



数据建模与分析-读书笔记

📖 数据建模与分析

📍 [1309.5549](#)

📖 优化基本理论与方法

Introduction

1951年，Robbins 和 Monro 提出经典 SA 算法，用噪声梯度替代精确梯度，且在强凸随机规划问题中达到“渐近最优”收敛速率，但实践中因步长策略难以实施而表现不佳。Polyak 与 Polyak & Juditsky 建议采用更长步长并对迭代结果做平均，使算法对步长选择更鲁棒，且仍能实现强凸情形下的渐近最优率。Nemirovski 等引入镜面下降 SA (Mirror Descent SA)，证明其对一般非光滑凸 SP 问题具有最优的 $\mathcal{O}(1/\varepsilon^2)$ 迭代复杂度，并在某些凸 SP 问题上显著优于样本均值近似法 (SAA)，Lan 考虑目标函数的光滑性，提出了一种适用于光滑、非光滑和随机优化的优化方法。但是这些方法中，函数的凸性对于 SA 算法的收敛性起到重要的作用。现存的大部分 SA 算法并不能解决非凸目标函数的问题。

考虑如下典型的无约束非线性规划问题：

$$f^* := \inf_{x \in \mathbb{R}^n} f(x)$$

要求 f 可微（不一定要凸函数）有下界，并且满足如下 Lipschitz 条件：

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$$

我们假定使用迭代算法来求解如上问题，在迭代算法中，我们通过随机调用一阶 Oracle (SFO) 来获取 f 的梯度

非凸随机规划 (SP) 有如下集中应用场景

机器学习的正则化目标函数:

$$f(x) = \int_{\Xi} L(x, \xi) dP(\xi) + r(x)$$

其中 $\Xi \subseteq \mathbb{R}^d$ 是分布 P 的样本空间, 损失函数 $L(x, \xi)$ 和正则化函数 $r(x)$ 都是非凸的

内生不确定性随机规划

$$f(x) = \int_{\Xi(x)} F(x, \xi) dP_x(\xi)$$

$f(x)$ 通常是非凸函数尽管 $F(x, \xi)$ 相对于 x 是一个凸函数

基于仿真的优化问题

$$f(x) = \mathbb{E}_{\xi}[F(x, \xi)]$$

$F(x, \xi)$ 并没有显式给出, 而是通过一个黑箱仿真过程, 因此, 我们并不知道 $f(x)$ 是不是一个凸函数。

在无噪声下, Nesterov 证明经典梯度下降在 $N = \mathcal{O}(1/\varepsilon)$ 步即可保证 $\min_{k \leq N} \|\nabla f(x_k)\|^2 \leq \varepsilon$, Cartis 证明其界是紧的。尽管如此, Nesterov 的证明并不适用于随机集, 即使我们能证明 $\min_{k \leq N} \|\nabla f(x_k)\|^2 \leq \varepsilon$, 因为不能确切知道 $\|\nabla f(x_k)\|$, 所以最优解并不是很容易就能找到。

Stochastic first-order methods

Randomized Stochastic Gradient

Assumption 1

$$\mathbb{E}[G(x_k, \xi_k)] = \nabla f(x_k)$$

$$\mathbb{E}[\|G(x_k, \xi_k) - \nabla f(x_k)\|^2] \leq \sigma^2$$

其中 $G(x_k, \xi_k)$ 是  [Borel 函数](#), 由 SFO 给出的。也就是说 $G(x_k, \xi_k)$ 是对 $\nabla f(x_k)$ 的无偏估计, 且随机变量 $\|G(x_k, \xi_k) - \nabla f(x_k)\|$ 的方差是有界。

假设 1 有如下增强形式:

Assumption 2

$$\mathbb{E}[\exp\{\|G(x, \xi_k) - g(x)\|^2/\sigma^2\}] \leq \exp\{1\}$$

A randomized stochastic gradient method

Input: 初始点 x_1 , 最大迭代次数 N , 步长 γ_k , 概率函数 $P_R(\cdot)$

1. 记 R 为由概率函数 P_R 给出的随机变量
2. 调用随机一阶 Oracle 计算 $G(x_k, \xi_k)$, 并取

$$x_{k+1} = x_k - \gamma_k G(x_k, \xi_k)$$

Output x_R

与经典的 SA 算法相比, 这里使用了随机的迭代次数 R 来停止 RSG。或者说, 我们运行算法 N 次, 从中挑选搜索点 x_R 作为算法的输出, 因此, 我们只需要运行算法前 R 次, 并不关心后 $N - R$ 次的内容, 值得强调的是, RSG 算法并不是为了减少计算开支, 而是为解决非凸 SP 问题, 当然以上的算法内容都只是概念性上的, 因为我们并没有给出如何挑选步长和概率函数。

Theorem

假定 RSG 算法中，步长和概率函数满足 $\gamma_k < \frac{2}{L}$ 和

$$P_R(k) := \text{Prob} \{R = k\} = \frac{2\gamma_k - L\gamma_k^2}{\sum_{k=1}^N (2\gamma_k - L\gamma_k^2)}$$

那么，根据假设 1，给出：

$$\frac{1}{L} \mathbb{E}[\|\nabla f(x_R)\|^2] \leq \frac{D_f^2 + \sigma^2 \sum_{k=1}^N \gamma_k^2}{\sum_{k=1}^N (2\gamma_k - L\gamma_k^2)}$$

其中

$$D_f := \left[\frac{2(f(x_1) - f^*)}{L} \right]^{1/2}$$

f^* 是问题的最优解。

如果无约束问题是凸的，则

$$\mathbb{E}[f(x_R) - f^*] \leq \frac{D_X^2 + \sigma^2 \sum_{k=1}^N \gamma_k^2}{\sum_{k=1}^N (2\gamma_k - L\gamma_k^2)}$$

其中

$$D_X := \|x_1 - x^*\|$$

x^* 是问题的最优解。

基于上面定理给出如下推论：

Corollary

对于 $\tilde{D} > 0$ ，步长由如下方式给出

$$\gamma_k = \min \left\{ \frac{1}{L}, \frac{\tilde{D}}{\sigma\sqrt{N}} \right\}$$

根据假设给出：

$$\frac{1}{L} \mathbb{E}[\|\nabla f(x_R)\|^2] \leq \mathcal{B}_N := \frac{LD_f^2}{N} + \left(\tilde{D} + \frac{D_f^2}{\tilde{D}} \right) \frac{\sigma}{\sqrt{N}}$$

同样，对于凸优化问题有：

$$\mathbb{E}[f(X_R) - f^*] \leq \frac{LD_X^2}{N} + \left(\tilde{D} + \frac{D_X^2}{\tilde{D}} \right) \frac{\sigma}{\sqrt{N}}$$

与其随便从 $\{x_1, \dots, x_N\}$ 任意给出一个解，另一种给出解的方式是满足：

$$\|\nabla f(\hat{x}_N)\| = \min_{k=1, \dots, N} \|\nabla f(x_k)\|$$

$\mathbb{E}\|\nabla f(\hat{x}_N)\|$ 随着定理和推论中相同的方式收敛到 0。

但是这种方法需要额外计算 $\|\nabla f(x_k)\|$ for all $k = 1, \dots, N$ ，但由于 $\|\nabla f(x_k)\|$ 无法被精确计算，所以需要采用蒙特卡洛模拟，会带来额外的误差。

关于 \tilde{D} 的选取，我们可以选择 D_f 求解非凸问题，选择 D_X 求解凸问题。

RSG 算法的一个缺点是需要选取一个合适的 L 来估计步长 γ_k 的上界，这个会影响到 P_R 的选取（因为其包含）。在一些确定性的梯度下降方法中，我们可以通过用线搜索来加强这些方法的表现以放松我们对 L 的要求，但是这个方法并不适用于随机集。不过，在 RSG 算法中，我们不需要很精确的估计 L ，实际上，当步长满足：

$$\min \left\{ \frac{1}{qL}, \frac{\tilde{D}}{\sigma\sqrt{N}} \right\}$$

其中 $q \in \{1, \sqrt{N}\}$ 时，RSG 算法以 $\mathcal{O}(1/\sqrt{N})$ 的速率收敛，因此，我们扩大了 L 的估计范围。我们采用在小数目样本点上计算随机梯度来估计 L 。

RSG 方法在凸问题上比其他方法有一定程度的优化

增长的步长方法：

$$\gamma_k = \min \left\{ \frac{1}{L}, \frac{\tilde{D}\sqrt{k}}{\sigma N} \right\}$$

减缓的步长方法:

$$\gamma_k = \min \left\{ \frac{1}{L}, \frac{\tilde{D}}{\sigma(kN)^{1/4}} \right\}$$

第二种有利于提早停止算法, 第一种有利于更好地获得梯度的本地信息。

方法单次运行时差生的偏差满足

$$\text{Prob} \{ \|\nabla f(x_R)\|^2 \geq \lambda L \mathcal{B}_N \} \leq \frac{1}{\lambda} \quad \forall \lambda > 0$$

调用 SFO 次数满足界:

$$\mathcal{O} \left\{ \frac{1}{\Lambda \varepsilon} + \frac{\sigma^2}{\Lambda^2 \varepsilon^2} \right\}$$

A two-phase randomized stochastic gradient method

优化阶段, 运行 RSG 生成数个候选方案; 后优化阶段, 从候选方案中选出一个方案

A two-phase RSG (2-RSG) method

Input: 初始点 x_1 , 运行次数 S , 迭代限制 N , 样本大小 T

Optimization phase:

在 $s = 1, \dots, S$, 在给定条件下调用 RSG 算法, 输出 \bar{x}_s

Post-optimization phase:

从 $\{\bar{x}_1, \dots, \bar{x}_S\}$ 中挑选解 \bar{x}^* 满足:

$$\|g(\bar{x}^*)\| = \min_{s=1, \dots, S} \|g(\bar{x}_s)\|, g(\bar{x}_s) := \frac{1}{T} \sum_{k=1}^T G(\bar{x}_s, \xi_k)$$

$G(x, \xi_k)$ 是由 SFO 算法给出的随机梯度。

或者我们可以选择满足入下条件的解:

$$\tilde{f}(\bar{x}^*) = \min_{1, \dots, S} \tilde{f}(\bar{x}_s), \tilde{f}(\bar{x}_s) = \frac{1}{T} \sum_{k=1}^T F(\bar{x}_s, \xi_k)$$

2-RSG 解的挑选方式与 Nesterov 的取平均的方法不同。

在 2-RSG 方法中, SFO 调用次数分别为 $S \times N$ (优化阶段), $S \times T$ (后优化阶段), 序列 $\{\xi_k\}$ 可以回收使用。

Lemma

给定一个具有 Borel 概率测度 μ 的 Polish 空间, 以及一个 Borel σ -代数 Ω 的递增序列 $\mathcal{F}_0 = \{\emptyset, \Omega\} \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots$, 设 $\zeta_i \in \mathbb{R}^n$ 时定义在 Ω 上的 Borel 函数的鞅差序列, 满足 ζ_i 是 \mathcal{F}_i 可测的, 且 $\mathbb{E}[\zeta_i | \mathcal{F}_{i-1}] = 0$, 其中 $\mathbb{E}[\cdot | \mathcal{F}_i]$ 表示相对于 \mathcal{F}_i 的条件期望, $\mathbb{E} = \mathbb{E}[\cdot | \mathcal{F}_0]$ 是相对于 μ 的期望。

1. 如果 $\mathbb{E}[\|\zeta_i\|^2] \leq \sigma_i^2$ 则 $\mathbb{E}[\|\sum_{i=1}^N \zeta_i\|^2] \leq \sum_{i=1}^N \sigma_i^2$, 有:

$$\forall N \geq 1, \lambda \geq 0 : \text{Prob} \left\{ \left\| \sum_{i=1}^N \zeta_i \right\| \geq \lambda \sum_{i=1}^N \sigma_i^2 \right\} \leq \frac{1}{\lambda}$$

2. 如果对于几乎所有的 $i \geq 1$ 有 $\mathbb{E} \{ \exp(\|\zeta_i\|^2 / \sigma_i^2) | \mathcal{F}_{i-1} \} \leq \exp(1)$

$$\forall N \geq 1, \lambda \geq 0 : \text{Prob} \left\{ \left\| \sum_{i=1}^N \zeta_i \right\| \geq \sqrt{2}(1 + \lambda) \sqrt{\sum_{i=1}^N \sigma_i^2} \right\} \leq \exp(-\lambda^2/3)$$

Theorem

$$1. \quad \text{Prob} \left\{ \|\nabla f(\bar{x}^*)\|^2 \geq 2 \left(4L\mathcal{B}_N + \frac{3\lambda\sigma^2}{T} \right) \right\} \leq \frac{S+1}{\lambda} + 2^{-S}$$

2. 给定 $\varepsilon > 0$ 和 $\Lambda \in (0, 1)$, 记:

$$S = S(\Lambda) := \lceil \log(2/\Lambda) \rceil$$

$$N = N(\varepsilon) := \left\lceil \max \left\{ \frac{32L^2 D_f^2}{\varepsilon} \right\}, \left\lceil 32L \left(\tilde{D} + \frac{D_f^2}{\tilde{D}} \right) \frac{\sigma}{\varepsilon} \right\rceil \right\rceil$$

$$T = T(\varepsilon, \Lambda) := \left\lceil \frac{24(S+1)\sigma^2}{\Lambda\varepsilon} \right\rceil$$

2-RSG 算法收敛最多要调用

$$S(\Lambda)[N(\varepsilon) + T(\varepsilon, \Lambda)]$$

次 SFO

Corollary

在假设 1 和 假设 2 情况下，定理可以加强为

$$1. \text{Prob} \left\{ \|\nabla f(\bar{x}^*)\|^2 \geq 4 \left(2L\mathcal{B}_N + 3(1 + \lambda) \frac{\sigma^2}{T} \right) \right\} \leq (S + 1) \exp(-\lambda^2/3) + 2^{-S}$$

2. 给定 $\varepsilon > 0$ 和 $\Lambda \in (0, 1)$, S, N 保持不变, T 改写为:

$$T = T'(\varepsilon, \Lambda) := \frac{24\sigma^2}{\varepsilon} \left[1 + \left(3 \ln \frac{2(S + 1)}{\Lambda} \right)^{1/2} \right]^2$$

则 2-RSG 算法收敛最多要调用

$$S(\Lambda)[N(\varepsilon) + T'(\varepsilon, \Lambda)]$$

次 SFO

复杂度:

$$\mathcal{O} \left\{ \frac{\log(1/\Lambda)}{\varepsilon} + \frac{\sigma^2}{\varepsilon^2} \log \frac{1}{\Lambda} + \frac{\log^2(1/\Lambda)\sigma^2}{\varepsilon} \right\}$$

Stochastic zeroth-order methods

考虑如下优化问题:

$$f^* := \inf_{x \in \mathbb{R}^n} \left\{ f(x) := \int_{\Xi} F(x, \xi) dP(\xi) \right\}$$

假定 $F(x, \xi) \in \mathcal{C}_L^{1,1}(\mathbb{R}^n)$, 同时有 $f(x) \in \mathcal{C}_L^{1,1}(\mathbb{R}^n)$ 。我们关心的问题是在只给出 f 的零阶 Oracle 的情况如何调用我们的 RSG 算法和 2-RSG 算法

The randomized stochastic gradient free method

在第 k 次迭代中, 我们输入 x_k 和 ξ_k , 随机零阶 Oracle (SZO) 会量化 $F(x_k, \xi_k)$

Assumption 3

对于任意的 $k \geq 1$, 有:

$$\mathbb{E}[F(x_k, \xi_k)] = f(x_k)$$

f 与任何非负、可测且有界函数 $\psi: \mathbb{R}^n \rightarrow \mathbb{R}$ 的卷积，且满足 $\int_{\mathbb{R}^n} \psi(u) du = 1$ ，是 f 的一个近似，其平滑程度至少与 f 相同，函数 ψ 的一个最重要的例子是概率密度函数。

考虑如下光滑近似：

$$f_\mu(x) = \frac{1}{(2\pi)^{n/2}} \int f(x + \mu u) e^{-1/2\|u\|^2} du = \mathbb{E}_u[f(x + \mu u)]$$

其中 $\mu > 0$ 是平滑参数， u 是一个 n 维标准高斯随机向量。

光滑近似满足如下特性 (Nesterov)

Theorem

1. f_μ 的梯度：

$$\nabla f_\mu(x) = \frac{1}{(2\pi)^{n/2}} \int \frac{f(x + \mu u) - f(x)}{\mu} u e^{-1/2\|u\|^2} du$$

满足 L_μ -Lipschitz 光滑，其中 $L_\mu \leq L$

2. 对于任意 $x \in \mathbb{R}^n$ 有：

$$\begin{aligned} |f_\mu(x) - f(x)| &\leq \frac{\mu^2}{2} Ln \\ \|\nabla f_\mu(x) - \nabla f(x)\| &\leq \frac{\mu}{2} L(n+3)^{3/2} \end{aligned}$$

3. 对于任意 $x \in \mathbb{R}^n$ 有：

$$\frac{1}{\mu^2} \mathbb{E}_u[\{f(x + \mu u) - f(x)\}^2 \|u\|^2] \leq \frac{\mu^2}{2} L^2(n+6)^3 + 2(n+4) \|\nabla f(x)\|^2$$

根据定理第二条的第二小条，可以得出：

$$\|\nabla f_\mu(x)\|^2 \leq 2\|\nabla f(x)\| + \frac{\mu^2}{2} L^2(n+3)^3$$

$$\|\nabla f(x)\| \leq 2\|\nabla f_\mu(x)\|^2 + \frac{\mu^2}{2} L^2(n+3)^3$$

记

$$f_\mu^* := \min_{x \in \mathbb{R}^n} f_\mu(x)$$

根据定理第二条的第一小条有：

$$-\mu^2 Ln \leq [f_\mu(x) - f_\mu^*] - [f(x) - f^*] \leq \mu^2 Ln$$

A randomized stochastic gradient free (RSGF) method

Input: 初始点 x_1 , 最大迭代次数 N , 步长 $\{\gamma_k\}_{k \geq 1}$ 概率函数 $P_{R(\cdot)}$

Step 0: 取 R 为 $P_{R(\cdot)}$ 给出随机变量

Step $k = 1, \dots, R$, 生成高斯随机向量 u_k , 调用随机零阶 Oracle 按如下方式计算 $G_\mu(x_k, \xi_k, u_k)$:

$$G_\mu(x_k, \xi_k, u_k) = \frac{F(x_k + \mu u_k, \xi_k) - F(x_k, \xi_k)}{\mu} u_k$$

设置:

$$x_{k+1} = x_k - \gamma_k G_\mu(x_k, \xi_k, u_k)$$

Output x_R

梯度估计满足:

$$\mathbb{E}_{\xi, u}[G_\mu] = \mathbb{E}_\mu[\mathbb{E}_\xi[G_\mu]] = \nabla f_\mu(x)$$

也就是说 G_μ 是 $\nabla f_\mu(x)$ 的一个无偏估计

Theorem

在 RSGF 算法中, 假定步长和概率函数满足: $\gamma_k < 1/[2(n+4)L]$ 和

$$P_R(k) := \text{Prob}\{R = k\} = \frac{\gamma_k - 2L(n+4)\gamma_k^2}{\sum_{k=1}^N [\gamma_k - 2L(n+4)\gamma_k^2]}$$

对于任何 $N \geq 1$, 有:

$$\frac{1}{L} \mathbb{E}[\|\nabla f(x_R)\|^2] \leq \frac{1}{\sum_{k=1}^N [\gamma_k - 2L(n+4)\gamma_k^2]} \left[D_f^2 + 2\mu^2(n+4) \left(1 + L(n+4)^2 \sum_{k=1}^N \left(\frac{\gamma_k}{4} + L\gamma_k^2 \right) \right) + 2(n+4)\sigma^2 \sum_{k=1}^N \gamma_k^2 \right]$$

如果问题是凸的, 则对于最优解 x^* 有:

$$\mathbb{E}[\|f(x_R) - f^*\|^2] \leq \frac{1}{2 \sum_{k=1}^N [\gamma_k - 2L(n+4)\gamma_k^2]} \left[D_X^2 + 2\mu^2 L(n+4) \sum_{k=1}^N [\gamma_k + L(n+4)^2 \gamma_k^2] + 2(n+4)\sigma^2 \sum_{k=1}^N \gamma_k^2 \right]$$

步长满足

$$\gamma_k = \frac{1}{\sqrt{n+4}} \min \left\{ \frac{1}{4L\sqrt{n+4}}, \frac{\tilde{D}}{\sigma\sqrt{N}} \right\}$$

平滑参数满足:

$$\mu \leq \frac{D_f}{(n+4)\sqrt{2N}}$$

在假设 1 和假设 3 条件下, 给出:

$$\frac{1}{L} \mathbb{E}[\|\nabla f(x_R)\|^2] \leq \bar{\mathcal{B}}_N := \frac{12(n+4)LD_f^2}{N} + \frac{4\sigma\sqrt{n+4}}{\sqrt{N}} \left(\tilde{D} + \frac{D_f^2}{\tilde{D}} \right)$$

此外, 如果问题是凸的, 最优解 x^* , 平滑参数满足:

$$\mu \leq \frac{D_X}{\sqrt{(n+4)}}$$

且

$$\mathbb{E}[f(x_R) - f^*] \leq \frac{5L(n+4)D_X^2}{N} + \frac{2\sigma\sqrt{n+4}}{\sqrt{N}} \left(\tilde{D} + \frac{D_X^2}{\tilde{D}} \right)$$

对于凸和非凸的 SP 问题, 在 RSGF 中我们都采用相同的步长和概率函数, RSGF 的复杂度界为 $\mathcal{O}(n/\varepsilon^2)$, 对于凸问题, 满足 $\mathbb{E}[f(\bar{x}) - f^*] \leq \varepsilon$ 的解 \bar{x} 可以在 $\mathcal{O}(n/\varepsilon)$ 次迭代中找到。RSGF 的依赖性比解决一般非光滑凸 SP 问题的依赖性弱, 这是因为我们处理的是一个更特殊的 SP 问题类别。在 $\sigma = 0$ 的情况下, RSGF 方法的复杂度降低到 $\mathcal{O}(n/\varepsilon)$, 与 Nesterov 的无导数随机搜索方法获得的复杂度相似。

同样的, 对于 \tilde{D} , 其非凸和凸情况下的最优选择分别是 D_f 和 D_X

$$\text{Prob} \left\{ \|\nabla f(x_R)\|^2 \geq \lambda L \bar{\mathcal{B}}_N \right\} \leq \frac{1}{\lambda}$$

所以找到一个 (ε, Λ) 解的界为:

$$\mathcal{O} \left\{ \frac{nL^2 D_f}{\Lambda \varepsilon} + \frac{nL^2}{\Lambda^2} \left(\tilde{D} + \frac{D_f^2}{\tilde{D}} \right) \frac{\sigma^2}{\varepsilon^2} \right\}$$

A two-phase RSGF(2-RSGF) method

Input : 初始点 x_1 , 运行次数 S , 迭代限制 N , 样本大小 T

Optimization phase :

在 $s = 1, \dots, S$, 在给定条件下调用 RSGF 算法, 输出 \bar{x}_s

Post-optimization phase:

从 $\{\bar{x}_1, \dots, \bar{x}_S\}$ 中挑选解 \bar{x}^* 满足:

$$\|g_\mu(\bar{x}^*)\| = \min_{s=1, \dots, S} \|g_\mu(\bar{x}_s)\| \quad g_\mu(\bar{x}_s) := \frac{1}{T} \sum_{k=1}^T G_\mu(\bar{x}_s, \xi_k, u_k)$$

收敛特性由如下定理给出:

Theorem

$$\begin{aligned} \text{Prob} \left\{ \|\nabla f(\bar{x}^*)\|^2 \geq 8L\bar{\mathcal{B}}_N + \frac{3(n+4)L^2D_f^2}{2N} + \right. \\ \left. \frac{24(n+4)\lambda}{T} \left[L\bar{\mathcal{B}}_N + \frac{(n+4)L^2D_f^2}{N} + \sigma^2 \right] \right\} \\ \leq \frac{S+1}{\lambda} + 2^{-S} \end{aligned}$$

给定 $\varepsilon > 0$ 和 $\Lambda \in (0, 1)$, S 由之前定义给出, 定义:

$$\begin{aligned} N = \hat{N}(\varepsilon) &:= \max \left\{ \frac{12(n+4)(6LD_f)^2}{\varepsilon}, \left[72L\sqrt{n+4} \left(\bar{D} + \frac{D_f^2}{\bar{D}} \right) \frac{\sigma}{\varepsilon} \right]^2 \right\} \\ T = \hat{T}(\varepsilon, \Lambda) &:= \frac{24(n+4)(S+1)}{\Lambda} \max \left\{ 1, \frac{6\sigma^2}{\varepsilon} \right\} \end{aligned}$$

2-RSGF 可以计算出一个 (ε, Λ) 的解最多调用

$$2S(\Lambda) \left[\hat{N}(\varepsilon) + \hat{T}(\varepsilon, \Lambda) \right]$$

次 SZO

2-RSGF 方法调用的总的 SZO 次数的界为:

$$\mathcal{O} \left\{ \frac{nL^2D_f^2 \log(1/\Lambda)}{\varepsilon} + nL^2 \left(\bar{D} + \frac{D_f^2}{\bar{D}} \right) \frac{\sigma^2}{\varepsilon^2} \log \frac{1}{\Lambda} + \frac{n \log^2(1/\Lambda)}{\Lambda} \left(1 + \frac{\sigma^2}{\varepsilon} \right) \right\}$$

比 RGSF 方法小了 $\mathcal{O}(1/[\Lambda^2 \log(1/\Lambda)])$ ，二者都是第二项在起主导作用

在本文中，我们提出了一类新的SA（随机逼近）方法，用于解决具有噪声一阶信息的经典无约束NLP（非线性规划）问题。我们建立了一些关于计算这类问题 ε -解的复杂度的新结果，并表明当问题是凸的时候，它们几乎是最优的。此外，我们引入了一个后优化阶段，以改善RSG（随机梯度）方法的大偏差性质。这些程序及其复杂度结果随后被专门用于仅提供随机零阶信息的基于模拟的优化问题。此外，我们表明，对于光滑凸SP（优化问题），无梯度方法的复杂度对维度 n 的依赖性比更一般的非光滑凸SP要弱得多。