Optimization for Data Science

Lec 16: Newton and Cubic Regularization

Yaoliang Yu

UNIVERSITY OF WATERLOO
FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

Problem

Smooth minimization:

$$\min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w})$$

- f is a sufficiently smooth and (non)convex function
- Can high-order derivatives improver convergece?

Gradient Descent Recalled

- First-order approximation:

$$f(\mathbf{w}) \le f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t, f'(\mathbf{w}_t) \rangle + \frac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_t\|_2^2$$

- Minimize the upper bound we obtain the familiar GD:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t f'(\mathbf{w}_t)$$

- If interested in maximizing f, use GA instead:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \eta_t f'(\mathbf{w}_t)$$

- For L-smooth functions, gradient norm converges at rate $O(1/\sqrt{t})$
- For convex and L-smooth functions, function value converges at rate $O(1/t)$

Newton's Algorithm

- With 2nd order derivative, we have

$$f(\mathbf{w}) \approx f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t, f'(\mathbf{w}_t) \rangle + \frac{1}{2\eta_t} \langle \mathbf{w} - \mathbf{w}_t, f''(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t) \rangle$$

- Similarly, minimize the approximation we obtain Newton's algorithm:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t [f''(\mathbf{w}_t)]^{-1} f'(\mathbf{w}_t)$$

  - often $\eta_t \equiv 1$, at least in later stages
  - require the Hessian f'' to be nondegenerate

- Backbone of interior-point methods

Affine Equivariance

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t [f''(\mathbf{w}_t)]^{-1} f'(\mathbf{w}_t)$$

- Consider the change-of-variable $\mathbf{w} = A\mathbf{z}$ for some invertible A:

$$(f \circ A)'(\mathbf{z}) = A^\top f'(A\mathbf{z})$$
$$(f \circ A)''(\mathbf{z}) = A^\top f''(A\mathbf{z}) A$$

- Newton update is affine equivalent:

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

- How about gradient descent?

Affine Invariance

- Consider changing the inner product with a positive definite matrix Q:

$$\langle \mathbf{w}, \mathbf{z} \rangle_Q := \langle \mathbf{w}, Q\mathbf{z} \rangle$$

- Under the new inner product, we have

$$\nabla f \to Q^{-1} \nabla f, \qquad \nabla^2 f \to Q^{-1} \nabla^2 f$$

- Newton's update remains again the same

$$f(\mathbf{w}) \approx f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t, f'(\mathbf{w}_t) \rangle + \frac{1}{2\eta_t} \langle \mathbf{w} - \mathbf{w}_t, f''(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t) \rangle$$
$$f(\mathbf{w}) \leq f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t, f'(\mathbf{w}_t) \rangle + \frac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_t\|_2^2$$

Newton's Indifference

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t [f''(\mathbf{w}_t)]^{-1} f'(\mathbf{w}_t)$$

- Consider scaling f to αf for any $\alpha \in \mathbb{R} \setminus \{0\}$
- Newton's update remains the same:

$$(\alpha f)' = \alpha f', \qquad (\alpha f)'' = \alpha f''$$

- In other words, minimizing f or maximizing f yields the same Newton update!
- Newton only cares to find a root: $f'(\mathbf{w}) = 0$

Local Quadratic Convergence

Theorem:

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:

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

where $\mathbf{w}_*$ is the unique minimizer of f and $\eta_t \equiv 1$.

- f is σ-strongly convex if $f'' \succeq \sigma \cdot \mathrm{Id}$
- f'' is L-Lipschitz continuous if $\|f'''\| \leq \mathsf{L}$
- $q < 1$ if initializer $\mathbf{w}_0$ is close to $\mathbf{w}_*$, i.e. $\|f'(\mathbf{w}_0)\|_2 < \frac{2\sigma^2}{\mathsf{L}}$

- L-Lipschitz continuity of f'' implies that

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

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

$$\begin{aligned}\|f'(\mathbf{w}_{t+1})\|_2 \leq \frac{\mathsf{L}}{2}\|[f''(\mathbf{w}_t)]^{-1} f'(\mathbf{w}_t)\|_2^2 &\leq \frac{\mathsf{L}}{2}\|[f''(\mathbf{w}_t)]^{-1}\|_{\mathrm{sp}}^2 \cdot \|f'(\mathbf{w}_t)\|_2^2 \\ &\leq \frac{\mathsf{L}}{2\sigma^2}\|f'(\mathbf{w}_t)\|_2^2\end{aligned}$$

- Therefore, telescoping yields for $t \geq 0$:

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

- Lastly, it follows from the strong convexity of f that

$$\|f'(\mathbf{w}_t)\|_2 = \|f'(\mathbf{w}_t) - f'(\mathbf{w}_*)\|_2 \geq \sigma\|\mathbf{w}_t - \mathbf{w}_*\|_2$$

Example: Newton may NOT converge faster than linearly

Let us consider the simple univariate function

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

- Clearly, we have

$$f'(w) = \frac{5}{2}\operatorname{sign}(w)|w|^{3/2}, \qquad f''(w) = \frac{15}{4}|w|^{1/2}$$

- f'' is not Lipschitz continuous and f is not strongly convex
- The Newton update is:

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

- Converges to 0, the unique minimizer, at a linear rate.

Example: Newton may cycle

Consider the simple univariate function

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

- Around 0, f is locally (strongly) convex and f'' is locally Lipschitz continuous
- The Newton update is:

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

$$1 \underset{t \leftarrow t+1}{\rightleftarrows} -1$$

- With $w_0 = 1$ we enter a cycle:
- 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: Newton can be chaotic

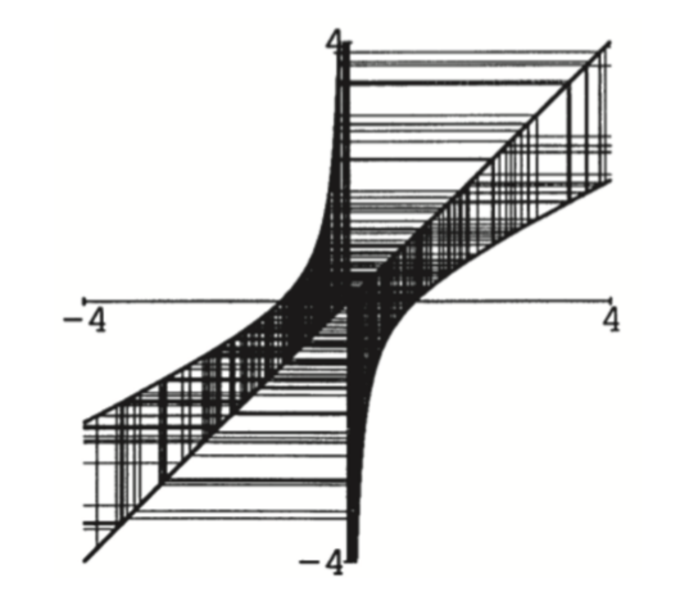
Consider the simple univariate function

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

- 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:

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

- $f' > 0$ hence Newton cannot find any root and goes crazy...
- Fixec point of the Newton update is $w^2 = -1$, i.e. $w = \pm\imath$


4
−4
4
−4

Dealing with Degeneracy

$$f(\mathbf{w}) \approx f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t, f'(\mathbf{w}_t) \rangle + \frac{1}{2\eta_t} \langle \mathbf{w} - \mathbf{w}_t, f''(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t) \rangle$$
$$f(\mathbf{w}) \leq f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t, f'(\mathbf{w}_t) \rangle + \frac{1}{2\eta_t} \|\mathbf{w} - \mathbf{w}_t\|_2^2$$

- Levenberg-Marquardt Regularization:

$$\min_{\mathbf{w}} \; f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t, f'(\mathbf{w}_t) \rangle + \frac{1}{2\eta_t} \langle \mathbf{w} - \mathbf{w}_t, f''(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t) \rangle + \frac{\alpha_t}{2\eta_t} \|\mathbf{w} - \mathbf{w}_t\|_2^2$$

- Interpolation between ideas:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \cdot [f''(\mathbf{w}_t) + \alpha_t \mathrm{Id}]^{-1} f'(\mathbf{w}_t)$$

  - $\alpha_t \to 0$: Newton's update
  - $\alpha_t \to \infty$: gradietn descent (upon normalization)

K. Levenberg. "A method for the solution of certain non-linear problems in least squares". *Quarterly of Applied Mathematics*, vol. 2, no. 2 (1944), pp. 164–168, D. W. Marquardt. "An Algorithm for Least-Squares Estimation of Nonlinear Parameters". *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, no. 2 (1963), pp. 431–441.

Cubic Regularization

$$\underbrace{f(\mathbf{w}_t) + \langle f'(\mathbf{w}_t), \mathbf{w} - \mathbf{w}_t \rangle + \frac{1}{2} \langle f''(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t), \mathbf{w} - \mathbf{w}_t \rangle + \frac{1}{6\eta_t} \|\mathbf{w} - \mathbf{w}_t\|_2^3}_{\bar{f}_t(\mathbf{w}) = \bar{f}_{\eta_t}(\mathbf{w}; \mathbf{w}_t)}$$

- Setting derivative to zero:

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

- Implicit update: $\mathbf{w}_{t+1} = \mathbf{w}_t - \left[f''(\mathbf{w}_t) + \frac{1}{2\eta_t} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \cdot \mathrm{Id} \right]^{-1} f'(\mathbf{w}_t)$
- Essentially Newton's update with adaptive Levenberg-Marquardt regularization
- Since $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 \to 0$ (hopefully), cubic regularization eventually behaves similarly to Newton's update

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

Convergence Guarantee

Theorem:

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

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

- If $\eta_t = \frac{1}{\mathsf{L}}$, we have $\sum_t \|\frac{f'(\mathbf{w}_{t+1})}{\mathsf{L}}\|_2^{3/2} \le \sum_t \|\mathbf{w}_t - \mathbf{w}_{t+1}\|_2^3 \le \frac{12(f_0 - f_\star)}{\mathsf{L}}$
- Gradient norm $\min_t \|f'(\mathbf{w}_t)\|_2$ converges to 0 at rate $O(t^{-2/3})$
- Descending, hence cannot converge to a local maxima or saddle point!

Theorem:

Suppose f is (star) convex, f'' is L-Lipschitz continuous, and the (sub)level set $[\![f \le f(\mathbf{w}_0)]\!]$ is bounded in diameter by ϱ. Then, the cubic regularization iterates satisfy:

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

provided that for all t, $\eta_{t+1} \le 3\eta_t$ and $\eta_t \le \frac{1}{\mathsf{L}}$.

- For constant step size (say) $\eta_t \equiv \frac{1}{\mathsf{L}}$, $f(\mathbf{w}_t) - f_\star \le \frac{9\varrho^3\mathsf{L}}{t^2}$
- Matches the rate of accelerated gradient; can be further accelerated
- Converges for open loop step size: $\eta_t \to 0$ and $\sum_t \sqrt{\eta_t} = \infty$

- Consider σ-strongly convex functions with L-Lipschitz continuous Hessian
- It follows that $\varrho := \inf\{\|\mathbf{w} - \mathbf{w}_\star\|_2 : f(\mathbf{w}) \le f(\mathbf{w}_0)\} \le \sqrt{\frac{2[f(\mathbf{w}_0) - f_\star]}{\sigma}}$
- We divide the progress of cubic regularization into three stages
- Stage 1: we have

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

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

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

- Stage 2: we have

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

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

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

- Stage 3: we have (the transition has happened)

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

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

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

- The total number of iterations is bounded by $6.4\sqrt{\varrho \mathsf{L}/\sigma} + \log_{\frac{3}{2}} \log_9 \frac{9\sigma^3}{8\epsilon \mathsf{L}^2}$
- In comparison, let $\mathsf{L}^{[1]} = \|f''(\mathbf{w}_\star)\|_{\text{sp}}$ and we estimate

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

- Thus, the accelerated gradient algorithm needs

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

  iterations to get an ϵ-approximate minimizer, which is substantially worse