GRADIENT METHODS WITH CHAOTIC-INSPIRED STEP SIZES
CONFERENCE SUBMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper investigates gradient methods with dynamically varying step sizes for convex quadratic optimization problems. We propose a family of step-size rules, named FC and FCA, motivated by the chaotic behavior of gradient iterations. The main theoretical contribution is a rigorous analysis of the two-dimensional convex quadratic setting. For this low-dimensional case, we derive closed-form characterizations of the iteration map and establish the stability properties of its fixed points under different parameter regimes. The emergence of chaotic dynamics is demonstrated numerically via positive Lyapunov exponents, initial sensitivity, and ergodicity, though a rigorous mathematical proof of chaos is not established in this work.

The theoretical findings in two dimensions inspire the construction of the FC/FCA step-size schemes for high-dimensional problems. We conduct extensive numerical experiments on large-scale convex quadratic test problems with a wide range of condition numbers, including ill-conditioned cases. Our numerical results demonstrate the empirical performance of FC/FCA compared with classical gradient methods such as the BB and CBB step-size rules.

We further state several conjectures regarding the high-dimensional dynamical behavior and iteration complexity of the proposed methods. Rigorous proofs for these high-dimensional claims remain open.

1 INTRODUCTION

We consider the convex quadratic optimization problem

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{2}x^\top Ax - b^\top x, \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite and $b \in \mathbb{R}^n$. First-order methods for Eq.(1) generate a sequence $\{x_k\}$ via

$$x_{k+1} = x_k - \alpha_k g_k, \tag{2}$$

where $g_k = \nabla f(x_k) = Ax_k - b$ and $\alpha_k > 0$ is the steplength. The classical steepest descent (SD) method, introduced by Cauchy A.Cauchy (1847) , selects α_k by exact minimization along the negative gradient:

$$\alpha_k^{SD} = \frac{g_k^\top g_k}{g_k^\top A g_k}. \tag{3}$$

The theoretical and practical limitations of SD for ill-conditioned problems are well documented Akaike (1959) Forsythe (1968) . To overcome the slow convergence of SD, Barzilai and Borwein J.BARZILAI & J.M.BORWEIN (1988) proposed a two-point steplength that has since become a cornerstone of practical gradient methods. Another direct line search method is the minimal gradient (MG) method YH (2003), which minimizes the gradient norm along the current gradient direction. Inspired by the BB method, subsequent research has proposed various improved strategies, including the alternating step (AS) method proposed by Dai YH (2006), the alternating minimization (AM) method proposed by Dai and Yuan, the step-size formula proposed by Yuan Yuan (2006) that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

alternates between the Yuan steplength and exact line search, the relaxed steepest descent (RSD) and cyclic Barzilai-Borwein (CBB) methods proposed by Raydan and Svaiter Raydan M (2002), the RSDA method proposed by De Asmundis et al. Serafino;F.Riccio;G.Toraldo (2013), and the randomized steepest descent method proposed by Kalousek Kalousek (2015).

First-order methods with dynamically varying step sizes, such as Barzilai–Borwein (BB) and its cyclic variant (CBB), are widely used for their efficiency on ill-conditioned problems . Non-monotone step-size schedules have a long history YH (2006); Jason Altschuler (2024), sharing with our approach the goal of intermittently contracting the slowest eigenmode, but typically relying on precomputed schedules or two-point secant approximations. Chaotic behavior in gradient descent on quadratics has also been observed Kees van den Doel (2012), where the Rayleigh quotient can follow a chaotic trajectory that accelerates convergence by escaping Akaike's 2-cycle trap.

Our work differs in introducing a multiplicative relaxation parameter t on the reciprocal Cauchy steplength, yielding a one-parameter family of rational maps whose bifurcation structure we characterize rigorously in two dimensions, and in connecting the chaotic dynamics to a practical algorithmic mechanism for high-dimensional problems. While not directly designed for a specific machine learning task, our framework contributes to the understanding of non-monotone step-size methods, which play an important role in modern machine learning optimization.

In this work, we study a different aspect of the SD iteration. Define the reciprocal of the Cauchy steplength as

$$r_k = \frac{1}{\alpha_k^{SD}} = \frac{g_k^\top A g_k}{g_k^\top g_k}. \tag{4}$$

For quadratic f, r_k satisfies $\lambda_{\min} \le r_k \le \lambda_{\max}$. We introduce a multiplicative relaxation parameter $t > 0$ and consider the modified iteration

$$x_{k+1} = x_k - \frac{1}{t r_k} g_k. \tag{5}$$

Our central observation is that the asymptotic behavior of $\{r_k\}$ depends critically on t, and that in a certain range of t, the projection of the high-dimensional Rayleigh quotient dynamics onto a low-dimensional scalar rational map follows an approximately deterministic trajectory that can be exploited computationally.

The core intuition motivating the repelling regime $t > 1$ is spectral. When r_k wanders chaotically, it intermittently visits neighborhoods of the smallest eigenvalue $\lambda_{\min}$. During such excursions, the step size $1/(tr_k)$ becomes large, and the slowest eigenmode—responsible for the plateau in ill-conditioned quadratic minimization—is contracted aggressively. This mechanism is fundamentally different from monotone step-size schedules, which must balance slow-mode contraction against fast-mode stability at every step. The chaotic wandering allows the algorithm to take occasional long steps that target the slow mode, while subsequent short steps quench the fast-mode amplification induced by those long steps. We make this intuition precise in Section 3 and generalize it into a high-dimensional complexity conjecture in Section 6.

1.1 Contributions

The main contributions of this paper are summarized as follows:

1. **Rigorous dynamical theory for two-dimensional convex quadratic problems.** We derive the iteration map associated with FC/FCA gradient updates, prove stability properties of its fixed points, and provide a full characterization of three distinct dynamical regimes for this discrete iteration system.
2. **Construction of FC/FCA step-size rules for high-dimensional convex quadratic optimization**, inspired by the two-dimensional dynamical insights.
3. **Formulation of several open conjectures about high-dimensional iteration dynamics and complexity.** These conjectures are supported by our numerical evidence, but rigorous mathematical proofs are left for future research.

We emphasize that all theoretical results in this work are established only for the two-dimensional setting. Extending these rigorous statements to general high-dimensional convex quadratic problems

is highly non-trivial and remains an open challenge. The high-dimensional behavior we observe numerically is presented as conjectures rather than proven theorems.

Our focus lies in understanding how the dynamical behavior of the reciprocal Cauchy steplength r_k can be exploited to construct practical gradient algorithms for large-scale quadratic problems.

The remainder of the paper is organized as follows. Section 2 provides preliminaries. Section 3 presents the rigorous two-dimensional dynamical analysis. Section 4 introduces the FC and FCA algorithms. Section 5 reports numerical experiments. Section 6 states the conjectures for the high-dimensional case. Section 7 concludes.

2 PRELIMINARIES

2.1 PROBLEM SETTING AND EIGEN-COORDINATE REPRESENTATION

We consider the convex quadratic problem equation 1 with A symmetric positive definite. Let the eigenvalues of A be ordered in descending order as

$$\lambda_{\max} = a^{(1)} \geq a^{(2)} \geq \cdots \geq a^{(n)} = \lambda_{\min} > 0, \tag{6}$$

and denote the condition number by

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{a^{(1)}}{a^{(n)}} \gg 1. \tag{7}$$

This is the regime of practical interest for ill-conditioned quadratic optimization.

Without loss of generality, we assume the coordinate origin is shifted to the unique minimizer $x^* = A^{-1}b$; equivalently, we work with the error coordinates $z_k = x_k - x^*$. Since A is symmetric positive definite, it admits the spectral decomposition

$$A = Q\Lambda Q^\top, \qquad \Lambda = \mathrm{diag}(a^{(1)}, \ldots, a^{(n)}), \tag{8}$$

where Q is orthogonal and $a^{(1)} \geq a^{(2)} \geq \cdots \geq a^{(n)} > 0$. In the eigen-coordinates $y_k = Q^\top z_k$, the quadratic objective takes the separable form

$$f(x_k) = \frac{1}{2} z_k^\top A z_k = \frac{1}{2} \sum_{i=1}^{n} a^{(i)} (y_k^{(i)})^2. \tag{9}$$

In particular, for the two-dimensional case $n = 2$ analyzed in Section 3, the objective reduces to

$$f(x_k) = \frac{1}{2} \left(a^{(1)} (y_k^{(1)})^2 + a^{(2)} (y_k^{(2)})^2 \right), \tag{10}$$

where $a^{(1)} = \lambda_{\max}$ and $a^{(2)} = \lambda_{\min}$. This simplified form eliminates the linear term $b^\top x$ and renders the dynamics of the iteration separable in the eigen-coordinates, which greatly facilitates the analysis. The gradient in eigen-coordinates is

$$g_k = A z_k = Q\Lambda Q^\top z_k = Q\Lambda y_k, \tag{11}$$

with componentwise expression $g_k^{(i)} = a^{(i)} y_k^{(i)}$.

2.2 STEPLENGTH, RELAXATION, AND COMPONENTWISE UPDATES

The classical steepest descent (SD) method selects α_k by exact minimization along the negative gradient, and we define the reciprocal of the Cauchy steplength as $r_k = 1/\alpha_k^{SD} = (g_k^\top A g_k)/(g_k^\top g_k)$. In the eigen-coordinates, r_k can be written as

$$r_k = \frac{\sum_{i=1}^{n} a^{(i)} {g_k^{(i)}}^2}{\sum_{i=1}^{n} {g_k^{(i)}}^2}. \tag{12}$$

Thus r_k is the weighted average of the eigenvalues $a^{(i)}$, with weights proportional to the squared gradient components.

We introduce a multiplicative relaxation parameter $t > 0$ and consider the modified iteration $x_{k+1} = x_k - \frac{1}{tr_k} g_k$. The standard SD method corresponds to $t = 1$; $t > 1$ corresponds to under-relaxation, while $t < 1$ corresponds to over-relaxation of the Cauchy step. Throughout this paper, we restrict attention to $t \in (1, 2)$. This range balances three considerations:

1. **Sufficient chaos.** For $t > 1$, the one-dimensional invariant map G is in the repelling (chaotic) regime analyzed in Section 3. A value of t slightly above 1 is already sufficient to generate chaos, as observed in the gradient-descent literature **?**.
2. **Bounded steplength dilution.** The effective steplength is $1/(tr_k)$. For t close to 1, this is close to the Cauchy step; for larger t, it is increasingly diluted, and the per-step contraction of both the slow and fast eigenmodes weakens. Our numerical results (Section 5) show that the asymptotic contraction constant $C_n^*(t)$ decreases with t, confirming that excessively large t is counterproductive.
3. **Preservation of the invariant interval.** The critical point r_3 of the mapping G (slightly larger than $a^{(2)} = \lambda_{\min}$) corresponds to a local maximum $G(r_3)$ near the lower end of the domain, which determines the maximum jump amplitude from $r_k \approx a^{(2)}$ to its next iterate. As t increases, $G(r_3)$ decreases monotonically, narrowing the range of chaotic wandering within the invariant interval and weakening the efficiency of chaotic exploration. Therefore $t \leq 2$ also helps to keep the invariant interval of G sufficiently wide. The explicit form of G and the closed-form expression for r_3 are given in Section 3, where the dependence of $G(r_3)$ on t is analyzed.

The sub-threshold regime $t \in (0, c)$ with $c \approx 1/2$, in which the optimization iterates diverge despite the convergence of r_k, is excluded by this restriction; see Appendix A.

Under equation **??**, the componentwise updates of y_k and g_k in the eigen-coordinates are

$$y_{k+1}^{(i)} = y_k^{(i)} \left(1 - \frac{a^{(i)}}{tr_k}\right), \qquad g_{k+1}^{(i)} = g_k^{(i)} \left(1 - \frac{a^{(i)}}{tr_k}\right). \tag{13}$$

Consequently, the reciprocal steplength satisfies the recursion

$$r_{k+1} = \frac{\sum_{i=1}^n a^{(i)} {g_k^{(i)}}^2 (tr_k - a^{(i)})^2}{\sum_{i=1}^n {g_k^{(i)}}^2 (tr_k - a^{(i)})^2}. \tag{14}$$

These componentwise relations will be used repeatedly in the dynamical analysis of Section 3. In particular, for the two-dimensional case, they lead directly to the mapping $G(r)$ studied in Section 3.1.

3 Two-Dimensional Rigorous Dynamical Analysis

In this section, we restrict attention to the two-dimensional case. All rigorous theorems, propositions, and lemmas in this paper are confined to this section.

3.1 Iteration Map and Stability

The mapping $G(r)$ is obtained by eliminating the gradient components $g_k^{(1)}, g_k^{(2)}$ under the two-dimensional quadratic setting. We assume $g_k^{(1)} \neq 0$ and $g_k^{(2)} \neq 0$, so that r_k is well-defined and lies in $(a^{(2)}, a^{(1)})$.

For a two-dimensional positive definite quadratic with eigenvalues $0 < a^{(2)} < a^{(1)}$, the update rule for r_k is given by

$$r_{k+1} = \frac{a^{(1)} {g_k^{(1)}}^2 (tr_k - a^{(1)})^2 + a^{(2)} {g_k^{(2)}}^2 (tr_k - a^{(2)})^2}{{g_k^{(1)}}^2 (tr_k - a^{(1)})^2 + {g_k^{(2)}}^2 (tr_k - a^{(2)})^2}. \tag{15}$$

Treating r_k as a continuous variable r, we define the mapping

$$G(r) = \frac{a^{(1)}(r - a^{(2)})(tr - a^{(1)})^2 - a^{(2)}(r - a^{(1)})(tr - a^{(2)})^2}{(tr - a^{(1)})^2 (r - a^{(2)}) - (tr - a^{(2)})^2 (r - a^{(1)})}, \tag{16}$$

where $r \in (a^{(2)}, a^{(1)})$ and $G(r) \in [a^{(2)}, G(r_3^*)]$. Figure 1 illustrates $G(r)$ for several values of t.The lower endpoint $a^{(2)}$ is attained at the critical point $r_1^* = a^{(1)}/t$, while the upper endpoint $G(r_3^*)$ is the local maximum attained at the critical point r_3^* .The local maximum $G(r_3^*)$ decreases monotonically with t, which is consistent with the third consideration in Section 2.2 for restricting $t < 2$

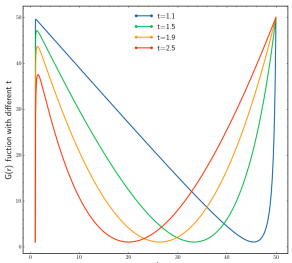

Figure 1: The function $G(r)$ for $a^{(1)} = 50$, $a^{(2)} = 1$, with $t = 1.1, 1.5, 1.9, 2.5$.

We note that the first iterate r_0 is not constrained by the parameter t; it may take any value in $(a^{(2)}, a^{(1)})$ depending on the initial point x_0. Since the dynamics are chaotic, r_k rapidly escapes from any initial location and contracts into the invariant interval $[a^{(2)}, G(r_3^*)]$, where all subsequent iterates remain. In particular, even if $r_0 > G(r_3^*)$, the iterates decrease below $G(r_3^*)$ within a small number of steps. Therefore, the chaotic trajectory of r_k is asymptotically confined to the invariant interval $[a^{(2)}, G(r_3^*)]$, whose width decreases monotonically with t.

Proposition 1 (Critical Points)**.** *Under the standing assumption $\kappa \gg 1$ and the restriction $t \in (1, 2)$, setting $G'(r) = 0$ yields four critical points: $r_1^* = \frac{a^{(1)}}{t}, \quad r_2^* = \frac{a^{(2)}}{t}$,*

$$r_{3,4}^* = \frac{a^{(1)} + a^{(2)}}{2(2-t)} \mp \frac{\sqrt{t^2(a^{(1)} + a^{(2)})^2 - 4t(2-t)(2t-1)a^{(1)}a^{(2)}}}{2t(2-t)}, \tag{17}$$

Under this parameter domain, only r_1^* and r_3^* lie in $(a^{(2)}, a^{(1)})$. The critical point $r_1^* = a^{(1)}/t$ gives $G(r_1^*) = a^{(2)}$, the lower endpoint of the invariant interval, while r_3^* gives the local maximum $G(r_3^*)$, the upper endpoint of the invariant interval. In particular, for $a^{(1)} = 50$, $a^{(2)} = 1$, $t = 1.2$, we have $r_3^* \approx 1.165$.

Proof. See Appendix A. □

Proposition 2 (Fixed Point)**.** *The unique interior fixed point of G is $r_e = \frac{a^{(1)}+a^{(2)}}{2t}$, and the derivative at this point is*

$$G'(r_e) = 1 + \frac{8\left(ta^{(1)}a^{(2)} + \frac{1}{t}\left(\frac{a^{(1)}+a^{(2)}}{2}\right)^2 - \frac{(a^{(1)}+a^{(2)})^2}{2}\right)}{(a^{(1)} - a^{(2)})^2}. \tag{18}$$

The closed-form expression for $G'(r_e)$ is algebraically involved; it simplifies substantially when evaluated for specific eigenvalue pairs.

Proof. See Appendix A. □

Theorem 1 (Stability Classification)**.** *The dynamical behavior of $\{r_k\}$ under G is determined by t as follows:*

(i) Stable regime *($t < 1$)**:** The parameter range can be divided into two subintervals, $(c, 1)$ and $(0, c)$, where*

$$c = \frac{1}{2} + \frac{a^{(2)}}{2a^{(1)}}. \tag{19}$$

If $t \in (c, 1)$, then $|G'(r_e)| < 1$, r_e is an attracting fixed point, and $r_k \to r_e$. If $t \in (0, c)$, then $r_e > a^{(1)}$ falls outside the valid range, and $a^{(1)}$ is the unique attracting fixed point, so $r_k \to a^{(1)}$. In particular, for large condition numbers $\kappa = a^{(1)}/a^{(2)} \gg 1$, we have $a^{(2)}/a^{(1)} \to 0$, so $c \to 1/2$.

***(ii) Critical regime** ($t = 1$): $G'(r_e) = -1$, so r_e is a neutral fixed point undergoing a period-doubling (flip) bifurcation. The sequence r_k alternates between two values satisfying*

$$r_k + r_{k+1} = a^{(1)} + a^{(2)}.$$

***(iii) Repelling regime** ($t > 1$): $G'(r_e) < -1$, so r_e is a repelling fixed point. The boundary point $a^{(1)}$ is also repelling, with $G'(a^{(1)}) \approx (t/t-1)^2 > 1$ Under the standing assumption $\kappa \gg 1$ and the restriction $t \in (1, 2)$, the threshold $t_2 = \frac{(\kappa+1)^2}{4\kappa}$ satisfies $t_2 \approx \kappa/4 \gg 2$, so that the entire interval of practical interest lies within the repelling regime $(1, t_2)$.*

(i) Stable regime of the invariant map ($t < 1$): ... If $t \in (0, c)$, then $r_e > a^{(1)}$ falls outside the valid range, and $a^{(1)}$ is the unique attracting fixed point, so $r_k \to a^{(1)}$. We note that in the latter case the optimization iterates diverge despite the convergence of r_k; see Appendix A for details.

Proof. See Appendix A. □

3.2 Chaos and Algorithmic Implications

The following results are based on numerical computation and are stated as numerical evidence, not as rigorously proven theorems.

Proposition 3 (Numerical Evidence of Chaos)**.** *For $t = 1.2$, $a^{(1)} = 50$, $a^{(2)} = 1$, the mapping G exhibits the following chaotic characteristics:*

***(i) Positive Lyapunov exponent.** Numerical computation yields*

$$\lambda = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \ln |G'(r_k)| \approx 0.4424 > 0, \tag{20}$$

indicating sensitive dependence on initial conditions.

***(ii) Initial sensitivity.** Taking two nearby initial values $r_0 = 20$ and $r_0 = 20.001$, the corresponding trajectories remain nearly identical for the first 15 steps and diverge significantly from the 16th iteration onward. This directly verifies the sensitive dependence on initial conditions.*

***(iii) Ergodicity.** We conduct two groups of tests. In the first group, we use the same initial value and run 10^4, 5×10^4, and 10^5 iterations, respectively. In the second group, we use three distinct initial values $r_0 = 2, 20, 45$ and run 10^4 iterations for each. In both groups, only the initial value r_0 is varied while all other problem parameters are held fixed. The resulting distributions of r_k are highly similar, indicating that the long-term statistical behavior of the system is independent of the initial condition and the iteration length.*

We note that these observations are numerical in nature and do not constitute a rigorous proof of topological chaos. A rigorous proof would require additional arguments, such as the existence of dense periodic orbits and topological transitivity.

A positive Lyapunov exponent, together with initial sensitivity and ergodicity, provides strong numerical evidence that the dynamics are genuinely chaotic in the repelling regime $t > 1$.

The three dynamical regimes have direct consequences for algorithm design:

1. **Stable regime ($t < 1$):** r_k converges to a fixed value, so the decay rate of the slowest eigencomponent is constant. Not advantageous for accelerating convergence.
2. **Critical regime ($t = 1$):** r_k alternates between two values. The decay rate is improved relative to $t < 1$ but remains limited.
3. **Repelling regime ($t > 1$):** In the repelling regime, r_k wanders chaotically, and its long-term distribution concentrates near the largest and smallest eigenvalues. The slow mode is the bottleneck of convergence: when r_k approaches $a^{(n)}$, the slow-mode multiplier $|1 -$

$a^{(n)}/(tr_k)| \to |1-1/t|$, which is smaller than its value when r_k is large, producing a strong contraction. The chaotic wandering causes this strong contraction to occur repeatedly, which is the core mechanism underlying the improved empirical convergence of FC. The fast mode is effectively contracted when r_k is large and amplified when r_k approaches $a^{(n)}$; in the distribution-weighted sense, the amplification is compensated on average by the subsequent short steps, and the fast mode does not constitute a long-term bottleneck. Consequently, the overall convergence rate is governed by the average contraction rate of the slow mode.

Conjecture on intermediate modes. We conjecture that intermediate eigenmodes decay rapidly, so that the iterates collapse onto the two-dimensional subspace spanned by $v_{\min}$ and $v_{\max}$. This would justify reducing the overall convergence analysis to the slow-mode contraction rate. Numerical evidence is provided in Figure 2 in Appendix D; a rigorous proof remains open.

4 The FC and FCA Algorithms for High-Dimensional Problems

Inspired by the two-dimensional dynamical insights, we propose the following strategy. Within each block of m iterations, we compute l exact values of r_k, and for the remaining $m - l$ steps, we predict r_k via the approximate mapping $G_n(r_{k-1})$. The iterate x_k is updated using the steplength determined by r_k. Both FC and FCA share the same prediction and safeguarding logic; they differ only in the update rule for x_{k+1}.

The scalar surrogate map $G_n(r)$ is defined as

$$G_n(r) = \frac{a^{(1)}(r - a^{(2)})(tr - a^{(1)})^2 - a^{(2)}(r - a^{(1)})(tr - a^{(2)})^2}{(tr - a^{(1)})^2(r - a^{(2)}) - (tr - a^{(2)})^2(r - a^{(1)})}, \tag{21}$$

which is the one-dimensional rational map derived in Section 3. The map G_n requires the extremal eigenvalues $a^{(1)} = \lambda_{\max}$ and $a^{(2)} = \lambda_{\min}$ as inputs.

4.1 Safeguard Thresholds

In the repelling regime $t > 1$, the effective steplength is $1/(tr_k)$. Although the chaotic wandering of r_k allows the trajectory to intermittently visit regions near $a^{(n)} = \lambda_{\min}$ and thereby enhance the contraction of the slow mode, letting r_k approach either endpoint too closely is problematic. When r_k approaches $a^{(n)}$ too closely, the effective steplength is diluted by the factor $t > 1$, and the slow-mode contraction efficiency degrades; this motivates the lower threshold θ_{low}. Similarly, when r_k approaches $a^{(1)} = \lambda_{\max}$, the decay rate of all components is reduced by the factor $t > 1$; this motivates the upper threshold θ_{high}.

Whether r_k is computed exactly or predicted by G_n, if it falls outside the interval $[\theta_{\text{low}}, \theta_{\text{high}}]$, we revert to a steplength that does not involve the parameter t. Specifically, the fallback update uses the exact Cauchy reciprocal steplength $1/r_k$, with either the SD step (for FC) or the CBB step (for FCA), thereby avoiding the dilution effect of $t > 1$.

4.2 Unified Algorithm for FC and FCA

The FC and FCA methods share the same prediction and safeguarding logic and differ only in the update rule. We present them as a single algorithm parameterized by a mode indicator.

Parameter selection. $t = 1.2$, $m = 10$, $l = 2$ provide a robust default setting. The empirical thresholds $\theta_{\text{low}} = 10$, $\theta_{\text{high}} = 8333$ are calibrated for problems with $a^{(2)} = 0.1$, $a^{(1)} = 10000$.

4.3 Computational Cost

Both FC and FCA substantially reduce the number of steplength evaluations. Within each block of m iterations, only l exact evaluations of r_k are required, while the remaining evaluations are obtained through the scalar surrogate map G_n, whose evaluation costs $O(1)$. From a computational standpoint, FC and FCA introduce only one additional scalar parameter t relative to the corresponding baseline methods. Details of the MVP and inner-product accounting are provided in Appendix B.

Algorithm 1 FC and FCA

Require: A, x_0, t, m, l, mode, θ_{low}, θ_{high}, tol

Compute the gradient g_0 at x_0
$r_0 \leftarrow \dfrac{g_0^\top A g_0}{g_0^\top g_0}$
for $k = 0, 1, \ldots$ until $\|g_k\|_2 < \text{tol}$ **do**
  if $k \bmod m < l$ **then**
    $r_k \leftarrow \dfrac{g_k^\top A g_k}{g_k^\top g_k}$; $\quad t_{\text{eff}} \leftarrow t$
  else
    $r_k \leftarrow G_n(r_{k-1})$
    if $r_k < \theta_{\text{low}}$ **or** $r_k > \theta_{\text{high}}$ **then**
      $r_k \leftarrow \dfrac{g_k^\top A g_k}{g_k^\top g_k}$; $\quad t_{\text{eff}} \leftarrow 1$
    else
      $t_{\text{eff}} \leftarrow t$
    end if
  end if
  if mode = FC **then**
    $x_{k+1} \leftarrow x_k - \dfrac{1}{t_{\text{eff}} r_k} g_k$
  else
    $x_{k+1} \leftarrow x_k - \dfrac{2}{t_{\text{eff}} r_k} g_k + \dfrac{1}{(t_{\text{eff}} r_k)^2} A g_k$
  end if
  Compute the gradient g_{k+1} at x_{k+1}
end for

5 Numerical Experiments

5.1 Comparison with Baseline Methods

We evaluate the proposed algorithms on synthetic quadratic problems with $n \in \{10, 100, 1000, 10000, 100000\}$, $\lambda_{\max} = 10^4$, $\lambda_{\min} = \lambda_{\max}/\kappa$, and $\kappa \in \{100, 1000, 10000, 100000\}$. Intermediate eigenvalues $a^{(i)}$ ($i = 2, \ldots, n-1$) are generated by either random sampling or uniform equidistant partitioning over $[a^{(n)}, a^{(1)}]$. We use two initialization strategies for x_0: "F" (all components set to 10000) and "R" (entries sampled uniformly from $(0, 1)$ and scaled by 10000). The stopping criterion is $\|g_k\|_2 < 10^{-40}$.

Table 1 summarizes the iteration counts for all methods. Across the tested quadratic instances, FCA generally yields the fastest convergence among the compared methods, while FC exhibits performance comparable to CBB. Both FC and FCA require only l exact steplength evaluations over every m iterations, substantially reducing computational overhead.The convergence curves for all methods are shown in Figure 3 in Appendix E.

The empirical convergence behavior of FC/FCA is observed to be competitive with BB and CBB on the tested ill-conditioned quadratic instances, with advantages on some large condition number problems. In particular, the empirical convergence speed of FCA is observed to be comparable to CBB in the early stage and faster in later iterations. We note that these observations are purely numerical and do not constitute a rigorous guarantee for all high-dimensional convex quadratic problems.

5.2 Convergence Rate Estimation

The convergence rate of FC and FCA is estimated via the empirical geometric-mean contraction factor of the slowest eigenmode. Full details, including the convergence guarantee for $t > 1$, the definition of the contraction metric, and numerical results, are provided in Appendix C.

Table 1: Iteration counts for different methods.The intermediate eigenvalues $a^{(i)}$ are generated by uniform equidistant partitioning over $[a^{(n)}, a^{(1)}]$, and the initial iterate x_0 is fixed with all components set to 10000. Parameters: $n = 10000$, $t = 1.2$, $m = 10$, $l = 2$.The symbol "–" indicates that the method did not reach the stopping criterion within a reasonable number of iterations.).

Method	$\kappa = 100$	$\kappa = 1000$	$\kappa = 10000$	$\kappa = 100000$
SD	5350	–	–	–
RSD	1357	5123	16098	–
BB	595	2224	8201	–
CBB	322	1229	4460	10267
FC	235	558	1621	6335
FCA	232	234	774	2320

6 Heuristics and Conjectures for the High-Dimensional Case

In this section, we present several conjectures about the high-dimensional behavior of FC/FCA iterations. These statements are strongly supported by our extensive numerical simulations. However, we do not provide rigorous proofs here, and verifying or disproving these claims remains an open research problem.

6.1 Approximate Mapping

Conjecture 1 (Approximate Mapping)**.** *For high-dimensional convex quadratic problems with well-spread eigenvalues, the evolution of r_k approximately follows the mapping*

$$G_n(r) = \frac{\lambda_{\max}(r-\lambda_{\min})(tr-\lambda_{\max})^2 - \lambda_{\min}(r-\lambda_{\max})(tr-\lambda_{\min})^2}{(r-\lambda_{\min})(tr-\lambda_{\max})^2 - (r-\lambda_{\max})(tr-\lambda_{\min})^2}, \tag{22}$$

where $\lambda_{\max} = a^{(1)}$ and $\lambda_{\min} = a^{(n)}$.

The approximation $G_n(r) \approx G(r)$ neglects cross-coupling between distinct eigencomponents, which is present in high-dimensional gradient dynamics. This decoupling approximation becomes more plausible when eigenvalues are well separated. This approximation is supported by numerical evidence (see Section 5) but has not been rigorously proven.

6.2 Stationary Density and Support

Conjecture 2 (Iteration Complexity)**.** *Let $\rho(r) = |1 - \lambda_{\min}/(tr)|$, where $\lambda_{\min} = a^{(n)}$, denote the one-step contraction factor of the slowest eigenmode. The average logarithmic decay rate of the slowest eigenmode satisfies $\mathbb{E}[\log \rho] \approx -C(t)/\sqrt{\kappa}$ as $\kappa \to \infty$, where $\kappa = \lambda_{\max}/\lambda_{\min}$ and $C(t) > 0$ depends only on t. For $t = 1.2$, numerical simulations give $C(1.2) \approx 1.08$ to 1.20. Consequently, the iteration complexity for the slowest eigenmode is conjectured to be $k = O(\sqrt{\kappa}\log(1/\varepsilon))$. A rigorous proof of this scaling and an analytical expression for $C(t)$ remain open problems.*

7 Conclusions

We studied first-order gradient methods with dynamically varying step sizes for unconstrained convex quadratic optimization. We established a rigorous two-dimensional dynamical analysis, proposed the FC and FCA algorithms, and benchmarked them on synthetic quadratic problems against BB and CBB baselines. The theoretical characterization of the high-dimensional case remains open and is formulated as conjectures.

Limitations: The rigorous dynamical analysis is confined to the two-dimensional case; a proof of the intermediate-mode collapse and of the FCA iteration complexity remains open.

REFERENCES

A.Cauchy. Méthode générale pour larésolution des systéms d'équations simultanées. *C.R. Acad. Sci. Paris*, 25:536–538, 1847.

H. Akaike. On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method. *Ann. Inst. Stat. Math. Tokyo*, 11:1–17, 1959.

E. Forsythe, G. On the asymptotic directions of the s-dimensional optimum gradient method. *[j]. Numerische Mathematik*, 11:57–76, 1968.

Pablo Parrilo. Jason Altschuler. Acceleration by stepsize hedging: Multi-step descent and the silver stepsize schedule. *Journal of the ACM*, 2024.

J.BARZILAI and J.M.BORWEIN. Two point step size gradient methods. *IMA J. Numer. Anal.*, 8: 141–148, 1988.

Z Kalousek. Steepest descent method with random steplengths. *[J].Foundations of Computational Mathematics*, 17(2):359–422, 2015.

Uri Ascher Kees van den Doel. The chaotic nature of faster gradient descent methods. *Journal of Scientific Computing*, 2012.

Svaiter B F. Raydan M. Relaxed steepest descent and cauchy-barzilai-borwein method. *[J]Computational Optimization and Applications*, 21:155–167, 2002.

R.De Asmundis;D.di Serafino;F.Riccio;G.Toraldo. On spectral properties of steepest descent methods. *IMA Journal of Numerical Analysis*, 4, 2013.

Dai YH. Alternate minimization gradient method. *IMA J. Numer. Anal*, 23(3):377–393, 2003.

Dai YH. Alternate step gradient method. *Optimization*, 52(4-5):395–415, 2006.

X. Yuan, Y. A new step size for the gradient method. *Journal of Computational Mathematics*, 24 (2):149–156, 2006.

A PROOFS

Proof of Proposition 1. Direct computation of $G'(r)$ and solution of the resulting quadratic equation. The restriction $t < 2$ ensures that the factors $2(2-t)$ and $2t(2-t)$ in the denominators of r_3^* and r_4^* are nonzero, so that these critical points are well-defined. The restriction $t < \kappa = a^{(1)}/a^{(2)}$ ensures that $r_1^* = a^{(1)}/t > a^{(2)}$, so that r_1^* lies in the interior $(a^{(2)}, a^{(1)})$. Since the standing assumption $\kappa \gg 1$ implies $\kappa > 2$, and the restriction $t \in (1,2)$ implies $t < 2$, both conditions are automatically satisfied.

Proof of Proposition 2. Setting $G(r) = r$ and solving yields the unique interior fixed point $r_e = (a^{(1)} + a^{(2)})/(2t)$. Substituting into $G'(r)$ yields the derivative formula.

Proof of Theorem 1. The derivative $G'(r_e)$ is a monotonically decreasing function of t on $(0, (\kappa+1)/(2\sqrt{\kappa}))$ and monotonically increasing thereafter. As $t \to 1$, $G'(r_e) \to -1$. For $t > 1$, $G'(r_e) < -1$; for $t < 1$, $|G'(r_e)| < 1$ when t exceeds the threshold c. Solving $G'(r_e) = -1$ for $t > 1$ yields two roots: $t = 1$ and $t_2 = (\kappa+1)^2/(4\kappa)$. For $t \in (t_2, (\kappa+1)/2)$, the fixed point r_e remains inside $(a^{(2)}, a^{(1)})$ but satisfies $|G'(r_e)| < 1$, making it locally attracting. However, under the standing assumption $\kappa \gg 1$, we have $t_2 \approx \kappa/4 \gg 2$, so the interval $t \in (1,2)$ of practical interest lies entirely within the repelling regime $(1, t_2)$. The boundary derivative follows from direct computation.

Divergence of optimization iterates in the sub-threshold regime $t \in (0,c)$. Theorem 1(i) establishes that the normalized Rayleigh quotient r_k converges to $a^{(1)}$ for $t \in (0,c)$, where $c = 1/2 + a^{(2)}/(2a^{(1)}) \approx 1/2$ in the large-condition-number regime. This convergence of the *invariant map* must be distinguished from the convergence of the *optimization iterates* $\{x_k\}$.

For $t \in (0,c)$, the asymptotic steplength satisfies

$$\alpha_k = \frac{1}{t\, r_k} \to \frac{1}{t\, a^{(1)}} > \frac{2}{a^{(1)}} = \frac{2}{\lambda_{\max}}. \tag{23}$$

Under this steplength, the per-iteration multiplier on the largest eigenmode satisfies

$$\left|1 - \frac{a^{(1)}}{t\, r_k}\right| \to \left|1 - \frac{1}{t}\right| > 1. \tag{24}$$

Consequently, the gradient component $g_k^{(1)}$ along the largest eigenmode diverges exponentially, and $\|g_k\| \to \infty$. The apparent convergence $r_k \to a^{(1)}$ is an artifact of the largest component $(g_k^{(1)})^2$ growing exponentially faster than the remaining components, not an indication that the optimization algorithm is stable.

True algorithmic stability of the optimization iterates is therefore restricted to $t \in (c,1)$, where the steplength does not exceed twice the Cauchy steplength in the limit and the per-iteration multipliers satisfy $|1 - a^{(i)}/(tr_k)| \le 1$ for all eigenmodes. The regime $t \in (0,c)$ is best understood as a *stable regime of the scalar invariant map* accompanied by *exponential divergence of the optimization iterates along the largest eigenmode*. This distinction motivates the algorithmic safeguards introduced in Section 4.

B COMPUTATIONAL COST ANALYSIS

The exact evaluation of the Cauchy reciprocal steplength $r_k = (g_k^\top A g_k)/(g_k^\top g_k)$ requires one matrix-vector product (MVP) Ag_k, in addition to two vector inner products ($O(n)$ flops). In contrast, the surrogate map $G_n(r_{k-1})$ is a scalar rational function whose evaluation costs $O(1)$. Thus, replacing exact evaluations by surrogate predictions substantially reduces the cost of computing r_k.

This reduction applies specifically to the cost of computing r_k. The gradient update still requires one MVP per iteration, and this cost is not affected by the prediction of r_k. For FC, the update requires one MVP per iteration; for FCA, the CBB-type update requires two MVPs per iteration.

C Convergence Rate Estimation

Convergence guarantee for $t > 1$. For $t > 1$, the iteration $x_{k+1} = x_k - \frac{1}{tr_k} g_k$ is guaranteed to converge. Indeed, along the negative gradient direction the quadratic objective is a parabola in the steplength α, minimized at $\alpha^* = 1/r_k$ and symmetric about it. For $t > 1$, the steplength satisfies $\alpha_k = \alpha^*/t < \alpha^* < 2\alpha^*$, so that $f(x_{k+1}) < f(x_k)$ unless $g_k = 0$.

Empirical contraction metric. We estimate the asymptotic contraction factor of the slowest eigenmode using the empirical geometric mean. The slowest eigenmode corresponds to the smallest eigenvalue $a^{(n)} = \lambda_{\min}$, and its empirical geometric-mean contraction factor along the iteration trajectory is defined as

$$\hat{\rho}_{\min} = \exp\left(\frac{1}{N}\sum_{k=1}^{N} \ln\left|1 - \frac{\lambda_{\min}}{tr_k}\right|\right). \tag{25}$$

Remarks. The decay of an eigenmode over N steps is multiplicative, so its asymptotic logarithmic decay rate is given by the geometric mean rather than the arithmetic mean. The absolute value ensures that the metric correctly represents the contraction magnitude in the repelling regime $t > 1$, where $1 - \lambda_{\min}/(tr_k)$ can be negative. We use two independent indices: $i \in \{1, \ldots, n\}$ indexes the spatial eigenmodes of A, while $k \in \{1, \ldots, N\}$ indexes the iteration steps.

Numerical results. For $a^{(n)} = 1$, $t = 1.2$, theoretical computation gives $\hat{\rho}_{\min} \approx 0.8589$ for $\kappa = 50$, $\hat{\rho}_{\min} \approx 0.8916$ for $\kappa = 100$, $\hat{\rho}_{\min} \approx 0.9634$ for $\kappa = 1000$, and $\hat{\rho}_{\min} \approx 0.9881$ for $\kappa = 10000$. As κ increases, $\hat{\rho}_{\min}$ approaches 1, indicating that the contraction of the slow mode becomes slower as the problem becomes more ill-conditioned. The corresponding scaling constant $C(1.2) = -\sqrt{\kappa} \ln \hat{\rho}_{\min}$ lies in the range 1.08 to 1.20, consistent with Conjecture 2. For a fixed κ, using different values of N (such as 10^4, 5×10^4, and 10^5) gives values of $\hat{\rho}_{\min}$ differing by approximately 0.001, indicating that the geometric mean has converged.

D High-Dimensional Scatter Plot of (r_k, r_{k+1})

Figure 2 shows the scatter plot of (r_k, r_{k+1}) for $a^{(1)} = 10000$, $a^{(n)} = 1$, $t = 1.2$, and $n = 10000$. The blue points represent the approximation of $G(r)$, the orange points represent the approximation of $H(r) = G^{-1}(r)$, and the green line is the identity function $Y(r) = r$. The points concentrate at both ends of the interval $[a^{(n)}, a^{(1)}]$, indicating that the long-term distribution of r_k is concentrated near the extreme eigenvalues and that intermediate eigenmodes decay rapidly.

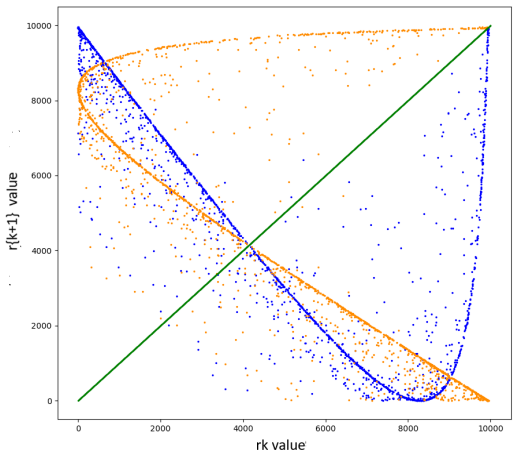

Figure 2: Scatter plot of (r_k, r_{k+1}) for $a^{(1)} = 10000$, $a^{(n)} = 1$, $t = 1.2$, $n = 10000$. Blue points: approximation of $G(r)$. Orange points: approximation of $H(r) = G^{-1}(r)$. Green line: identity function $Y(r) = r$.

E CONVERGENCE CURVES

Figure 3 shows the convergence curves of all methods for $n = 10000$, $\kappa = 10000$, $t = 1.2$, $m = 10$, and $l = 2$. The vertical axis shows $\log_{10} \|g_k\|$. FCA converges fastest, followed by CBB and FC, while SD and RSD converge very slowly.

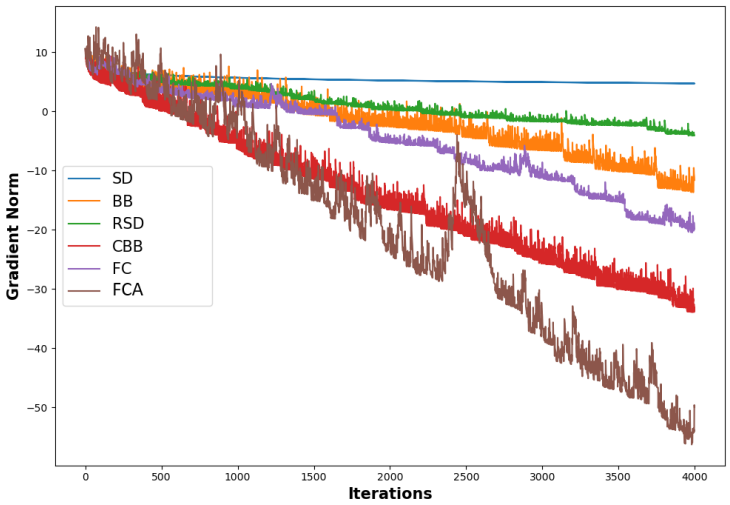

Figure 3: Convergence comparison of six methods ($n = 10000$, $\kappa = 10000$, $t = 1.2$, $m = 10$, $l = 2$). The vertical axis shows $\log_{10} \|g_k\|$.

F COMPLETE ITERATION COUNTS

Table 2: Iteration number for different methods across all tested configurations ($t = 1.2$, $m = 10$, $l = 2$)

n	Knd	$a^{(i)}$	x_0	SD	RSD	BB	CBB	FC	FCA
10	100	F	F	5304	1486	570	191	222	181
10	1000	F	F	-	4590	1656	557	557	441
10	10000	F	F	-	12662	5683	2031	1429	742
10	100000	F	F	-	-	18160	5124	5209	970
10	100	R	F	5264	1366	405	213	200	142
10	1000	R	F	-	5294	1539	687	664	441
10	10000	R	F	-	14486	4555	1505	1492	532
10	100000	R	F	-	-	-	4036	7333	1250
10	100	F	R	5203	1440	565	233	252	182
10	1000	F	R	-	5078	1413	642	529	341
10	10000	F	R	-	16292	4782	2187	1963	413
10	100000	F	R	-	-	17732	2805	3074	642
10	100	R	R	5243	1425	539	275	277	291
10	1000	R	R	-	4182	1304	419	491	282
10	10000	R	R	-	14843	9768	2936	3790	742
10	100000	R	R	-	-	-	5909	5910	1769
100	100	F	F	5308	1239	628	340	286	302
100	1000	F	F	-	4094	1790	919	964	661
100	10000	F	F	-	14843	6393	2274	3600	1652
100	100000	F	F	-	-	-	12296	8691	2742
100	100	R	F	5272	1537	616	304	317	304
100	1000	R	F	-	4938	2047	889	1023	734
100	10000	R	F	-	18881	5521	2959	3015	1622
100	100000	R	F	-	-	-	8440	7617	4997
100	100	F	R	5318	1390	618	299	312	338
100	1000	F	R	-	4878	2026	858	947	731
100	10000	F	R	-	13957	5741	3608	3230	1684
100	100000	F	R	-	-	-	10662	10577	4561
100	100	R	R	5284	1285	607	310	286	334
100	1000	R	R	-	4240	1702	780	811	941
100	10000	R	R	-	13869	7164	3153	2718	1281
100	100000	R	R	-	-	-	9919	8977	4142
1000	100	F	F	5308	1167	663	315	321	331
1000	1000	F	F	-	4573	2107	1002	1139	1025
1000	10000	F	F	-	17590	7675	3449	3303	2366
1000	100000	F	F	-	-	-	9610	9803	6217
1000	100	R	F	5344	1401	629	336	301	372
1000	1000	R	F	-	4697	2050	945	1042	931
1000	10000	R	F	-	15928	7579	3477	3720	2439
1000	100000	R	F	-	-	-	13011	10046	6273
1000	100	F	R	5296	1443	649	307	306	304
1000	1000	F	R	-	4458	2032	1102	1037	924
1000	10000	F	R	-	16385	6920	3285	3020	2862
1000	100000	F	R	-	-	-	13146	9985	6350
1000	100	R	R	5198	1310	612	318	309	361
1000	1000	R	R	-	4584	2200	993	999	1057
1000	10000	R	R	-	16090	6694	2758	2998	2797
1000	100000	R	R	-	-	-	10549	13160	5375
10000	100	F	F	5350	1357	595	322	235	232
10000	1000	F	F	-	5123	2224	1229	558	234
10000	10000	F	F	-	16098	8201	4460	1621	774
10000	100000	F	F	-	-	28551	10267	6335	2320
10000	100	R	F	5376	1345	615	321	323	392
10000	1000	R	F	-	5160	2288	1022	1053	1114
10000	10000	R	F	-	12766	7291	3389	3457	2901
10000	100000	R	F	-	-	-	12760	10215	6661
10000	100	F	R	5244	1517	693	326	303	312
10000	1000	F	R	-	5051	1954	1015	1057	1038
10000	10000	F	R	-	17516	7731	3676	3371	3642
10000	100000	F	R	-	-	-	13185	13281	9621
10000	100	R	R	5360	1215	646	321	342	382
10000	1000	R	R	-	4881	2241	1014	967	1011
10000	10000	R	R	-	15326	6549	3550	3438	2853
10000	100000	R	R	-	-	-	13332	12517	8363
100000	100	F	F	5350	1481	595	322	319	332
100000	1000	F	F	-	4783	2224	1229	1071	1031
100000	10000	F	F	-	15268	8201	4460	3419	3125
100000	100000	F	F	-	-	-	10267	11771	9655
100000	100	R	F	5406	1470	633	335	318	365
100000	1000	R	F	-	4798	2392	1068	1046	1062
100000	10000	R	F	-	15723	7386	3722	3759	3144
100000	100000	R	F	-	-	-	11752	9968	8144
100000	100	F	R	5328	1302	625	321	320	329
100000	1000	F	R	-	5125	2112	1035	947	954
100000	10000	F	R	-	14625	7485	3005	3481	2973
100000	100000	F	R	-	-	-	11353	13782	9281
100000	100	R	R	5370	1535	672	344	339	384
100000	1000	R	R	-	5040	2142	1060	1047	1120
100000	10000	R	R	-	17463	8246	3447	3396	3053
100000	100000	R	R	-	-	-	13021	11337	8141