(T/\beta)} \\ &\geq \frac{\alpha}{\ln(1/\alpha)} - \frac{2\beta}{\ln(T/\beta)} && (1-x \leq \ln(\tfrac{1}{x})) \\ &= \frac{\alpha T}{\ln(T/\beta)} - \frac{2\beta}{\ln(T/\beta)} \end{aligned}$$

□

引理 D.7. 对于 $\alpha = \left(\frac{\beta}{T}\right)^{1/T}$ 和任意 $\kappa > 0$,

$$\sum_{t=1}^{T} \alpha^{2t} \exp\left(-\frac{1}{2\kappa}\sum_{i=t+1}^{T}\alpha^i\right) \leq \frac{16\kappa^2 c_2 (\ln(T/\beta))^2}{e^2\alpha^2 T}$$

其中 $c_2 =$ 指数 $\left(\frac{1}{2\kappa} \frac{2\beta}{\ln(T/\beta)}\right)$

证明。 首先，注意到,

$$\sum_{i=t+1}^{T}\alpha^i = \frac{\alpha^{t+1} - \alpha^{T+1}}{1-\alpha}$$

我们有

$$\frac{\alpha^{T+1}}{1-\alpha} = \frac{\alpha\beta}{T(1-\alpha)} = \frac{\beta}{T}\cdot\frac{1}{1/\alpha - 1} \leq \frac{\beta}{T}\cdot\frac{2}{\ln(1/\alpha)} = \frac{\beta}{T}\cdot\frac{2}{\frac{1}{T}\ln(T/\beta)} = \frac{2\beta}{\ln(T/\beta)}$$

在该不等式中，我们使用了 D.4 以及以下事实： $1/\alpha > 1$。这些关系意味着,

$$\begin{aligned} &\sum_{i=t+1}^{T}\alpha^i \geq \frac{\alpha^{t+1}}{1-\alpha} - \frac{2\beta}{\ln(T/\beta)} \\ &\implies \exp\left(-\frac{1}{2\kappa}\sum_{i=t+1}^{T}\alpha^i\right) \leq \exp\left(-\frac{1}{2\kappa}\frac{\alpha^{t+1}}{1-\alpha} + \frac{1}{2\kappa}\frac{2\beta}{\ln(T/\beta)}\right) = c_2\exp\left(-\frac{1}{2\kappa}\frac{\alpha^{t+1}}{1-\alpha}\right) \end{aligned}$$

于是有

$$\begin{aligned} \sum_{t=1}^{T}\alpha^{2t}\exp\left(-\frac{1}{2\kappa}\sum_{i=t+1}^{T}\alpha^i\right) &\leq c_2\sum_{t=1}^{T}\alpha^{2t}\exp\left(-\frac{1}{2\kappa}\frac{\alpha^{t+1}}{1-\alpha}\right) \\ &\leq c_2\sum_{t=1}^{T}\alpha^{2t}\left(\frac{2(1-\alpha)2\kappa}{e\alpha^{t+1}}\right)^2 && \text{(Lemma D.5)} \\ &= \frac{16\kappa^2 c_2}{e^2\alpha^2}T(1-\alpha)^2 \\ &\leq \frac{16\kappa^2 c_2}{e^2\alpha^2}T(\ln(1/\alpha))^2 \\ &= \frac{16\kappa^2 c_2(\ln(T/\beta))^2}{e^2\alpha^2 T} \end{aligned}$$

□

引理 D.8. 对于 $\alpha = \left(\frac{\beta}{T}\right)^{1/T}$ 和任意 $\zeta > 0$,

$$\sum_{t=1}^{T} \exp\left(-\frac{1}{\zeta}\sum_{i=t}^{T}\alpha^i\right) \leq \exp\left(\frac{2\beta\ln(T)}{\zeta\ln(T/\beta)}\right)$$

证明。 首先，注意到,

$$\sum_{i=t}^{T}\alpha^i = \frac{\alpha^t - \alpha^{T+1}}{1-\alpha} \geq -\frac{\alpha^{T+1}}{1-\alpha}$$

我们有

$$\frac{\alpha^{T+1}}{1-\alpha} = \frac{\alpha\beta}{T(1-\alpha)} = \frac{\beta}{T}\cdot\frac{1}{1/\alpha - 1} \leq \frac{\beta}{T}\cdot\frac{2}{\ln(1/\alpha)} = \frac{\beta}{T}\cdot\frac{2}{\frac{1}{T}\ln(T/\beta)} = \frac{2\beta}{\ln(T/\beta)}$$

其中，不等式中用到了引理 D.4 以及以下事实： $1/\alpha > 1$。利用上述界，我们有

$$\sum_{t=1}^{T} \exp\left(-\frac{1}{\zeta}\sum_{i=t}^{T}\alpha^i\right) \leq \sum_{t=1}^{T}\exp\left(\frac{2\beta}{\zeta\ln(T/\beta)}\right) = \exp\left(\frac{2\beta\ln(T)}{\zeta\ln(T/\beta)}\right)$$

□

引理 D.9。 对于任意序列 α_t

$$\prod_{t=1}^{T}(1-\alpha_t) + \sum_{t=1}^{T}\alpha_t\prod_{i=t+1}^{T}(1-\alpha_i) = 1$$

证明。 我们对 T 使用归纳法来证明这一点。对于 $T = 1$,

$$(1-\alpha_1) + \alpha_1 = 1$$

归纳步骤:

$$\begin{aligned}
\prod_{t=1}^{T+1}(1-\alpha_t) + \sum_{t=1}^{T+1}\alpha_t\prod_{i=t+1}^{T+1}(1-\alpha_i) &= (1-\alpha_{T+1})\prod_{t=1}^{T}(1-\alpha_t) + \left(\alpha_{T+1} + \sum_{t=1}^{T}\alpha_t\prod_{i=t+1}^{T+1}(1-\alpha_i)\right) \\
&= (1-\alpha_{T+1})\prod_{t=1}^{T}(1-\alpha_t) + \left(\alpha_{T+1} + (1-\alpha_{T+1})\sum_{t=1}^{T}\alpha_t\prod_{i=t+1}^{T}(1-\alpha_i)\right) \\
&= (1-\alpha_{T+1})\left(\underbrace{\prod_{t=1}^{T}(1-\alpha_t) + \sum_{t=1}^{T}\alpha_t\prod_{i=t+1}^{T}(1-\alpha_i)}_{=1}\right) + \alpha_{T+1} \\
&\qquad\qquad \text{(Induction hypothesis)} \\
&= (1-\alpha_{T+1}) + \alpha_{T+1} = 1
\end{aligned}$$

□

引理 D.10。*(Schmidt 等人，2011年，引理 1)*。假设非负数列 $\{v_\tau\}$ 对于 $\tau \geq 1$ 满足以下递推：

$$v_\tau^2 \leq S_\tau + \sum_{t=1}^{\tau} \lambda_t v_t \,,$$

其中，$\{S_\tau\}$ 是一个递增数列，使得 $S_0 \geq v_0^2$ 且 $\lambda_t \geq 0$。那么对于所有 $\tau \geq 1$,

$$v_\tau \leq \frac{1}{2}\sum_{t=1}^{\tau} \lambda_t + \left(S_\tau + \left(\frac{1}{2}\sum_{t=1}^{\tau} \lambda_t \right)^2 \right)^{1/2} .$$

证明。我们通过归纳法证明这个引理。对于 $\tau = 1$，有

$$v_1^2 \leq S_1 + \lambda_1 v_1 \implies (v_1 - \frac{\lambda_1}{2})^2 \leq S_1 + \frac{\lambda_1^2}{4} \implies v_1 \leq \left(S_1 + \left(\frac{1}{2}\lambda_1 \right)^2 \right)^{1/2} + \frac{1}{2}\lambda_1$$

归纳假设：假设结论对 $\tau \in \{1, \ldots, k\}$成立。具体而言，对于 $\tau = k$,

$$v_k \leq \frac{1}{2}\sum_{t=1}^{k} \lambda_t + \left(S_k + \left(\frac{1}{2}\sum_{t=1}^{k} \lambda_t \right)^2 \right)^{1/2} .$$

现在证明结论对 $\tau = k+1$成立。定义 $\psi_k := \sum_{t=1}^{k} \lambda_t$。注意，由于对于所有 t， $\lambda_t \geq 0$，因此 $\psi_{k+1} \geq \psi_k$。于是，$v_k \leq \frac{1}{2}\psi_k + \left(S_k + \left(\frac{1}{2}\psi_k \right)^2 \right)^{1/2}$。定义 $k^* := \arg\max_{\tau \in \{0,\ldots,k\}} v_\tau$。利用关于 $\tau = k+1$的主要假设,

$$\begin{aligned}
& v_{k+1}^2 \leq S_{k+1} + \sum_{t=1}^{k+1} \lambda_t v_t \\
\implies & v_{k+1}^2 - \lambda_{k+1} v_{k+1} \leq S_{k+1} + \sum_{t=1}^{k} \lambda_t v_t \\
\implies & \left(v_{k+1} - \frac{\lambda_{k+1}}{2} \right)^2 \leq S_{k+1} + \frac{\lambda_{k+1}^2}{4} + \sum_{t=1}^{k} \lambda_t v_t \\
& \leq S_{k+1} + \frac{\lambda_{k+1}^2}{4} + v_{k^*} \sum_{t=1}^{k} \lambda_t && \text{（因为 } v_{k^*} \text{ 是最大值）} \\
& = S_{k+1} + \frac{\lambda_{k+1}^2}{4} + v_{k^*}\psi_k && \text{(based on the definition for } \psi_k\text{)}
\end{aligned}$$

由于 $k^* \leq k$，根据归纳假设,

$$\begin{aligned}
\left(v_{k+1} - \frac{\lambda_{k+1}}{2} \right)^2 & \leq S_{k+1} + \frac{\lambda_{k+1}^2}{4} + \psi_k \left\{ \frac{1}{2}\psi_{k^*} + \left(S_{k^*} + \left(\frac{1}{2}\psi_{k^*} \right)^2 \right)^{1/2} \right\} \\
& = S_{k+1} + \frac{\lambda_{k+1}^2}{4} + \frac{1}{2}\psi_k\psi_{k^*} + \psi_k \left(S_{k^*} + \left(\frac{1}{2}\psi_{k^*} \right)^2 \right)^{1/2} \\
& \leq S_{k+1} + \frac{\lambda_{k+1}^2}{4} + \frac{1}{2}\psi_k^2 + \psi_k \left(S_{k^*} + \left(\frac{1}{2}\psi_{k^*} \right)^2 \right)^{1/2} && \text{(Since } \{\psi_\tau\} \text{ is non-decreasing and } k^* \leq k\text{)}
\end{aligned}$$

此外，由于 $\{S_\tau\}$ 是递增的，且 $k^* < k+1$,

$$
\begin{aligned}
\left(v_{k+1} - \frac{\lambda_{k+1}}{2}\right)^2 &\le S_{k+1} + \frac{\lambda_{k+1}^2}{4} + \frac{1}{2}\psi_k^2 + \psi_k \left(S_{k+1} + \left(\frac{1}{2}\psi_{k^*}\right)^2\right)^{1/2} \\
&\le S_{k+1} + \frac{\lambda_{k+1}^2}{4} + \frac{1}{2}\psi_k^2 + \psi_k \left(S_{k+1} + \left(\frac{1}{2}\psi_{k+1}\right)^2\right)^{1/2} \\
&\qquad \text{(Since } \{\psi_\tau\} \text{ is non-decreasing and } k^* < k+1) \\
&= S_{k+1} + \frac{\lambda_{k+1}^2}{4} + \frac{1}{4}\psi_k^2 + \psi_k \left(S_{k+1} + \left(\frac{1}{2}\psi_{k+1}\right)^2\right)^{1/2} + \frac{1}{4}\psi_k^2 \\
&\le S_{k+1} + \frac{1}{4}\psi_{k+1}^2 + \psi_k \left(S_{k+1} + \left(\frac{1}{2}\psi_{k+1}\right)^2\right)^{1/2} + \frac{1}{4}\psi_k^2 \\
&\qquad (a^2 + b^2 \le (a+b)^2 \text{ for } a = \psi_k > 0 \text{ and } b = \lambda_{k+1} > 0) \\
&= \underbrace{S_{k+1} + \frac{1}{4}\psi_{k+1}^2}_{:=x^2} + \underbrace{\psi_k \left(S_{k+1} + \left(\frac{1}{2}\psi_{k+1}\right)^2\right)^{1/2}}_{=2xy} + \underbrace{\frac{1}{4}\psi_k^2}_{:=y^2} \\
&= \left(\left(S_{k+1} + \left(\frac{1}{2}\psi_{k+1}\right)^2\right)^{1/2} + \frac{1}{2}\psi_k\right)^2 \\
\implies v_{k+1} - \frac{\lambda_{k+1}}{2} &\le \left(S_{k+1} + \left(\frac{1}{2}\psi_{k+1}\right)^2\right)^{1/2} + \frac{1}{2}\psi_k \\
\implies v_{k+1} &\le \left(S_{k+1} + \left(\frac{1}{2}\psi_{k+1}\right)^2\right)^{1/2} + \frac{1}{2}\psi_k + \frac{\lambda_{k+1}}{2} \\
&= \left(S_{k+1} + \left(\frac{1}{2}\psi_{k+1}\right)^2\right)^{1/2} + \frac{1}{2}\psi_{k+1}
\end{aligned}
$$

将 ψ_{k+1} 替换为其定义，即可得到所需结果。 □

E. 其他实验结果

监督学习： 我们在 LibSVM 基准测试（Chang 和 Lin，2011年）上评估我们的框架。该基准测试套件包含一组标准的凸优化问题。这里考虑两个数据集——`mushrooms`和`rcv1`、两个损失——平方损失和逻辑损失，以及四种批量大小——{25, 125, 625, 全批量}，用于线性参数化（$f = X^\top \theta$）。每种优化算法均运行500 个训练轮次（对数据进行完整遍历）。

E.1. 随机代理优化

在三个 SVMLib 基准数据集 mushrooms、ijcnn 和 rcv1、两个损失——平方损失和逻辑损失，以及四种批量大小——{25, 125, 625, 全批量}上，对 SGD、SLS、Adam、Adagrad 和 SSO 进行了比较。每次运行均按照相同的初始化方案，在三个随机种子下进行评估。所有图均采用双对数空间，以便更清晰地呈现不同优化算法之间的趋势。所有算法和批量大小均训练 500 个训练轮次，性能表示为总优化步数的函数。我们将随机代理优化（SSO）与采用标准理论 $1/2L_\theta$ 步长的 SGD、根据随机线搜索设置步长的 SGD（Vaswani 等，2019b，即 SLS），以及使用默认超参数的 Adam（Kingma 和 Ba，2015）进行比较。由于 SSO 等价于目标空间中的投影随机梯度下降，因此我们将 η（在代理定义中）设为 $1/2L$，其中 L 是 ℓ 相对于 z的光滑性。对于平方损失，L 因此设为 1；而对于逻辑损失，则设为 2。这些图表表明：（i）当步长按照理论设定时，SSO 的表现优于 SGD；（ii）SSO 与 SLS 或 Adam 具有竞争力；（iii）随着 m 增大，平均而言，随着投影误差减小，SSO 的表现得到改善。更多详情请参见随附的代码仓库。下面我们列出三种不同的步长计划：恒定步长、$\frac{1}{\sqrt{t}}$（Orabona，2019），以及 $(1/T)^{t/T}$（Vaswani 等人，2022 年）。

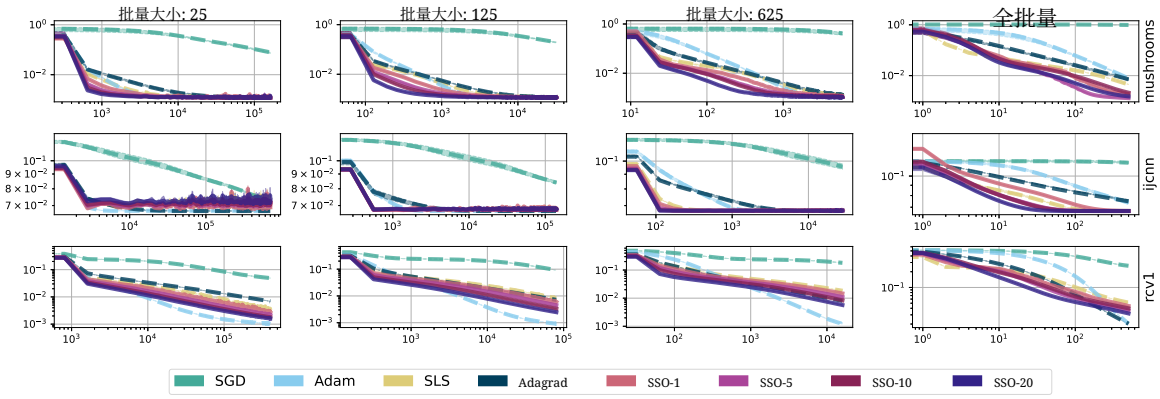


图 4：**恒定步长：**在**均方误差损失**下对比优化算法。我们注意到，SSO 的表现显著优于其参数化对应方法，且与 SLS 和 Adam 不相上下。此外，我们还注意到，在代理函数中采取更多步通常能提升表现。

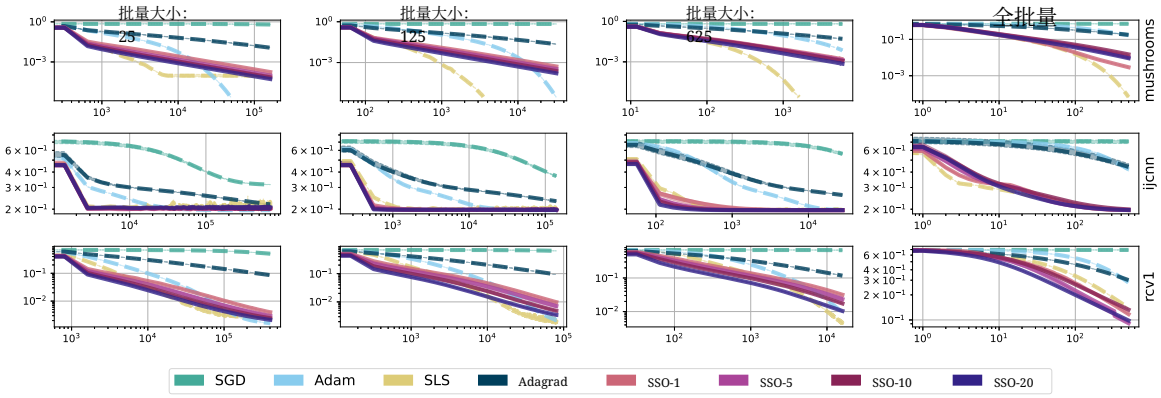


图 5：**恒定步长：**在平均**逻辑损失**下对比优化算法。我们注意到，SSO 显著优于其参数化对应方法，并且性能与 SLS 和 Adam 持平。此外，我们注意到，在代理函数中采取额外步长通常能够提升性能

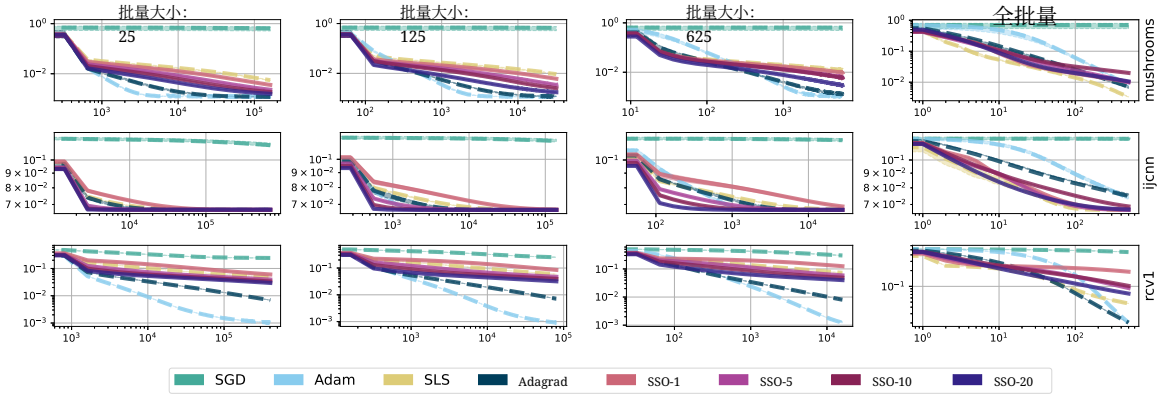


图 6：**递减步长：**在**均方误差损失**下对比优化算法。我们比较了包括采用衰减步长 $\frac{1}{\sqrt{t}}$ 的示例，以及 SSO、SGD 和 SLS。Adam（以及 Adagrad）。同样，我们注意到，在代理函数中采取额外步长通常能够提升性能。此外，递减步长似乎有助于保持严格的单调改进。

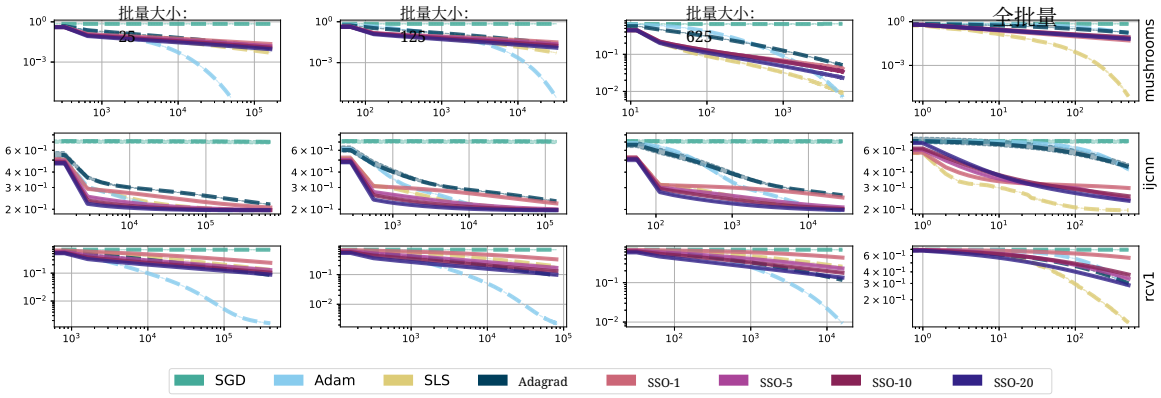


图 7：**递减步长：** 在**逻辑损失**下比较优化算法。我们比较了采用衰减步长 $\frac{1}{\sqrt{t}}$ 的示例，以及 SSO、SGD 和 SLS。Adam（以及 Adagrad）。同样，我们注意到，在代理函数上采取额外的步骤通常能提升性能。此外，递减步长似乎有助于保持严格的单调改进。

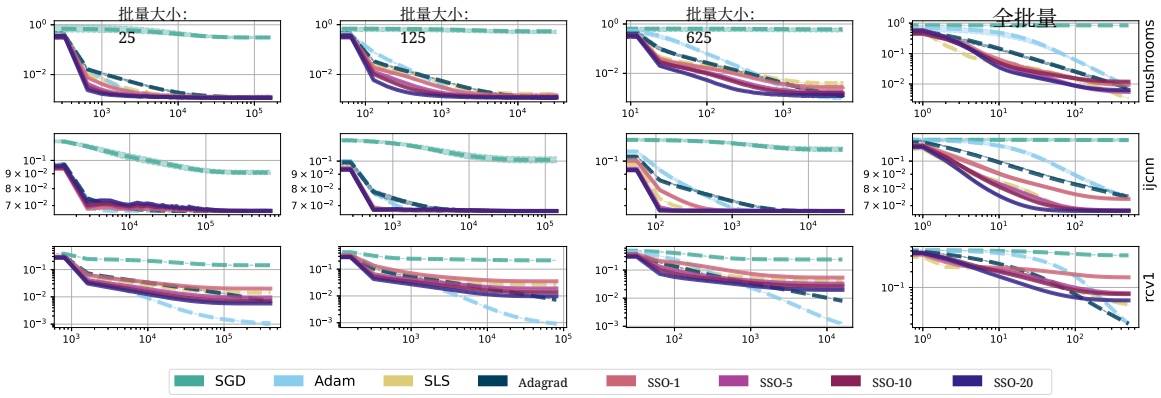


图 8：**指数步长：** 在**均方误差损失**下比较优化算法。我们比较了采用衰减步长 $(\frac{1}{T})^{t/T}$ 的示例，以及 SSO、SGD 和 SLS。Adam。同样，我们注意到，在代理函数上采取额外的步骤通常能提升性能。此外，递减步长似乎有助于保持严格的单调改进。最后，由于步长衰减较为缓和，这些优化算法比其随机 $\frac{1}{\sqrt{t}}$ 对应算法取得了更多进展。

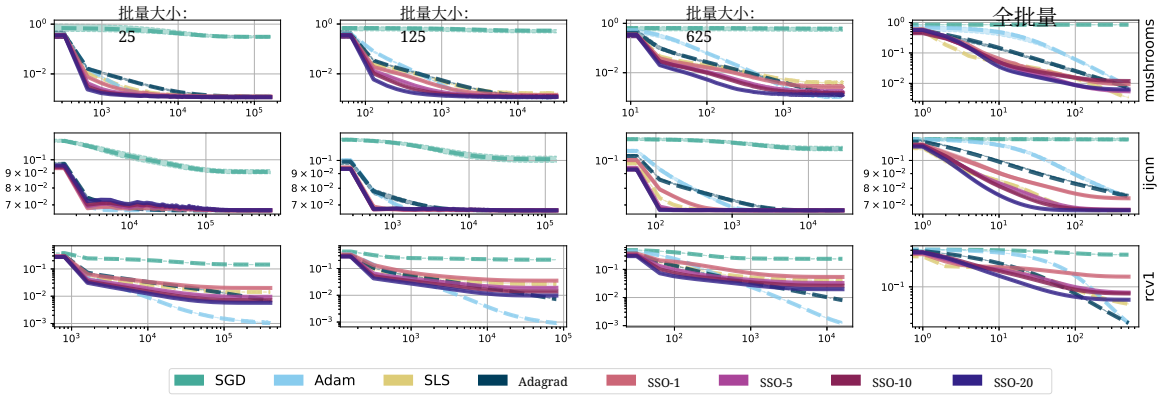


图9：**指数步长：**在**逻辑损失**下对比优化算法。我们比较了包含递减步长（$\frac{1}{T}$）$^{t/T}$的示例，并将其与 SSO、随机梯度下降和 SLS 进行对比。Adam 与上文相同。再次注意，在代理函数上采取更多步通常能提升性能。此外，递减步长似乎有助于保持严格的单调改进。最后，由于步长衰减较缓，优化算法比其随机 $\frac{1}{\sqrt{t}}$ 对应方法取得了更多进展。

E.2. 使用线搜索的随机代理优化

对 随机梯度下降、`SLS`、`Adam`、`Adagrad`和 `SSO-SLS` 在三个 SVMLib 基准数据集上的比较：`mushrooms`、`ijcnn`和 `rcv1`。每次运行均采用相同的初始化方案，并使用三个随机种子进行评估。所有图表均采用双对数空间，以便更清晰地展现不同优化算法的趋势。与之前一样，在所有设置中，算法均使用其理论步长（若有），否则使用（Paszke 等人，2019）所定义的默认值。内层优化循环的设置依据 Vaswani 等人（2019b）提出的线搜索参数和启发式方法。所有算法和批量大小均评估了 500 个训练轮次，性能以总优化步数为函数表示。

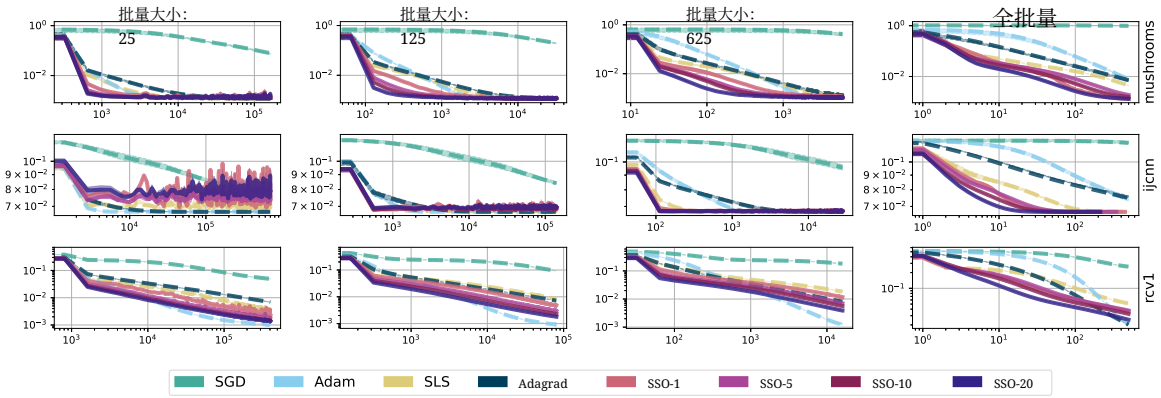


图10：**恒定步长：**在**均方误差损失**下对比优化算法。我们注意到，SSO-SLS 的表现优于其参数化对应方法，且性能与 SLS 和 Adam 不相上下。此外，我们还注意到，在代理函数上采取更多步通常能提升性能，尤其是在噪声较少的设置下（全批量和批量大小625）。

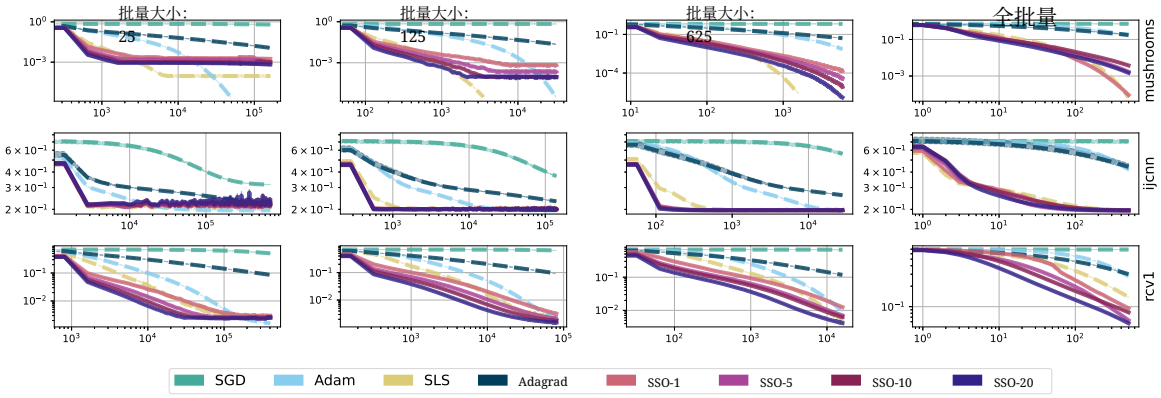


图 11：**恒定步长：** 在平均**逻辑损失**下对比优化算法。我们注意到，SSO-SLS 的表现优于其参数化对应方法，且与 SLS 和 Adam 的表现相当。此外，我们还注意到，在代理函数上执行更多步通常能提升性能。

E.3. 将随机代理优化与自适应梯度方法相结合

对 随机梯度下降、`SLS`、`Adam`、`Adagrad`和 `SSO-Adagrad` 在三个 SVMLib 基准数据集上的比较：`mushrooms`、`ijcnn`和 `rcv1`。每次运行均采用相同的初始化方案，并使用三个随机种子进行评估。所有图表均采用双对数空间，以便更清晰地展现不同优化算法的趋势。与之前一样，在所有设置中，算法均使用其理论步长（若有），否则使用（Paszke 等人，2019）所定义的默认值。内层优化循环的设置依据 Vaswani 等人（2019b）提出的线搜索参数和启发式方法。所有算法和批量大小均评估了 500 个训练轮次，性能以总优化步数为函数表示。此处，我们按照与标量 η 相同的计划进行更新 `Adagrad` （在Ward 等人（2020）中称为 AdaGrad-Norm）。由于 Adagrad 的最优理论步长不易计算，在我们的设置中，我们将对数学习率（即 log η的相反数）设为2。更多详情请参见随附的代码仓库。

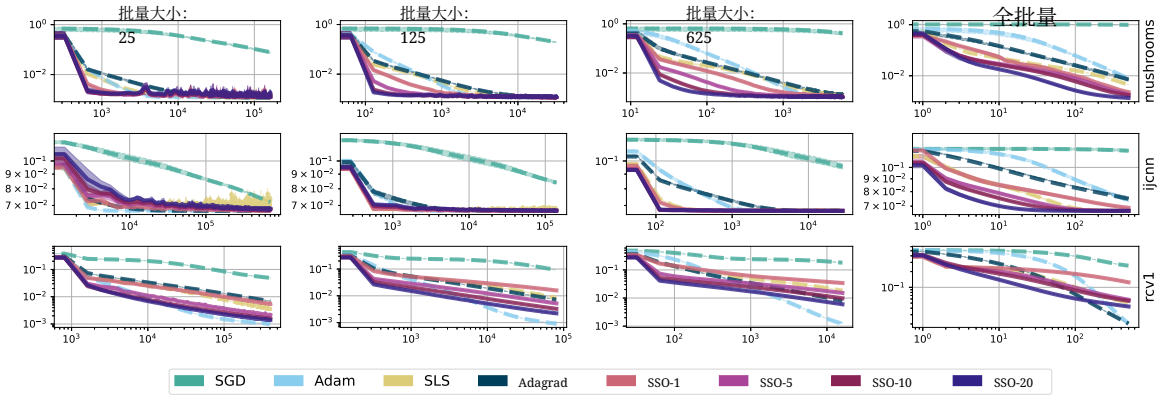


图 12：在**均方误差损失**下评估 SGD、SLS、Adam 和 SSO-Adagrad 的平均均方误差损失对比。这些图表显示，SSO-Adagrad 的表现优于其参数化对应方法，且与 SLS 和 Adam 的表现相当。此外，我们再次发现，在代理函数上执行更多步通常能提升性能。

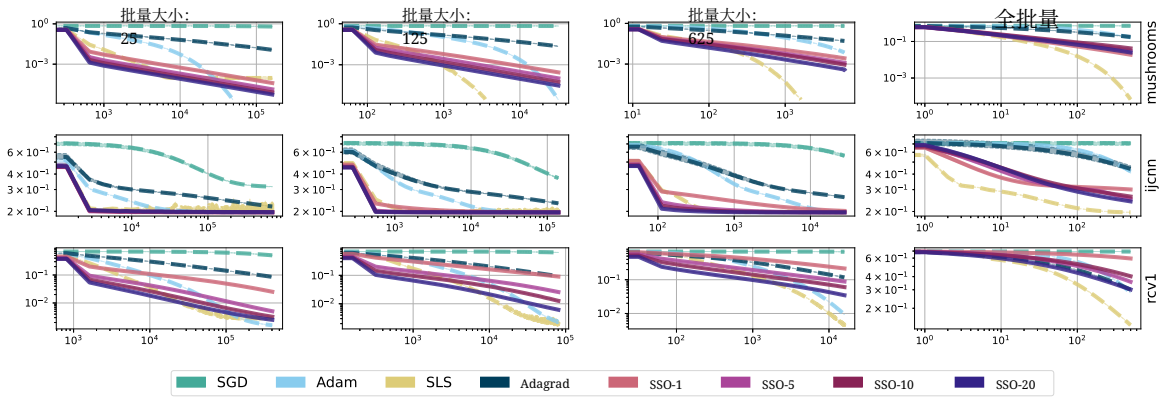


图13：在**逻辑损失**下评估随机梯度下降、SLS、Adam 和 SSO-Adagrad 的平均均方误差损失对比。这些图表明，SSO-Adagrad 的表现优于其参数化对应方法，且与 SLS 和 Adam 不相上下。此外，我们再次发现，在代理函数中采取更多步通常可以提升性能。

E.4. 将随机代理优化与在线牛顿步相结合

在 mushrooms、ijcnn 和 rcv1 这三个 SVMLib 基准数据集上，对 SGD、SLS、Adam、Adagrad 和 SSO-牛顿法进行了比较。每次运行均采用相同的初始化方案，并使用三个随机种子进行评估。所有图表均采用双对数空间，以便更清晰地呈现不同优化算法之间的趋势。与之前一样，在所有设定中，算法在有理论步长时采用理论步长，否则采用 (Paszke 等，2019) 定义的默认值。内层优化循环根据 Vaswani 等人（2019b）提出的线搜索参数和启发式方法进行设置。所有算法和批量大小均评估了 500 个训练轮次，性能以总优化步数为函数进行展示。这里，我们按照与在线牛顿法相同的计划更新 η。我们省略了均方误差示例，因为在此设定下，SSO-牛顿法等同于 SSO。不过，在下方展示的逻辑损失设定中，我们将正则化项重新缩放为 $(1-p)p$，其中 $p=\sigma(f(x))$，σ 是此 sigmoid 函数，f 是目标空间。此操作针对每个数据点分别执行；如下所示，即使在随机设定中，它通常也能带来极佳的性能。在下方图表中，如果曲线在达到最大优化步数之前消失，表示该算法已收敛至最小值，不再继续运行。值得注意的是，SSO-牛顿法在多个数据集和批量大小下都达到了这一点。

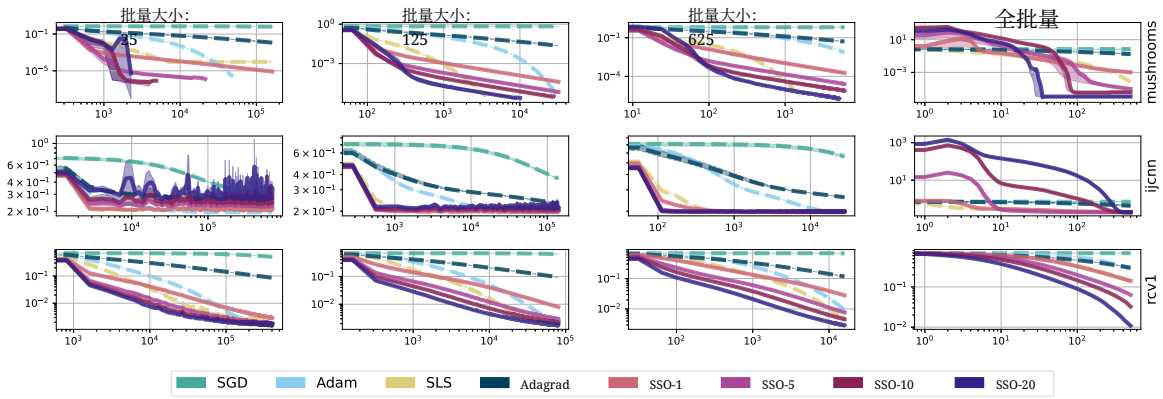


图14：在**逻辑损失**上评估随机梯度下降、SLS、Adam 和 SSO-牛顿法的平均逻辑损失对比。该图显示，通过正确重新缩放正则化项，无需额外成本即可取得显著改进。请注意，在 mushrooms 的情况下，对于 $m=20$，SSO-牛顿法在所有情况下都能在第 500 个 epoch 之前达到停止标准。其次，即使在许多随机设定下，SSO-牛顿法的表现也优于 SLS 和 Adam。

E.5. 与随机方差缩减梯度的比较

下面我们给出一个简单的监督学习示例，将标准方差缩减算法 SVRG（Johnson 和 Zhang，2013年）与 *SSO* 算法在监督学习设定下进行比较。与上面的示例一样，我们使用 Chang 和 Lin (2011年) rcv1 数据集，并采用均方误差损失。该图表明，即使批量大小较小，SSO 也能优于 SVRG。关于实现和所用超参数的更多详情，请参见 https://github.com/WilderLavington/Target-Based-Surrogates-For-Stochastic-Optimization。我们指出，对于足够小的批量大小，SVRG 可能具有竞争力，甚至能与 SSO 相媲美；不过，这类比较留待未来工作。

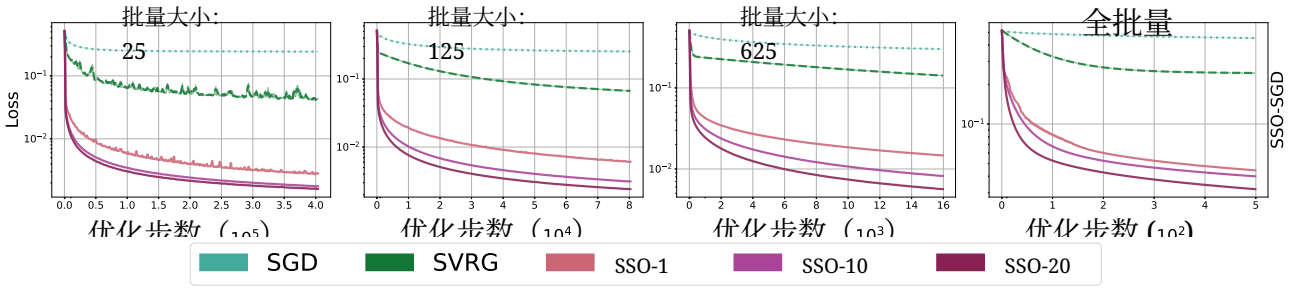
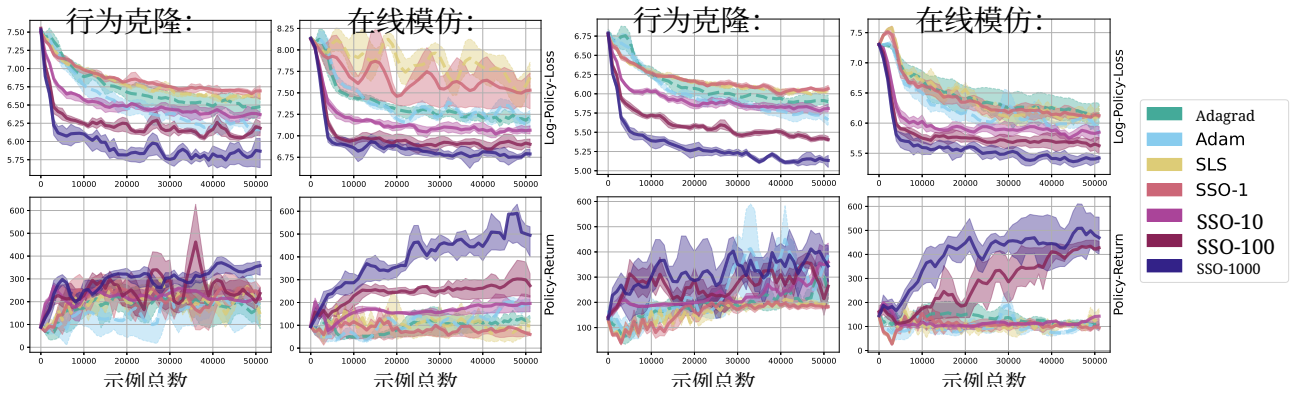


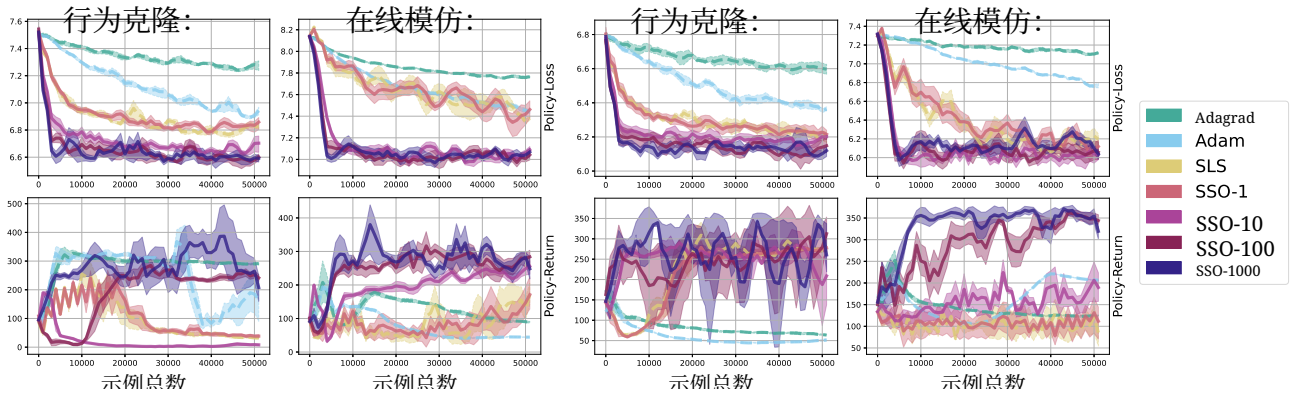
图 15：随机梯度下降、随机方差缩减梯度和 SSO-SGD 的平均均方误差比较。该图显示，即使在批量大小较小的情况下，SSO 也能优于随机方差缩减梯度等方差缩减技术。

E.6. 模仿学习



(a) Walker2d-v2 环境　　(b) Hopper-v2 环境

图 16：比较 随机梯度下降、SLS、Adam、Adagrad 和 SSO 随总交互次数变化而产生的策略回报和对数策略损失。与第 5 节不同，策略的均值由一个 **神经网络模型** 参数化。在两种环境中，对于两种行为策略，SSO 的表现均优于所有其他在线优化算法。此外，随着 m in 增大，SSO 在回报和损失方面的表现也会提升。



(a) Walker2d-v2 环境　　(b) Hopper-v2 环境

图17：比较随机梯度下降、SLS、Adam、Adagrad和SSO 随总交互次数变化所产生的策略回报和对数策略损失。与第5节不同，策略均值由一个**线性模型**参数化。在两个环境中，对于两种行为策略，SSO 的表现均优于所有其他在线优化算法。此外，随着 m 的增加，SSO在回报和损失方面的表现也随之提升。

在 Adagrad、SLS、Adam 和SSO 在两个 Mujoco（Todorov 等人，2012年）模仿学习基准（Lavington 等人，2022年）上的评估结果进行比较，Hopper-v2 和 Walker-v2。在这种设置中，训练和评估分轮进行。每一轮中，行为策略从环境中采集数据，专家为这些数据添加标签。目标是在下一阶段（以采样得到的状态为条件）猜测专家会为收集到的样本标注什么。与监督学习设置不同的是，我们会持续收到新的数据点，这些数据点可能彼此相关，并且会随时间从不同分布中抽取。从理论上说，这使优化问题变得更加困难；而且，由于我们必须与模拟器交互，查询随机梯度的成本可能很高。与附录 E中的示例一样，在这种设置下，我们既会依据专家的策略分布进行交互（行为克隆），也会依据智能体所诱导的策略分布进行交互（在线模仿）。我们将策略参数化为一个正态分布，其均值通过专家标签与智能体策略均值之间的均方误差损失进行学习。同样，专家使用软演员-评论家进行训练。所有实验均使用 NVIDIA GeForce RTX 2070 显卡和 AMD Ryzen 9 3900 12核处理器运行。

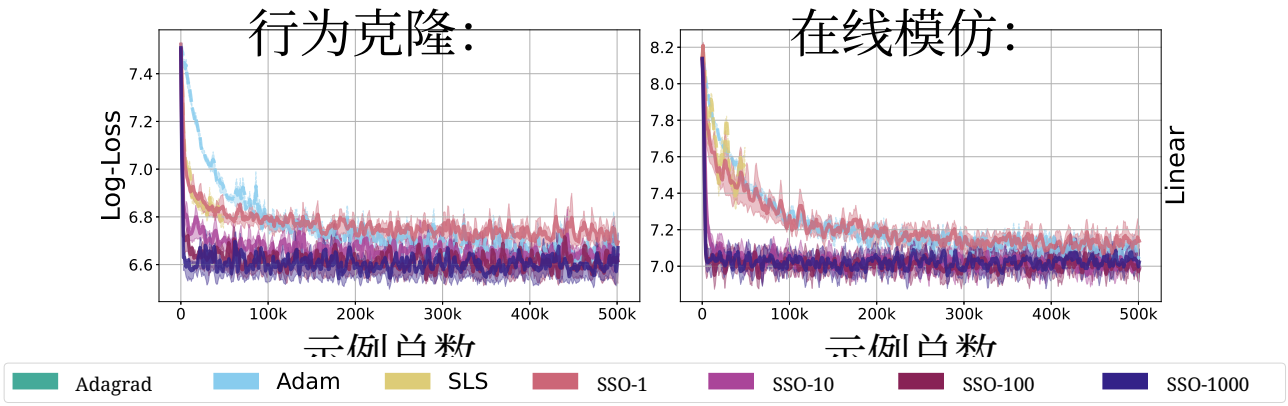


图 18：在 Walker2d Gym 环境中，随机梯度下降、SLS、Adam、Adagrad和 SSO 产生的对数策略损失随总交互次数的变化对比。与第 5 节不同，策略均值由一个**线性模型**参数化。与图 17 和图 16不同，此图展示了 50 万次迭代，而非仅 50 次。我们再次看到，SSO优于所有其他在线优化算法，即使在迭代次数非常多时也是如此。此外，随着 m in 增大，我们再次看到SSO 的对数损失也有所改善。

在这种设置中，我们评估两项指标：每轮的对数策略损失和策略回报。对数策略损失如上所述，而回报用于衡量模仿学习策略实际完成任务的效果。在 Mujoco 基准测试中，奖励定义为智能体在空间中移动的速度以及为此所施加的功率的函数。模仿学习通常通过采用高奖励策略（例如，能够以极小的扭矩在空间中移动），并直接模仿该策略来实现，而不是像强化学习中那样尝试学习代价到达函数（Sutton 等人，2000）。

每种算法均采用相同的初始化方案，在三个随机种子下进行评估。与之前一样，在所有设置中，算法在有理论步长可用时使用理论步长，否则使用由（Paszke 等人，2019）定义的默认值。内层优化循环根据线搜索参数和启发式方法进行设置，遵循Vaswani 等人（2019b）。所有算法和批量大小均进行 50 轮交互评估（图 18的情况除外，该图评估了 500 轮），性能表示为与环境的总交互次数的函数。下面我们学习一个由具有 256 个隐藏单元和 ReLU 激活的两层感知网络参数化的策略（与正文中的设置相同）。更多详情请参阅随附的代码仓库。