



一阶优化方法专题

动量、加速、性能估计、大步长与随机梯度

梁经纬 **Jingwei Liang**

Institute of Natural Sciences
School of Mathematical Sciences

Lecture Supplement: 一阶优化方法专题

2026 年秋季学期

饮水思源 · 爱国荣校

本讲内容

频谱视角
梯度法在不同特征方向上的动力学



惯性与加速
Heavy-ball 与 Nesterov 方法



性能估计
Performance estimation problem

大步长序列
Chebyshev 与 Silver stepsizes



随机优化
SGD、mini-batch、平均与 SVRG



延伸
Proximal/FISTA、CG、adaptive methods

两组贯穿实验分别比较确定性方法与随机方法；所有参数和随机种子在讲义末尾给出。

本讲结构；与 Lecture 03 的梯度法和预条件部分衔接

两条分析主线

确定性一阶方法

每次可获得完整梯度 $\nabla f(\mathbf{x}^{(k)})$ 。主要问题是

曲率差异如何影响收敛？

频谱、多项式和最坏情形分析提供统一语言。

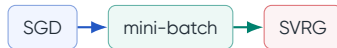


随机一阶方法

每次只能获得随机估计 $\hat{\mathbf{g}}^{(k)}$ ，满足

$$\mathbb{E}[\hat{\mathbf{g}}^{(k)} \mid \mathbf{x}^{(k)}] = \nabla F(\mathbf{x}^{(k)}).$$

计算预算、方差和步长必须同时考虑。

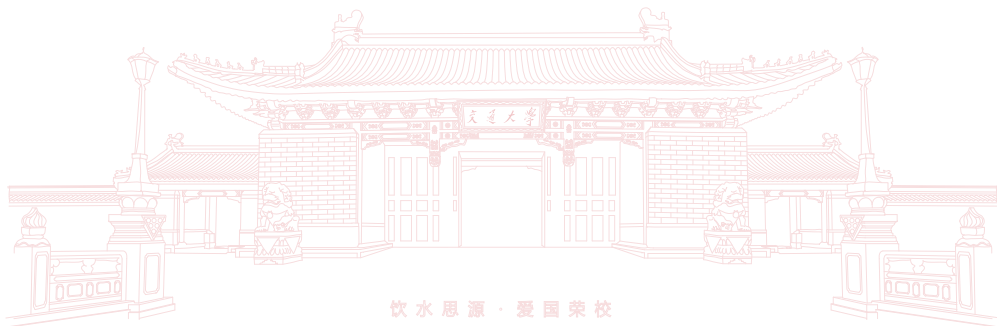


随机优化改变了信息模型（oracle model），不能只把它理解为“梯度法换了一个参数”。

Robbins & Monro (1951); Nesterov (1983); Polyak (1964)

01

梯度下降的频谱视角



函数类与条件数

设 $f: \mathbb{R}^n \rightarrow \mathbb{R}$ 可微, 且 $0 \leq m \leq M$ 。若对任意 x, y 有

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{M}{2} \|y - x\|_2^2,$$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2} \|y - x\|_2^2,$$

则称 $f \in \mathcal{F}_{m,M}$ 。当 $m > 0$ 时, 定义条件数

$$\kappa = \frac{M}{m}.$$

若最优解存在, 沿用 Lecture 03 的记号

$$x^* \in \arg \min f, \quad p^* = f(x^*).$$

M 控制局部变化

梯度不可能比 M 所允许的速度变化得更快; $1/M$ 因此成为自然的步长尺度。

m 控制全局曲率

强凸性排除过于平坦的方向; κ 衡量曲率最大与曲率最小方向之间的差距。

Nesterov, *Introductory Lectures on Convex Optimization*, Chapters 2–3

二次型是频谱显微镜

考虑

$$f(x) = \frac{1}{2}x^\top Qx - b^\top x, \quad Q \succ 0, \quad x^* = Q^{-1}b.$$

令误差 $e^{(k)} = x^{(k)} - x^*$, 梯度下降 $x^{(k+1)} = x^{(k)} - t\nabla f(x^{(k)})$ 化为

$$e^{(k+1)} = (I - tQ)e^{(k)}.$$

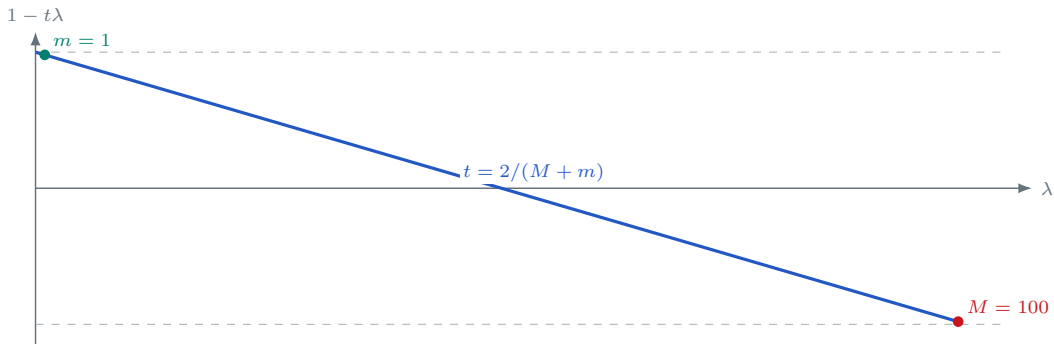
若 $Q = V \operatorname{diag}(\lambda_1, \dots, \lambda_n)V^\top$, 则

$$e^{(k)} = \sum_{i=1}^n c_i^{(k)} v_i, \quad c_i^{(k+1)} = (1 - t\lambda_i)c_i^{(k)}.$$

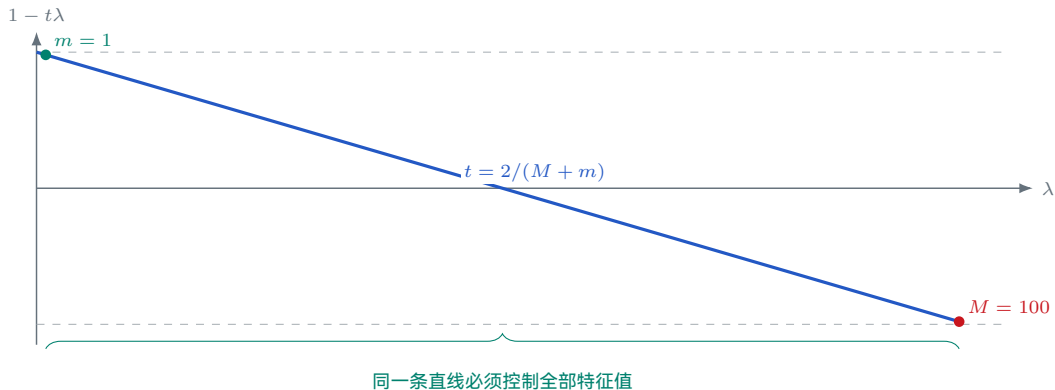
解释

每个特征方向独立演化。算法需要用同一个 t 同时处理整个谱区间 $[m, M]$ 。

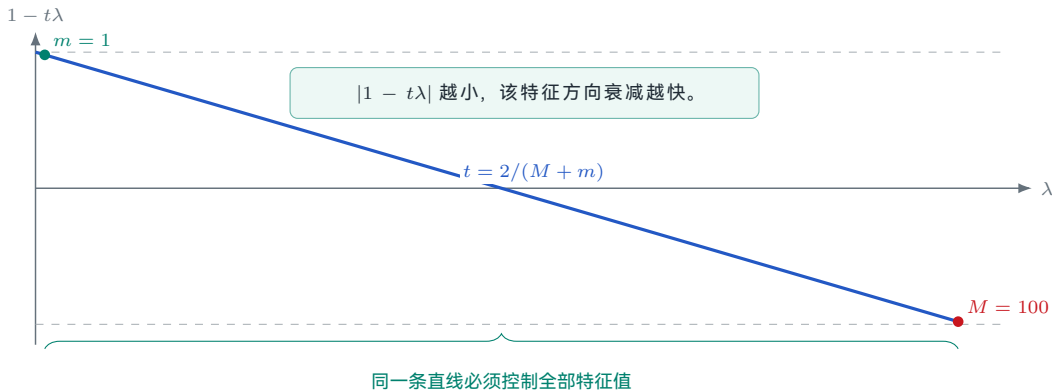
特征方向的衰减因子



特征方向的衰减因子



特征方向的衰减因子



二次型谱分解; 图示取 $m = 1, M = 100$, 纵坐标为无量纲衰减因子

步长的两个边界

对 M -smooth 凸函数, descent lemma 给出

$$f(\mathbf{x} - t\nabla f(\mathbf{x})) \leq f(\mathbf{x}) - t \left(1 - \frac{Mt}{2}\right) \|\nabla f(\mathbf{x})\|_2^2.$$



$1/M$ 是标准安全尺度; $2/M$ 是上式中下降系数保持为正的边界, 在 SPD 二次型上也正好是固定步长的谱稳定边界。

Descent lemma; quadratic stability condition $\rho(I - tQ) < 1$

最优固定步长

在 Q 的谱位于 $[m, M]$ 时，最坏方向的误差因子为

$$\rho(t) = \max_{\lambda \in [m, M]} |1 - t\lambda|.$$

最小化这个上包络，需要让区间两端等幅反号：

$$1 - tm = -(1 - tM).$$

因此

$$t_{\star} = \frac{2}{M + m}, \quad \rho_{\star} = \frac{M - m}{M + m} = \frac{\kappa - 1}{\kappa + 1}.$$

含义

当 $\kappa \gg 1$ 时， $\rho_{\star} \approx 1 - 2/\kappa$ 。即使固定步长已经最优，病态问题仍需要 $O(\kappa \log(1/\varepsilon))$ 次迭代。

实验 A: 病态二次问题

数值比较使用

$$f(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^{60} \lambda_i x_i^2, \quad \lambda_i \text{ 在 } [1, 1000] \text{ 上按对数均匀分布.}$$

初值取 $x_i^{(0)} \propto 1/\sqrt{\lambda_i}$, 使每个特征方向的初始目标值贡献相同。

问题参数

$$m = 1, \quad M = 1000, \quad \kappa = 1000.$$

所有方法从同一个 $\mathbf{x}^{(0)}$ 出发, 每次迭代调用一次完整梯度。

数据由随讲义提供的脚本生成, 未使用人工平滑或手工修改曲线。

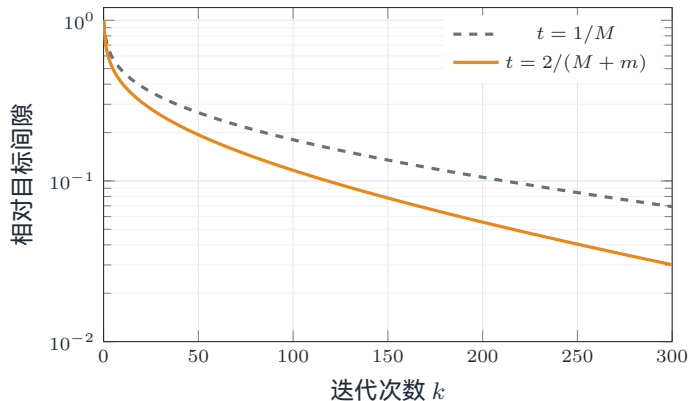
比较指标

$$\frac{f(\mathbf{x}^{(k)}) - p^*}{f(\mathbf{x}^{(0)}) - p^*}.$$

半对数坐标展示数量级; 另行记录目标值是否出现暂时上升。

自编可复现实验; `first-order-data/generate_first_order_data.py`

固定步长的数值比较



达到 10^{-2}

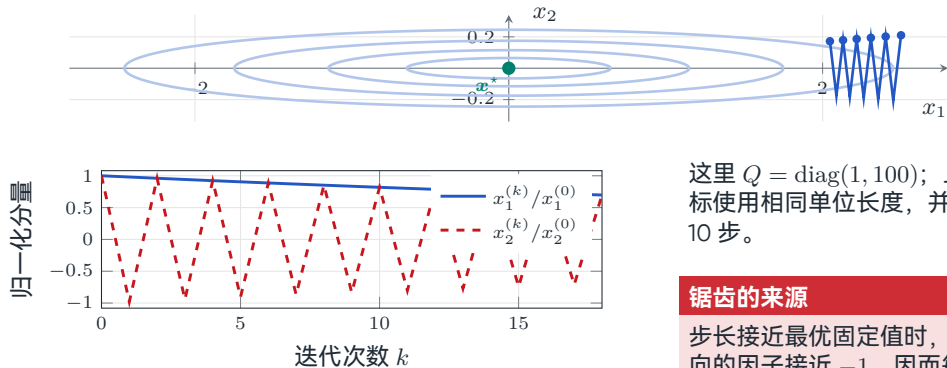
$1/M$ 923

$2/(M + m)$ 508

最优固定步长改善了最坏方向，却没有改变对 κ 的线性依赖。

自编数值实验; $n = 60, \kappa = 1000$

二维轨迹与锯齿



这里 $Q = \text{diag}(1, 100)$; 上图横纵坐标使用相同单位长度, 并只标出前 10 步。

锯齿的来源

步长接近最优固定值时, 高曲率方向的因子接近 -1 , 因而每一步都会换侧; 低曲率方向仍缓慢衰减。

自编二维二次例; 等比例坐标

固定步长的结构性限制

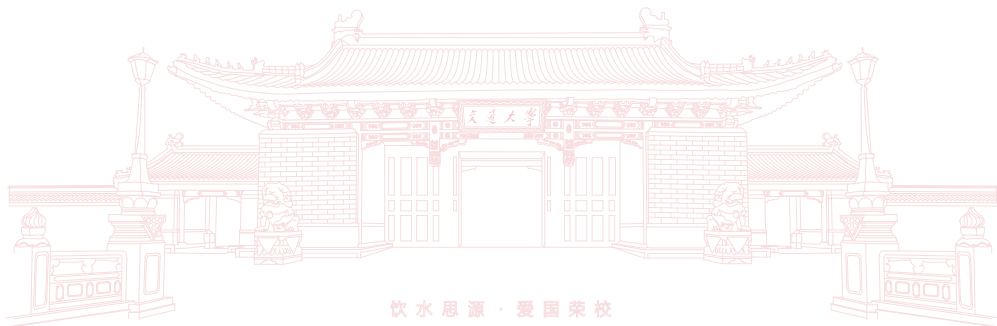


固定步长只能选择一条一次多项式 $1 - t\lambda$ 。加速方法通过跨迭代组合，构造更有利的高阶多项式。

谱滤波解释；Polyak (1964), Nesterov (1983)

02

Heavy-ball 与 Nesterov 加速



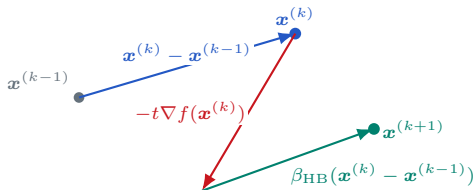
饮水思源 · 爱国荣校

Heavy-ball 方法

Polyak 的 heavy-ball (HB) 方法加入上一步位移:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - t \nabla f(\mathbf{x}^{(k)}) + \beta_{\text{HB}}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}).$$

这里 β_{HB} 表示动量系数, 与 Lecture 03 回溯线搜索中的缩减因子 β 无关。



新增状态

HB 是二阶递推。当前位置不足以决定下一步, 还需要保存上一位置或速度。

Polyak, "Some Methods of Speeding Up the Convergence of Iteration Methods," 1964

每个特征方向成为二阶系统

对 $f(x) = \frac{1}{2}x^\top Qx$, 在特征值 λ 对应的方向上, HB 满足

$$c^{(k+1)} = (1 + \beta_{\text{HB}} - t\lambda)c^{(k)} - \beta_{\text{HB}}c^{(k-1)}.$$

特征根由

$$r^2 - (1 + \beta_{\text{HB}} - t\lambda)r + \beta_{\text{HB}} = 0$$

给出。收敛要求对每个 $\lambda \in [m, M]$, 两根都位于单位圆内。

实根区间

两个衰减模态叠加; 参数改变根的位置和主导时间尺度。

复根区间

$|r| = \sqrt{\beta_{\text{HB}}}$ 时出现振荡衰减; 轨迹和目标值可以暂时上升。

Polyak (1964); scalar characteristic-polynomial analysis

HB 的二次型最优参数

若 $Q \succ 0$ 且 $\sigma(Q) \subseteq [m, M]$, 经典参数为

$$t_{\text{HB}} = \frac{4}{(\sqrt{M} + \sqrt{m})^2}, \quad \beta_{\text{HB}} = \left(\frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}} \right)^2.$$

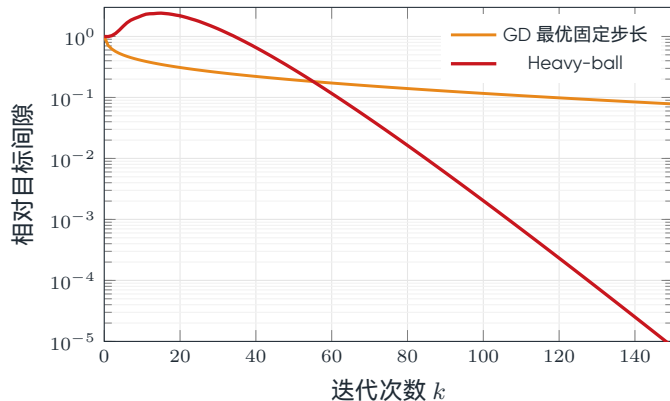
对应的误差模态渐近收敛因子为

$$\rho_{\text{HB}} = \sqrt{\beta_{\text{HB}}} = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \approx 1 - \frac{2}{\sqrt{\kappa}}.$$

适用范围

这个参数公式首先是 SPD 二次型的全局结论，也是一般光滑函数在非退化极小点附近的局部线性化结论。不能直接把它解释为任意 $f \in \mathcal{F}_{m,M}$ 上的统一全局加速保证。

HB 的暂态振荡



HB 的目标值最高达到初值的 2.41 倍，但随后进入明显更快的衰减阶段。

判读

非单调不等于发散；必须结合包络和最终指标判断。

实验 A；HB 使用二次型最优参数

参数失配的代价

对于给定 $\beta_{\text{HB}} \in [0, 1)$, SPD 二次型上一个基本稳定条件是

$$0 < t < \frac{2(1 + \beta_{\text{HB}})}{M}.$$

但“稳定”只说明特征根位于单位圆内，并不说明速度接近最优。

M 被低估

实际最大曲率超过模型值时，高曲率方向最先失稳；动量会放大持续振荡。

m 被高估

动量参数过于激进，曲率最小方向的预测不再准确；暂态可能显著延长。

工程实现通常结合回溯、阻尼、restart 或在线谱估计，而不是永久固定理论参数。

HB scalar recurrence and Jury stability test; quadratic setting

Nesterov 的前瞻梯度

强凸情形的一种常用写法为

$$\mathbf{y}^{(k)} = \mathbf{x}^{(k)} + \beta_{\text{NAG}}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}), \quad \mathbf{x}^{(k+1)} = \mathbf{y}^{(k)} - \frac{1}{M} \nabla f(\mathbf{y}^{(k)}).$$

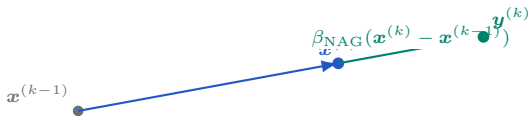


HB 在 $\mathbf{x}^{(k)}$ 计算梯度；Nesterov 方法在前瞻点 $\mathbf{y}^{(k)}$ 计算梯度。这一差异决定了证明结构和稳健性。

Nesterov 的前瞻梯度

强凸情形的一种常用写法为

$$\mathbf{y}^{(k)} = \mathbf{x}^{(k)} + \beta_{\text{NAG}}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}), \quad \mathbf{x}^{(k+1)} = \mathbf{y}^{(k)} - \frac{1}{M} \nabla f(\mathbf{y}^{(k)}).$$

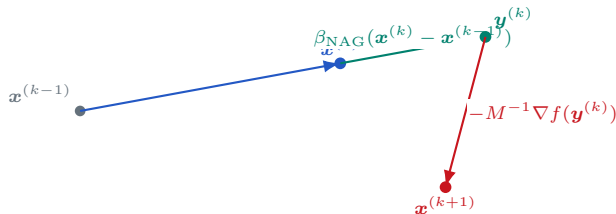


HB 在 $\mathbf{x}^{(k)}$ 计算梯度; Nesterov 方法在前瞻点 $\mathbf{y}^{(k)}$ 计算梯度。这一差异决定了证明结构和稳健性。

Nesterov 的前瞻梯度

强凸情形的一种常用写法为

$$\mathbf{y}^{(k)} = \mathbf{x}^{(k)} + \beta_{\text{NAG}}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}), \quad \mathbf{x}^{(k+1)} = \mathbf{y}^{(k)} - \frac{1}{M} \nabla f(\mathbf{y}^{(k)}).$$



HB 在 $\mathbf{x}^{(k)}$ 计算梯度；Nesterov 方法在前瞻点 $\mathbf{y}^{(k)}$ 计算梯度。这一差异决定了证明结构和稳健性。

Nesterov, "A Method of Solving a Convex Programming Problem with Convergence Rate $O(1/k^2)$," 1983

光滑凸情形的加速梯度

令 $\theta^{(0)} = 1$ 、 $\mathbf{y}^{(0)} = \mathbf{x}^{(0)}$ 。一种标准 fast gradient method (FGM) 为

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \mathbf{y}^{(k)} - \frac{1}{M} \nabla f(\mathbf{y}^{(k)}), \\ \theta^{(k+1)} &= \frac{1 + \sqrt{1 + 4(\theta^{(k)})^2}}{2}, \\ \mathbf{y}^{(k+1)} &= \mathbf{x}^{(k+1)} + \frac{\theta^{(k)} - 1}{\theta^{(k+1)}} (\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}).\end{aligned}$$

若 $f \in \mathcal{F}_{0,M}$ 且 $\mathbf{x}^*$ 存在, 则

$$f(\mathbf{x}^{(k)}) - p^* \leq \frac{2M \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2^2}{(k+1)^2}.$$

复杂度变化

GD 的典型函数值界为 $O(1/k)$; FGM 达到 $O(1/k^2)$ 。这不是常数优化, 而是阶的改变。

强凸情形的常动量

若 $f \in \mathcal{F}_{m,M}$ 且 $m > 0$, 取

$$t = \frac{1}{M}, \quad \beta_{\text{NAG}} = \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}} = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}.$$

典型函数值收敛因子为

$$f(\mathbf{x}^{(k)}) - p^* \leq C \left(1 - \sqrt{\frac{m}{M}}\right)^k,$$

其中 C 由初始状态决定。

梯度下降

$$O\left(\kappa \log \frac{1}{\varepsilon}\right)$$

加速梯度

$$O\left(\sqrt{\kappa} \log \frac{1}{\varepsilon}\right)$$

Nesterov (1983, 2004); smooth strongly convex acceleration

HB 与 NAG 的模态递推

对特征值 λ ，两种方法分别给出

$$\text{HB: } c^{(k+1)} = (1 + \beta_{\text{HB}} - t\lambda)c^{(k)} - \beta_{\text{HB}}c^{(k-1)},$$

$$\text{NAG: } c^{(k+1)} = (1 - t\lambda)(1 + \beta_{\text{NAG}})c^{(k)} - \beta_{\text{NAG}}(1 - t\lambda)c^{(k-1)}.$$

共同点

都使用两步记忆，把一次多项式滤波扩展为二阶递推；参数依赖曲率区间。

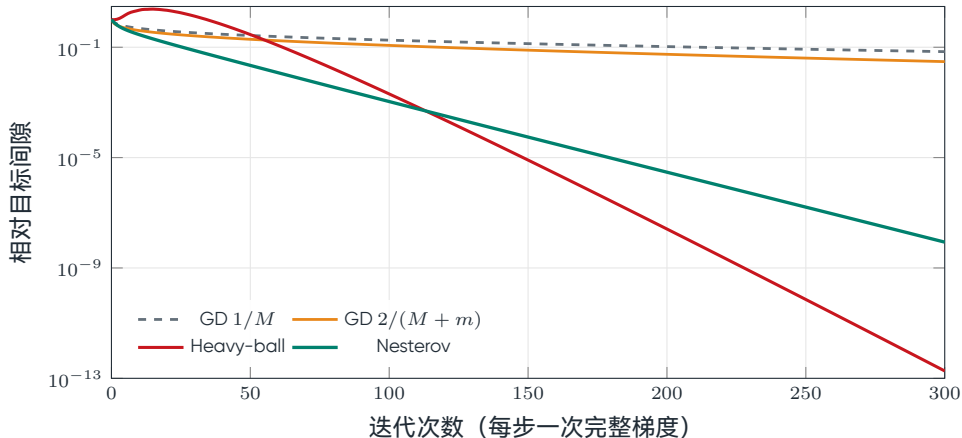
差别

NAG 的两个历史系数都含 $1 - t\lambda$ ，反映梯度在前瞻点计算。

相似的代码形式不意味着相同的动力学或相同的全局保证。

Quadratic-mode recurrences derived from the HB and NAG updates

确定性方法的数值比较



实验 A; $n = 60, m = 1, M = 1000$; 相同初值与梯度调用成本

速度与暂态需要同时报告

方法	达到 10^{-2}	达到 10^{-4}	达到 10^{-6}	最大相对目标值
GD, $t = 1/M$	923	2864	5018	1.00
GD, $t = 2/(M + m)$	508	1535	2644	1.00
Heavy-ball	85	128	169	2.41
Nesterov	63	141	219	1.00

HB 在本例的高精度阶段最快，但有明显过冲。

Nesterov 方法前期更快，且这组参数下目标值没有超过初值。

一条曲线不能替代参数、成本和暂态信息。

实验 A； 阈值按相对目标间隙首次低于给定值统计

Restart 与 阻尼

加速方法可能因参数失配或局部几何变化出现无效振荡。常见修正包括：

函数值 restart

若

$$f(\mathbf{x}^{(k)}) > f(\mathbf{x}^{(k-1)}),$$

则把动量参数重置为零；下一步以 $\mathbf{y}^{(k)} = \mathbf{x}^{(k)}$ 重新计算，不再沿用上一段外推状态。

梯度 restart

若

$$\langle \nabla f(\mathbf{y}^{(k-1)}), \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)} \rangle > 0,$$

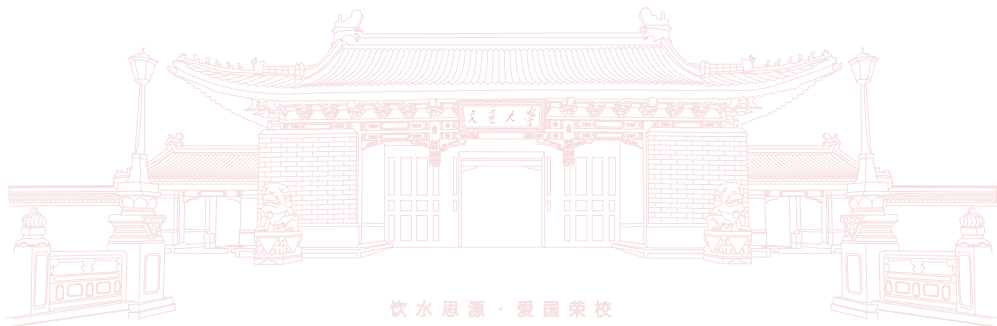
则动量方向已与下降需求不一致，可重启动量。

Restart 保留一阶计算成本，并减少对 m 精确估计的依赖；它改变的是算法，而不只是停止准则。

O'Donoghue & Candès, "Adaptive Restart for Accelerated Gradient Schemes," 2015

03

Performance Estimation Problem



饮水思源 · 爱国荣校

经验曲线与最坏情形

一组数值实验回答

“这个方法在这个问题上表现怎样？”

复杂度分析则希望回答

“在整个函数类上，它最坏能有多差？”

传统证明

选择若干已知不等式，逐步组合，得到一个对所有问题都成立的上界。证明中的松弛可能累积。

Performance Estimation Problem

把最坏函数、最坏初值和算法轨迹一起作为优化变量，直接最大化给定性能指标。

PEP 优化的是算法的最坏情形；它不负责求解原来的 $\min f(x)$ 。

PEP 的基本形式

固定函数类 $\mathcal{F}$ 、算法 $\mathcal{A}$ 、迭代次数 N 与初始半径 R 。以最终目标残差为指标，定义

$$\begin{aligned}\mathcal{W}_N(\mathcal{A}, \mathcal{F}, R) &:= \sup_{f, \mathbf{x}^{(0)}, \dots, \mathbf{x}^{(N)}, \mathbf{x}^*} f(\mathbf{x}^{(N)}) - f(\mathbf{x}^*) \\ &\text{subject to } f \in \mathcal{F}, \\ &\quad \mathbf{x}^{(k+1)} = \mathcal{A}_k(\mathbf{x}^{(0)}, \dots, \mathbf{x}^{(k)}; f_0, \dots, f_k; \mathbf{g}^{(0)}, \dots, \mathbf{g}^{(k)}), \\ &\quad \mathbf{x}^* \in \arg \min f, \quad \left\| \mathbf{x}^{(0)} - \mathbf{x}^* \right\|_2^2 \leq R^2.\end{aligned}$$

算法求解原问题

对一个固定的 f ，算法希望让 $f(\mathbf{x}^{(k)})$ 尽可能小。

PEP 评估最坏情形

允许的 f 与初值充当“对手”，使第 N 步误差尽可能大。

上确界的含义

$\mathcal{W}_N$ 是对整个函数类都成立的最小统一上界。对本讲的光滑凸函数类， $\mathbf{x}^* \in \arg \min f$ 等价于 $\nabla f(\mathbf{x}^*) = \mathbf{0}$ 。

Drori & Teboulle (2014), DOI: 10.1007/s10107-013-0653-0

有限插值条件

PEP 只需要记录 oracle transcript

$$(\mathbf{x}^{(i)}, f_i, \mathbf{g}^{(i)}), \quad f_i = f(\mathbf{x}^{(i)}), \quad \mathbf{g}^{(i)} = \nabla f(\mathbf{x}^{(i)}).$$

一组有限数据可由某个 $f \in \mathcal{F}_{0,M}$ 插值得到，当且仅当对所有 i, j 有

$$f_i - f_j - \langle \mathbf{g}^{(j)}, \mathbf{x}^{(i)} - \mathbf{x}^{(j)} \rangle \geq \frac{1}{2M} \left\| \mathbf{g}^{(i)} - \mathbf{g}^{(j)} \right\|_2^2.$$

左侧同时包含凸性的一阶下界和两个采样点之间的函数值一致性。

右侧是光滑性带来的额外曲率约束；只写凸性会得到更大的函数类。

这一步把“所有光滑凸函数”转化为有限个代数不等式。

从内积到半正定规划



把 z_a 取为初始位移和各次梯度后，距离、梯度范数和算法更新中的内积都成为 G 的线性函数。因此这里讨论的固定系数一阶算法，其 PEP 可化为有限维 semidefinite program (SDP)。

Primal SDP: 最坏实例

$$\begin{aligned} &\text{maximize } \Phi_N. \\ &G \succeq 0, \{f_i\} \end{aligned}$$

原始可行解描述一个具体轨迹，最优值给出最坏性能。

Dual SDP: 统一上界

$$\begin{aligned} &\text{minimize } \tau R^2. \\ &\tau, \lambda \end{aligned}$$

对偶权重组合出 $\Phi_N \leq \tau R^2$ 的证明证书。

在强对偶成立时，最坏实例的性能与最小统一上界相等。

GD 的经典界与紧界

对 $f \in \mathcal{F}_{0,M}$ 、 $t = 1/M$ 和 $\|x^{(0)} - x^*\|_2 \leq R$ ，常见教材证明给出

$$f(x^{(N)}) - p^* \leq \frac{MR^2}{2N}.$$

PEP 揭示并可证明更紧的最坏情形界

$$f(x^{(N)}) - p^* \leq \frac{MR^2}{4N+2}.$$

N	1	2	5	10
经典界 $/(MR^2)$	.5000	.2500	.1000	.0500
紧界 $/(MR^2)$	.1667	.1000	.0455	.0238

PEP 在这里改进的是分析常数；GD 的 $O(1/N)$ 阶并未改变。

Drori & Teboulle (2014), Theorem 3.1; tight bound for $t = 1/M$

一个可计算的 PEP workflow

```
problem = PEP()
f = problem.declare_function(
    SmoothConvexFunction, L=M)
x0 = problem.set_initial_point()
xs = f.stationary_point()
problem.set_initial_condition(
    (x0 - xs) ** 2 <= R ** 2)

x = x0
for k in range(N):
    x = x - (1 / M) * f.gradient(x)

problem.set_performance_metric(f(x) - f(xs))
tau_N = problem.solve()
```

建模顺序

1. 声明函数类;
2. 写出初始归一化;
3. 按原样写算法;
4. 指定最终性能指标;
5. 求解 SDP 并检查证书。

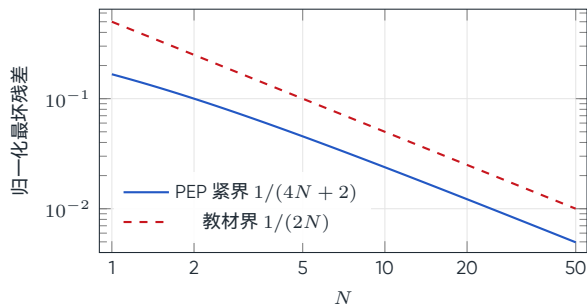


Goujaud et al., "PEPit: computer-assisted worst-case analyses," 2022, arXiv:2201.04040

由有限步 PEP 得到收敛曲线

固定函数类、算法参数和归一化条件，对不同的迭代次数分别求解

$$\hat{\tau}_N = \sup \left\{ \frac{f(\mathbf{x}^{(N)}) - p^*}{MR^2} \mid \begin{array}{l} f \in \mathcal{F}_{0,M}, \mathbf{x}^* \in \arg \min f, \|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2 \leq R, \\ \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - M^{-1} \mathbf{g}^{(k)}, \quad k = 0, \dots, N-1 \end{array} \right\}.$$



数值步骤

对 $N = 1, 2, \dots, N_{\max}$ 重复建立并求解 SDP，记录 $\hat{\tau}_N$ ，再比较候选收敛律。

收敛阶的数值诊断

log-log 斜率 ≈ -1	$O(1/N)$
log-log 斜率 ≈ -2	$O(1/N^2)$
semi-log 近似直线	$O(\rho^N)$

这里是在整个函数类上求有限步最坏值，而不是随机抽取测试函数；有限个 N 的曲线用于诊断候选收敛律，严格的全 N 结论仍需对偶证书或解析证明。

Taylor, Hendrickx & Glineur (2017); Goujaud et al. (2022)

PEP 也可以用于方法设计

固定 h 时，内层 PEP 沿用 $f \in \mathcal{F}$ 、 $\mathbf{x}^* \in \arg \min f$ 和 $\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2 \leq R$ 。把算法系数作为设计变量时，问题成为

$$\underset{h \in \mathcal{H}}{\text{minimize}} \tau_N(h), \quad \tau_N(h) := \sup_{f, \mathbf{x}^{(0)}, \mathbf{x}^*} \Phi_N(h; f, \mathbf{x}^{(0)}, \mathbf{x}^*)$$

外层：设计算法

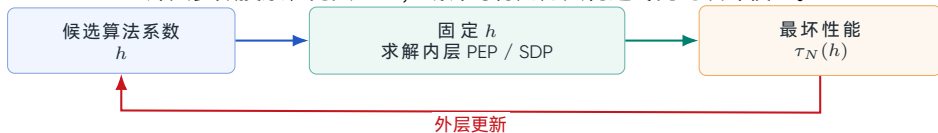
$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(0)} - \frac{1}{M} \sum_{j=0}^k h_{k+1,j} \mathbf{g}^{(j)}$$

选择 $h_{k,j}$ 、 $t^{(k)}$ 或动量系数 $\beta_{\text{mom}}^{(k)}$ ，压低最坏界。

内层：评估最坏情形

固定一组系数后，PEP 在指定函数类上寻找最坏函数、初值与算法轨迹。

固定 h 的内层可化为 SDP；联合优化时， h 与梯度或 Gram 变量发生乘积，整体通常非凸。实践中可采用外层参数搜索、内层 PEP，或针对特定方法构造专门的设计模型。



Drori & Teboulle (2014); Kim & Fessler, *Math. Program.* 159 (2016)

PEP 结论的边界

它能回答

- ▣ 给定函数类上的紧或数值紧最坏界;
- ▣ 一个方法的最坏实例;
- ▣ 对偶权重对应的证明证书;
- ▣ 有限时域内的参数设计。

建模时仍需判断

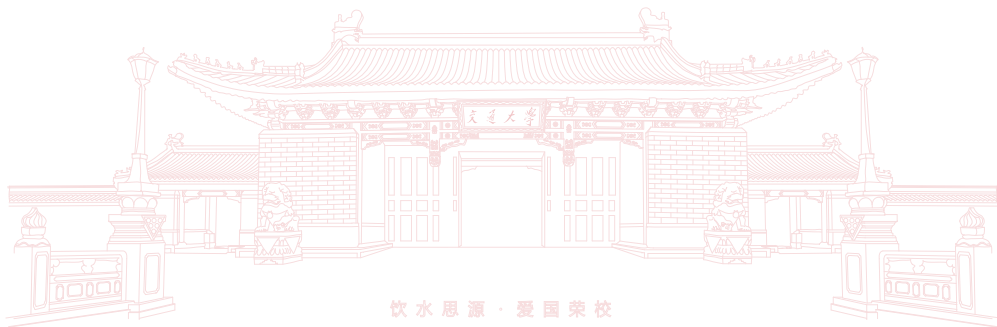
- ▣ M, m 是否可信;
- ▣ oracle 是否精确;
- ▣ 指标是函数值、梯度范数还是距离;
- ▣ 数值 SDP 的容差和对偶残差是否足够小。

PEP 的结论与所声明的函数类和信息模型同样严格，也同样有限。

Taylor, Hendrickx & Glineur (2017); Goujaud et al. (2022)

04

大步长梯度下降



“大步长”的三种含义

语境	代表尺度	结论范围
标准单步尺度	$t \leq 1/M$	一般 M -smooth 函数; 下降余量至少为 $t \ \nabla f\ _2^2 / 2$
单步下降边界	$0 < t < 2/M$	一般 M -smooth 函数仍下降; 在 SPD 二次型上等价于谱稳定
多步协同	某些 $t^{(k)} > 2/M$	依赖完整步长序列与全程证书; 单步可上升

大步长策略研究的是多步组合, 不能简化为“把学习率调得更大”。

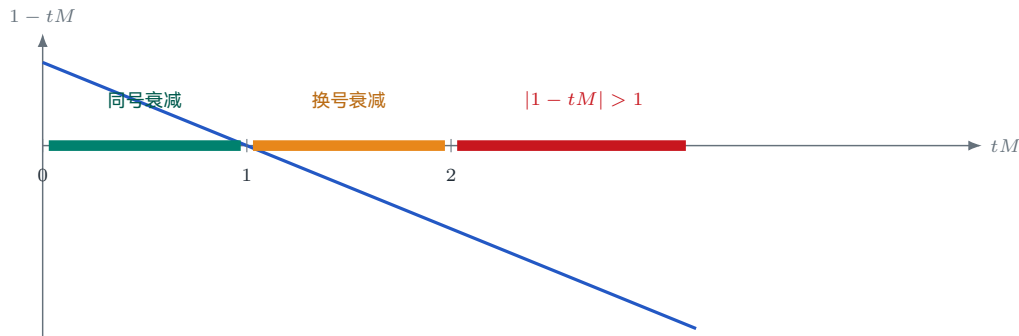
本节顺序

先用二次型的多项式解释有限步协同, 再介绍对一般光滑凸函数成立的 Silver stepsize schedule。

重复使用长步长会失稳

对一维二次函数 $f(x) = \frac{M}{2}x^2$ ，梯度法为

$$x^{(k+1)} = (1 - tM)x^{(k)}.$$



若每一步都取同一个 $t > 2/M$ ，最大曲率方向必然发散。长步长加速必须依靠不同步长之间的补偿。

有限步谱滤波

对二次型使用步长序列 $t^{(0)}, \dots, t^{(N-1)}$, 有

$$e^{(N)} = p_N(Q)e^{(0)}, \quad p_N(\lambda) = \prod_{k=0}^{N-1} (1 - t^{(k)}\lambda), \quad p_N(0) = 1.$$

有限时域的理想问题为

$$\underset{p_N(0)=1, \deg p_N \leq N}{\text{minimize}} \quad \max_{\lambda \in [m, M]} |p_N(\lambda)|.$$

固定步长只能使用 $(1 - t\lambda)^N$, 每个根都相同。

变步长允许把根分布在谱区间内, 直接压低最终的最坏放大因子。

这一定义只控制 p_N 的终点值; 中间前缀多项式仍可能很大。

Chebyshev 步长序列

令

$$c = \frac{M+m}{2}, \quad d = \frac{M-m}{2}.$$

二次型有限时域 minimax 多项式的根给出

$$t^{(j)} = \frac{1}{c - d \cos \frac{(2j-1)\pi}{2N}}, \quad j = 1, \dots, N.$$

其中部分步长可能远大于 $2/M$ 。最终收缩满足

$$\max_{\lambda \in [m, M]} |p_N(\lambda)| \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^N.$$

数值实现

步长的排列不改变最终多项式，却会显著改变中间放大和舍入误差。需要专门排序或使用稳定的 Chebyshev recurrence。

Silver stepsize schedule

令银比 (silver ratio) $\rho = 1 + \sqrt{2}$, $\nu_2(s)$ 表示正整数 s 的 2-进赋值 (2-adic valuation)。对 $j = 0, 1, 2, \dots$ 定义

$$\alpha^{(j)} = 1 + \rho^{\nu_2(j+1)-1}, \quad \mathbf{x}^{(j+1)} = \mathbf{x}^{(j)} - \frac{\alpha^{(j)}}{M} \nabla f(\mathbf{x}^{(j)}).$$

前 15 个归一化步长为

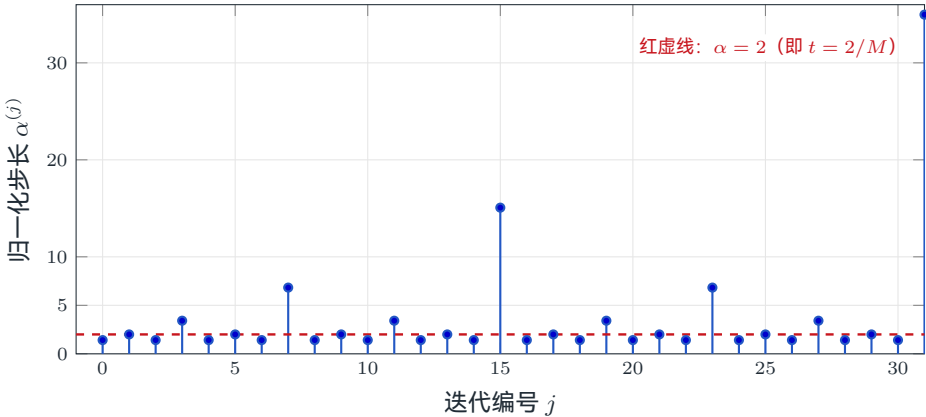
$$\sqrt{2}, 2, \sqrt{2}, 2 + \sqrt{2}, \sqrt{2}, 2, \sqrt{2}, 1 + \rho^2, \dots$$

结构

序列非单调并呈递归结构；越来越大的尖峰出现得越来越稀疏。算法仍是纯梯度下降，没有动量状态。

Altschuler & Parrilo, "Acceleration by Stepsize Hedging II," arXiv:2309.16530, Definition 2.1

Silver 步长的递归尖峰



尖峰超过二次型单步稳定阈值，但整段序列具有多步下降证书。

Altschuler & Parrilo (2023–2024); first 32 normalized Silver steps

Silver schedule 的一般光滑凸保证

设 $f \in \mathcal{F}_{0,M}$ 。令 $N_q = 2^q - 1$ 。Silver schedule 满足

$$f(\mathbf{x}^{(N_q)}) - p^* \leq r_q M \left\| \mathbf{x}^{(0)} - \mathbf{x}^* \right\|_2^2, \quad r_q = \frac{1}{1 + \sqrt{4\rho^{2q} - 3}}.$$

进一步有

$$r_q \leq \frac{1}{2N_q^{\log_2 \rho}} \approx \frac{1}{2N_q^{1.2716}}.$$

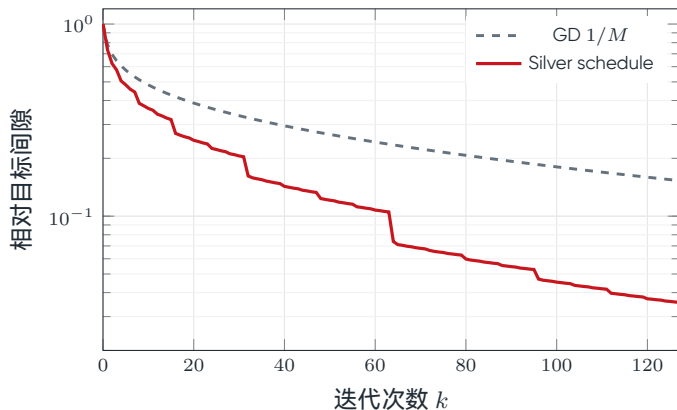
沿 $N_q = 2^q - 1$ 这些 horizon (或在不超过目标迭代数的最近此类 horizon 停止), 达到精度 ε 所需迭代次数为

$$O\left(\varepsilon^{-\log_\rho 2}\right) \approx O(\varepsilon^{-0.7864}).$$

位置

该速率快于标准 GD 的 $O(1/N)$, 但仍慢于 Nesterov 加速的 $O(1/N^2)$ 。它说明仅改变步长序列也能得到严格的渐近加速。

Silver steps 的数值表现



实验 A 在 $k = 127$ 时得到

$$\begin{aligned}\text{GD:} & 1.53 \times 10^{-1}, \\ \text{Silver:} & 3.56 \times 10^{-2}.\end{aligned}$$

这是一个强凸二次实例；一般保证来自前页定理，而非这条曲线。

自编实验 A; Silver schedule 使用 $\alpha^{(j)} / M$

多步证书取代逐步下降

对 M -smooth 凸函数，任意两个 oracle 样本满足

$$Q_{ij} := 2M \left(f_i - f_j - \langle \mathbf{g}^{(j)}, \mathbf{x}^{(i)} - \mathbf{x}^{(j)} \rangle \right) - \left\| \mathbf{g}^{(i)} - \mathbf{g}^{(j)} \right\|_2^2 \geq 0.$$

长步长分析寻找非负权重 λ_{ij} 和由递归证书确定的标量 c_q ，把多个 Q_{ij} 组合成

$$\sum_{i \neq j} \lambda_{ij} Q_{ij} = M \left\| \mathbf{x}^{(0)} - \mathbf{x}^* \right\|_2^2 - M \left\| \mathbf{x}^{(N_q)} - c_q \mathbf{g}^{(N_q)} - \mathbf{x}^* \right\|_2^2 + \frac{p^* - f_{N_q}}{r_q}.$$

由于左侧非负，移项后得到

$$\frac{f_{N_q} - p^*}{r_q} \leq M \left\| \mathbf{x}^{(0)} - \mathbf{x}^* \right\|_2^2 - M \left\| \mathbf{x}^{(N_q)} - c_q \mathbf{g}^{(N_q)} - \mathbf{x}^* \right\|_2^2 \leq M \left\| \mathbf{x}^{(0)} - \mathbf{x}^* \right\|_2^2.$$

关键变化

证明不再要求每一步都下降，而是利用相隔多步的插值一致性；这正是 PEP 语言与长步长序列的连接。

长步长策略的适用边界

需要检查

- ▣ M 的估计是否足够可靠;
- ▣ 目标是否满足所用函数类假设;
- ▣ 大中间值是否造成溢出或舍入误差;
- ▣ 梯度是否含噪声;
- ▣ 证书控制的是最终点还是最优遍历点。

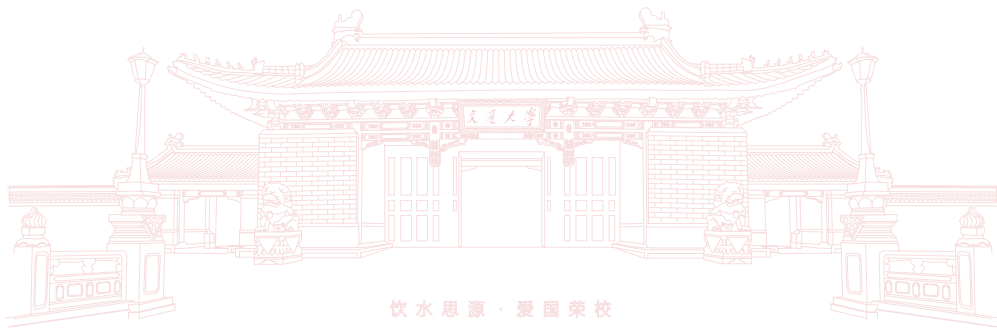
不要混同

深度学习中的 edge of stability 描述非凸训练时局部 Hessian 与学习率共同演化的经验现象。它与本节光滑凸函数上的可证明长步长序列并非同一结论。

Chebyshev、Silver 与 Nesterov 都能用谱滤波理解，但保证范围和状态变量不同。

05

随机优化



饮水思源 · 爱国荣校

两类随机性

随机目标

$$F(\boldsymbol{x}) = \mathbb{E}_{\xi}[\ell(\boldsymbol{x}; \xi)].$$

期望本身可能无法精确计算，只能从分布或数据流中抽样。

有限和的随机求解

$$F(\boldsymbol{x}) = \frac{1}{N_{\text{data}}} \sum_{i=1}^{N_{\text{data}}} f_i(\boldsymbol{x}).$$

目标是确定的，但完整梯度成本为 N_{data} 个样本梯度，随机抽样可降低单步成本。

本节重点使用有限和模型。若 i_k 均匀抽样，则

$$\hat{\boldsymbol{g}}^{(k)} = \nabla f_{i_k}(\boldsymbol{x}^{(k)}), \quad \mathbb{E}[\hat{\boldsymbol{g}}^{(k)} \mid \boldsymbol{x}^{(k)}] = \nabla F(\boldsymbol{x}^{(k)}).$$

无偏性保证方向平均正确，却不消除单步方差。下文仍记 $\boldsymbol{x}^* \in \arg \min F$ ，并令 $p^* = F(\boldsymbol{x}^*)$ 。

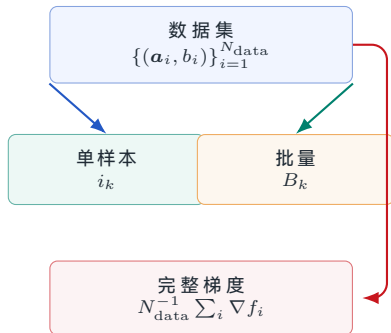
Robbins & Monro, *Annals of Mathematical Statistics* 22 (1951), 400–407

有限和模型的实际来源

经验风险最小化

$$F(\mathbf{x}) = \frac{1}{N_{\text{data}}} \sum_{i=1}^{N_{\text{data}}} \ell(\mathbf{a}_i^\top \mathbf{x}, b_i) + \frac{\lambda}{2} \|\mathbf{x}\|_2^2.$$

每个 f_i 对应一个样本，常见于回归、分类和参数估计。



在流式传感、在线控制和大规模机器学习中，“一次完整遍历数据”比“更新一次参数”更适合作为计算成本单位。

随机梯度下降

stochastic gradient descent (SGD) 每步只使用一个随机梯度:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - t^{(k)} \hat{\mathbf{g}}^{(k)}.$$

单步成本

若一个样本梯度成本为 C ，完整梯度约为 $N_{\text{data}}C$ ，SGD 一步约为 C 。因此相同迭代次数并不代表相同工作量。

停止指标

训练中常同时监控

$$F(\mathbf{x}^{(k)}), \quad \left\| \nabla F(\mathbf{x}^{(k)}) \right\|_2,$$

以及验证集指标 (validation metric)。随机单步损失不能替代完整目标值。

理论分析必须说明抽样方式、条件无偏性、方差假设以及步长是否依赖历史信息。

Robbins & Monro (1951); standard SGD recursion

方差产生噪声地板

假设

$$\mathbb{E}[\hat{\mathbf{g}}^{(k)} \mid \mathbf{x}^{(k)}] = \nabla F(\mathbf{x}^{(k)}), \quad \mathbb{E}[\|\hat{\mathbf{g}}^{(k)} - \nabla F(\mathbf{x}^{(k)})\|_2^2 \mid \mathbf{x}^{(k)}] \leq \sigma^2,$$

且 $F \in \mathcal{F}_{m,M}$, 其中 $0 < m \leq M$ 。令 $\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}^*$ 。当 $0 < t^{(k)} \leq 2/(M+m)$ 时, 强凸光滑插值不等式给出

$$\mathbb{E}[\|\mathbf{e}^{(k+1)}\|_2^2 \mid \mathbf{x}^{(k)}] \leq \left(1 - \frac{2t^{(k)}mM}{m+M}\right) \|\mathbf{e}^{(k)}\|_2^2 + (t^{(k)})^2 \sigma^2.$$

常步长

前期近似线性下降, 随后均方距离停在 $O(t\sigma^2/m)$ 的量级; 对应半径为其平方根量级。

递减步长

逐步降低噪声影响, 可趋近精确最优点; 后期更新也会越来越慢。

Standard strongly convex SGD recursion under bounded conditional variance

步长与平均迭代

对凸问题, 若 $\|x^{(0)} - x^*\|_2 \leq R$, 随机次梯度二阶矩有界为 G^2 (有约束时采用投影 SGD), 平均迭代可满足典型界

$$\mathbb{E}[F(\bar{x}^{(T)}) - p^*] \leq \frac{R^2}{2tT} + \frac{tG^2}{2}.$$

取 $t = R/(G\sqrt{T})$ 得 $O(1/\sqrt{T})$ 。

Robbins–Monro 条件

$$\sum_{k=0}^{\infty} t^{(k)} = \infty, \quad \sum_{k=0}^{\infty} (t^{(k)})^2 < \infty.$$

它平衡持续移动与累计噪声。

Polyak–Ruppert 平均

$$\bar{x}^{(T)} = \frac{1}{T - T_0} \sum_{k=T_0+1}^T x^{(k)}.$$

平均轨迹可降低末端方差并改善统计效率。

Robbins & Monro (1951); Polyak & Juditsky, SIAM J. Control Optim. 30 (1992)

Mini-batch 梯度

对大小为 B 的随机批量 $\mathcal{B}_k$ ，定义

$$\hat{\mathbf{g}}^{(k)} = \frac{1}{B} \sum_{i \in \mathcal{B}_k} \nabla f_i(\mathbf{x}^{(k)}).$$

独立抽样下，若单样本方差为 σ^2 ，则近似有

$$\mathbb{E} \left\| \hat{\mathbf{g}}^{(k)} - \nabla F(\mathbf{x}^{(k)}) \right\|_2^2 \leq \frac{\sigma^2}{B}.$$

收益

方差下降，矩阵运算和硬件并行度提高，轨迹更平滑。

代价

每次更新处理 B 个样本。超过硬件或统计临界批量后，继续增大 B 未必减少总耗时。

比较不同 batch size 时，横轴应使用 sample-gradient evaluations 或 effective passes。

Mini-batch variance under independent sampling

实验 B: 正则化最小二乘

使用合成有限和问题

$$F(\mathbf{x}) = \frac{1}{2N_{\text{data}}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \frac{\lambda}{2} \|\mathbf{x}\|_2^2, \quad N_{\text{data}} = 2000, \quad n = 40, \quad \lambda = 0.01.$$

数据

$\mathbf{A}$ 的列尺度从 1 到 0.1 按对数变化;
 $\mathbf{b} = \mathbf{A}\mathbf{x}_{\text{true}} + 0.2\xi$; 随机种子为 6015。

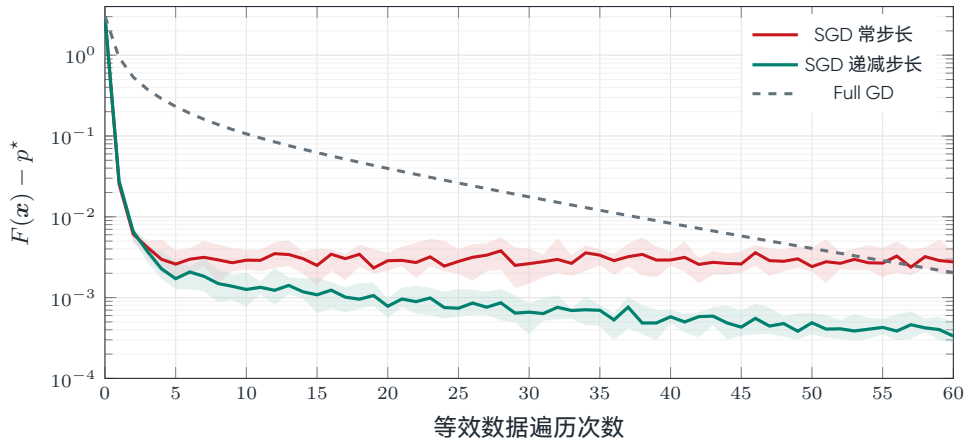
曲率

$m = 0.019599, \quad M = 0.989440, \quad \kappa = 50.484.$
 $\mathbf{x}^*$ 由正规方程高精度求得。

随机方法使用 12 个独立抽样种子; 实线为中位数, 阴影为 10%–90% 分位区间。

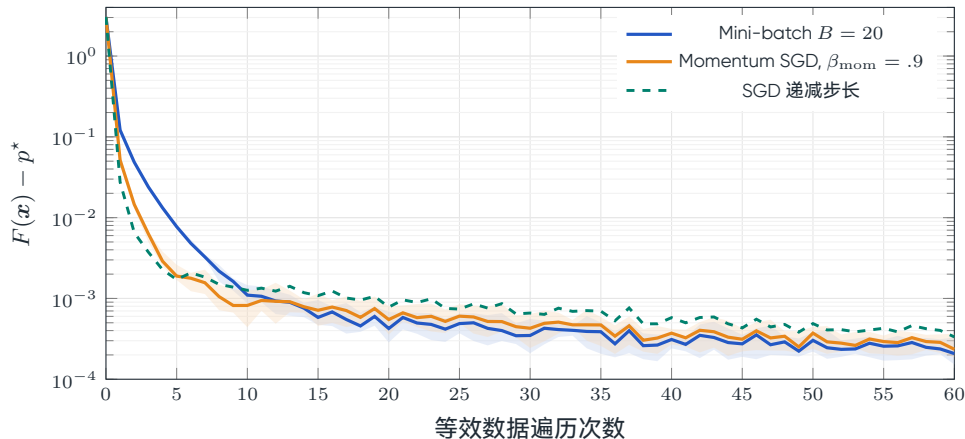
自编可复现实验; `first-order-data/generate_first_order_data.py`

常步长与递减步长 SGD



实验 B; 常步长 $0.015/M$; 递减步长 $(0.015/M)/(1 + 0.1 \text{ passes})$

Mini-batch 与随机动量



实验 B; 阴影为 12 次独立抽样的 10%-90% 分位带

随机动量需要单独分析

一种常见 momentum SGD 写法为

$$\mathbf{v}^{(k+1)} = \beta_{\text{mom}} \mathbf{v}^{(k)} + (1 - \beta_{\text{mom}}) \hat{\mathbf{g}}^{(k)}, \quad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - t^{(k)} \mathbf{v}^{(k+1)}.$$

潜在收益

速度状态对高频抽样噪声进行时间平均，并可在持续方向上积累移动。

新增困难

噪声与速度状态相关；确定性 HB 的参数和收敛因子不能直接搬到随机情形。

实验 B 中，动量曲线比单样本常步长更平滑，但在等效数据成本下并未自动优于 mini-batch。算法名不能替代调参和公平比较。

Stochastic momentum recursion; experiment B

SVRG 的控制变量

对 $F(\mathbf{x}) = N_{\text{data}}^{-1} \sum_i f_i(\mathbf{x})$, 在 snapshot $\tilde{\mathbf{x}}$ 计算完整梯度, 并在内循环使用

$$\hat{\mathbf{g}}_{\text{vr}}^{(k)} = \nabla f_{i_k}(\mathbf{x}^{(k)}) - \nabla f_{i_k}(\tilde{\mathbf{x}}) + \nabla F(\tilde{\mathbf{x}}).$$

无偏性

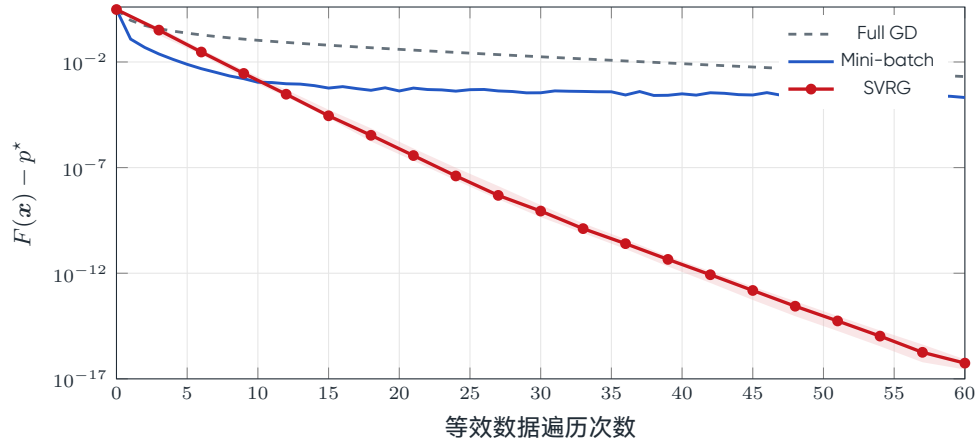
$$\mathbb{E}[\hat{\mathbf{g}}_{\text{vr}}^{(k)} \mid \mathbf{x}^{(k)}, \tilde{\mathbf{x}}] = \nabla F(\mathbf{x}^{(k)}).$$

方差缩减

当 $\mathbf{x}^{(k)}$ 接近 $\tilde{\mathbf{x}}$ 时, 两个样本梯度高度相关, 其差值方差变小。

SVRG 周期性支付完整梯度成本, 以换取内循环中不断减小的估计方差。

SVRG 的等效成本比较



实验 B; SVRG 一个 outer epoch 按 $N_{\text{data}} + 2N_{\text{data}} = 3N_{\text{data}}$ 个样本梯度计为 3 次遍历

公平比较随机方法

方法	单次更新成本	主要误差来源	合适场景
Full GD	N_{data} 个样本梯度	条件数	数据可完整遍历、需要稳定高精度
SGD	1 个样本梯度	抽样方差	数据流、大规模低到中等精度
Mini-batch	B 个样本梯度	约 σ^2/B	并行硬件、希望轨迹更平滑
SVRG	周期性完整梯度	snapshot 更新间隔	有限和、光滑强凸、高精度

至少同时报告：样本梯度调用数、batch size、步长规则、随机重复次数及分位带。壁钟时间还依赖内存、并行度和数据访问模式。

Complexity accounting used in experiment B

流式传感器标定

Lecture 03 的标定模型可写成样本损失

$$f_i(c, g) = \frac{1}{2}(c + gt_i - z_i)^2 + \frac{\lambda}{2}(c^2 + g^2).$$

当测量 (t_i, z_i) 持续到达时, SGD 更新为

$$\begin{bmatrix} c^{(k+1)} \\ g^{(k+1)} \end{bmatrix} = \begin{bmatrix} c^{(k)} \\ g^{(k)} \end{bmatrix} - t^{(k)} \left[(c^{(k)} + g^{(k)}t_i - z_i) \begin{bmatrix} 1 \\ t_i \end{bmatrix} + \lambda \begin{bmatrix} c^{(k)} \\ g^{(k)} \end{bmatrix} \right].$$

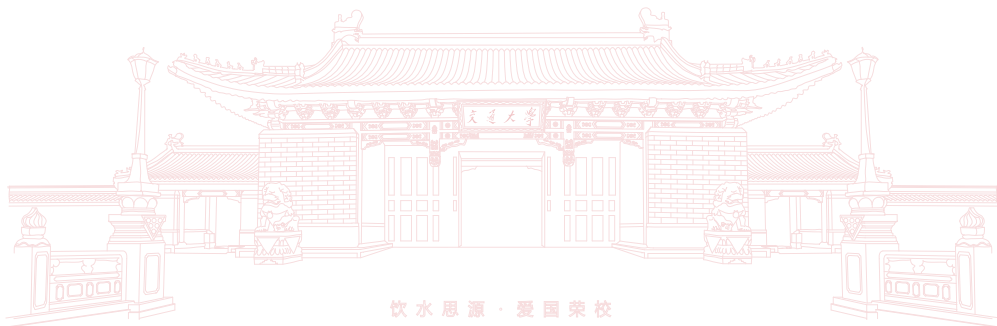
新的观测只触发 $O(1)$ 次向量运算, 不需要重新求解全部正规方程。

若数据分布随时间变化, 常步长能持续跟踪; 此时噪声地板也成为响应速度与稳态误差的权衡。

自编应用; 延续 Lecture 03 传感器标定模型

06

延伸与方法选择



Proximal gradient 与 FISTA

若目标含非光滑正则项

minimize _{x} $f(x) + g(x)$, f 光滑, g 为闭的 proper 凸函数, 且近端映射可计算,

proximal gradient 为

$$x^{(k+1)} = \text{prox}_{tg} \left(x^{(k)} - t \nabla f(x^{(k)}) \right).$$

例: Lasso

$g(x) = \lambda \|x\|_1$ 的 proximal operator 是逐分量 soft-thresholding。

FISTA

把 Nesterov 型外推与 $t \leq 1/M$ 的 proximal step (或回溯线搜索) 结合, 在标准凸设定下得到 $O(1/k^2)$ 函数值界。

这是从“光滑一阶方法”进入复合优化的自然下一步。

Beck & Teboulle, “A Fast Iterative Shrinkage-Thresholding Algorithm,” SIAM J. Imaging Sci. 2009

共轭梯度、Chebyshev 与 BB 步长

Conjugate gradient	Chebyshev	Barzilai–Borwein
针对 SPD 线性系统，在 Krylov 子空间中自适应构造误差多项式；不需要预先给出完整谱。	根据已知谱区间预先设计 minimax 多项式；无需内积正交化，但依赖可靠的 $[m, M]$ 。	用最近的位移与梯度差估计标量曲率；实践中常与非单调线搜索结合。

三者都在利用曲率信息，但获取方式分别是 Krylov 正交性、谱先验和 secant 近似。

Classical CG, Chebyshev semi-iteration, and Barzilai–Borwein methods

AdaGrad、Adam 与 Edge of Stability

坐标自适应

AdaGrad 以历史平方梯度构造对角度量；Adam 再加入一阶与二阶指数滑动平均。它们改变了每个坐标的有效步长。

$$G_{ii}^{(k)} = \sum_{j=0}^k (g_i^{(j)})^2,$$
$$x_i^{(k+1)} = x_i^{(k)} - \frac{t}{\sqrt{G_{ii}^{(k)}} + \varepsilon} g_i^{(k)}.$$

这些主题值得单独讨论；把它们附加在 SGD 参数表后面会掩盖各自不同的理论问题。

非凸训练现象

Edge of stability 研究神经网络训练时步长与局部最大 Hessian 特征值共同演化的现象。其经验解释不能替代凸问题中的全局收敛证明。

Duchi et al. (2011); Kingma & Ba (2015); Cohen et al. (ICLR 2021)

方法地图

信息与结构	代表方法	主要先验	主要风险
完整梯度, 光滑凸	GD / NAG	M , 强凸时还需 m	条件数、参数失配
SPD 二次型	HB / CG / Chebyshev	谱区间或线性系统结构	暂态放大、有限精度
光滑凸, 有限时域	PEP / Silver schedule	函数类和性能指标	证书适用范围
有限和或数据流	SGD / mini-batch / SVRG	无偏性、方差、数据规模	噪声地板、成本记账
复合非光滑	Proximal / FISTA	proximal operator 可计算	单步子问题成本

算法选择先从信息模型和问题结构出发, 再讨论步长、动量和实现细节。

本讲结论

1. 二次型频谱把条件数、固定步长、振荡和加速放在同一动力学模型中。
2. Heavy-ball 与 Nesterov 都使用两步记忆，但梯度评估位置和保证范围不同。
3. PEP 把算法的最坏情形写成优化问题，可用于核查证明常数和设计有限时域系数。
4. 长步长序列依赖多步补偿；单步超过 $2/M$ 与整段最终收敛并不矛盾。
5. 随机优化改变了信息模型 (oracle model)。方差、梯度成本和抽样重复必须进入数值比较。

贯穿观点

迭代公式只描述“怎样走”；函数类、信息模型和评价指标共同决定“为什么有效”以及“比较是否公平”。

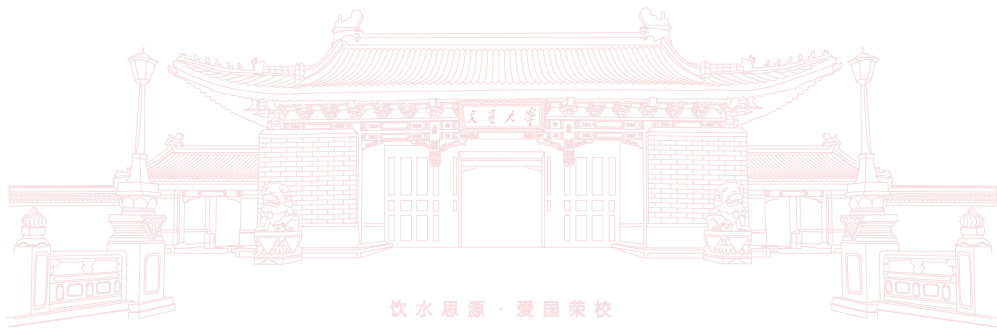
主要文献

Polyak (1964)	<i>Some Methods of Speeding Up the Convergence of Iteration Methods.</i>
Nesterov (1983)	<i>A Method of Solving a Convex Programming Problem with Convergence Rate $O(1/k^2)$.</i>
Drori–Teboulle (2014)	<i>Performance of First-Order Methods for Smooth Convex Minimization.</i>
Taylor–Hendrickx–Glineur (2017)	<i>Smooth Strongly Convex Interpolation and Exact Worst-Case Performance of First-Order Methods.</i>
Goujaud et al. (2022)	<i>PEPit: Computer-Assisted Worst-Case Analyses of First-Order Optimization Methods.</i>
Altschuler–Parrilo (2023–2024)	<i>Acceleration by Stepsize Hedging I–II.</i>
Robbins–Monro (1951)	<i>A Stochastic Approximation Method.</i>
Polyak–Juditsky (1992)	<i>Acceleration of Stochastic Approximation by Averaging.</i>
Johnson–Zhang (2013)	<i>Accelerating Stochastic Gradient Descent using Predictive Variance Reduction.</i>

正文各页的右下角给出对应出处；数值实验的完整参数与脚本见附录。

01

附录



饮水思源 · 爱国荣校

附录：HB 参数的端点平衡

对模态 λ ，HB 特征多项式为

$$r^2 - (1 + \beta_{\text{HB}} - t\lambda)r + \beta_{\text{HB}}.$$

选择参数使 $\lambda = m$ 与 $\lambda = M$ 的判别式同时达到临界：

$$1 + \beta_{\text{HB}} - tm = 2\sqrt{\beta_{\text{HB}}},$$

$$1 + \beta_{\text{HB}} - tM = -2\sqrt{\beta_{\text{HB}}}.$$

两式相减并求解得到

$$\sqrt{\beta_{\text{HB}}} = \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}}, \quad t = \frac{4}{(\sqrt{M} + \sqrt{m})^2}.$$

此时所有端点模态的根模均为 $\sqrt{\beta_{\text{HB}}}$ ；区间内部模态也受同一圆盘控制。

附录：强凸光滑插值条件

对 $0 \leq m < M < \infty$ ，有限数据 $(\mathbf{x}^{(i)}, f_i, \mathbf{g}^{(i)})$ 可由某个 $f \in \mathcal{F}_{m,M}$ 插值，当且仅当对任意 i, j 有

$$f_i - f_j - \langle \mathbf{g}^{(j)}, \mathbf{x}^{(i)} - \mathbf{x}^{(j)} \rangle \geq \frac{1}{2(1 - m/M)} \left(\frac{1}{M} \|\mathbf{g}^{(i)} - \mathbf{g}^{(j)}\|_2^2 + m \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|_2^2 - \frac{2m}{M} \langle \mathbf{g}^{(i)} - \mathbf{g}^{(j)}, \mathbf{x}^{(i)} - \mathbf{x}^{(j)} \rangle \right).$$

令 $m = 0$ 即退化为正文使用的光滑凸插值条件。该等价条件使 $\mathcal{F}_{m,M}$ 上固定步长一阶方法的 PEP 可写成 SDP。

附录：数值实验参数

实验	参数	方法参数
确定性二次型	$n = 60; \lambda_i \in [1, 1000]$ 对数均匀; $x_i^{(0)} \propto \lambda_i^{-1/2}$	GD $1/M$; GD $2/(M+m)$; HB 经典二次参数; NAG 强凸参数; Silver $\alpha^{(j)}/M$
二维轨迹	$Q = \text{diag}(1, 100)$; $\mathbf{x}^{(0)} = (2.5, 0.21)$; 横纵等比例	展示最优固定步长 GD 的前 18 步
有限和 ridge	$N_{\text{data}} = 2000, n = 40, \lambda = .01$; 数据种子 6015; $m = .019599, M = .989440$	SGD 常步 $.015/M$; 递减步长 $(.015/M)/(1 + .1p)$; mini-batch $B = 20, (.10/M)/(1 + .05p)$; momentum $\beta_{\text{mom}} = .9, (.010/M)/(1 + .10p)$; SVRG 步长 $.02$ 、内循环 N_{data} 步
随机重复	12 个抽样种子	各 SGD 变体复用同一组种子流作配对比较; 报告中位数及 10%-90% 分位带
成本	一个样本梯度为一单位	N_{data} 单位为一次 pass; SVRG outer epoch 按 $3N_{\text{data}}$ 计

所有曲线从 CSV 读取并由 PGFPlots 绘制; 纵轴下限只用于显示, 不修改原始数据。

本讲可复现实验设置

附录：数据生成与重建

讲义目录中的文件为

`first-order-data/generate_first_order_data.py`.

重新生成数据:

```
python3.14 first-order-data/generate_first_order_data.py
```

重新编译讲义:

```
latexmk -xelatex -interaction=nonstopmode \  
-halt-on-error -file-line-error \  
math6015-supplement-first-order-methods.tex
```

脚本同时生成 `threshold_summary.csv` 与 `chebyshev_diagnostics.csv`, 分别对应正文阈值表和 Chebyshev 附录; 所有曲线、分位带、公式和标注在 Beamer 中保持可编辑。

附录：Chebyshev 的中间放大

对实验 A 的谱区间 $[1, 1000]$ ，长度 $N = 16$ 的 Chebyshev 步长满足

$$\max_j t^{(j)} M \approx 293.7,$$

远大于单步稳定阈值 2。按根的自然顺序直接执行时，最终目标间隙约为初值的 2.01×10^{-1} ，但中间最大值可超过 10^{13} 倍。

解析结论

最终多项式只由根集合决定，与步长排列无关。

数值结论

中间前缀多项式、溢出和舍入误差高度依赖排列与实现。

因此正文把 Chebyshev 用作谱多项式原理，并用具有一般光滑凸证书的 Silver schedule 做主要长步数值比较。

自编精确谱实验；Chebyshev $N = 16$, $[m, M] = [1, 1000]$

附录：SGD 无偏性与 SVRG 无偏性

若 i 在 $\{1, \dots, N_{\text{data}}\}$ 上均匀抽样，则

$$\mathbb{E}_i[\nabla f_i(\mathbf{x})] = \frac{1}{N_{\text{data}}} \sum_{i=1}^{N_{\text{data}}} \nabla f_i(\mathbf{x}) = \nabla F(\mathbf{x}).$$

对 SVRG 估计量

$$\hat{\mathbf{g}}_{\text{vr}} = \nabla f_i(\mathbf{x}) - \nabla f_i(\tilde{\mathbf{x}}) + \nabla F(\tilde{\mathbf{x}}),$$

条件期望为

$$\mathbb{E}_i[\hat{\mathbf{g}}_{\text{vr}}] = \nabla F(\mathbf{x}) - \nabla F(\tilde{\mathbf{x}}) + \nabla F(\tilde{\mathbf{x}}) = \nabla F(\mathbf{x}).$$

方差缩减并不来自偏差，而来自相关差值 $\nabla f_i(\mathbf{x}) - \nabla f_i(\tilde{\mathbf{x}})$ 在两点接近时变小。



上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

自然科学研究院
INSTITUTE OF NATURAL SCIENCES

Fin

