实用最优化算法　课堂论文汇报

谱梯度法的离散动力系统分析

收敛性、最坏收敛因子与 Lyapunov 函数

杨书泰

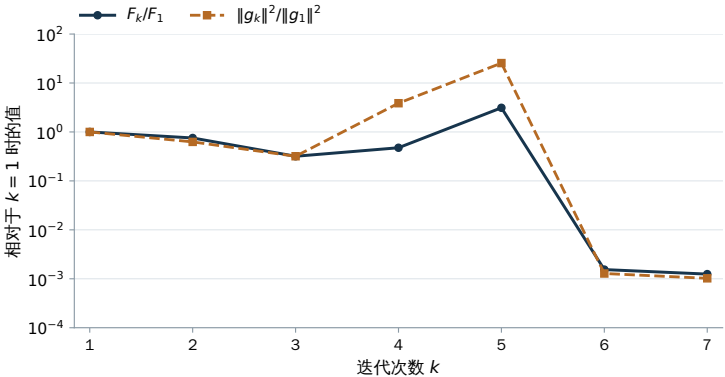
2026 年 10 月

精读论文：J. Barzilai & J. M. Borwein (1988)

Two-point step size gradient methods

IMA Journal of Numerical Analysis, 8(1):141–148.
DOI: 10.1093/imanum/8.1.141

一条非单调的 BB 迭代轨道



$H = \mathrm{diag}(1, 10)$，BB1。

函数值误差和梯度范数都会反弹。

步长怎样依赖历史梯度？有没有一直下降的量？

算例取自 Yang–Yuan, arXiv:2609.12084v1 的 Table 1：$g_0 = (1, 2)^{\mathsf{T}}$，首步 $\eta_0 = 5/41$。原表第 0 行（首步之后）在这里记为 $k = 1$；图中只画出前 7 行。

用连续模型理解加速

背景：用微分方程研究加速方法

Nesterov 加速梯度法（NAG）的小步长极限（$t = k\sqrt{s}$）：

$$\ddot{X} + \frac{3}{t}\dot{X} + \nabla f(X) = 0, \qquad E = t^2(f(X) - f_*) + 2\left\| X + \frac{t}{2}\dot{X} - x_* \right\|^2.$$

由 $E' \le 0$ 即得 $O(t^{-2})$ 的收敛速度。但离散迭代逼近方程的解，是在 $s \to 0$、时间区间有限的条件下证明的。

高分辨率模型保留 $\sqrt{s}$ 阶的修正项，如强凸情形下 NAG 的高分辨率方程中的

$$\sqrt{s}\,\nabla^2 f(X)\dot{X}.$$

借助这一项，就能区分低分辨率模型分不清的算法，如 NAG 与重球法。

Su–Boyd–Candès, JMLR 17(153), 2016, Thms. 2–3；Shi–Du–Jordan–Su, Mathematical Programming 195, 2022。[17, 19]

从连续模型回到离散算法

模型	保留的信息
低分辨率 ODE	小步长极限下的主导行为
高分辨率 ODE	关于步长的高阶修正项及其耗散作用
离散迭代本身	步长不必趋于零，可以直接研究长期行为

要由连续能量得到离散算法的结论，还须估计差分余项和离散修正项，并限制步长的范围。

对谱梯度法，我们直接研究离散迭代：步长由历史梯度决定，又反过来改变梯度。

Wibisono–Wilson–Jordan, PNAS 2016（参考的是 arXiv v1）；Wilson–Recht–Jordan, JMLR 2021；Shi 等，NeurIPS 2019；Ascher 等，ESAIM:M2AN 2009。[2, 16, 20, 21]

BB 步长：两点割线条件

精读 BB1988　1/4

令 $s_{k-1} = x_k - x_{k-1}$，$y_{k-1} = g_k - g_{k-1}$。

$$\min_{\tau} \left\| \tau s_{k-1} - y_{k-1} \right\|^2 \quad \Longrightarrow \quad \eta_k^{\mathrm{BB1}} = \frac{s_{k-1}^{\mathsf{T}} s_{k-1}}{s_{k-1}^{\mathsf{T}} y_{k-1}},$$

$$\min_{\eta} \left\| s_{k-1} - \eta y_{k-1} \right\|^2 \quad \Longrightarrow \quad \eta_k^{\mathrm{BB2}} = \frac{s_{k-1}^{\mathsf{T}} y_{k-1}}{y_{k-1}^{\mathsf{T}} y_{k-1}}.$$

不做精确线搜索：只用相邻两点的信息，以数量矩阵近似 Hesse 矩阵或其逆。

原文式 (5) 是今天所说的 BB2，式 (6) 是 BB1。本次汇报中步长都记为 η，步长的倒数记为 τ。[4]

二次情形：步长滞后一步

精读 BB1988　2/4

设 $H = H^{\mathsf{T}} \succ 0$，$g_k = H(x_k - x_*)$。由 $s_{k-1} = -\eta_{k-1} g_{k-1}$ 得

$$\eta_k^{\mathrm{BB1}} = \frac{\|g_{k-1}\|^2}{g_{k-1}^{\mathsf{T}} H g_{k-1}}, \qquad \eta_k^{\mathrm{BB2}} = \frac{g_{k-1}^{\mathsf{T}} H g_{k-1}}{g_{k-1}^{\mathsf{T}} H^2 g_{k-1}}.$$

$$g_{k-1} \xrightarrow{\text{决定}} \eta_k \xrightarrow{\text{作用于当前梯度}} g_{k+1} = (I - \eta_k H) g_k$$

只知道 g_k 还不能确定下一步，状态应取相邻两个梯度 (g_{k-1}, g_k)。

从 g_0 出发，先用任意正步长 η_0 得到 g_1；BB 步长从 $k = 1$ 开始使用。[4]

二维情形：分量之比的递推

精读 BB1988　3/4

$H = \mathrm{diag}(a, b)$，$0 < a < b$。在两个分量都不为零时，令 $\zeta_k = g_{k,a}^2 / g_{k,b}^2$。

$$\text{BB1：}\ \zeta_{k+1} = \frac{\zeta_k}{\zeta_{k-1}^2} \qquad \text{BB2：}\ \zeta_{k+1} = \left(\frac{b}{a}\right)^2 \frac{\zeta_k}{\zeta_{k-1}^2}.$$

BB1 取 $z_k = \log \zeta_k$；BB2 取 $z_k = \log \zeta_k - \log(b/a)$。

$$\boxed{z_{k+1} = z_k - 2z_{k-1}}, \qquad \lambda_\pm = \frac{1 \pm i\sqrt{7}}{2}, \quad |\lambda_\pm| = \sqrt{2}.$$

z_k 不恒为零时振幅越来越大，梯度范数却可以很快趋于零。

BB1988 分析的是今天所说的 BB2，并假设第一步取最速下降步长；Dai (2013) 直接分析了 BB1，对首步的要求也更弱。$\sqrt{2}$ 描述的是 z_k 的增长，不是梯度的收缩因子。[4, 6]

二维情形：超线性收敛与平衡轨道

精读 BB1988　4/4

两个分量都不为零时，除平衡轨道外，迭代可以是 $\sqrt{2}$ 阶 R-超线性收敛的。

但对 BB1，若 $\|P_a g_0\| = \|P_b g_0\|$，且首步取 $\eta_0 = 2/(a+b)$，则

$$\eta_k = \frac{2}{a+b}, \qquad \|g_k\| = \left(\frac{b-a}{b+a}\right)^k \|g_0\| .$$

BB2 的相应条件是 $a\|P_a g_0\|^2 = b\|P_b g_0\|^2$。

由此引出一般维数下的问题：

是否对任意初值都收敛？能否一致地估计收敛速度？
最坏的渐近收敛因子是否就是这个常数？

BB2 的超线性收敛见 BB1988（第一步取最速下降步长），BB1 的二维结论见 Dai (2013)；平衡轨道见 BB1988，Li–Sun (2021) 的 Proposition 2 也给出了这条轨道。R-超线性收敛并不意味着 Q-超线性收敛。[4, 6, 11]

相关研究的三条线索

步长只依赖当前梯度的方法

Akaike 1959：把最速下降法化为概率分布的迭代

Forsythe 1968：渐近方向　Pronzato 等 2006：*P*-梯度方法

BB 方法：步长滞后一步

BB1988 → Raydan 1993 全局收敛 → Dai–Liao 2002 R-线性收敛

→ Dai 2013 二维情形 → Li–Sun 2021 显式收敛因子 → 最坏收敛因子

有限记忆最速下降法（LMSD）：利用上一组的梯度

Fletcher 2012 → Curtis–Guo 2018 附加条件下的 R-线性收敛

Ferrandi–Hochstenbach 2025：多割线条件、子空间与实现

我们：研究一组步长的乘子多项式，补全收敛证明

箭头表示问题上的承接，各文的算法和假设不尽相同；Pronzato 等研究的是使用当前梯度的方法。综述见 Zou–Magoulès (2022)。[1, 5, 7–10, 13, 14, 26]

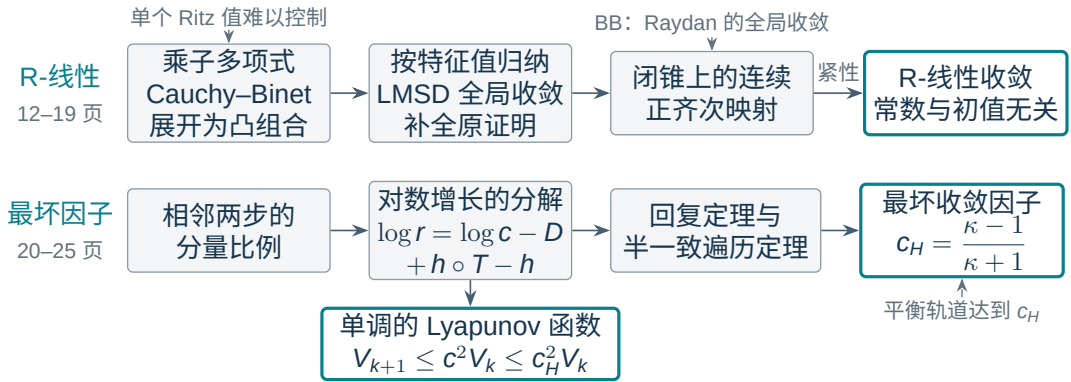
已有结果与我们的结果

严格凸二次函数，$c_H = (\kappa-1)/(\kappa+1)$　横线以下为我们的结果

文献	方法	结论
BB1988（精读）	BB，二维	BB2：R-超线性；平衡轨道的因子为 c_H
Raydan 1993	BB	任意维数全局收敛
Dai–Liao 2002	BB	R-线性
Li–Sun 2021 等	BB 及一类非单调方法	R-线性，因子 $1-1/\kappa$
Fletcher 2012	LMSD	全局收敛，证明中有一步缺少依据
Curtis–Guo 2018	LMSD	R-线性，需归一化的历史矩阵一致非退化
袁亚湘 2026	二、三维子空间步长	全局收敛；R-线性
我们	LMSD（含加权）	R-线性无需非退化假设；补全 Fletcher 证明
我们	BB1，BB2	最坏收敛因子恰为 c_H；单调 Lyapunov 函数

BB1988 的超线性分析假设第一步取最速下降步长；“等”指 Li–Huang (2025)，其结果适用于满足 Property B 的一类方法；袁亚湘 (2026) 的二维步长在相邻两次位移接近共线时改用 BB 步长。[4, 5, 7, 9, 11, 12, 14, 22–25, 27]

两项结果的证明路线



共同思路：先选好状态，再用齐次性、紧性或遍历理论得到与初值无关的估计。

$c = (b-a)/(b+a)$，其中 $a < b$ 是 g_0 的非零分量所对应的最小、最大特征值。上行：Yang–Yuan, arXiv:2609.12100v1 与 2609.12114v1；下行：arXiv:2608.07839v1 与 2609.12084v1。[22–25]

BB 迭代的状态锥

我们的工作　状态取相邻两个梯度，不做归一化

令 $m = \lambda_{\min}(H)$，$L = \lambda_{\max}(H)$。取

$$\Omega_{\mathrm{BB}} = \{(u, v) : v = (I - \eta H)u,\ \eta \in [1/L, 1/m]\},$$

$$T(u, v) = \big(v, (I - \eta^{\mathrm{BB1}}(u)H)v\big), \qquad T(0, 0) = (0, 0).$$

- $u = 0$ 时必有 $v = 0$：算法不会产生的状态 $(0, v \neq 0)$ 不在锥中。
- Rayleigh 商在原点没有极限，但 T 在原点仍然连续。
- 锥中每个状态都出现在 BB1 的某条迭代轨道上。

Yang–Yuan, arXiv:2609.12100v1。这个锥是闭的，但不一定是凸的。首步 $\eta_0 > 0$ 任取时，(g_0, g_1) 未必在锥中，但 (g_1, g_2) 一定在。[22]

齐次性：由每条轨道收敛得到一致的几何收敛

已有的稳定性定理　我们的工作：选取状态、处理边界

有限维闭锥 Ω 上，T 连续且 $T(tz) = tT(z)$（$t \geq 0$）。若

$$\text{每个} z \in \Omega \text{ 都有} T^k z \longrightarrow 0,$$

则存在 $C < \infty$、$\rho < 1$ 使

$$\boxed{\left\| T^k z \right\| \leq C\rho^k \left\| z \right\| \qquad \forall z \in \Omega.}$$

证明思路：锥中每个单位向量在某一步后范数减半 $\rightarrow$ 由连续性和有限覆盖，所需步数有一致上界 $\rightarrow$ 由齐次性，可以反复减半；中间的反弹另有一致上界。

对 BB，每条轨道收敛来自 Raydan 定理和上页第三条；对 LMSD 需另行证明。

这是 Shen–Hu (2012) 定理 5.1 在单个映射、一次齐次时的特例；用于 BB 见 Yang–Yuan (2026)。对一般的非线性系统，渐近稳定推不出指数稳定。[14, 15, 22]

LMSD：用上一组的梯度确定下一组的步长

每组至多 p 步（不终止时恰好 p 步）。第 j 组的历史矩阵由该组的前 ℓ_j 个梯度组成，组末的梯度就是下一组的组首梯度：

$$G_j = [g_{j,0}, \dots, g_{j,\ell_j-1}], \qquad g_{j,\ell_j} = g_{j+1,0}.$$

第 j 组的组首梯度 u $\mathcal{K}_p(H,u)$ → 在 G_j 的列空间上求出全部 Ritz 值 → 第 $j{+}1$ 组从 v 出发依次以 Ritz 值的倒数为步长

$$v^+ = \Psi_u(H)v, \qquad \Psi_u(t) = \prod_{r=1}^{d}\left(1 - \frac{t}{\theta_r(u)}\right), \qquad d = \operatorname{rank} G_j.$$

一组走完后的梯度与 Ritz 值的使用顺序无关；组内各步的梯度则与顺序有关。

Fletcher (2012)。这里考虑的 LMSD：精确算术，记忆长度 p 固定，每组把每个 Ritz 值的倒数各用一次；历史矩阵降秩时，在它的列空间上求出全部 Ritz 值。[9, 23]

历史梯度趋于线性相关时的 Ritz 值

Fletcher 附录中有一步极限推理，需要控制取极限之前的全部 Ritz 值。例如：

$$H = \mathrm{diag}(1, 2, 3), \quad p = 2, \quad u_\varepsilon \propto (\varepsilon, \varepsilon^2, 1)^{\mathsf{T}} \longrightarrow e_3.$$

$$\frac{(H - 3I)u_\varepsilon}{\|(H - 3I)u_\varepsilon\|} = \frac{(-2, -\varepsilon, 0)^{\mathsf{T}}}{\sqrt{4 + \varepsilon^2}} \longrightarrow -e_1.$$

因而 $\mathcal{K}_2(H, u_\varepsilon) \to \mathrm{span}\{e_3, e_1\}$，Ritz 值 $\to (3, 1)$。

u_ε 趋于最大特征值的方向，Ritz 值却不都趋于大的特征值。

Curtis–Guo 为证明 R-线性收敛，假设归一化的历史矩阵一致非退化；我们不作这个假设，而是直接控制一组步长的乘子多项式。

Fletcher (2012) 附录式 (36)；Curtis–Guo (2018) Assumption 3.4；Yang–Yuan, arXiv:2609.12114v1。这里只是一族向量取极限，不是 LMSD 的迭代轨道，更不是发散的反例。[5, 9, 23]

乘子多项式的凸组合表示

设迭代不终止，这时 $d = p$。用 Cauchy–Binet 公式展开，得

$$\Psi_u(\lambda_r) = \sum_{|I|=p} \pi_I(u)\phi_I^{(r)}, \quad \pi_I = \frac{\beta_I}{\sum_J \beta_J}, \quad \beta_I = \Delta_I^2 \prod_{i\in I} \lambda_i u_i^2,$$

其中 $\phi_I^{(r)} = \prod_{i\in I}(1-\lambda_r/\lambda_i)$ 只依赖 H 和 p，权 $\pi_I \geq 0$，$\sum_I \pi_I = 1$。

- $r \in I$ 时 $\phi_I^{(r)} = 0$；I 中下标都大于 r 时 $|\phi_I^{(r)}| < 1$。
- 其余的 I 含某个 $i < r$。若 $\beta_I > 0$，把 i 换成 r，记 $J = (I \setminus \{i\}) \cup \{r\}$，则

$$\frac{\beta_I}{\beta_J} = C_I \left(\frac{u_i}{u_r}\right)^2 \longrightarrow 0 \quad \text{当} u_i \to 0 \text{ 且} |u_r| \geq \varepsilon \text{ 时.}$$

这些项的权趋于零；再用数列引理处理“滞后一组”，按特征值从小到大归纳。

Yang–Yuan, arXiv:2609.12114v1。Δ_I 为 Vandermonde 行列式。$\phi_I^{(r)}$ 可负，绝对值也可大于 1，凸组合本身不保证收缩。[23]

降秩时映射仍然连续

允许的 p 步多项式为 $q(t)=\prod_{r=1}^{p}(1-\eta_r t)$，$\eta_r\in[1/L,1/m]$。

$$\Omega_p=\{(u,v):v=q(H)u,\ q\text{ 为允许的}p\text{ 步多项式}\}.$$

记 u 的非零分量对应的互异特征值有 $s(u)$ 个，则 $\operatorname{rank}\mathcal{K}_p(H,u)=\min(p,s(u))$。

$$T_p(u,v)=\begin{cases}(v,\Psi_u(H)v), & s(u)>p,\\ (v,0), & s(u)\le p.\end{cases}$$

- $s(u)\le p$ 时，下一组迭代会在组内终止；$s(u)=p$ 时历史矩阵仍然满秩。
- 分 $s>p$、$s=p$、$1\le s<p$ 和 $u=0$ 四种情形，可以证明 T_p 连续。
- 锥中每个状态都出现在 LMSD 的某条迭代轨道上。

Yang–Yuan, arXiv:2609.12114v1。这里证明的是 T_p 本身的连续性，不需要单独控制每个 Ritz 值。[23]

BB 与 LMSD 的 R-线性收敛：统一的证明

	BB	LMSD
状态	相邻两个梯度	相邻两组的组首梯度
每条轨道收敛	Raydan 的定理	按特征值归纳（第 16 页）
共同结构	闭锥上的连续正齐次映射；锥中每个状态都在某条迭代轨道上	
结论	由紧性得到常数与初值无关的 R-线性收敛	

LMSD 先得到每组的收缩因子 $\rho_{H,p} < 1$，

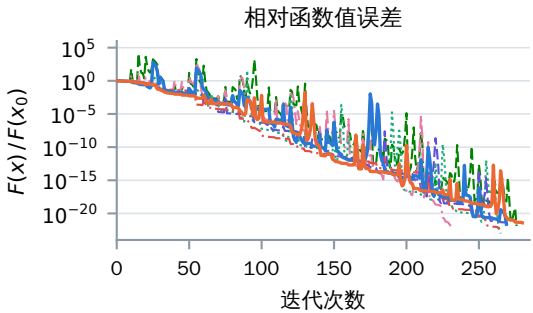
再用组内放大倍数的上界，换算成每步的因子 $\rho_{H,p}^{1/p}$。

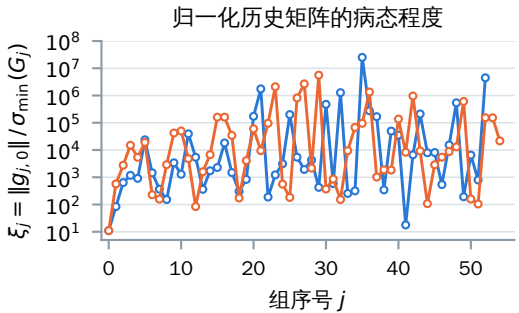
权取为 $W = \omega(H)$ 时，令 $\widetilde{g} = W^{1/2} g$，就化为同一个 H 对应的不加权方法。

这里只证明了收敛因子存在；$p > 1$ 时最优的显式因子还没有确定。

Yang–Yuan, arXiv:2609.12100v1 与 2609.12114v1。上界从初始组之后的第一组算起；初始组的步长可以任意大，所以若以 g_0 为基准，常数必然依赖这些步长。[22, 23]

数值例子：满秩的历史矩阵也可能非常病态





W	H^{-1}	$H^{-1/2}$	I	$H^{1/2}$	H	$H^{3/2}$	H^{2}
$\max_j \xi_j$	1.30×10^{8}	2.35×10^{8}	2.51×10^{7}	1.59×10^{7}	5.59×10^{6}	1.99×10^{7}	1.48×10^{7}
$\max_j A_j$（组内放大）	1.33×10^{3}	1.55×10^{5}	2.26×10^{5}	6.93×10^{4}	3.30×10^{4}	2.21×10^{4}	6.99×10^{1}

Yang–Yuan, arXiv:2609.12114v1，据原图重绘；$n = 60$，$p = 5$，第 0 组用精确线搜索；七种权在 231–280 步内停止。[23]

BB 方法的最坏渐近收敛因子

我们的工作　严格凸二次函数，全程使用 BB1 或全程使用 BB2

$$\rho_g(g_0,\eta_0) := \limsup_k \left(\frac{\|g_k\|}{\|g_0\|}\right)^{1/k} \le \frac{b-a}{b+a} \le \boxed{c_H = \frac{\kappa-1}{\kappa+1}}.$$

a、b 为 g_0 的非零分量所对应的最小、最大特征值；首步 $\eta_0 > 0$ 可以任取。

Li–Sun 2021	BB1 的显式上界 $1-1/\kappa$（首步也按 BB 公式计算）
已知的平衡轨道	因子恰为 c_H，BB1988 中已经给出
我们的结果	所有轨道的因子都不超过 c_H，与下界吻合
函数值误差	最坏渐近收敛因子为 c_H^2

Yang–Yuan, arXiv:2608.07839v1。首步也按 BB 公式由 g_0 计算时，对每个 $\gamma > c_H$，有与初值无关的 C_γ 使 $\|g_k\| \le C_\gamma \gamma^k \|g_0\|$；$\gamma = c_H$ 时是否也有这样的常数，目前还没有结论。[11, 25]

分量比例与范数分开考虑

以下以 BB1 为例，并设迭代不终止；BB2 可通过变量替换化为 BB1。

$$w_i^k = \frac{\|P_{\lambda_i} g_k\|^2}{\|g_k\|^2}, \qquad \chi_k = (w^k, w^{k-1}) \in \mathcal{P} \times \mathcal{P}.$$

令 $\tau(z) = \sum_i \lambda_i z_i$，$\phi_i(z) = (1 - \lambda_i/\tau(z))^2$，则

$$r(w,z)^2 = \sum_i \phi_i(z) w_i, \qquad T(w,z) = \left(\frac{\phi(z) \odot w}{r(w,z)^2}, w \right),$$

$$\log \frac{\|g_n\|}{\|g_1\|} = \sum_{k=1}^{n-1} \log r(\chi_k).$$

最坏收敛因子由 $\log r$ 沿轨道的长期平均决定。

Yang–Yuan, arXiv:2608.07839v1。$r = 0$ 表示迭代终止；补上孤立的终止点后，T 是 Borel 可测映射，但不处处连续。[25]

端点不等式与对数增长的分解

设沿轨道始终不为零的那些分量所对应的特征值中，最小、最大的为 $a < b$，$c = (b-a)/(b+a)$。由加权算术–几何平均不等式，对 $\tau \in [a,b]$ 有

$$\left(\frac{\tau - a}{\tau}\right)^{b/(a+b)} \left(\frac{b-\tau}{\tau}\right)^{a/(a+b)} \le c.$$

其余分量变为零以后，得到恒等式

$$\boxed{\log r = \log c - D + h \circ T - h}, \qquad D \ge 0,$$

$$h(w,z) = -\frac{b}{2(a+b)} \log w_a - \frac{a}{2(a+b)} \log w_b.$$

一步中的增长都体现在差分项 $h \circ T - h$ 里；但端点分量的比例趋于零时 h 无界，所以不能直接对差分项积分，也不能想当然地认为 $h(T^n\chi)/n \to 0$。

Yang–Yuan, arXiv:2608.07839v1。D 是上面不等式两边取对数后之差，$D = 0$ 当且仅当 $\tau = (a+b)/2$。[25]

遍历理论的两步论证

第一步　对不变测度估计平均

对遍历测度，先找出几乎处处不为零的分量。Birkhoff 遍历定理给出时间平均的极限；Poincaré 回复定理给出一列时刻 n_ℓ，使 $h(T^{n_\ell}\chi)$ 有界，于是差分项除以 n_ℓ 后趋于零。

$$\int \log r \, d\mu \le \log c.$$

第二步　从测度回到每一条轨道

若在很长的轨道段上 $\log r$ 的平均值偏高，取其经验测度的弱极限，截断后可知

$$\text{积分有限且大于}\log c \implies \mu(\text{奇异集}) = 0 \implies \text{极限测度不变}.$$

这与第一步矛盾。由此得到半一致估计，再开 n 次方，就得到对所有初值成立的上界。

Yang–Yuan, arXiv:2608.07839v1。半一致遍历定理见 Sturman–Stark (2000)，这里推广到 Borel 映射和取值 $-\infty$ 的函数。回复定理只用在测度意义下几乎所有的点上，不假设每条 BB 轨道都回复。[18, 25]

由分解式得到单调的 Lyapunov 函数

设 $a < b$ 为 g_0 的非零分量所对应的最小、最大特征值。若 g_1 在 a、b 上的分量都不为零（于是此后也一直不为零），令

$$\boxed{V(g) = \|P_a g\|^{\frac{2b}{a+b}} \, \|P_b g\|^{\frac{2a}{a+b}}}, \qquad \sqrt{V(g)} = \|g\| \; e^{-h}.$$

差分项恰好消去：

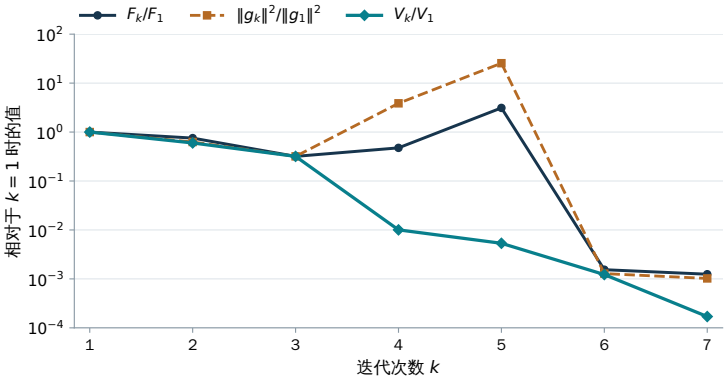
$$\boxed{V(g_{k+1}) = c^2 e^{-2D(\tau_k)} V(g_k) \le c^2 V(g_k)}.$$

这也可以直接由两个端点分量的更新式和第 22 页的不等式验证。

在形如 $V_\beta = \|P_a g\|^{2\beta} \, \|P_b g\|^{2(1-\beta)}$ 的函数中，只有取 $\beta = b/(a+b)$ 时，一步因子的最大值才最小。

Yang–Yuan, arXiv:2609.12084v1。这里的最优性只限于上述函数族，并不是说 V 在所有 Lyapunov 函数中最优。[24]

同一条轨道上的单调量



F_k 和 $\|g_k\|^2$ 都有反弹，V_k 一直下降。

但 V 不是正定的：某个端点分量为零时，$g \neq 0$ 也可能有 $V = 0$。

要得到 $\|g_k\|$ 的收敛，还须处理端点分量趋于零的情形。

算例、坐标尺度和编号与第 2 页相同。用 V 对特征值集合归纳，可以另外证明 R-线性收敛；最坏因子的精确值仍要靠前面的遍历论证。[24, 25]

推广：Hilbert 空间与非线性问题

Hilbert 空间

算子有界、自伴、一致正定。状态取相邻两步的谱测度，状态空间弱紧；端点不一定是特征值，改用端点附近一小段谱所占的比例。结论：给出每条轨道渐近收敛因子的上界；首步步长按 BB 公式计算时，确定一致上界中最优的因子。

非线性问题（局部）

梯度在 x_* 处严格 Fréchet 可微，导数一致正定；全程使用同一种 BB 步长。对任意 $\gamma \in (c_*, 1)$，两个初始点都足够接近 x_* 时，有一致的上界 $C_\gamma \gamma^k$。证明分三步：给出二次模型的一致上界、在有限步内跟踪轨道、分段重新启动。

两项推广各有自己的条件，详见读书报告第 9 节。

Hilbert 空间中 BB 方法的早期工作见 Azmi–Kunisch (2020)；以上两项见 Yang–Yuan, arXiv:2608.07839v1。$c_* = (L_* - m_*)/(L_* + m_*)$，$m_*$、$L_*$ 为 $\nabla^2 f(x_*)$ 谱的端点。[3, 25]

进一步的问题

BB1988 的二维递推告诉我们：步长对历史的依赖值得专门分析。

- 状态锥与齐次性：由每条轨道收敛，得到一致的几何收敛。
- 一组步长的乘子多项式：绕开单个 Ritz 值在降秩时的极限问题。
- 长期平均与端点不等式：确定最坏因子，得到单调的 Lyapunov 函数。

可以继续研究的问题

我们的结论能否推广到按数值秩截断、权随迭代变化的情形？

$p > 1$ 时 LMSD 的最坏因子是多少？

能否找到不需要端点特征空间上的投影、可以实际计算的 Lyapunov 函数？

证明的细节、条件和文献见读书报告。

附录：闭锥上一致几何收敛的证明

锥在单位球面上的部分 $\Sigma = \Omega \cap \{\|z\| = 1\}$ 是紧集。对每个 $z \in \Sigma$，有 n_z 使 $\|T^{n_z} z\| < 1/2$；由连续性，这对 z 的一个邻域也成立；取有限覆盖，得共同上界 N_0：

$$\forall z \neq 0 \quad \exists n \in [1, N_0] : \ \left\|T^n z\right\| \leq q \|z\|, \qquad q < 1.$$

对每次收缩后的状态（归一化后）重复这一步；两次收缩之间的迭代，用前 N_0 步的上界 M 控制。若第 k 步之前已完成 r 次收缩，则 $r \geq k/N_0 - 1$，

$$\left\|T^k z\right\| \leq M q^r \|z\| \leq M q^{-1} (q^{1/N_0})^k \|z\|.$$

再取足够大的 N 使 $C\rho^N < 1$，就得到固定步数 N 上的一致收缩。各段的长度可以不同，段内也允许反弹。

这一定理见 Shen–Hu (2012)；用于 BB 与 LMSD 见 Yang–Yuan, arXiv:2609.12100v1 与 2609.12114v1。

附录：加权方法化为不加权的方法

设 $W=\omega(H)\succ 0$ 固定，$S=W^{1/2}$，则 $SH=HS$。

$$\eta_W(u)=\frac{u^{\mathsf T}Wu}{u^{\mathsf T}WHu}=\frac{\|Su\|^2}{(Su)^{\mathsf T}H(Su)}.$$

$$S(I-\eta_W(u)H)v=(I-\eta_I(Su)H)Sv.$$

对 LMSD，加权的 Ritz 值同样等于 SG_j 上不加权的 Ritz 值。因此加权与不加权的迭代通过 S 线性共轭，渐近收敛因子不变；范数上界中的常数至多再乘以 $\sqrt{\kappa_2(W)}$。

这里比较的是两个相互对应的初值；同一个 g_0 在不同权下的迭代过程可以很不一样。权随迭代变化，或在不同规则之间切换时，需要另外证明。

BB2 对应 $W=H$。见 Yang–Yuan, arXiv:2609.12100v1、2609.12114v1、2608.07839v1。

附录：LMSD 收敛证明中的数列引理

令 $a_j = \|P_r u_j\|$ 为第 j 组的组首梯度在第 r 个特征值上的分量。由前面的归纳，对任意 $\varepsilon > 0$，当 j 足够大时

$$a_j \geq \varepsilon \implies a_{j+2} \leq c_r a_{j+1}, \quad c_r < 1, \qquad a_{j+2} \leq M a_{j+1}.$$

若 a_j 始终不小于 δ，则它几何衰减，矛盾；所以总有某个 $a_N < \delta$。此后对每个 n：若 $a_{n-1} < \delta$，连用两次简单上界，得 $a_{n+1} < M^2\delta$；若 $a_{n-1} \geq \delta$，则 $a_{n+1} \leq c_r a_n$。归纳可知 $j \geq N$ 时 $a_j < M^2\delta$，再令 $\delta \to 0$。

条件看的是 a_j，收缩却发生在 $a_{j+1} \to a_{j+2}$：步长滞后一组，证明时必须考虑到这一点。

Yang–Yuan, arXiv:2609.12114v1。

附录：T_p 在四种情形下的连续性

$T_p(u,v)=(v,F_p(u,v))$，只需看第二个分量 F_p（$s(u)\le p$ 时 $F_p=0$）：

情形	连续的理由
$s(u)>p$	Gram 行列式（分母）为正，Ψ_u 连续地依赖于 u。
$s(u)=p$	Ψ_u 在 u 的非零分量所对应的特征值上等于零；其余分量上 v 趋于零，由乘子的一致上界控制。
$1\le s(u)<p$	不含 r 的 p 元下标集必含某个 i，u 在 i 上的分量趋于零；交换下标后，两项的权之比为 $C(u_i/u_r)^2\to 0$。
$u=0$	这时 $v=0$，由同一个上界即得。

这里按精确秩取历史矩阵的全部列空间；按数值阈值删去某些列是另一种算法，需要另作扰动分析。

Yang–Yuan, arXiv:2609.12114v1。

附录：数值例子的设置

H 为 60 阶对角矩阵：在 $[1,2]$、$[10,15]$、$[50,60]$、$[200,250]$、$[800,1000]$ 中各取 4 个等距点（含端点），共 20 个互异特征值，各重复 3 次。初始梯度为

$$(g_0)_i = 1 + \sin\big(0.37(i+7)\big) + 0.3\cos\big(1.17(i+7)\big) + 0.5\sin\big(0.11(i+7)\big), \quad i = 1, \dots, 60.$$

$p = 5$；第 0 组用精确线搜索步长；$W = H^a$，$a = -1, -0.5, \dots, 2$；Ritz 值按从大到小的顺序使用。数值秩的相对阈值为 10^{-12}，停止条件为 $\|g\| / \|g_0\| \le 10^{-11}$。

- 七种权下，历史矩阵的数值秩始终为 5，阈值没有删去任何一列。
- $A_j \le (\kappa - 1)^p = 999^5 \approx 10^{15}$ 只是简单的上界，实际最大为 2.26×10^5。
- Curtis–Guo 的试验中，同一比值最大约为 2×10^{16}，即 $\sigma_{\min} / \|g\| \approx 5 \times 10^{-17}$，已小于机器精度。

Yang–Yuan, arXiv:2609.12114v1；Curtis–Guo (2018), Table 1。有限步的试验不能判断 ξ_j 是否有界。[5, 23]

附录：半一致遍历定理的条件

X 为紧度量空间，T 为 Borel 可测映射，集合 S 包含 T 的全部不连续点；f 上半连续、有上界，且在 S 上取值 $-\infty$。若每个使 $\int f\,d\mu > -\infty$ 的不变概率测度都满足 $\int f\,d\mu \le \ell_0$，则对每个 $\varepsilon > 0$，存在 $B_\varepsilon < \infty$，使所有不经过 S 的轨道段都满足

$$\sum_{j=0}^{n-1} f(T^j x) \le n(\ell_0 + \varepsilon) + B_\varepsilon .$$

用反证法：取平均值偏高的轨道段，其经验测度有弱极限 μ；对 $\max(f, -M)$ 用 Portmanteau 定理，再令 $M \to \infty$，知 $\int f\,d\mu$ 有限，从而 $\mu(S) = 0$。于是对连续函数 q，$q \circ T$ 只在一个 μ 零测集上不连续，可以取极限，证得 μ 不变，矛盾。

Yang–Yuan, arXiv:2608.07839v1。对一般的 Borel 映射，经验测度的弱极限不一定是不变测度。

附录：渐近收敛因子与一致上界

结论	条件
每条轨道 $\rho_g \le c_H$	任意 $g_0 \neq 0$，任意首步 $\eta_0 > 0$；全程使用 BB1（或 BB2）。
最坏情形 $\sup \rho_g = c_H$	首步步长按 BB 公式计算的平衡轨道达到 c_H；新的上界说明不会更大。
一致上界 $C_\gamma \gamma^k$	对每个 $\gamma > c_H$ 成立；要求首步步长也按 BB 公式计算，或限制在固定的区间 $[\eta_-, \eta_+] \subset (0, \infty)$ 内。
一致上界 $C c_H^k$	尚未证明，不能由渐近收敛因子直接推出。

函数值误差的渐近收敛因子是梯度渐近收敛因子的平方；上界 $C_\gamma \gamma^k$ 并不要求每一步梯度都下降。

Yang–Yuan, arXiv:2608.07839v1。

附录：Lyapunov 函数中指数的选取

$$V_\beta(g) = \|P_a g\|^{2\beta} \, \|P_b g\|^{2(1-\beta)}, \qquad 0 < \beta < 1.$$

对 $\tau \in [a, b]$ 求 $\sqrt{V_\beta}$ 一步因子的最大值，得到

$$\Gamma(\beta) = \frac{(b-a)\beta^\beta (1-\beta)^{1-\beta}}{b^\beta a^{1-\beta}}.$$

$$\frac{d}{d\beta} \log \Gamma = \log \frac{a\beta}{b(1-\beta)}, \qquad \frac{d^2}{d\beta^2} \log \Gamma = \frac{1}{\beta} + \frac{1}{1-\beta} > 0.$$

唯一的最小点是 $\beta_* = b/(a+b)$，且 $\Gamma(\beta_*) = c$。这一最优性只针对上述函数族；V 在端点分量为零处仍然退化。

Yang–Yuan, arXiv:2609.12084v1。

附录：连续谱情形也能达到最坏因子

在 $\mathcal{H} = L^2([a,b])$ 上取 $(Ag)(\lambda) = \lambda g(\lambda)$，$g_0 \equiv 1$，BB1 的首步也按 BB 公式计算。

$$g_k(\lambda) = \left(1 - \frac{2\lambda}{a+b}\right)^k, \qquad \eta_k = \frac{2}{a+b}.$$

$$\frac{\|g_k\|}{\|g_0\|} = \frac{c^k}{\sqrt{2k+1}}, \qquad c = \frac{b-a}{b+a}.$$

A 没有特征向量，渐近收敛因子仍为 c。这只是一个例子，并不说明每个连续谱算子都有达到 c 的轨道；一般情形下，“一致上界中的因子不能小于 c”是利用端点附近的谱、在有限步内逼近平衡情形而得到的。

Yang–Yuan, arXiv:2608.07839v1，Hilbert 空间附录。

附录：非线性问题的局部结果

$G = \nabla f$ 在 x_* 处严格 Fréchet 可微：

$$\sup_{\substack{x \neq y \\ x,y \in B_r(x_*)}} \frac{\|G(x) - G(y) - H_*(x-y)\|}{\|x-y\|} \to 0, \qquad 0 \prec H_* = H_*^*.$$

记 $m_* = \min \operatorname{spec}(H_*) > 0$，$L_* = \max \operatorname{spec}(H_*) < \infty$，$c_* = (L_* - m_*)/(L_* + m_*)$。全程使用 BB1（或 BB2）。对 $\gamma \in (c_*, 1)$，存在 $r_\gamma > 0$，使从球 $B_{r_\gamma}(x_*)$ 内任意两个不同的初始点出发，都有一致的上界 $C_\gamma \gamma^k$。证明分三步：先对首步有界的情形证明二次模型的一致上界；再在梯度变小之前跟踪二次模型；最后分段收缩并重新启动。

仅有 x_* 处的 Fréchet 可微性，不足以一致地控制两点割线的误差；$f \in C^2$（Hesse 矩阵按算子范数连续）即可，不需要 Hesse 矩阵局部 Lipschitz 连续。$c_* = 0$ 时渐近收敛因子为 0，但不能断言 Q-超线性收敛。

Yang–Yuan, arXiv:2608.07839v1，局部非线性部分。

参考文献 I

[1] H. Akaike. On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method. *Annals of the Institute of Statistical Mathematics*, 11(1):1–16, 1959. doi: 10.1007/BF01831719.

[2] U. M. Ascher, K. van den Doel, H. Huang, and B. F. Svaiter. Gradient descent and fast artificial time integration. *ESAIM: Mathematical Modelling and Numerical Analysis*, 43(4):689–708, 2009. doi: 10.1051/m2an/2009025.

[3] B. Azmi and K. Kunisch. Analysis of the Barzilai–Borwein step-sizes for problems in Hilbert spaces. *Journal of Optimization Theory and Applications*, 185(3):819–844, 2020. doi: 10.1007/s10957-020-01677-y.

[4] J. Barzilai and J. M. Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1): 141–148, 1988. doi: 10.1093/imanum/8.1.141.

[5] F. E. Curtis and W. Guo. *R*-linear convergence of limited memory steepest descent. *IMA Journal of Numerical Analysis*, 38(2):720–742, 2018. doi: 10.1093/imanum/drx016.

[6] Y.-H. Dai. A new analysis on the Barzilai–Borwein gradient method. *Journal of the Operations Research Society of China*, 1(2):187–198, 2013. doi: 10.1007/s40305-013-0007-x.

[7] Y.-H. Dai and L.-Z. Liao. *R*-linear convergence of the Barzilai and Borwein gradient method. *IMA Journal of Numerical Analysis*, 22(1):1–10, 2002. doi: 10.1093/imanum/22.1.1.

参考文献 II

[8] G. Ferrandi and M. E. Hochstenbach. Limited memory gradient methods for unconstrained optimization. *Numerical Algorithms*, 99:735–764, 2025. doi: 10.1007/s11075-024-01895-9.

[9] R. Fletcher. A limited memory steepest descent method. *Mathematical Programming*, 135:413–436, 2012. doi: 10.1007/s10107-011-0479-6.

[10] G. E. Forsythe. On the asymptotic directions of the *s*-dimensional optimum gradient method. *Numerische Mathematik*, 11(1):57–76, 1968. doi: 10.1007/BF02165472.

[11] D. Li and R. Sun. On a faster *R*-linear convergence rate of the Barzilai–Borwein method, 2021. URL https://arxiv.org/abs/2101.00205v2. Preprint, arXiv:2101.00205v2.

[12] X.-R. Li and Y.-K. Huang. A note on *R*-linear convergence of nonmonotone gradient methods. *Journal of the Operations Research Society of China*, 13(1):313–325, 2025. doi: 10.1007/s40305-023-00468-2.

[13] L. Pronzato, H. P. Wynn, and A. A. Zhigljavsky. Asymptotic behaviour of a family of gradient algorithms in $\mathbb{R}^d$ and Hilbert spaces. *Mathematical Programming*, 107(3):409–438, 2006. doi: 10.1007/s10107-005-0602-7.

[14] M. Raydan. On the Barzilai and Borwein choice of steplength for the gradient method. *IMA Journal of Numerical Analysis*, 13(3):321–326, 1993. doi: 10.1093/imanum/13.3.321.

参考文献 III

[15] J. Shen and J. Hu. Stability of discrete-time switched homogeneous systems on cones and conewise homogeneous inclusions. *SIAM Journal on Control and Optimization*, 50(4):2216–2253, 2012. doi: 10.1137/110845215.

[16] B. Shi, S. S. Du, W. J. Su, and M. I. Jordan. Acceleration via symplectic discretization of high-resolution differential equations. In *Advances in Neural Information Processing Systems*, volume 32, pages 5745–5753, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/a9986cb066812f440bc2bb6e3c13696c-Abstract.html.

[17] B. Shi, S. S. Du, M. I. Jordan, and W. J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, 195:79–148, 2022. doi: 10.1007/s10107-021-01681-8.

[18] R. Sturman and J. Stark. Semi-uniform ergodic theorems and applications to forced systems. *Nonlinearity*, 13(1):113–143, 2000. doi: 10.1088/0951-7715/13/1/306.

[19] W. Su, S. Boyd, and E. J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016. URL https://jmlr.org/papers/v17/15-084.html.

参考文献 IV

[20] A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences of the United States of America*, 113(47): E7351–E7358, 2016. doi: 10.1073/pnas.1614734113. URL https://arxiv.org/abs/1603.04245v1.

[21] A. C. Wilson, B. Recht, and M. I. Jordan. A Lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34, 2021. URL https://jmlr.org/papers/v22/20-195.html.

[22] S. Yang and Y.-x. Yuan. *R*-linear convergence of Barzilai–Borwein methods on strictly convex quadratics via closed-cone homogeneous dynamics, 2026. URL https://arxiv.org/abs/2609.12100v1. Preprint, arXiv:2609.12100v1, 10 September 2026.

[23] S. Yang and Y.-x. Yuan. Global and unconditional *R*-linear convergence of rank-compressed weighted LMSD sweeps for strictly convex quadratics, 2026. URL https://arxiv.org/abs/2609.12114v1. Preprint, arXiv:2609.12114v1, 10 September 2026.

[24] S. Yang and Y.-x. Yuan. Lyapunov functions and *R*-linear convergence for quadratic Barzilai–Borwein dynamics, 2026. URL https://arxiv.org/abs/2609.12084v1. Preprint, arXiv:2609.12084v1, 10 September 2026.

参考文献 V

[25] S. Yang and Y.-X. Yuan. The sharp worst-case asymptotic rate of the Barzilai–Borwein method in $\mathbb{R}^d$ and Hilbert spaces, 2026. URL https://arxiv.org/abs/2608.07839v1. Preprint, arXiv:2608.07839v1, 8 August 2026.

[26] Q. Zou and F. Magoulès. Delayed gradient methods for symmetric and positive definite linear systems. *SIAM Review*, 64(3):517–553, 2022. doi: 10.1137/20M1321140.

[27] 袁亚湘. 梯度法步长基于子空间逼近的选取. 计算数学, 48(3):395–404, 2026. doi: 10.12286/jssx.j2026-1387.