Second-Order Optimization

Optimization Techniques (ENGG*6140)

School of Engineering,
University of Guelph, ON, Canada

Course Instructor: Benyamin Ghojogh
Winter 2023

Newton's Method

Newton-Raphson root finding method

$$\frac{\tanh(x)}{\sqrt{1+\sinh(x)}}\, x^2 + 3x = 0$$

- We can find the root of a function $f : \boldsymbol{x} \mapsto f(\boldsymbol{x})$ by solving the equation $f(\boldsymbol{x}) \overset{\text{set}}{=} 0$.
- The root of function can be found iteratively where we get closer to the root over iterations.
- One of the iterative root-finding methods is the **Newton-Raphson method** [1]. In every iteration, it finds the next solution as:

$$\boldsymbol{x}^{(k+1)} := \boldsymbol{x}^{(k)} - \frac{f(\boldsymbol{x}^{(k)})}{\nabla f(\boldsymbol{x}^{(k)})}, \tag{1}$$

where $\nabla f(\boldsymbol{x}^{(k)})$ is the derivative of function w.r.t. $\boldsymbol{x}$.

Univariate Newton's method

- Recall Eq. (1): We saw that Newton-Raphson method solves $f(\boldsymbol{x}) \overset{\text{set}}{=} 0$ by:

$$\boldsymbol{x}^{(k+1)} := \boldsymbol{x}^{(k)} - \frac{f(\boldsymbol{x}^{(k)})}{\nabla f(\boldsymbol{x}^{(k)})}.$$

- In unconstrained optimization, we can find the extremum (minimum or maximum) of the function by setting its derivative to zero, i.e.,

$$\nabla f(\boldsymbol{x}) \overset{\text{set}}{=} 0.$$

- Therefore, the root of $\nabla f(\boldsymbol{x}) \overset{\text{set}}{=} 0$ can be found by Newton-Raphson method. We replace $f(\boldsymbol{x})$ with $\nabla f(\boldsymbol{x})$ in Eq. (1):

$$\boldsymbol{x}^{(k+1)} := \boldsymbol{x}^{(k)} - \eta^{(k)} \frac{\nabla f(\boldsymbol{x}^{(k)})}{\nabla^2 f(\boldsymbol{x}^{(k)})}, \tag{2}$$

where $\nabla^2 f(\boldsymbol{x}^{(k)})$ is the second derivative of function w.r.t. $\boldsymbol{x}$ and we have included a step size at iteration k denoted by $\eta^{(k)} > 0$. This step size can be either fixed or adaptive.

Multivariate Newton's method

- Recall Eq. (2):

$$\boldsymbol{x}^{(k+1)} := \boldsymbol{x}^{(k)} - \eta^{(k)} \frac{\nabla f(\boldsymbol{x}^{(k)})}{\nabla^2 f(\boldsymbol{x}^{(k)})}.$$

- If $\boldsymbol{x}$ is multivariate, i.e. $\boldsymbol{x} \in \mathbb{R}^d$, Eq. (2) is written as:

$$\boldsymbol{x}^{(k+1)} := \boldsymbol{x}^{(k)} - \eta^{(k)} \left(\nabla^2 f(\boldsymbol{x}^{(k)})\right)^{-1} \nabla f(\boldsymbol{x}^{(k)}), \tag{3}$$

  where $\nabla f(\boldsymbol{x}^{(k)}) \in \mathbb{R}^d$ is the gradient of function w.r.t. $\boldsymbol{x}$ and $\nabla^2 f(\boldsymbol{x}^{(k)}) \in \mathbb{R}^{d \times d}$ is the Hessian matrix w.r.t. $\boldsymbol{x}$.

- Because of the second derivative or the Hessian, this optimization method is a second-order method. The name of this method is the **Newton's method**.

Newton's Method for Unconstrained Optimization

Newton's Method for Unconstrained Optimization

- Consider the following optimization problem:

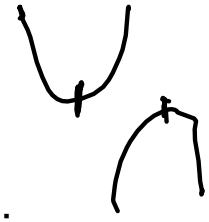
$$\underset{\boldsymbol{x}}{\text{minimize}} \quad f(\boldsymbol{x}). \tag{4}$$

  where $f(.)$ is a convex function.
- Iterative optimization can be first-order or second-order. Iterative optimization updates solution iteratively:

$$\boldsymbol{x}^{(k+1)} := \boldsymbol{x}^{(k)} + \Delta \boldsymbol{x}, \tag{5}$$

  The update continues until $\Delta \boldsymbol{x}$ becomes very small which is the convergence of optimization.
- Recall that in the first-order optimization, the step of updating is $\Delta \boldsymbol{x} := -\nabla f(\boldsymbol{x})$.

Newton's Method for Unconstrained Optimization



- Near the optimal point $\boldsymbol{x}^*$, gradient is very small so the second-order Taylor series expansion of function becomes:

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}^*) + \underbrace{\nabla f(\boldsymbol{x}^*)^\top}_{\approx 0}(\boldsymbol{x} - \boldsymbol{x}^*) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^\top \nabla^2 f(\boldsymbol{x}^*)(\boldsymbol{x} - \boldsymbol{x}^*)$$

$$\approx f(\boldsymbol{x}^*) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^*)^\top \nabla^2 f(\boldsymbol{x}^*)(\boldsymbol{x} - \boldsymbol{x}^*). \tag{6}$$

  This shows that the function is almost quadratic near the optimal point.

- Following this intuition, Newton's method uses Hessian $\nabla^2 f(\boldsymbol{x})$ in its updating step:

$$\Delta \boldsymbol{x} := -\nabla^2 f(\boldsymbol{x})^{-1} \nabla f(\boldsymbol{x}). \tag{7}$$

- In the literature, this equation is sometimes restated to:

$$\nabla^2 f(\boldsymbol{x})\, \Delta \boldsymbol{x} := -\nabla f(\boldsymbol{x}). \tag{8}$$

Newton's Method for Equality Constrained Optimization

Newton's method for equality constrained optimization

- The optimization problem may have equality constraints:

$$\begin{aligned} &\underset{\boldsymbol{x}}{\text{minimize}} && f(\boldsymbol{x}) \\ &\text{subject to} && \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \end{aligned} \tag{9}$$

- After a step of update by $\boldsymbol{p} = \Delta\boldsymbol{x}$, this optimization becomes:

$$\begin{aligned} &\underset{\boldsymbol{x}}{\text{minimize}} && f(\boldsymbol{x}+\boldsymbol{p}) \\ &\text{subject to} && \boldsymbol{A}(\boldsymbol{x}+\boldsymbol{p}) = \boldsymbol{b}. \end{aligned} \tag{10}$$

- The Lagrangian of this optimization problem is:

$$\mathcal{L} = f(\boldsymbol{x}+\boldsymbol{p}) + \boldsymbol{\nu}^\top(\boldsymbol{A}(\boldsymbol{x}+\boldsymbol{p}) - \boldsymbol{b}),$$

  where $\boldsymbol{\nu}$ is the dual variable.

- The second-order Taylor series expansion of function $f(\boldsymbol{x}+\boldsymbol{p})$ is:

$$f(\boldsymbol{x}+\boldsymbol{p}) \approx f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^\top \nabla^2 f(\boldsymbol{x})\, \boldsymbol{p}. \tag{11}$$

- Substituting this into the Lagrangian gives:

$$\mathcal{L} = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^\top \nabla^2 f(\boldsymbol{x})\, \boldsymbol{p} + \boldsymbol{\nu}^\top(\boldsymbol{A}(\boldsymbol{x}+\boldsymbol{p}) - \boldsymbol{b}).$$

Newton's method for equality constrained optimization

- We found:

$$\mathcal{L} = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^\top \nabla^2 f(\boldsymbol{x})\, \boldsymbol{p} + \boldsymbol{\nu}^\top (\boldsymbol{A}(\boldsymbol{x}+\boldsymbol{p}) - \boldsymbol{b}).$$

- According to KKT conditions, the primal and dual residuals must be zero:

$$\nabla_{\boldsymbol{x}} \mathcal{L} = \nabla f(\boldsymbol{x}) + \nabla^2 f(\boldsymbol{x})^\top \boldsymbol{p} + \boldsymbol{p}^\top \underbrace{\nabla^3 f(\boldsymbol{x})}_{\approx \mathbf{0}}\, \boldsymbol{p} + \boldsymbol{A}^\top \boldsymbol{\nu} \overset{\text{set}}{=} \mathbf{0}$$

$$\Longrightarrow \nabla^2 f(\boldsymbol{x})^\top \boldsymbol{p} + \boldsymbol{A}^\top \boldsymbol{\nu} = -\nabla f(\boldsymbol{x}), \tag{12}$$

$$\nabla_{\boldsymbol{\nu}} \mathcal{L} = \boldsymbol{A}(\boldsymbol{x}+\boldsymbol{p}) - \boldsymbol{b} \overset{(a)}{=} \boldsymbol{A}\boldsymbol{p} \overset{\text{set}}{=} \mathbf{0}, \tag{13}$$

where we have $\nabla^3 f(\boldsymbol{x}) \approx 0$ because the third-order gradient is usually very small compared to the first and second gradients and (a) is because of the constraint $\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} = \mathbf{0}$ in problem (9).

- Eqs. (12) and (13) can be written as a system of equations:

$$\begin{bmatrix} \nabla^2 f(\boldsymbol{x})^\top & \boldsymbol{A}^\top \\ \boldsymbol{A} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{p} \\ \boldsymbol{\nu} \end{bmatrix} = \begin{bmatrix} -\nabla f(\boldsymbol{x}) \\ \mathbf{0} \end{bmatrix}. \tag{14}$$

Solving this system of equations gives the desired step $\boldsymbol{p}$ (i.e., $\Delta \boldsymbol{x}$) for updating the solution at the iteration.

Starting with non-feasible initial point

- Newton's method can even start with a non-feasible point which does not satisfy all the constraints.
- If the initial point for optimization is not a feasible point, i.e., $\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \neq \boldsymbol{0}$, Eq. (13) becomes:

$$\nabla_{\nu}\mathcal{L} = \boldsymbol{A}(\boldsymbol{x} + \boldsymbol{p}) - \boldsymbol{b} \overset{\text{set}}{=} \boldsymbol{0} \implies \boldsymbol{A}\boldsymbol{p} = -(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}). \tag{15}$$

- Therefore, for the first iteration, we solve the following system rather than Eq. (14):

$$\begin{bmatrix} \nabla^2 f(\boldsymbol{x})^{\top} & \boldsymbol{A}^{\top} \\ \boldsymbol{A} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{p} \\ \boldsymbol{\nu} \end{bmatrix} = - \begin{bmatrix} \nabla f(\boldsymbol{x}) \\ \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \end{bmatrix}, \tag{16}$$

and we use Eq. (16) for the rest of iterations because the next points will be in the feasibility set (because we force the solutions to satisfy $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$).

Interior-Point and Barrier Methods: Newton's Method for Inequality Constrained Optimization

Interior-Point Method

- The optimization problem may have inequality constraints:

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & y_i(\boldsymbol{x}) \leq 0, \quad i \in \{1, \ldots, m_1\}, \\ & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \end{aligned} \tag{17}$$

- We can solve constrained optimization problems using **Barrier methods**, also known as **interior-point methods** [2, 3, 4, 5].
- Interior-point methods were first proposed in 1967 [6].
- The interior-point method is also referred to as the **Unconstrained Minimization Technique (UMT)** or **Sequential UMT (SUMT)** [7] because it converts the problem to an unconstrained problem and solves it iteratively.

Interior-Point Method

- The barrier methods or the interior-point methods, convert inequality constrained problems to equality constrained or unconstrained problems. Ideally, we can do this conversion using the indicator function $\mathbb{I}(.)$ which is zero if its input condition is satisfied and is infinity otherwise:

$$\mathbb{I}(\boldsymbol{x} \in \mathcal{S}) = \begin{cases} 0 & \text{if } \boldsymbol{x} \in \mathcal{S} \\ \infty & \text{if } \boldsymbol{x} \notin \mathcal{S}. \end{cases} \tag{18}$$

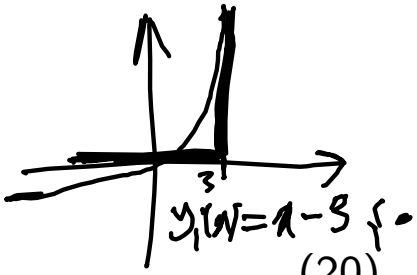
- The problem

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & y_i(\boldsymbol{x}) \leq 0, \quad i \in \{1, \ldots, m_1\}, \\ & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \end{aligned}$$

is converted to:

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & f(\boldsymbol{x}) + \sum_{i=1}^{m_1} \mathbb{I}(y_i(\boldsymbol{x}) \leq 0) \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \end{aligned} \tag{19}$$

Interior-Point Method



- The indicator function is not differentiable because it is not smooth:

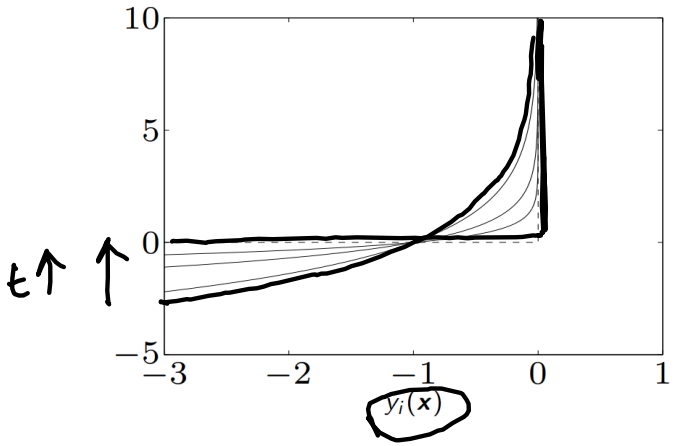
$$\mathbb{I}(y_i(\boldsymbol{x}) \leq 0) := \begin{cases} 0 & \text{if } y_i(\boldsymbol{x}) \leq 0 \\ \infty & \text{if } y_i(\boldsymbol{x}) > 0. \end{cases} \tag{20}$$

  Hence, we can approximate it with differentiable functions called the **barrier functions** [4, 8].

- One of the barrier functions is logarithm, named the **logarithmic barrier** or **log barrier** in short. It approximates the indicator function by:

$$\mathbb{I}(y_i(\boldsymbol{x}) \leq 0) \approx -\frac{1}{t} \log(-y_i(\boldsymbol{x})), \tag{21}$$

  where $t > 0$ (usually a large number such as $t = 10^6$) and the approximation becomes more accurate by $t \to \infty$.



Interior-Point Method

- The problem had become:

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & f(\boldsymbol{x}) + \sum_{i=1}^{m_1} \mathbb{I}(y_i(\boldsymbol{x}) \leq 0) \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \end{aligned}$$

- The log barrier:

$$\mathbb{I}(y_i(\boldsymbol{x}) \leq 0) \approx -\frac{1}{t} \log(-y_i(\boldsymbol{x})).$$

- It changes the problem to:

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & f(\boldsymbol{x}) - \frac{1}{t} \sum_{i=1}^{m_1} \log(-y_i(\boldsymbol{x})) \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \end{aligned} \tag{22}$$

This optimization problem is an equality constrained optimization problem which we already explained how to solve.

- Note that there exist many approximations for the barrier. One of mostly used methods is the logarithmic barrier.

Interior-Point Method

- If the problem is convex, the iterative solutions of the interior-point method satisfy:

$$\begin{aligned} &\{\boldsymbol{x}^{(0)}, \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots\} \to \boldsymbol{x}^*, \\ &\{\boldsymbol{\nu}^{(0)}, \boldsymbol{\nu}^{(1)}, \boldsymbol{\nu}^{(2)}, \dots\} \to \boldsymbol{\nu}^*, \\ &f(\boldsymbol{x}^{(0)}) \geq f(\boldsymbol{x}^{(1)}) \geq f(\boldsymbol{x}^{(2)}) \geq \cdots \geq f(\boldsymbol{x}^*), \\ &g(\boldsymbol{\nu}^{(0)}) \leq g(\boldsymbol{\nu}^{(1)}) \leq \cdots \leq g(\boldsymbol{\nu}^*). \end{aligned} \tag{23}$$

- If the optimization problem is a convex problem, the solution of interior-point method is the global solution; otherwise, the solution is local.
- The interior-point and barrier methods are used in many optimization toolboxes such as CVX [9].

Accuracy of the log barrier method

Theorem (On the sub-optimality of log-barrier method)

Let the optimum of problems (17) and (22) be denoted by f^* and f_r^*, respectively. We have:

$$f^* - \frac{m_1}{t} \leq f_r^* \leq f^*, \tag{24}$$

meaning that the optimum of problem (22) is no more than m_1/t from the optimum of problem (17).

Proof.

See our tutorial [10] for proof. Also explained in the next slide. □

- The above theorem indicates that by $t \to \infty$, the log-barrier method is more accurate; i.e., the solution of problem (22) is more accurately close to the solution of problem (17).
- This is expected since the approximation in Eq. (21) gets more accurate by increasing t.
- Note that by increasing t, optimization gets more accurate but harder to solve and slower to converge.

Acknowledgement

- Some slides of this slide deck are inspired by the lectures of Prof. Stephen Boyd at the Stanford University.
- Our tutorial also has the materials of this slide deck: [10]

References

[1] J. Stoer and R. Bulirsch, *Introduction to numerical analysis*, vol. 12. Springer Science & Business Media, 2013.

[2] Y. Nesterov and A. Nemirovskii, *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

[3] F. A. Potra and S. J. Wright, "Interior-point methods," *Journal of computational and applied mathematics*, vol. 124, no. 1-2, pp. 281–302, 2000.

[4] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[5] M. Wright, "The interior-point revolution in optimization: history, recent developments, and lasting consequences," *Bulletin of the American mathematical society*, vol. 42, no. 1, pp. 39–56, 2005.

[6] I. Dikin, "Iterative solution of problems of linear and quadratic programming," in *Doklady Akademii Nauk*, vol. 174, pp. 747–748, Russian Academy of Sciences, 1967.

[7] A. V. Fiacco and G. P. McCormick, "The sequential unconstrained minimization technique (SUMT) without parameters," *Operations Research*, vol. 15, no. 5, pp. 820–827, 1967.

[8] Y. Nesterov, *Lectures on convex optimization*, vol. 137. Springer, 2018.

References (cont.)

[9] M. Grant, S. Boyd, and Y. Ye, "CVX: Matlab software for disciplined convex programming," 2009.

[10] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, "KKT conditions, first-order and second-order optimization, and distributed optimization: Tutorial and survey," *arXiv preprint arXiv:2110.01858*, 2021.