27 Newton's Algorithm

Goal

Newton's algorithm, conjugate gradient, auto-differentiation, local quadratic rate of convergence, affine invariance, cubic regularization

Alert 27.1: Convention

Nice surveys for Newton's algorithm include Polyak (2006) and Ypma (1995).

Gray boxes are not required hence can be omitted for unenthusiastic readers.

This note is likely to be updated again soon.

Polyak, B. T. (2006). "Newton-Kantorovich Method and Its Global Convergence". *Journal of Mathematical Sciences*, vol. 133, pp. 1513–1523.

Ypma, T. J. (1995). "Historical Development of the Newton-Raphson Method". *SIAM Review*, vol. 37, no. 4, pp. 531–551.

Definition 27.2: Problem

In this lecture we consider solving the smooth minimization problem:

$$\min_{\mathbf{w}} \ f(\mathbf{w}),$$

where f is sufficiently smooth (twice or thrice continuously differentiable).

Algorithm 27.3: Newton

Newton's algorithm iteratively minimizes the second order Taylor series of the smooth objective f:

$$\begin{aligned}\mathbf{w}_{t+1} &= \operatorname*{argmin}_{\mathbf{w}} \ f(\mathbf{w}_t) + \langle f'(\mathbf{w}_t), \mathbf{w} - \mathbf{w}_t\rangle + \tfrac{1}{2\eta_t} \langle f''_t(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t), \mathbf{w} - \mathbf{w}_t\rangle \\ &= \mathbf{w}_t - \eta_t [f''_t(\mathbf{w}_t)]^{-1} \cdot f'_t(\mathbf{w}_t), \qquad (27.1)\end{aligned}$$

where the step size η_t is often set to the constant 1 (at least in later stages when we are sufficiently close to the local optimum).

Often, we also add some regularization (e.g. Levenberg-Marquardt) to the Hessian so that its inverse is well-behaved.

History 27.4: Early studies of Newton's algorithm

To mention just a few: Fine (1916), Bennett (1916), Kantorovich (1948, 1949, 1957), Cheney and Goldstein (1959), Goldstein (1965), and Goldfeld et al. (1966).

Fine, H. B. (1916). "On Newton's method of approximation". *Proceedings of National Academy of Sciences*, vol. 2, no. 9, pp. 546–552.

Bennett, A. A. (1916). "Newton's Method in General Analysis". *Proceedings of National Academy of Sciences*, vol. 2, no. 10, pp. 592–598.

Kantorovich, L. V. (1948). "On Newton's method for functional equations". *Dokl. Akad. Nauk SSSR*, vol. 59, pp. 1237–1240.

— (1949). "On the Newton method". *Proceedings of the Steklov Institute of Mathematics*, vol. 28, pp. 104–144.

— (1957). "Some further applications of the Newton method for functional equations". *Vestn. LGU, Ser. Math. Mech.*, vol. 7, pp. 68–103.

Cheney, E. W. and A. A. Goldstein (1959). "Newton's Method for Convex Programming and Tchebycheff Approximation". *Numerische Mathematik*, vol. 1, pp. 253–268.

Goldstein, A. A. (1965). "On Newton's method". *Numerische Mathematik*, vol. 7, pp. 391–393.
Goldfeld, S. M., R. E. Quandt, and H. F. Trotter (1966). "Maximization by Quadratic Hill-Climbing". *Econometrica*, vol. 34, no. 3, pp. 541–551.

Alert 27.5: Affine equivariance and invariance

Newton's Algorithm 27.3 is affine equivariant. Indeed, consider the change-of-variable $\mathbf{w} = A\mathbf{z}$ for any invertible linear map A, we have

$$(f \circ A)'(\mathbf{z}) = A^\top f'(A\mathbf{z}), \quad (f \circ A)''(\mathbf{z}) = A^\top f''(A\mathbf{z})A, \implies \mathbf{z}_{t+1} = \mathbf{z}_t - \eta_t A^{-1}[f''(A\mathbf{z}_t)]^{-1} f'(A\mathbf{z}_t),$$

and hence $\mathbf{w}_{t+1} = A\mathbf{z}_{t+1}$ if we start with $\mathbf{w}_0 = A\mathbf{z}_0$.

Newton's Algorithm 27.3 is also affine invariant w.r.t. the Euclidean metric. Indeed, if we change the inner product to $\langle \mathbf{w}, \mathbf{z} \rangle_Q := \langle Q\mathbf{w}, \mathbf{z} \rangle$ for some (symmetric) positive definite Q, we have

$$f' \to Q^{-1} f' \quad \text{and} \quad f'' \to Q^{-1} f''$$

so that the Newton update remains the same.

Alert 27.6: Scaling invariance

Perhaps more surprisingly, Newton's Algorithm 27.3 is also scaling invariant: if we change f to αf for any $\alpha \in \mathbb{R}$, Newton's update still remains the same.

This simple observation actually reveals the true nature of Newton's algorithm: it merely aims to solve the nonlinear equation

$$f'(\mathbf{w}) = \mathbf{0},$$

but does not care if $\mathbf{w}$ is a (local) minimizer or maximizer.

Remark 27.7: Semismooth Newton

For semismooth functions, the Hessian, like a subdifferential, is a set $\partial^2 f$ so that

$$\left[\sup_{H \in \partial^2 f(\mathbf{w})} \|f'(\mathbf{w} + \mathbf{z}) - f'(\mathbf{w}) - H\mathbf{z}\| \right] = o(\|\mathbf{z}\|).$$

We may then still apply Newton's algorithm with an arbitrary $H \in \partial^2 f$. The resulting convergence rate is often superlinear, see Ulbrich (2011).

Ulbrich, M. (2011). "Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces". SIAM.

Theorem 27.8: Local quadratic rate under strong convexity

Suppose f is σ-strongly convex and f'' is L-Lipschitz continuous (w.r.t. the ℓ_2 norm), and $q = \frac{\mathsf{L}}{2\sigma^2}\|f'(\mathbf{w}_0)\|_2 < 1$, then for all t:

$$\|\mathbf{w}_t - \mathbf{w}_*\|_2 \le \tfrac{1}{\sigma}\|f'(\mathbf{w}_t)\|_2 \le \tfrac{2\sigma}{\mathsf{L}} q^{2^t}, \tag{27.2}$$

where $\mathbf{w}_*$ is the unique minimizer of f.

Proof: According to Proposition 2.12, the L-Lipschitz continuity of f'' implies that

$$\|f'(\mathbf{w}_t + \mathbf{z}) - f'(\mathbf{w}_t) - f''(\mathbf{w}_t)\mathbf{z}\|_2 \le \tfrac{\mathsf{L}}{2}\|\mathbf{z}\|_2^2.$$

Taking $\mathbf{z} = -[f''(\mathbf{w}_t)]^{-1}f'(\mathbf{w}_t) =: \mathbf{w}_{t+1} - \mathbf{w}_t$ we obtain

$$\|f'(\mathbf{w}_{t+1})\|_2 \le \tfrac{\mathsf{L}}{2}\|[f''(\mathbf{w}_t)]^{-1}f'(\mathbf{w}_t)\|_2^2 \le \tfrac{\mathsf{L}}{2}\|[f''(\mathbf{w}_t)]^{-1}\|_{\mathrm{sp}}^2 \cdot \|f'(\mathbf{w}_t)\|_2^2 \le \tfrac{\mathsf{L}}{2\sigma^2}\|f'(\mathbf{w}_t)\|_2^2.$$

Therefore, telescoping yields for $t \ge 0$:

$$\tfrac{\mathsf{L}}{2\sigma^2}\|f'(\mathbf{w}_{t+1})\|_2 \le \left(\tfrac{\mathsf{L}}{2\sigma^2}\|f'(\mathbf{w}_t)\|_2\right)^2 \le \cdots \le \left(\tfrac{\mathsf{L}}{2\sigma^2}\|f'(\mathbf{w}_0)\|_2\right)^{2^{t+1}}.$$

Lastly, it follows from the strong convexity of f that (see Proposition 6.22)

$$\|f'(\mathbf{w}_t)\|_2 = \|f'(\mathbf{w}_t) - f'(\mathbf{w}_*)\|_2 \ge \sigma\|\mathbf{w}_t - \mathbf{w}_*\|_2. \qquad \blacksquare$$

The condition $q = \frac{\mathsf{L}}{2\sigma^2}\|f'(\mathbf{w}_0)\|_2 < 1$ implies that

$$\|\mathbf{w}_0 - \mathbf{w}_*\| < \tfrac{2\sigma}{\mathsf{L}},$$

i.e., the starting point $\mathbf{w}_0$ is sufficiently close to the minimizer $\mathbf{w}_*$. Inspecting (27.2) we observe that once the iterate $\mathbf{w}_t$ enters a small ball around $\mathbf{w}_*$ (such that $f'(\mathbf{w}_t) < \frac{2\sigma^2}{\mathsf{L}}$, implying the radius is less than $\frac{2\sigma}{\mathsf{L}}$), it will remain there and converge to $\mathbf{w}_*$ at a qudratic rate. Thus, the constants L and σ can be relativized, as long as we initialize carefully.

Example 27.9: Newton may **not** converge faster than linearly

Let us consider the simple univariate function

$$f(w) := |w|^{5/2}.$$

Clearly, we have $f'(w) = \frac{5}{2}\operatorname{sign}(w)|w|^{3/2}$ and $f''(w) = \frac{15}{4}|w|^{1/2}$. Note that f'' is not Lipschitz continuous and f is not strongly convex. The Newton update is:

$$w_{t+1} = w_t - \tfrac{4}{15}|w_t|^{-1/2} \cdot \tfrac{5}{2}\operatorname{sign}(w_t)|w_t|^{3/2} = w_t - \tfrac{2}{3}w_t = \tfrac{1}{3}w_t,$$

which converges to 0, the unique minimizer, at a linear rate.

Example 27.10: Newton may cycle

Consider the simple univariate function

$$f(w) = -\tfrac{1}{4}w^4 + \tfrac{5}{2}w^2, \qquad f'(w) = -w^3 + 5w, \quad f''(w) = -3w^2 + 5,$$

which, around 0, is locally (strongly) convex and f'' is locally Lipschitz continuous. The Newton update is:

$$w_{t+1} = w_t - \frac{-w_t^3 + 5w_t}{-3w_t^2 + 5} = \frac{2w_t^3}{3w_t^2 - 5}.$$

Thus, with $w_0 = 1$ we enter a cycle

$$\begin{tikzcd} 1 \arrow[r, bend left] & -1 \arrow[l, bend left, "t \leftarrow t+1"] \end{tikzcd}$$

. We verify that restricted to the unit ball around the origin, $\mathsf{L} = 6$ and $\sigma = 2$, so that $q = \frac{\mathsf{L}}{2\sigma^2}\|f'(w_0)\|_2 = 6 \times 4/2^3 = 3 \not< 1$.

Example 27.11: Newton can be chaotic

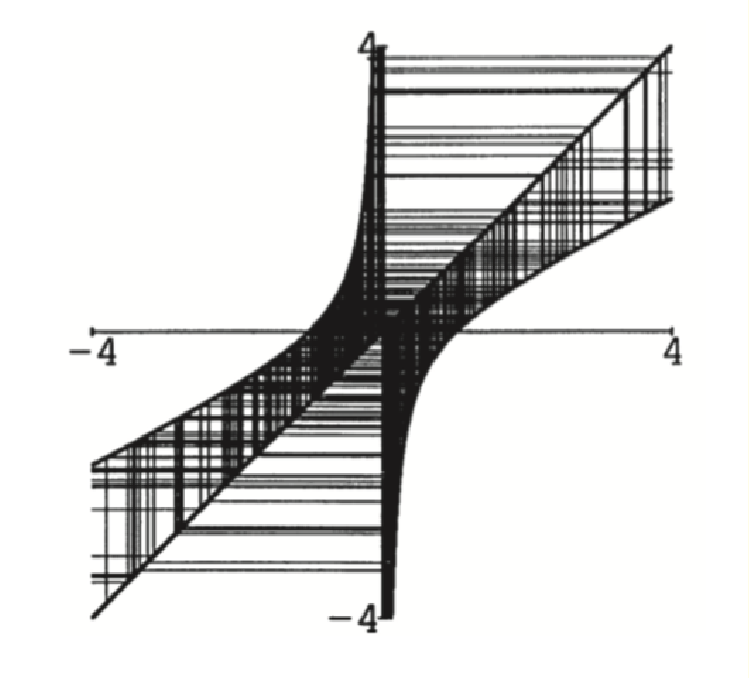
Consider the simple univariate function

$$f(w) = \tfrac{1}{3}w^3 + w, \qquad f'(w) = w^2 + 1, \qquad f''(w) = 2w.$$

Note that f, being nonconvex, tends to $-\infty$ as $w \to -\infty$ while f'' is 2-Lipschitz continuous and vanishes at $w = 0$. The Newton update is:

$$w_{t+1} = w_t - \frac{w_t^2 + 1}{2w_t} = \tfrac{1}{2}(w_t - \tfrac{1}{w_t}),$$

which behaves chaotically:



Algorithm 27.12: Cubic regularization (Nesterov and Polyak 2006)

Since gradient descent minimizes a quadratic upper bound of our function, it is natural to consider minimizing the following cubic upper bound as an alternative to Newton's update (27.1):

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}} \underbrace{f(\mathbf{w}_t) + \langle f'(\mathbf{w}_t), \mathbf{w} - \mathbf{w}_t\rangle + \tfrac{1}{2}\langle f''(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t), \mathbf{w} - \mathbf{w}_t\rangle + \tfrac{1}{6\eta_t}\|\mathbf{w} - \mathbf{w}_t\|_2^3}_{\bar{f}_t(\mathbf{w}) = \bar{f}_{\eta_t}(\mathbf{w};\mathbf{w}_t)}. \tag{27.3}$$

Setting the derivative at $\mathbf{w}_{t+1}$ to zero we obtain:

$$f'(\mathbf{w}_t) + f''(\mathbf{w}_t)(\mathbf{w}_{t+1} - \mathbf{w}_t) + \tfrac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \cdot (\mathbf{w}_{t+1} - \mathbf{w}_t) = \mathbf{0}, \tag{27.4}$$

$$\text{also } \langle f'(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \langle f''(\mathbf{w}_t)(\mathbf{w}_{t+1} - \mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \tfrac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^3 = 0.$$

In other words,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \Big[f''(\mathbf{w}_t) + \tfrac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \cdot \mathrm{Id}\Big]^{-1} f'(\mathbf{w}_t),$$

which is essentially Newton's update with an adaptive Levenberg-Marquardt regularization. In particular, noting that usually $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \to 0$, the regularization automatically dies down as we progress, so cubic regularization eventually behaves similarly to Newton's update.

Nesterov, Y. and B. T. Polyak (2006). "Cubic regularization of Newton method and its global performance". *Mathematical Programming*, vol. 108, pp. 177–205.

Exercise 27.13: Nitpicking the proof

Before we continue, let us revisit our proof in Theorem 4.21 for gradient descent (set $r \equiv 0$ there). Can you recycle and adapt the proof for cubic regularization (27.3)?

[You may assume f is convex if it helps, although this is not really needed below.]

Alert 27.14: Strong duality

We point out that the subproblem in (27.3) does not appear to be convex, since we do not assume f to be convex. However, we may consider the standard semidefinite program (SDP) relaxation:

$$\min_{Z \succeq \mathbf{0}, Z_{11}=1, \langle I, Z\rangle = \zeta+1} f(\mathbf{w}_t) + \left\langle \frac{1}{2}\begin{bmatrix} 0 & f'(\mathbf{w}_t)^\top \\ f'(\mathbf{w}_t) & f''(\mathbf{w}_t) \end{bmatrix}, Z \right\rangle + \tfrac{1}{6\eta_t}\zeta^{3/2}, \tag{27.5}$$

where we think of $Z = \begin{bmatrix} 1 \\ \mathbf{w} - \mathbf{w}_t \end{bmatrix}\begin{bmatrix} 1 \\ \mathbf{w} - \mathbf{w}_t \end{bmatrix}^\top$. Since we only have two linear constraints on Z, and the objective is linear (and hence concave) in Z, it follows that an optimal Z can be chosen as an extreme point of the constraint set, which is known to have rank exactly 1 (Pataki (1998), see also Barvinok (1995) and Polyak (1998)). In other words, the SDP relaxation is equivalent to the original, seemingly nonconvex problem (27.3)!

We now make two important observations from the convex equivalent (27.5):

- From the KKT conditions we have

$$f''(\mathbf{w}_t) + \tfrac{1}{2\eta_t}\|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2 \cdot \mathrm{Id} \succeq \mathbf{0}, \tag{27.6}$$

  whereas the second-order necessary condition for (27.3) would lose the factor $\frac{1}{2}$ in the second term.

- Setting $\mathbf{w} = \mathbf{w}_t$ and $\mathbf{w} = \mathbf{w}_{t+1}$ in Z respectively we conclude

$$0 \geq \langle f'(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \tfrac{1}{2}\langle f''(\mathbf{w}_t)(\mathbf{w}_{t+1} - \mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \tfrac{1}{4\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^3,$$

  or put it succinctly,

$$f(\mathbf{w}_t) \geq \bar{f}_t(\mathbf{w}_{t+1}) + \tfrac{1}{12\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^3. \tag{27.7}$$

  Note that we do not (need to) know if the above inequality holds for all $\mathbf{w}$ in place of $\mathbf{w}_t$.

See also Barvinok (2001, 2014)

Polyak (2001, 2003) and Uderzo (2013, 2019)

Ben-Tal and Teboulle (1996), Bohnenblust (1948), Dymarsky (2016), Hiriart-Urruty (2001), and Hiriart-Urruty and Torki (2002)

Sturm and Zhang (2003), Luo et al. (2004), So et al. (2008), Ai et al. (2008), Gruber (2009), and Jiang et al. (2017)

Pataki, G. (1998). "On the Rank of Extreme Matrices in Semidefinite Programs and the Multiplicity of Optimal Eigenvalues". *Mathematics of Operations Research*, vol. 23, no. 2, pp. 339–358.

Barvinok, A. I. (1995). "Problems of distance geometry and convex properties of quadratic maps". *Discrete & Computational Geometry*, vol. 13, pp. 189–202.

Polyak, B. T. (1998). "Convexity of Quadratic Transformations and Its Use in Control and Optimization". *Journal of Optimization Theory and Applications*, vol. 99, pp. 553–583.

Barvinok, A. I. (2001). "A Remark on the Rank of Positive Semidefinite Matrices Subject to Affine Constraints". *Discrete & Computational Geometry*, vol. 25, pp. 23–31.

Barvinok, A. I. (2014). "Convexity of the image of a quadratic map via the relative entropy distance". *Beiträge zur Algebra und Geometrie / Contributions to Algebra and Geometry*, vol. 55, pp. 577–593.

Polyak, B. T. (2001). "Convexity of Nonlinear Image of a Small Ball with Applications to Optimization". *Set-Valued Analysis*, vol. 9, pp. 159–168.

— (2003). "The convexity principle and its applications". *Bulletin of the Brazilian Mathematical Society, New Series*, vol. 34, no. 1, pp. 59–75.

Uderzo, A. (2013). "On the Polyak convexity principle and its application to variational analysis". *Nonlinear Analysis: Theory, Methods & Applications*, vol. 91, pp. 60–71.

— (2019). "An extension of the Polyak convexity principle with application to nonconvex optimization". *Pure and Applied Functional Analysis*, vol. 4, no. 2, pp. 427–445.

Ben-Tal, A. and M. Teboulle (1996). "Hidden convexity in some nonconvex quadratically constrained quadratic programming". *Mathematical Programming*, vol. 72, pp. 51–63.

Bohnenblust, F. (1948). "Joint positiveness of matrices". Tech. rep. California Institute of Technology.

Dymarsky, A. (2016). "Convexity of a small ball under quadratic map". *Linear Algebra and its Applications*, vol. 488, pp. 109–123.

Hiriart-Urruty, J.-B. (2001). "Global Optimality Conditions in Maximizing a Convex Quadratic Function under Convex Quadratic Constraints". *Journal of Global Optimization*, vol. 21, pp. 443–453.

Hiriart-Urruty, J.-B. and M. Torki (2002). "Permanently Going Back and Forth between the "Quadratic World" and the "Convexity World" in Optimization". *Applied Mathematics and Optimization*, vol. 45, pp. 169–184.

Sturm, J. F. and S. Zhang (2003). "On cones of nonnegative quadratic functions". *Mathematics of Operations Research*, vol. 28, no. 2, pp. 246–267.

Luo, Z.-Q., J. F. Sturm, and S. Zhang (2004). "Multivariate Nonnegative Quadratic Mappings". *SIAM Journal on Optimization*, vol. 14, no. 4, pp. 1140–1162.

So, A. M.-C., Y. Ye, and J. Zhang (2008). "A Unified Theorem on SDP Rank Reduction". *Mathematics of Operations Research*, vol. 33, no. 4, pp. 910–920.

Ai, W., Y. Huang, and S. Zhang (2008). "On the Low Rank Solutions for Linear Matrix Inequalities". *Mathematics of Operations Research*, vol. 33, no. 4, pp. 965–975.

Gruber, P. M. (2009). "Geometry of the cone of positive quadratic forms". *Forum Mathematicum*, vol. 21, no. 1, pp. 147–166.

Jiang, B., Z. Li, and S. Zhang (2017). "On Cones of Nonnegative Quartic Forms". *Foundations of Computational Mathematics*, vol. 17, pp. 161–197.

Exercise 27.15: Filling in the details

Prove (27.6) and (27.7).

[Hint: for the latter, apply Proposition 4.20 with $\mathbf{w} = \mathbf{w}_t$ and $\mathbf{w}_\star = \mathbf{w}_{t+1}$ in Z, noting that the Bregman divergence

$$\mathsf{D}_{(\cdot)^{3/2}}(\alpha, \beta) = \alpha^{3/2} + \tfrac{1}{2}\beta^{3/2} - \tfrac{3}{2}\alpha\beta^{1/2},$$

and at optimality $\zeta = \|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2^2$.] Indeed, the Lagrangian multiplier associated with the linear constraint $\langle I, Z\rangle = \gamma - 1$ can be shown to be $\frac{\sqrt{\gamma}}{4\eta_t}$, by solving γ and recall that γ at optimality is $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2$.

Exercise 27.16: Tighter bound under convexity

Suppose f is convex. Prove that

$$f(\mathbf{w}_t) \geq \bar{f}_t(\mathbf{w}_{t+1}) + \tfrac{1}{3\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^3,$$

which is a factor of 4 improvement compared to (27.7). Needless to say, the inequality (27.6) is now trivial.

Proposition 27.17: Sandwiching cubic regularization

Suppose f'' is $\mathsf{L} = \mathsf{L}^{[2]}$-Lipschitz continuous (w.r.t. the ℓ_2 norm). Then, the cubic regularization iterates $\{\mathbf{w}_t\}$ in (27.3) satisfy the following sandwich inequality:

$$f(\mathbf{w}_{t+1}) - \tfrac{1}{6}(\mathsf{L} - \tfrac{1}{\eta_t})\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^3 \le \bar{f}_t(\mathbf{w}_{t+1}) \le f(\mathbf{w}_t) - \tfrac{1}{12\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^3.$$

In particular,

- *if $\eta_t \le \frac{3}{2\mathsf{L}}$, then $f(\mathbf{w}_{t+1}) \le f(\mathbf{w}_t)$, i.e., cubic regularization is descending;*
- *if $\eta_t \le \frac{1}{\mathsf{L}}$, then $f(\mathbf{w}_{t+1}) \le \bar{f}_t(\mathbf{w}_{t+1}) \le f(\mathbf{w}_t)$.*

Proof: The first inequality follows from Theorem 2.13:

$$f(\mathbf{w}_{t+1}) \le f(\mathbf{w}_t) + \langle f'(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \tfrac{1}{2}\langle f''(\mathbf{w}_t)(\mathbf{w}_{t+1} - \mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \tfrac{\mathsf{L}}{6}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^3,$$

while the second inequality was already established in (27.7). ■

For comparison, gradient descent requires $\eta_t \le \frac{2}{\mathsf{L}^{[1]}}$ to guarantee descending (see Theorem 2.17).

Exercise 27.18: Bigger step size under convexity

If f is additionally convex in Proposition 27.17, then

$$f(\mathbf{w}_{t+1}) - \tfrac{1}{6}(\mathsf{L} - \tfrac{1}{\eta_t})\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^3 \le \bar{f}_t(\mathbf{w}_{t+1}) \le f(\mathbf{w}_t) - \tfrac{1}{3\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^3.$$

In particular, if $\eta_t \le \frac{3}{\mathsf{L}}$, then $f(\mathbf{w}_{t+1}) \le f(\mathbf{w}_t)$, i.e., cubic regularization is descending.

Proposition 27.19: Relating progress

Suppose f'' is $\mathsf{L} = \mathsf{L}^{[2]}$-Lipschitz continuous (w.r.t. the ℓ_2 norm). Then, the iterates $\{\mathbf{w}_t\}$ in (27.3) satisfy:

$$\|f'(\mathbf{w}_{t+1})\|_2 \le \tfrac{1}{2}(\mathsf{L} + \tfrac{1}{\eta_t})\|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2^2 \tag{27.8}$$

$$f''(\mathbf{w}_{t+1}) \succeq -(\mathsf{L} + \tfrac{1}{2\eta_t})\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \cdot \mathrm{Id}$$

$$\bar{f}_t(\mathbf{w}_{t+1}) \le \min_{\mathbf{w}}\ f(\mathbf{w}) + \tfrac{1}{6}(\mathsf{L} + \tfrac{1}{\eta_t})\|\mathbf{w} - \mathbf{w}_t\|_2^3, \tag{27.9}$$

where the right-hand side is the cubic proximal point.

Proof: For the first inequality, apply (27.4) and Theorem 2.13. For the second inequality, apply (27.6) and Lipschitz continuity of f''. The last inequality again follows from Theorem 2.13. ■

Theorem 27.20: Sublinear rate of cubic regularization (Nesterov and Polyak 2006)

Suppose f'' is $\mathsf{L} = \mathsf{L}^{[2]}$-Lipschitz continuous (w.r.t. the ℓ_2 norm) and f is bounded from below by $f_\star$. Then, assuming $\eta_t \in [0, \frac{3}{2\mathsf{L}}]$, the cubic regularization iterates $\{\mathbf{w}_t\}$ in (27.3) satisfy:

$$\sum_{t=0}^{\infty}(\tfrac{1}{4\eta_t} - \tfrac{\mathsf{L}}{6})(\tfrac{2\eta_t}{1+\eta_t\mathsf{L}})^{3/2}\|f'(\mathbf{w}_{t+1})\|_2^{3/2} \le \sum_{t=0}^{\infty}(\tfrac{1}{4\eta_t} - \tfrac{\mathsf{L}}{6})\|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2^3 \le f(\mathbf{w}_0) - f_\star.$$

In particular, if $\eta_t = \frac{1}{\mathsf{L}}$*, we have* $\sum_t \|\frac{f'(\mathbf{w}_{t+1})}{\mathsf{L}}\|_2^{3/2} \le \sum_t \|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2^3 \le \frac{12(f_0 - f_\star)}{\mathsf{L}}$.

Proof: We simply telescope the sandwich inequality in Proposition 27.19:

$$f(\mathbf{w}_0) - f_\star \ge \sum_{t=0}^{\infty} (f(\mathbf{w}_t) - f(\mathbf{w}_{t+1})) \ge \sum_{t=0}^{\infty} (\tfrac{1}{4\eta_t} - \tfrac{\mathsf{L}}{6}) \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^3.$$

When $\eta_t \le \frac{3}{2\mathsf{L}}$, we further apply (27.8). ■

Of course, when f is convex, we can replace the factor $\frac{1}{4\eta_t} - \frac{\mathsf{L}}{6}$ with $\frac{1}{2\eta_t} - \frac{\mathsf{L}}{6}$.

It follows from Proposition 27.19 that if $\mathbf{w}_*$ is a limit point of $\mathbf{w}_t$, then we must have

$$f'(\mathbf{w}_*) = 0, \qquad f''(\mathbf{w}_*) \succeq \mathbf{0}.$$

In other words, $\mathbf{w}_*$ cannot be a local maximizer (or a strict saddle, namely the Hessian has both a positive and negative eigenvalue), which is in sharp contrast to Newton's algorithm (see Alert 27.6)!

Nesterov, Y. and B. T. Polyak (2006). "Cubic regularization of Newton method and its global performance". *Mathematical Programming*, vol. 108, pp. 177–205.

Remark 27.21: Making gradient small

Theorem 27.20 implies that the (minimum) gradient of cubic regularization decays at the rate of $O(t^{-2/3})$, which is faster than gradient descent, see Theorem 2.17 and Nesterov (2012). See also (Kim and Fessler 2021; Allen-Zhu 2018; Carmon et al. 2018; Foster et al. 2019; Ito and Fukuda 2021).

Adapt to the min-max nonconvex setting through regularization? Compare to Daskalakis et al. (2021) and Fearnley et al. (2021) proved PPAD-completeness for a non-product domain: is it possible to extend to a product domain, for (regularized) gradient iterates? or maybe $poly(d/\epsilon)$ complexity bound is possible?

Nesterov, Y. (2012). "How to make the gradients small". In: *Optima*. Vol. 88, pp. 10–11.

Kim, D. and J. A. Fessler (2021). "Optimizing the Efficiency of First-Order Methods for Decreasing the Gradient of Smooth Convex Functions". *Journal of Optimization Theory and Applications*, vol. 188, pp. 192–219.

Allen-Zhu, Z. (2018). "How To Make the Gradients Small Stochastically: Even Faster Convex and Nonconvex SGD". In: *Advances in Neural Information Processing Systems*.

Carmon, Y., J. C. Duchi, O. Hinder, and A. Sidford (2018). "Accelerated Methods for NonConvex Optimization". *SIAM Journal on Optimization*, vol. 28, no. 2, pp. 1751–1772.

Foster, D. J., A. Sekhari, O. Shamir, N. Srebro, K. Sridharan, and B. Woodworth (2019). "The Complexity of Making the Gradient Small in Stochastic Convex Optimization". In: *Proceedings of the Thirty-Second Conference on Learning Theory*, pp. 1319–1345.

Ito, M. and M. Fukuda (2021). "Nearly Optimal First-Order Methods for Convex Optimization under Gradient Norm Measure: An Adaptive Regularization Approach". *Journal of Optimization Theory and Applications*, vol. 188, pp. 770–804.

Daskalakis, C., S. Skoulakis, and M. Zampetakis (2021). "The Complexity of Constrained Min-Max Optimization". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 1466–1478.

Fearnley, J., P. W. Goldberg, A. Hollender, and R. Savani (2021). "The Complexity of Gradient Descent: CLS = PPAD ∩ PLS". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 46–59.

Proposition 27.22: Quantitative behaviour around a strict saddle or local maximizer

Let $\mathbf{w}_*$ *be a strict saddle or local maximizer. Then, there exist* $\epsilon > 0$ *and* $\delta > 0$ *such that*

$$\|\mathbf{w}_t - \mathbf{w}_*\|_2 \le \epsilon \implies f(\mathbf{w}_{t+1}) \le f(\mathbf{w}_*) - \delta,$$

as long as $\eta_t \ge \underline{\eta} > 0$*, where* $\{\mathbf{w}_t\}$ *are the cubic regularization iterates in* (27.3).

Proof: Let $\mathbf{d}$ be a (normalized) direction such that $\langle f''(\mathbf{w}_*)\mathbf{d}, \mathbf{d}\rangle =: -\sigma < 0$. Applying (27.9) we have

$$\begin{aligned} f(\mathbf{w}_{t+1}) &\le f(\mathbf{w}_* + \tau\mathbf{d}) + \tfrac{1}{6}(\mathsf{L} + \tfrac{1}{\eta_t})\|\mathbf{w}_* + \tau\mathbf{d} - \mathbf{w}_t\|_2^3 \\ &\le f(\mathbf{w}_*) - \tfrac{\sigma\tau^2}{2} + \tfrac{\mathsf{L}}{6}|\tau|^3 + \tfrac{1}{6}(\mathsf{L} + \tfrac{1}{\eta_t})\big[\epsilon^2 + 2\tau\,\langle \mathbf{d}, \mathbf{w}_* - \mathbf{w}_t\rangle + \tau^2\big]^{3/2}. \end{aligned}$$

Since we are free to choose the sign of τ, upon setting $|\tau| = \epsilon$ we obtain

$$f(\mathbf{w}_{t+1}) \le f(\mathbf{w}_*) \underbrace{- \tfrac{\sigma\epsilon^2}{2} + \tfrac{\mathsf{L}}{6}\epsilon^3 + \tfrac{1}{6}(\mathsf{L} + \tfrac{1}{\eta_t})2^{3/2}\epsilon^3}_{:=-\delta},$$

where $\delta > 0$ as long as ϵ is sufficiently small and η_t remains away from 0. ■

Theorem 27.23: Local quadratic rate of cubic regularization (Nesterov and Polyak 2006)

Let $q_t := \frac{\mathsf{L}\|f'(\mathbf{w}_t)\|_2}{\sigma_t^2}$, where σ_t is the minimum eigenvalue of $f''(\mathbf{w}_t)$. Suppose f'' is $\mathsf{L} = \mathsf{L}^{[2]}$-Lipschitz continuous and $\sigma_0 > 0$. Then, the cubic regularization iterates (27.3) satisfy:

$$q_{t+1} \le \tfrac{1+\alpha}{2}\left(\tfrac{q_t}{1-q_t}\right)^2 \le \tfrac{1+\alpha}{2(1-\beta)^2}q_t^2 \le \tfrac{(1+\alpha)\beta}{2(1-\beta)^2}q_t,$$

as long as $\eta_t \ge \frac{1}{\alpha\mathsf{L}}$, $q_0 \le \beta < 1$ and $\frac{(1+\alpha)\beta}{2(1-\beta)^2} < 1$. Moreover, the gradient norm $\|f'(\mathbf{w}_t)\|_2$ converges to 0 and the iterate $\mathbf{w}_t$ converges to a local minimizer at a quadratic rate.

Proof: Using (27.6), we first note that

$$\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 = \left\|\Big[f''(\mathbf{w}_t) + \tfrac{1}{\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \cdot \mathrm{Id}\Big]^{-1} f'(\mathbf{w}_t)\right\|_2 \le \frac{\|f'(\mathbf{w}_t)\|_2}{\sigma_t + \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2/\eta_t},$$

and hence assuming $\sigma_t > 0$,

$$\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \le \frac{2\|f'(\mathbf{w}_t)\|_2}{\sqrt{\sigma_t^2 + 4\|f'(\mathbf{w}_t)\|_2/\eta_t} + \sigma_t} \le \frac{\|f'(\mathbf{w}_t)\|_2}{\sigma_t}. \tag{27.10}$$

Therefore, applying the L-Lipschitz continuity of f'':

$$\sigma_{t+1} \ge \sigma_t - \mathsf{L}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \ge \sigma_t - \frac{\mathsf{L}\|f'(\mathbf{w}_t)\|_2}{\sigma_t} = (1 - q_t)\sigma_t, \tag{27.11}$$

$$\sigma_{t+1} \le \sigma_t + \mathsf{L}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \le \sigma_t + \frac{\mathsf{L}\|f'(\mathbf{w}_t)\|_2}{\sigma_t} = (1 + q_t)\sigma_t. \tag{27.12}$$

Moreover, applying (27.8) and assuming $q_t \le 1$ we know

$$q_{t+1} := \frac{\mathsf{L}\|f'(\mathbf{w}_{t+1})\|_2}{\sigma_{t+1}^2} \le \frac{\mathsf{L}(\mathsf{L} + \frac{1}{\eta_t})\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2}{2\sigma_{t+1}^2} \le \frac{\mathsf{L}(\mathsf{L} + \frac{1}{\eta_t})\|f'(\mathbf{w}_t)\|_2^2}{2(1 - q_t)^2\sigma_t^4} = \tfrac{1}{2}\big(1 + \tfrac{1}{\eta_t\mathsf{L}}\big)\big(\tfrac{q_t}{1-q_t}\big)^2.$$

With our choice on η_t and β, $q_{t+1} \le q_t \le \beta \le 1$ and hence $\sigma_{t+1} \ge (1 - q_t)\sigma_t > 0$ are satisfied recursively.

Clearly, $\sum_t q_t < \infty$ and hence it follows from (27.11) and (27.12) that σ_t remains bounded and away from 0. Therefore, the (local) quadratic convergence rate of q_t translates to the same for the gradient $\|f'(\mathbf{w}_t)\|_2$ (by the definition of q_t) and the iterate $\{\mathbf{w}_t\}$ (due to (27.10)). ■

Setting $\eta_t = \infty$, $\alpha = 0$ and $\beta < \frac{1}{2}$ yields a similar result for the Newton's Algorithm 27.3 as Theorem 27.8. This is not surprising, after all as we commented in Algorithm 27.12, cubic regularization gradually reduces to Newton's update.

Nesterov, Y. and B. T. Polyak (2006). "Cubic regularization of Newton method and its global performance". *Mathematical Programming*, vol. 108, pp. 177–205.

Algorithm 27.24: Solving cubic regularization

Let us examine how we can solve the cubic regularization iterate (27.3). Denote $A := \begin{bmatrix} 0 & \mathbf{g}^\top \\ \mathbf{g} & H \end{bmatrix}$, we derive:

$$
\begin{aligned}
\min_{\mathbf{z}} \ 2\langle \mathbf{z}, \mathbf{g}\rangle + \langle \mathbf{z}, H\mathbf{z}\rangle + \tfrac{1}{3\eta}\|\mathbf{z}\|_2^3 \quad &= \quad \min_{Z \succeq \mathbf{0}, Z_{11}=1, \langle I, Z\rangle = \zeta+1} \ \langle A, Z\rangle + \tfrac{1}{3\eta}\zeta^{3/2} \\
&= \quad \sup_{\lambda,\mu} \ \min_{\gamma, Z \succeq \mathbf{0}} \ \mu(\langle \mathbf{e}_1\mathbf{e}_1^\top, Z\rangle - 1) + \lambda(\langle I, Z\rangle - \zeta - 1) + \langle A, Z\rangle + \tfrac{1}{3\eta}\zeta^{3/2} \\
&= \quad \sup_{A + \lambda I + \mu \mathbf{e}_1\mathbf{e}_1^\top \succeq \mathbf{0}} \ \min_{\zeta} \ -\mu - \lambda(\zeta+1) + \tfrac{1}{3\eta}\zeta^{3/2}, \\
[\sqrt{\zeta} = 2\eta\lambda_+] \quad &= \quad \sup_{\lambda \geq 0, \mu} \ -(\lambda+\mu) - \tfrac{4}{3}\eta^2\lambda^3 \quad \text{s.t.} \quad \begin{bmatrix} \lambda+\mu & \mathbf{g}^\top \\ \mathbf{g} & H + \lambda I \end{bmatrix} \succeq \mathbf{0} \\
[\text{Schur's complement}] \quad &= \quad -\inf_{\lambda \geq (-\lambda_{\min}(H))_+} \ \langle (H+\lambda I)^\dagger \mathbf{g}, \mathbf{g}\rangle + \tfrac{4}{3}\eta^2\lambda^3, \quad \text{s.t.} \quad \mathbf{g} \in \operatorname{rge}(H + \lambda I),
\end{aligned}
$$

which is a univariate convex minimization problem. Setting derivative to $\mathbf{0}$ we obtain the fixed point equation:

$$2\eta\bar{\lambda} = \|(H + \bar{\lambda} I)^\dagger \mathbf{g}\|_2.$$

The optimal λ_* is given by (see Theorem 3.4)

$$\lambda_\star = \bar{\lambda} \vee (-\lambda_{\min}(H))_+$$

and it follows from (27.4) that (recall $\sqrt{\zeta} = 2\eta\lambda_+ = \|\mathbf{z}\|_2$)

$$(H + \lambda_\star I)\mathbf{z} + \mathbf{g} = \mathbf{0}, \tag{27.13}$$

which has a unique solution if $\lambda_\star = \bar{\lambda} > (-\lambda_{\min}(H))_+$ or $\lambda_{\min}(H) > 0$. When $\lambda_\star = -\lambda_{\min}(H) \geq 0$, we pick any solution of (27.13) such that $\|\mathbf{z}\|_2 = 2\eta\lambda_\star$.

For faster algorithms, see Paternain et al. (2019), Carmon and Duchi (2020), Lieder (2020), and Jiang et al. (2021).

Paternain, S., A. Mokhtari, and A. Ribeiro (2019). "A Newton-Based Method for Nonconvex Optimization with Fast Evasion of Saddle Points". *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 343–368.

Carmon, Y. and J. C. Duchi (2020). "First-Order Methods for Nonconvex Quadratic Minimization". *SIAM Review*, vol. 62, no. 2, pp. 395–436.

Lieder, F. (2020). "Solving Large-Scale Cubic Regularization by a Generalized Eigenvalue Problem". *SIAM Journal on Optimization*, vol. 30, no. 4, pp. 3345–3358.

Jiang, R., M.-C. Yue, and Z. Zhou (2021). "An accelerated first-order method with complexity analysis for solving cubic regularization subproblems". *Computational Optimization and Applications*, vol. 79, pp. 471–506.

Example 27.25: Solving cubic regularization

Consider the following instance:

$$\mathbf{g} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \quad H = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}, \quad \eta = 1.$$

We have the primal problem

$$\min_{\mathbf{z} \in \mathbb{R}^2} \ -2z_1 - z_2^2 + \tfrac{1}{3}\|\mathbf{z}\|_2^3,$$

whose first-order necessary condition is:

$$\left.\begin{aligned} -2 \quad + \|\mathbf{z}\|_2 z_1 &= 0 \\ -2z_2 + \|\mathbf{z}\|_2 z_2 &= 0 \end{aligned}\right\} \implies \mathbf{z} = \begin{bmatrix} 1 \\ \pm\sqrt{3} \end{bmatrix}.$$

(The other possibility $\mathbf{z} = \begin{bmatrix}\sqrt{2}\\ 0\end{bmatrix}$ is a strict saddle after checking the second-order condition (27.6).)

The dual problem is

$$-\inf_{\lambda \geq 1} \lambda^{-1} + \tfrac{4}{3}\lambda^3 \implies \lambda_\star = 1.$$

Solving the linear system (27.13) we obtain $z_1 = 1$ and we find $z_2 = \pm\sqrt{3}$ so that $\|\mathbf{z}\|_2 = 2$.

Lemma 27.26: Recursive estimate (e.g. Polyak 1987, Lemma 6, p. 46)

Consider nonnegative sequences u_t and ξ_t. Let $p > 0$ and $q \in (0, 1]$. Then,

- $u_{t+1} \leq u_t(1 - \xi_t u_t^p) \implies u_{t+1} \leq u_0 \left(1 + p u_0^p \sum_{\tau=0}^t \xi_\tau\right)^{-1/p}$.
- $u_{t+1} \leq u_t(1 - \xi_t u_t^{-q}) \implies u_{t+1}^q \leq u_t^q - q\xi_t$.

Proof: For the first claim, we deduce

$$u_{t+1}^{-p} \geq u_t^{-p}(1 - \xi_t u_t^p)^{-p} \geq u_t^{-p}(1 + p\xi_t u_t^p) = u_t^{-p} + p\xi_t,$$

where we applied the convex inequality $(1 - x)^{-p} \geq 1 + px$. Telescoping completes the proof.

For the second claim, we deduce

$$u_{t+1}^q \leq u_t^q(1 - \xi_t u_t^{-q})^q \leq u_t^q(1 - q\xi_t u_t^{-q}) = u_t^q - q\xi_t,$$

where we applied the concave inequality $(1 - x)^q \leq 1 - qx$. ■

Polyak, B. T. (1987). "Introduction to Optimization". Optimization Software.

Theorem 27.27: Global sublinear rate under star convexity (Nesterov and Polyak 2006)

Suppose f is convex, f'' is $\mathsf{L} = \mathsf{L}_2^{[2]}$-Lipschitz continuous, and the (sub)level set $[\![f \leq f(\mathbf{w}_0)]\!]$ is bounded in diameter by ϱ. Then, the cubic regularization iterates (27.3) satisfy for all $t \geq 0$:

$$f(\mathbf{w}_{t+1}) - f_\star \leq \frac{f(\mathbf{w}_1) - f_\star}{\left(1 + \sqrt{f(\mathbf{w}_1) - f_\star}\sum_{\tau=1}^t \sqrt{\frac{2}{9(\mathsf{L}+1/\eta_\tau)\varrho^3}}\right)^2} \leq \frac{9\varrho^3\mathsf{L}}{2\left(\sum_{\tau=0}^t \sqrt{\frac{\eta_\tau \mathsf{L}}{1+\eta_\tau\mathsf{L}}}\right)^2},$$

provided that for all t, $\frac{1}{\eta_t} \leq 2\mathsf{L} + \frac{3}{\eta_{t+1}}$, in particular, if $\eta_{t+1} \leq 3\eta_t$, and inequality (27.14) holds, in particular if $\eta_t \leq \frac{1}{\mathsf{L}}$.

Proof: Using the condition on the step size η_t and applying (27.9) we have

$$\begin{aligned}
f(\mathbf{w}_{t+1}) - f_\star &\leq \bar{f}_t(\mathbf{w}_{t+1}) - f_\star \qquad (27.14)\\
&\leq \inf_{\mathbf{w}} f(\mathbf{w}) - f_\star + \tfrac{1}{6}(\mathsf{L} + \tfrac{1}{\eta_t})\|\mathbf{w} - \mathbf{w}_t\|_2^3\\
&\leq \min_{\beta_t \in [0,1]} f\big((1 - \beta_t)\mathbf{w}_t + \beta_t\mathbf{w}_\star\big) - f_\star + \tfrac{\beta_t^3}{6}(\mathsf{L} + \tfrac{1}{\eta_t})\|\mathbf{w}_\star - \mathbf{w}_t\|_2^3\\
&\leq \min_{\beta_t \in [0,1]} f(\mathbf{w}_t) - f_\star - \beta_t[f(\mathbf{w}_t) - f_\star] + \tfrac{\beta_t^3\varrho^3}{6}(\mathsf{L} + \tfrac{1}{\eta_t}).
\end{aligned}$$

(Note that inequality (27.14) implies that $f(\mathbf{w}_{t+1}) \leq f(\mathbf{w}_t) \leq \cdots \leq f(\mathbf{w}_0)$, i.e. $\{\mathbf{w}_t\}$ remains bounded in diameter ϱ.) Setting the derivative w.r.t. β_t to zero we obtain (see Theorem 3.4):

$$\beta_t = 1 \wedge \sqrt{\frac{2(f(\mathbf{w}_t) - f_\star)}{(\mathsf{L} + 1/\eta_t)\varrho^3}}.$$

If $\beta_t = 1$ (which we deduce below can happen only at $t = 0$), then

$$f(\mathbf{w}_{t+1}) - f_\star \leq \frac{\varrho^3(\mathsf{L} + 1/\eta_t)}{6}. \tag{27.15}$$

Provided that $\frac{1}{\eta_t} \leq 2\mathsf{L} + \frac{3}{\eta_{t+1}}$, we will have $\beta_{t+1} \leq 1$ since, and hence for all $t \geq 1$:

$$f(\mathbf{w}_{t+1}) - f_\star \leq f(\mathbf{w}_t) - f_\star - (f(\mathbf{w}_t) - f_\star)^{3/2}\sqrt{\tfrac{8}{9(\mathsf{L}+1/\eta_t)\varrho^3}}.$$

Apply Lemma 27.26 and note that $f(\mathbf{w}_1)$ satisfies (27.15). ■

For constant step size $\eta_t \equiv \eta$, the function value decreases at the rate of $O(1/t^2)$, matching that of the accelerated gradient Theorem 10.8. Moreover, the open loop step size

$$\eta_t \to 0, \quad \sum_t \sqrt{\eta_t} = \infty$$

suffices to guarantee convergence. It is, however, not possible to achieve the rate $O(1/t^2)$ with an open loop step size (at least based on our current bounds).

Nesterov, Y. and B. T. Polyak (2006). "Cubic regularization of Newton method and its global performance". *Mathematical Programming*, vol. 108, pp. 177–205.

Remark 27.28: Digesting the proof

Inspecting the proof of Theorem 27.27 carefully leads to the following observations:

- Amijo's backtracking for the step size η_t applies (see Remark 2.20), so we need only search η_t so that (27.14) holds at each iteration. Moreover, we could also increase η_t by a factor of 3 if η_t is found small.
- The function f need only be star convex w.r.t. some global minimizer, i.e., for some $\mathbf{w}_\star$ such that $f(\mathbf{w}_\star) = \inf_{\mathbf{w}} f(\mathbf{w})$, we have for all $\mathbf{w}$:

  $$f\big((1-\beta)\mathbf{w} + \beta\mathbf{w}_\star\big) \leq (1-\beta)f(\mathbf{w}) + \beta f(\mathbf{w}_\star).$$

  For example, the functions $f(w) = |w|(1 - \exp(-|w|))$ and $f(w, z) = w^2z^2 + w^2 + z^2$ are star-convex but not convex.
- There is a small cost to the above generality: Theorem 27.27 is only about the gap between $f(\mathbf{w}_t)$ and the minimum value $f_\star$ while recall that in Theorem 4.21 or Theorem 10.8 we are able to prove a convergence rate for the gap between $f(\mathbf{w}_t)$ and any $f(\mathbf{w})$.

Exercise 27.29: New gun for old battles

Can you adapt the proof of Theorem 27.27 to the gradient descent Algorithm 2.4 and establish its rate of convergence for star-convex functions?

See Hinder et al. (2020).

Hinder, O., A. Sidford, and N. Sohoni (2020). "Near-Optimal Methods for Minimizing Star-Convex Functions and Beyond". In: *Proceedings of Thirty Third Conference on Learning Theory*, pp. 1894–1938.

Theorem 27.30: Global superlinear rate under (γ, p)-growth (Nesterov and Polyak 2006)

Let F denote the (global) minimizer(s) of f and suppose f have (γ, p)-growth, i.e.,

$$f(\mathbf{w}) - f_\star \ge \tfrac{\gamma}{p} \cdot \mathrm{dist}^p(\mathbf{w}, \mathsf{F}), \quad \textit{where} \quad p \in [1, 2].$$

Suppose f is star-convex and f'' is $\mathsf{L} = \mathsf{L}_2^{[2]}$-Lipschitz continuous. If the step size $\eta_t \ge \underline{\eta} > 0$ always satisfies (27.17) (e.g. $\eta_t \le \frac{1}{\mathsf{L}}$), we have at first

$$\big(f(\mathbf{w}_{t+1}) - f_\star\big)^{(3-p)/(2p)} \le \big(f(\mathbf{w}_t) - f_\star\big)^{(3-p)/(2p)} - \tfrac{3-p}{p}\left(\tfrac{\gamma}{p}\right)^{3/(2p)} \cdot \sqrt{\tfrac{2\eta_t}{1+\eta_t \mathsf{L}}}$$

and then

$$f(\mathbf{w}_{t+1}) - f_\star \le \tfrac{1}{6}(\mathsf{L} + \tfrac{1}{\eta_t}) \left(\tfrac{p}{\gamma}\right)^{3/p} [f(\mathbf{w}_t) - f_\star]^{3/p}, \tag{27.16}$$

where the transition happens when (at the latest)

$$f(\mathbf{w}_t) - f_\star \le \left[\tfrac{2\underline{\eta}}{1+\underline{\eta}\mathsf{L}} \left(\tfrac{\gamma}{p}\right)^{3/p}\right]^{p/(3-p)}.$$

Proof: Using the condition on the step size η_t and applying (27.9) we have

$$\begin{aligned} f(\mathbf{w}_{t+1}) - f_\star &\le \bar{f}_t(\mathbf{w}_{t+1}) - f_\star \qquad (27.17)\\ &\le \inf_{\mathbf{w}} \; f(\mathbf{w}) - f_\star + \tfrac{1}{6}(\mathsf{L} + \tfrac{1}{\eta_t})\|\mathbf{w} - \mathbf{w}_t\|_2^3 \\ &\le \min_{\beta_t \in [0,1]} f\big((1-\beta_t)\mathbf{w}_t + \beta_t \mathbf{w}_\star\big) - f_\star + \tfrac{\beta_t^3}{6}(\mathsf{L} + \tfrac{1}{\eta_t})\|\mathbf{w}_\star - \mathbf{w}_t\|_2^3 \\ &\le \min_{\beta_t \in [0,1]} f(\mathbf{w}_t) - f_\star - \beta_t[f(\mathbf{w}_t) - f_\star] + \tfrac{\beta_t^3}{6}(\mathsf{L} + \tfrac{1}{\eta_t})[p(f(\mathbf{w}_t) - f_\star)/\gamma]^{3/p}. \end{aligned}$$

Setting the derivative w.r.t. β_t to zero we obtain (see Theorem 3.4):

$$\beta_t = 1 \wedge \sqrt{\tfrac{2(\frac{\gamma}{p})^{3/p}(f(\mathbf{w}_t) - f_\star)^{1-3/p}}{(\mathsf{L}+1/\eta_t)}}.$$

If $\beta_t < 1$, we have

$$f(\mathbf{w}_{t+1}) - f_\star \le f(\mathbf{w}_t) - f_\star - \sqrt{\tfrac{8(\frac{\gamma}{p})^{3/p}}{(\mathsf{L}+1/\eta_t)}}(f(\mathbf{w}_t) - f_\star)^{1-(3-p)/(2p)}.$$

Applying Lemma 27.26 we deduce

$$\big(f(\mathbf{w}_{t+1}) - f_\star\big)^{(3-p)/(2p)} \le \big(f(\mathbf{w}_t) - f_\star\big)^{(3-p)/(2p)} - \tfrac{3-p}{p} \cdot \sqrt{\tfrac{2(\frac{\gamma}{p})^{3/p}}{(\mathsf{L}+1/\eta_t)}}.$$

Since $\eta_t \ge \underline{\eta} > 0$, after a constant number of iterations, we must have $\beta_t \equiv 1$, i.e., the transition to the superlinear rate (27.16) happens. ■

In other words, after (at most) a constant number of iterations, cubic regularization settles to a superlinear rate.

For instance, σ-strongly convex functions are of $(\sigma, 2)$-growth. However, the (global) superlinear rate obtained here (for $p = 2$) is slower than the (local) quadratic rate in Theorem 27.23.

Nesterov, Y. and B. T. Polyak (2006). "Cubic regularization of Newton method and its global performance". *Mathematical Programming*, vol. 108, pp. 177–205.

Lemma 27.31: Recursive estimate II

Consider nonnegative *sequences u_t and ξ_t, where $\bar\xi \geq \xi_t \geq \underline{\xi} > 0$. Let $q \in (0, 1]$. Then,*

- *the recursion $u_t \geq u_{t+1}(1 + \xi_t u_{t+1}^q)$ implies that*

$$\begin{cases} \ln(u_t) & \leq \left(\frac{1}{q+1}\right)^t \ln(u_0 \underline{\xi}^{1/q}) - \ln(\underline{\xi}^{1/q}) \\ u_t^{-q} & \leq u_{t+1}^{-q} - \theta\xi_t, \quad \text{if } u_t \leq (\mu/\underline{\xi})^{1/q} \text{ for some } \mu > 1 \end{cases},$$

where $\theta := [1 - (1+\delta)^{-q}]/\delta$ and $\delta := \mu\bar\xi/\underline{\xi}$.

- *the recursion $u_t \geq u_{t+1}(1 + \xi_t u_{t+1}^{-q})$ implies that*

$$u_{t+1} \leq \frac{1}{1+\underline{\xi} u_0^{-q}} \cdot u_t \quad \wedge \quad \underline{\xi}^{-1/(1-q)} \cdot u_t^{1/(1-q)}.$$

Proof: For the first claim, we note that

$$u_t \geq \xi_t u_{t+1}^{q+1} \iff \ln(u_{t+1}) \leq \tfrac{1}{q+1}\ln(u_t) - \tfrac{1}{q+1}\ln(\xi_t) \implies \ln(u_t) \leq \left(\tfrac{1}{q+1}\right)^t \ln(u_0) - \sum_{\tau=0}^{t-1} \left(\tfrac{1}{q+1}\right)^{t-\tau} \ln(\xi_\tau)$$
$$\leq \left(\tfrac{1}{q+1}\right)^t \ln(u_0) - \tfrac{1}{q}\left[1 - \left(\tfrac{1}{q+1}\right)^t\right] \ln(\underline{\xi}).$$

Thus, u_t decreases below $\mu^{1/q}\underline{\xi}^{-1/q}$ for any $\mu > 1$ at a linear rate, after which we switch to the following bound:

$$u_t^{-q} \leq u_{t+1}^{-q}(1 + \xi_t u_{t+1}^q)^{-q} \leq u_{t+1}^{-q}(1 - \theta\xi_t u_{t+1}^q) = u_{t+1}^{-q} - \theta\xi_t,$$

where we applied the inequality $(1+x)^{-q} \leq 1 - \theta x$ for $x \in [0, \delta]$, with $\theta := [1 - (1+\delta)^{-q}]/\delta$ and $\delta := \mu\bar\xi/\underline{\xi}$.

The second claim is obvious once we note that $u_t \geq u_{t+1}$ is monotone. ■

Definition 27.32: (γ, p)-gradient growth (Polyak 1963)

Recall that a function is of (γ, p)-gradient growth if for all $\mathbf{w}$:

$$f(\mathbf{w}) - f_\star \leq \tfrac{\gamma}{p} \cdot \|f'(\mathbf{w})\|_2^p,$$

where $\gamma > 0$ and $p \in [1, 2]$. For instance, σ-strongly convex functions are of $(\frac{1}{\sigma}, 2)$-gradient growth.
extend PL to the composite setting

Polyak, B. T. (1963). "Gradient methods for the minimization of functionals". *USSR Computational Mathematics and Mathematical Physics*, vol. 3, no. 4, pp. 643–653.

Theorem 27.33: Global convergence rate under gradient growth (Nesterov and Polyak 2006)

Suppose f is of (γ, p)-gradient growth and f'' is $\mathsf{L} = \mathsf{L}_2^{[2]}$-Lipschitz continuous. Suppose $0 < \underline{\eta} \leq \eta_t \leq \bar\eta$ and η_t satisfies (27.18) for some $\alpha > 0$ (e.g., if $\eta_t \leq \frac{3(1-4\alpha)}{2\mathsf{L}}$). Let $q := \frac{3}{2p} - 1$ and $\xi_t := \sqrt{\frac{8\eta_t\alpha^2}{(1+\eta_t\mathsf{L})^3}} \left(\frac{p}{\gamma}\right)^{3/(2p)}$ with upper and lower bound $\bar\xi$ and $\underline{\xi}$, respectively.

- *If $p \in [1, \frac{3}{2})$, $q \in (0, \frac{1}{2}]$ and we have*

$$\begin{cases} \ln[f(\mathbf{w}_t) - f_\star] & \leq \left(\frac{1}{q+1}\right)^t \ln\left([f(\mathbf{w}_0) - f_\star]\underline{\xi}^{1/q}\right) - \ln(\underline{\xi}^{1/q}) \\ [f(\mathbf{w}_t) - f_\star]^{-q} & \leq [f(\mathbf{w}_{t+1}) - f_\star]^{-q} - \theta\xi_t, \quad \text{if } [f(\mathbf{w}_t) - f_\star] \leq (\mu/\underline{\xi})^{1/q} \text{ for some } \mu > 1 \end{cases},$$

  where $\theta := [1 - (1+\delta)^{-q}]/\delta$ and $\delta := \mu\bar{\xi}/\underline{\xi}$.

- *If $p \in [\frac{3}{2}, 2]$, $q \in [-\frac{1}{4}, 0]$ and we have*

$$[f(\mathbf{w}_{t+1}) - f_\star] \leq \tfrac{1}{1+\underline{\xi}[f(\mathbf{w}_0)-f_\star]^{-q}} \cdot [f(\mathbf{w}_t) - f_\star] \quad \wedge \quad \underline{\xi}^{-1/(1-q)} \cdot [f(\mathbf{w}_t) - f_\star]^{1/(1-q)}.$$

Proof: Applying Proposition 27.17, Proposition 27.19 and gradient growth, we have

$$\begin{aligned} f(\mathbf{w}_t) - f(\mathbf{w}_{t+1}) &\geq \tfrac{\alpha}{\eta_t}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^3 \qquad (27.18) \\ &\geq \tfrac{\alpha}{\eta_t}\left[\tfrac{2\eta_t}{1+\eta_t\mathsf{L}}\right]^{3/2} \|f'(\mathbf{w}_{t+1})\|_2^{3/2} \\ &\geq \tfrac{\alpha}{\eta_t}\left[\tfrac{2\eta_t}{1+\eta_t\mathsf{L}}\right]^{3/2} \left(\tfrac{p}{\gamma}\right)^{3/(2p)} [f(\mathbf{w}_{t+1}) - f_\star]^{3/(2p)}. \end{aligned}$$

Applying Lemma 27.31 completes the proof. ■

Thus, we see some sharp transitions in the convergence rate:

- When $p \in [1, \frac{3}{2})$, cubic regularization first converges superlinearly and then settles into the sublinear rate $O(t^{-(2p)/(3-2p)})$. In particular, when $p = 1$, we obtain the familiar rate $O(t^{-2})$.
- When $p \in [\frac{3}{2}, 2]$, cubic regularization first converges linearly and then superlinearly (with exponent $2p/3$).

For convenience, we may set $\alpha = \frac{1}{12}$, in which case we need only perform backtracking to guarantee $f(\mathbf{w}_{t+1}) \leq \bar{f}_t(\mathbf{w}_{t+1})$, see Proposition 27.17.

Nesterov and Polyak (2006) mentioned that it is possible to embed 1-gradient growth into 2-gradient growth, but the convergence rates would become worse by doing so.

Nesterov, Y. and B. T. Polyak (2006). "Cubic regularization of Newton method and its global performance". *Mathematical Programming*, vol. 108, pp. 177–205.

Exercise 27.34: (γ, p)-gradient growth

Prove the following:

- A convex function restricted to a domain of diameter ϱ is of $(\varrho, 1)$-gradient growth.
- Suppose φ is of (γ, p)-gradient growth, $(\mathbf{s}')^\top\mathbf{s}' \succeq \sigma\mathrm{Id}$ and $\inf\varphi = \inf\varphi \circ \mathbf{s}$. Then, $\varphi \circ \mathbf{s}$ is of $(\sigma^{p/2}\gamma, p)$-gradient growth.

Alert 27.35: Adaptation

We remark that the fast rates in both Theorem 27.30 and Theorem 27.33 are achieved without the knowledge of (γ, p), i.e. the step size of cubic regularization does not even depend on them!

Remark 27.36: Composition with a homeomorphism

All of our results immediately extends to the composite function $f \circ \mathbf{s}$ with the same conditions on f, provided that $\mathbf{s}$ is a homeomorphism whose inverse is 1-Lipschitz continuous:

$$\|\mathbf{w} - \mathbf{z}\|_2 \leq \|\mathbf{s}(\mathbf{w}) - \mathbf{s}(\mathbf{z})\|_2.$$

We remark that triangular maps form a natural family of homeomorphisms.

Example 27.37: Comparison with first-order algorithms

Let us now compare cubic-regularization with first-order gradient algorithms. Consider the class of σ-strongly convex functions with $\mathsf{L} = \mathsf{L}^{[2]}$-Lipschitz continuous Hessian. It follows that

$$\varrho := \inf\{\|\mathbf{w} - \mathbf{w}_\star\|_2 : f(\mathbf{w}) \leq f(\mathbf{w}_0)\} \leq \sqrt{\tfrac{2[f(\mathbf{w}_0) - f_\star]}{\sigma}}.$$

Let $p = 2$, $\gamma = \sigma$, and $\eta_t = \frac{1}{\mathsf{L}}$. We divide the progress of cubic regularization into three stages:

- Stage 1: using Theorem 27.27 we have

  $$f(\mathbf{w}_t) - f_\star \leq \tfrac{9\varrho^3 \mathsf{L}}{t^2}.$$

  Thus, after $t_1 \leq 3\sqrt{\varrho \mathsf{L}/\sigma}$ iterations we arrive at:

  $$f(\mathbf{w}_{t_1}) - f_\star \leq \sigma \varrho^2.$$

- Stage 2: using Theorem 27.30 we have

  $$\sqrt[4]{f(\mathbf{w}_{t+1}) - f_\star} \leq \sqrt[4]{f(\mathbf{w}_t) - f_\star} - \frac{1}{2}\left(\frac{\sigma}{2}\right)^{3/4} \cdot \sqrt{\frac{1}{\mathsf{L}}}.$$

  Thus, after another $t_2 \leq 2^{7/4}\sqrt{\varrho \mathsf{L}/\sigma} \leq 3.4\sqrt{\varrho \mathsf{L}/\sigma}$ iterations we arrive at:

  $$f(\mathbf{w}_{t_1+t_2}) - f_\star \leq \tfrac{\sigma^3}{8\mathsf{L}^2}.$$

- Stage 3: using Theorem 27.30 again we then have (the transition has happened)

  $$f(\mathbf{w}_{t+1}) - f_\star \leq \tfrac{\mathsf{L}}{3}\left(\tfrac{2}{\sigma}\right)^{3/2} [f(\mathbf{w}_t) - f_\star]^{3/2}.$$

  Thus, after another $t_3 \leq \log_{\frac{3}{2}} \log_9 \frac{9\sigma^3}{8\epsilon \mathsf{L}^2}$ we finally obtain

  $$f(\mathbf{w}_{t_1+t_2+t_3}) - f_\star \leq \epsilon.$$

The total number of iterations is bounded by $6.4\sqrt{\varrho \mathsf{L}/\sigma} + \log_{\frac{3}{2}} \log_9 \frac{9\sigma^3}{8\epsilon \mathsf{L}^2}$ (which is by no means optimized). In comparison, let $\mathsf{L}^{[1]} = \|f''(\mathbf{w}_\star)\|_{\mathrm{sp}}$ and we estimate

$$\sigma \cdot \mathrm{Id} \leq f''(\mathbf{w}) \leq (\mathsf{L}^{[1]} + \varrho \mathsf{L}^{[2]}) \cdot \mathrm{Id}.$$

Thus, the accelerated gradient Line 7 needs

$$O\left(\sqrt{\tfrac{\mathsf{L}^{[1]} + \varrho \mathsf{L}^{[2]}}{\sigma}} \log \tfrac{(\mathsf{L}^{[1]} + \varrho \mathsf{L}^{[2]})\varrho^2}{\epsilon}\right)$$

iterations to get an ϵ-approximate minimizer, which is substantially worse than that of cubic regularization. (For gradient descent, remove the square root to get an even worse bound.)