CS 4501: Optimization

Shangtong Zhang SHANGTONG@VIRGINIA.EDU
Department of Computer Science
University of Virginia

Roadmap

1. What's Optimization?

$$\min_\theta f(\theta)$$

A typical setting in AI and ML is that f is parameterized by θ.

- Regression, e.g., x_i is a zip code, y_i is house price

$$f(\theta) \doteq \sum_{i=1}^{N} (g_\theta(x_i) - y_i)^2$$

- Classification, e.g., x_i is an image, y_i is a probability distribution over possible objects, cat or dog?

$$f(\theta) \doteq \sum_{i=1}^{N} L(g_\theta(x_i), y_i)$$

- Unsupervised learning, e.g., Large Language Model (LLM)

$$f(\theta) \doteq \sum_{i=1}^{N} L(h_{\theta_2}(g_{\theta_1}(x_i)), x_i)$$

- Reinforcement learning, e.g., AlphaGo

$$f(\theta) \doteq \mathbb{E}\left[\sum_{t=1}^{T} r(S_t, \pi_\theta(S_t))\right]$$

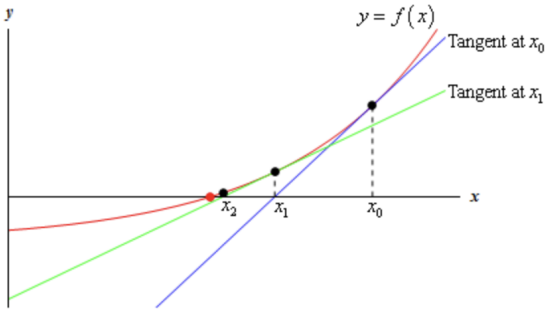
2. Warmup with Newton's Method

Find some x_* such that $f(x_*) = 0$ with Newton's Method.
How is Newton's method related to optimization?

Let's suppose that we want to approximate the solution to $f(x) = 0$ and let's also suppose that we have somehow found an initial approximation to this solution say, x_0. This initial approximation is probably not all that good, in fact it may be nothing more than a quick guess we made, and so we'd like to find a better approximation. This is easy enough to do. First, we will get the tangent line to$f(x)$at x_0.

$$y = f(x_0) + f'(x_0)(x - x_0)$$

Now, take a look at the graph below.



The blue line (if you're reading this in color anyway…) is the tangent line at x_0. We can see that this line will cross the x-axis much closer to the actual solution to the equation than x_0 does. Let's call this point where the tangent at x_0 crosses the x-axis x_1 and we'll use this point as our new approximation to the solution.

So, how do we find this point? Well we know it's coordinates, $(x_1, 0)$, and we know that it's on the tangent line so plug this point into the tangent line and solve for x_1 as follows,

$$\begin{aligned} 0 &= f(x_0) + f'(x_0)(x_1 - x_0) \\ x_1 - x_0 &= -\frac{f(x_0)}{f'(x_0)} \\ x_1 &= x_0 - \frac{f(x_0)}{f'(x_0)} \end{aligned}$$

So, we can find the new approximation provided the derivative isn't zero at the original approximation.

Now we repeat the whole process to find an even better approximation. We form up the tangent line to $f(x)$at x_1 and use its root, which we'll call x_2, as a new approximation to the actual solution. If we do this we will arrive at the following formula.

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

This point is also shown on the graph above and we can see from this graph that if we continue following this process will get a sequence of numbers that are getting very close the actual solution. This process is called Newton's Method.

Figure 1: Intuition behind Newton's method (source)

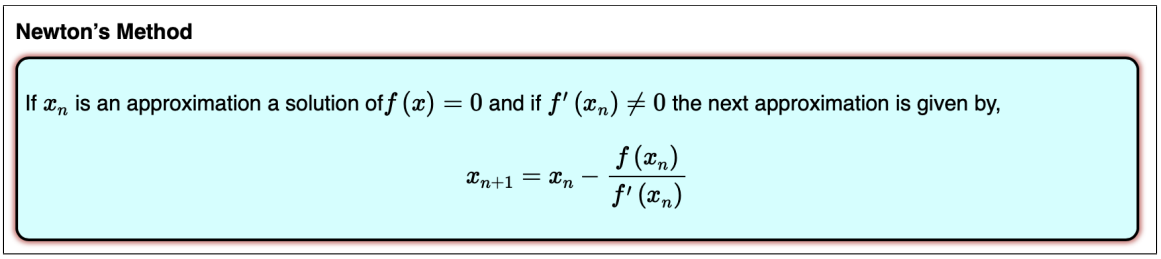


Figure 2: Newton's method (source)

Newton's method applied to the optimization problem

$$\min_{\theta} f(\theta)$$

$$\theta_{n+1} = \theta_n - \frac{f'(\theta_n)}{f''(\theta_n)}$$

Proof of quadratic convergence for Newton's iterative method [edit]

According to Taylor's theorem, any function $f(x)$ which has a continuous second derivative can be represented by an expansion about a point that is close to a root of $f(x)$. Suppose this root is α. Then the expansion of $f(\alpha)$ about x_n is:

$$f(\alpha) = f(x_n) + f'(x_n)(\alpha - x_n) + R_1 \quad (1)$$

where the Lagrange form of the Taylor series expansion remainder is

$$R_1 = \frac{1}{2!} f''(\xi_n)(\alpha - x_n)^2 ,$$

where ξ_n is in between x_n and α.

Since α is the root, (1) becomes:

$$0 = f(\alpha) = f(x_n) + f'(x_n)(\alpha - x_n) + \tfrac{1}{2} f''(\xi_n)(\alpha - x_n)^2 \quad (2)$$

Dividing equation (2) by $f'(x_n)$ and rearranging gives

$$\frac{f(x_n)}{f'(x_n)} + (\alpha - x_n) = \frac{-f''(\xi_n)}{2f'(x_n)}(\alpha - x_n)^2 \quad (3)$$

Remembering that x_{n+1} is defined by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} , \quad (4)$$

one finds that

$$\underbrace{\alpha - x_{n+1}}_{\varepsilon_{n+1}} = \frac{-f''(\xi_n)}{2f'(x_n)} (\underbrace{\alpha - x_n}_{\varepsilon_n})^2 .$$

That is,

$$\varepsilon_{n+1} = \frac{-f''(\xi_n)}{2f'(x_n)} \cdot \varepsilon_n^2 . \quad (5)$$

Taking the absolute value of both sides gives

$$|\varepsilon_{n+1}| = \frac{|f''(\xi_n)|}{2|f'(x_n)|} \cdot \varepsilon_n^2 . \quad (6)$$

Equation (6) shows that the order of convergence is at least quadratic if the following conditions are satisfied:

1. $f'(x) \neq 0$; for all $x \in I$, where I is the interval $[\alpha - |\varepsilon_0|, \alpha + |\varepsilon_0|]$;
2. $f''(x)$ is continuous, for all $x \in I$;
3. $M|\varepsilon_0| < 1$

where M is given by

$$M = \frac{1}{2} \left(\sup_{x \in I} |f''(x)| \right) \left(\sup_{x \in I} \frac{1}{|f'(x)|} \right) .$$

If these conditions hold,

$$|\varepsilon_{n+1}| \leq M \cdot \varepsilon_n^2 .$$

Figure 3: Convergence proof (source)

What if θ is not a scalar?

- Computing Hessian is expensive → first-order methods
- Write things in a compact way → linear algebra

3. Background for Projected Gradient Descent

- extended real function
- motivation of linear algebra (Section 3.1 of Gallier and Quaintance (2019))
- vector space (Section 3.2 of Gallier and Quaintance (2019))
- linear independence, subspaces (Section 3.4 of Gallier and Quaintance (2019))
- norms, inner product
- matrices (Section 3.6 of Gallier and Quaintance (2019))
- Euclidean space
- linear transformation
- extended real function
- convex set
- convex function
- differentiability
- closedness
- projection into convex set
- Lipschitz continuity
- orthogonal projections (Section 6.4)

4. Projected Gradient Descent

Problem:

$$\min_{x \in C} f(x)$$

Assumption 4.1 *f is strictly convex and differentiable at every $x \in C$*

Assumption 4.2 *C is closed, and convex*

Assumption 4.3 *f is Lipschitz continous with a Lipschitz constant L_f*

Denote the optimal solution as x_* and the optimal value as f_*. Define

$$f_{*,k} \doteq \min_{k=0,\dots,k} f(x_k).$$

Algorithm:

$$x_{k+1} = P_C\left(x_k - t_k f'(x_k)\right)$$

Lemma 1 (fundamental inequality)

$$\|x_{k+1} - x_*\|^2 \le \|x_k - x_*\|^2 - 2t_k\left(f(x_k) - f_*\right) + t_k^2\left\|f'(x_k)\right\|^2$$

Proof

$$\begin{aligned}
\|x_{k+1} - x_*\|^2 \le& \left\|P_C(x_k - t_k f'(x_k)) - P_C(x_*)\right\|^2 \\
\le& \left\|x_k - t_k f'(x_k) - x_*\right\|^2 \\
\le& \|x_k - x_*\|^2 - 2t_k\left\langle f'(x_k), x_k - x_*\right\rangle + t_k^2\left\|f'(x_k)\right\|^2 \\
\le& \|x_k - x_*\|^2 - 2t_k(f(x_k) - f_*) + t_k^2\left\|f'(x_k)\right\|^2
\end{aligned}$$

■

Polyak's step size:

$$t_k \doteq \begin{cases} \frac{f(x_k) - f_*}{\|f'(x_k)\|^2}, & f'(x_k) \neq 0, \\ 1, & f'(x_k) = 0 \end{cases}$$

Theorem 2 (convergence with Polyak's step size)

$$\begin{aligned}
\|x_{k+1} - x_*\|^2 \le& \|x_k - x_*\|^2, \\
\lim_{k\to\infty} f(x_k) =& f_*, \\
f_{*,k} - f_* \le& \frac{L_f\|x_0 - x_*\|}{\sqrt{k+1}}
\end{aligned}$$

Proof

$$\begin{aligned}
\|x_{k+1} - x_*\|^2 \le& \|x_k - x_*\|^2 - 2t_k\left(f(x_k) - f_*\right) + t_k^2\left\|f'(x_k)\right\|^2 \\
=& \|x_k - x_*\|^2 - \frac{(f(x_k) - f_*)^2}{\|f'(x_k)\|^2} \\
=& \|x_k - x_*\|^2 - \frac{(f(x_k) - f_*)^2}{L_f^2}
\end{aligned}$$

Telescoping yields

$$\frac{1}{L_f^2}\sum_{n=0}^{k}(f(x_k)-f_*)^2 \leq \|x_0-x_*\|^2 - \|x_{k+1}-x_*\|^2.$$

Moreover,

$$\sum_{n=0}^{k}(f(x_k)-f_*)^2 \geq (k+1)(f_{*,k}-f_*)^2$$

■

Fejer monotonicity (Definition 8.15, Theorem 8.16, and Theorem 8.17 of Beck (2017)) implies $\lim_{k\to\infty} x_k = x_*$.

1. convergence with dynamic step sizes (Section 8.2.4 of Beck (2017))
2. strongly convex case (Section 8.2.5 of Beck (2017))

5. Background for Mirror Descent

- dual space
- conjugate functions
  - Fenchel's inequality
  - $f = f^**$
  - conjugate gradient theorem
  - Theorem 4.23
- L-smoothness
  - Theorem 5.4
  - descent lemma
  - first order characterization
  - second order characterization*
- strong convexity
  - first order characterization
  - conjugate correspondence theorem

The conjugate of f is

$$f^*(y) = \max_x \langle y, x\rangle - f(x).$$

Theorem 3 (first order characterization of strong convexity) *The following claims are equivalent.*

(i) f *is* σ*-strongly convex*

(ii) $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\sigma}{2}\|y - x\|^2$

(iii) $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \sigma \|x - y\|^2$

Theorem 4 (first order characterization of L-smoothness) *The following claims are equivalent.*

(i) f *is* L*-smooth*

(ii) $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2$

(iii) $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_*^2$

(iv) $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_*^2$

(v) $f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda) f(y) - \frac{L}{2}\lambda(1-\lambda)\|x - y\|^2$

Theorem 5 (conjugate gradient theorem) *Suppose f is convex, then the following claims are equivalent.*

(i) $\langle x, y \rangle = f(x) + f^*(y)$

(ii) $y = \nabla f(x)$

(iii) $x = \nabla f^*(y)$

Proof

$$\begin{aligned} & (ii) \\ \iff & f(z) \geq f(x) + \langle y, z - x \rangle \, \forall z \\ \iff & \langle y, x \rangle - f(x) \geq \langle y, z \rangle - f(z) \, \forall z \\ \iff & \langle y, x \rangle - f(x) \geq f^*(y) \\ \iff & (i) \end{aligned}$$

■

$$\nabla f^*(y) = \arg\max_x \{\langle y, x \rangle - f(x)\}$$

Theorem 6 (conjugate correspondence theorem) *Suppose f is convex. f is σ-strongly convex w.r.t. $\|\cdot\|$ if and only if f^* is $\frac{1}{\sigma}$-smooth w.r.t. $\|\cdot\|_*$.*

Proof

$$\begin{aligned} \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq & \sigma \|x - y\|^2 \\ \langle x - y, \nabla f^*(x) - \nabla f^*(y) \rangle \geq & \sigma \|\nabla f^*(x) - \nabla f^*(y)\|_*^2 \end{aligned}$$

■

6. Mirror Descent

Problem:

$$\min_{x\in C} f(x)$$

Assumption 6.1 *f is strictly convex and differentiable at every $x \in C$*

Assumption 6.2 *C is closed and convex*

Bregman distance

$$B_w(x,y) = w(x) - w(y) - \langle \nabla w(y), x - y\rangle$$

Assumption 6.3 *1. w is convex*

2. w is differentiable

3. $w + \delta_C$ is σ-strongly convex

Projected Gradient Descent:

$$\begin{aligned}
x_{k+1} =& P_C(x_k - t_k f'(x_k)) \\
=& \arg\min_{x\in C}\left\{\frac{1}{2}\big\|x - (x_k - t_k f'(x_k))\big\|^2\right\} \\
=& \arg\min_{x\in C}\left\{\frac{1}{2}\|x - x_k\|^2 + \frac{1}{2}t_k^2\big\|f'(x_k)\big\|^2 + t_k\big\langle f'(x_k), x - x_k\big\rangle\right\} \\
=& \arg\min_{x\in C}\left\{\frac{1}{2}\|x - x_k\|^2 + t_k\big\langle f'(x_k), x - x_k\big\rangle\right\}
\end{aligned}$$

Mirror Descent:

$$\begin{aligned}
x_{k+1} =& \arg\min_{x\in C}\big\{B_w(x, x_k) + t_k\big\langle f'(x_k), x - x_k\big\rangle\big\} \\
=& \arg\min_{x\in C}\{w(x) - w(x_k) - \langle \nabla w(x_k), x - x_k\rangle + t_k\langle \nabla f(x_k), x - x_k\rangle\} \\
=& \arg\min_{x\in C}\{w(x) - \langle \nabla w(x_k), x\rangle + t_k\langle \nabla f(x_k), x\rangle\} \\
=& \arg\min_{x\in C}\{\langle t_k\nabla f(x_k) - \nabla w(x_k), x\rangle + w(x)\} \\
=& \nabla w^*(\nabla w(x_k) - t_k\nabla f(x_k))
\end{aligned}$$

Theorem 7 (non-Euclidean second prox theorem) *Suppose w and ψ are convex and differentiable and $dom(\psi) \subseteq dom(w)$, and $w + \delta_{dom(\psi)}$ is σ-strongly convex. For any b, define*

$$a \doteq \arg\min_x \{\psi(x) + B_w(x, b)\}.$$

Then for any u, we have

$$\langle \nabla w(b) - \nabla(a), u - a\rangle \le \psi(u) - \psi(a).$$

Proof

$$\begin{aligned}\nabla\psi(a) + \nabla w(a) - \nabla(b) =& 0,\\ \langle \nabla w(b) - \nabla w(a), u - a\rangle =& \langle \nabla\psi(a), u - a\rangle \le \psi(u) - \psi(a).\end{aligned}$$

■

Lemma 8 (fundamental inequality for mirror descent)

$$t_k(f(x_k) - f_*) \le B_w(x_*, x_k) - B_w(x_*, x_{k+1}) + \frac{t_k^2}{2\sigma}\left\|f'(x_k)\right\|_*^2.$$

Proof

$$\langle \nabla w(x_k) - \nabla w(x_{k+1}), u - x_{k+1}\rangle \le t_k\left\langle f'(x_k), u - x_{k+1}\right\rangle.$$

By the three-points lemma (cf. $\frac{1}{2}\|x-y\|^2 = \frac{1}{2}\|x-z\|^2 + \frac{1}{2}\|z-y\|^2 + \langle x-z, z-y\rangle$), we have

$$B_w(u, x_{k+1}) + B_w(x_{k+1}, x_k) - B_w(u, x_k) \le t_k\left\langle f'(x_k), u - x_{k+1}\right\rangle.$$

■

Convergence with fixed number of iterations (Section 9.2.2 of Beck (2017)) and dynamic stepsize (Section 9.2.3 of Beck (2017)).

7. Background for Proximal Gradient Descent

- second prox theorem
- nonexpansivity

The proximal of x w.r.t. f is

$$\operatorname{prox}_f(x) \doteq \arg\min_u \left\{ f(u) + \frac{1}{2}\|x - u\|^2 \right\}$$

8. Proximal Gradient Descent

$$\min_x \{F(x) \doteq f(x) + g(x)\}$$

Projected Gradient Descent:

$$\begin{aligned}x_{k+1} =& P_C(x_k - t_k f'(x_k))\\ =& \arg\min_{x\in C}\left\{\frac{1}{2}\left\|x - (x_k - t_k f'(x_k))\right\|^2\right\}\end{aligned}$$

Proximal Gradient Descent:

$$\begin{aligned} x_{k+1} =& \arg\min_x \left\{ t_k g(x) + \frac{1}{2}\left\|x - (x_k - t_k f'(x_k))\right\|^2 \right\} \\ =& \text{prox}_{t_k g}(x_k - t_k f'(x_k)) \end{aligned}$$

Define the prox-grad operator

$$T_L^{f,g}(x) \doteq \text{prox}_{\frac{1}{L}g}\left(x - \frac{1}{L}f'(x)\right).$$

Then

$$\begin{aligned} x_{k+1} =& T^{f,g}_{\frac{1}{t_k}}(x_k) \\ =& x_k - t_k \frac{1}{t_k}\left(x_k - T^{f,g}_{\frac{1}{t_k}}(x_k)\right) \\ =& x_k - t_k G^{f,g}_{\frac{1}{t_k}}(x_k), \end{aligned}$$

where the gradient mapping operator $G_L^{f,g}$ is defined as

$$G_L^{f,g}(x) \doteq L(x - T_L^{f,g}(x)).$$

- Sufficent decrease (Section 10.3.1)
- Gradient mapping (Section 10.3.2)
- Nonconvex case (Section 10.3.3.)
- Convex case (Section 10.4)
- Strongly convex case (Section 10.6)

Theorem 9 (fundamental prox-grad inequality) *Suppose*

$$f(T_L(y)) \le f(y) + \left\langle f'(y), T_L(y) - y \right\rangle + \frac{L}{2}\|T_L(y) - y\|^2.$$

Then

$$F(x) - F(T_L(y)) \ge \frac{L}{2}\|x - T_L(y)\|^2 - \frac{L}{2}\|x - y\|^2 + \ell_f(x, y),$$

where

$$\ell_f(x, y) = f(x) - f(y) - \left\langle f'(y), x - y \right\rangle.$$

Proof Consider the function

$$\phi(u) = f(y) + \left\langle f'(y), u - y \right\rangle + g(u) + \frac{L}{2}\|u - y\|^2.$$

Recall that

$$\begin{aligned}
T_L(y) =& \operatorname{prox}_{\frac{1}{L}g}\left(y - \frac{1}{L}f'(y)\right) \\
=& \arg\min_x \left\{ \frac{1}{L}g(x) + \frac{1}{2}\left\| x - (y - \frac{1}{L}f'(y)) \right\|^2 \right\} \\
=& \arg\min_x \left\{ \frac{1}{L}g(x) + \frac{1}{2}\|x-y\|^2 + \frac{1}{2L^2}\left\|f'(y)\right\|^2 + \frac{1}{L}\left\langle f'(y), x-y\right\rangle \right\} \\
=& \arg\min_x \left\{ \frac{1}{L}g(x) + \frac{1}{2}\|x-y\|^2 + \frac{1}{L}\left\langle f'(y), x-y\right\rangle \right\} \\
=& \arg\min_x \left\{ g(x) + \frac{L}{2}\|x-y\|^2 + \left\langle f'(y), x-y\right\rangle \right\} \\
=& \arg\min_x \phi(x)
\end{aligned}$$

Then

$$\phi(x) - \phi(T_L(y)) \geq \frac{L}{2}\|x - T_L(y)\|^2.$$

■

References

Beck, A. (2017). *First-order methods in optimization*. SIAM.

Gallier, J. and Quaintance, J. (2019). *Algebra, Topology, Differential Calculus, and Optimization Theory For Computer Science and Engineering*. Philadelphia.