What Caps the Step Size Is Not the Size of the Gradient but the Curvature in the Steepest Direction

— For a quadratic with condition number $\kappa = 100$, the cap on the step is $2/\kappa = 0.02000000$, and at 1.01 times that, 0.02020000, one step multiplies by 1.020000 and diverges — the separator is whether progress is limited by the size of the gradient or by the spread of curvature — [Paper 734]

Yuuki YAMAGISHI (山岸 勇輝)
Yamagishi Complexity Torus Lab, Japan
September 2026

— From thinking to syncing. —

Abstract

Raising the step size in gradient descent makes it faster. But there is a cap, and the cap is set by the shape of the problem.The cap, the best step and the number of steps are all derived from one number.No new theorem or law is claimed. First, the cap is $2/\kappa$——0.02000000 at $\kappa = 100$, with neither the gradient nor the starting point in it (Section 2). **Second, a little above the cap it diverges——at 1.01 times, 0.02020000, one step multiplies by 1.020000** (Section 2). **Third, and this is the core. Even at the best step the contraction is close to one——$2/(1+\kappa) = 0.01980198$ contracts by only 0.980198, agreeing with $(\kappa-1)/(\kappa+1)$ to a difference of 0** (Section 3). **Fourth, so steps are needed——230.250834 of them to reach a hundredth, and 2302.584325 at $\kappa = 1000$** (Section 3). **Fifth, the two curvatures do different work——the larger sets the cap and the smaller only the number of steps** (Section 3). **Sixth, the separator is whether progress is limited by the size of the gradient or by the spread of curvature——it is the latter; the former places no limit on the step at all** (Section 4). **Keywords**: gradient descent, step size, condition number, rate of convergence, a cap set by curvature, one phrase two quantities.

Scope of this paper (scope note): **No new theorem or law is claimed——the convergence condition of gradient descent on a quadratic, the optimal step $2/(L+\mu)$ and the rate $(\kappa-1)/(\kappa+1)$ are all standard (Cauchy 1847, Nesterov 2004).Restricted to quadratics——non-convex surfaces, saddle points, stochastic gradient descent, momentum and adaptive methods are not treated.No real problem is claimed——particular learners and data are not treated; only the condition number is varied.Preconditioning is not designed——means of lowering the condition number are pointed to but not built.A caution at $\kappa = 1$——that one step suffices there holds because the function is quadratic, and does not hold in general.Relation to earlier papers**: **Paper 434 showed that what makes it cheaper is the count, not the calendar ——the phrase "learning rate" appears there seventeen times, but it means the slope of the learning curve, the fraction by which cost falls as cumulative output doubles. The "learning rate" here is a step size in optimisation: a different dimension and a different meaning. One phrase naming two quantities.**What is added is **deriving the cap as $2/\kappa$ and confirming it on both sides**, **showing the best step's contraction to match $(\kappa-1)/(\kappa+1)$**, **tabulating the number of steps against the condition number**, **separating the two roles of the two curvatures**, and **placing the separator on gradient against spread of curvature**.

1. Notation and the Question

Setting: **take a quadratic whose two directions have curvatures 1 and κ. Gradient descent then multiplies each direction by a fixed factor per step.**

$$f(x) = \tfrac{1}{2}\left(x_1^2 + \kappa x_2^2\right), \qquad \text{contraction} = \max\left(|1-\eta|,\ |1-\eta\kappa|\right)$$

η is the step and κ the condition number. A contraction of one or more diverges.

Question: **how far can the step be raised?**

2. The Cap Is Set by the Steepest Direction

Computation 2.1: **at $\kappa = 100$, the contraction per step was derived for each step size.**

step	contraction per step	outcome
0.00100000	0.999000	contracts
0.00500000	0.995000	contracts
0.01980000	0.980200	contracts
0.01980198	0.980198	contracts
0.02500000	1.500000	diverges

Observation 2.2

the cap is $2/\kappa = 0.02000000$.At 1.01 times, 0.02020000, one step multiplies by 1.020000; at 0.99 times, 0.01980000, by 0.980200——the sign was confirmed on both sides of the boundary.

Observation 2.3

the size of the gradient is not in that cap.What is in it is the curvature in the steepest direction alone——however large the gradient, the ceiling on the step does not move.

Observation 2.4

the cap goes inversely with the curvature.Ten times the κ lowers the cap tenfold, from 0.02000000 to 0.00200000——one steep direction is enough to bind the whole step.

3. Even at the Best Step, the Contraction Is Close to One

Computation 3.1: for each condition number, the best step, its contraction and the steps to reach a hundredth were derived.

κ	best step	contraction	steps needed
1	1.00000000	0.000000	1
10	0.18181818	0.818182	22.948892
100	0.01980198	0.980198	230.250834
1000	0.00199800	0.998002	2302.584325

Observation 3.2

this is the core: even at the best step, $\kappa = 100$ contracts by only 0.980198.That agrees with $(\kappa - 1)/(\kappa + 1)$ to a difference of 0——choose the step as well as possible and the contraction still sticks to one when the condition number is large.

Observation 3.3

so the number of steps is nearly proportional to the condition number.22.948892 at $\kappa = 10$, 230.250834 at 100, 2302.584325 at 1000——tenfold each time, so lowering the condition number beats adjusting the step.

Observation 3.4

the two curvatures do different work.The larger sets the cap; the smaller sets the number of steps——asking "what learning rate?" touches only the first.

Observation 3.5

one step suffices at $\kappa = 1$ because the function is quadratic.With one curvature, a step of 1 lands exactly at the bottom——this does not hold for general functions, so that row belongs inside the quadratic.

4. The Separator

Observation 4.1

the separator is whether progress is limited by the size of the gradient or by the spread of curvature.It is the latter; the former places no limit on the step——“the slope is steep, so go carefully” is not a reason inside this expression.

Observation 4.2

beside Paper 434.Paper 434 showed that what makes it cheaper is the count, not the calendar——“learning rate” appears there too, meaning the fall of cost against cumulative output, a different dimension and a different meaning from the step here. One phrase naming two quantities.

Observation 4.3

what is not said.Non-convex surfaces, saddle points, stochastic methods and the design of preconditioners are not treated, and the account is restricted to quadratics——what is shown is which quantity sets the cap and which sets the count.

5. Roll Call and Closing

1. **cap** $\mathbf{0.02000000}$ (Observation 2.2): $2/\kappa$**.**
2. $\mathbf{0.02020000}$ **gives** $\mathbf{1.020000}$ (Observation 2.2): **it diverges.**
3. $\mathbf{0.01980000}$ **gives** $\mathbf{0.980200}$ (Computation 2.1): **inside the boundary.**
4. 0.02500000 **gives** $\mathbf{1.500000}$ (Computation 2.1): **plainly divergent.**
5. **best step** $\mathbf{0.01980198}$ (Computation 3.1): $2/(1+\kappa)$**.**
6. **contraction** $\mathbf{0.980198}$ (Observation 3.2): **the core.**
7. $\mathbf{230.250834}$ **steps** (Computation 3.1): **at** $\kappa = 100$**.**
8. $\mathbf{22.948892}$ **at** $\kappa = 10$ (Computation 3.1): **a tenth.**
9. $\mathbf{2302.584325}$ **at** $\kappa = 1000$ (Computation 3.1): **ten times.**
10. **gradient or spread of curvature** (Observation 4.1): **the separator; Paper 434.**

Closing: **What caps the step size is not the size of the gradient but the curvature in the steepest direction.For a quadratic with $\kappa = 100$ the cap is $2/\kappa = 0.02000000$, and at 1.01 times that, 0.02020000, one step multiplies by 1.020000 and diverges, while at 0.99 times, 0.01980000, it contracts by 0.980200——neither the gradient nor the starting point is in that cap; only the curvature in the steepest direction is.The best step is $2/(1+\kappa) = 0.01980198$, yet its contraction is only 0.980198, agreeing with $(\kappa-1)/(\kappa+1)$ to a difference of 0——choose the step as well as possible and the contraction still sticks to one when the condition number is large.So steps are needed——22.948892 to reach a hundredth at $\kappa = 10$, 230.250834 at 100 and 2302.584325 at 1000, tenfold each time.The two curvatures do different work: the larger sets the cap and the smaller sets the number of steps——asking "what learning rate?" touches only the first, so lowering the condition number beats adjusting the step.The separator is whether progress is limited by the size of the gradient or by the spread of curvature.Placed among the earlier papers——"learning rate" appears in Paper 434 as well, meaning the fall of cost against cumulative output: a different dimension and a different meaning.To be honest——the account is restricted to quadratics, and that one step suffices at $\kappa = 1$ holds only because of that.**

References

1. Cauchy, A. (1847). Methode generale pour la resolution des systemes d'equations simultanees. *Comptes Rendus de l'Academie des Sciences*.
2. Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization*. Springer.
3. Yamagishi, Y. (2026). *What Makes It Cheaper Is the Count, Not the Calendar* [Paper 434]. Zenodo. 10.5281/zenodo.22708735

Yuuki YAMAGISHI (山岸 勇輝)

© 2026 Yamagishi Complexity Torus Lab, Japan · Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)