Algorithms

Bolei Zhang

December 23, 2024

Problem 1: Quadratic minimization and least-squares

The general convex quadratic minimization problem has the form

$$\text{minimize} \quad \frac{1}{2}x^T P x + q^T x + r, \tag{1}$$

where $P \in S_+^n, q \in \mathbb{R}^n$, and $r \in \mathbb{R}$. This problem can be solved via the optimality conditions, $Px^* + q = 0$, which is a set of linear equations. When $P \succ 0$, there is a unique solution, $x^* = -P^{-1}q$. In the more general case when P is not positive definite, any solution of $Px^* = -q$ is optimal for (1); if $Px^* = -q$ does not have a solution, then the problem (1) is unbounded below.

One special case of the quadratic minimization problem that arises very frequently is the least-squares problem

$$\text{minimize} \quad |Ax - b|_2^2 = x^T(A^T A)x - 2(A^T b)^T x + b^T b.$$

The optimality conditions

$$A^T A x^\star = A^T b$$

are called the normal equations of the least-squares problem.

Problem 2: Unconstrained geometric programming

Consider an unconstrained geometric program in convex form,

$$\text{minimize} \quad f(x) = \log\left(\sum_{i=1}^m \exp(a_i^T x + b_i)\right).$$

The optimality condition is

$$\nabla f(x^\star) = \frac{1}{\sum_{j=1}^m \exp(a_j^T x^\star + b_j)} \sum_{i=1}^m \exp(a_i^T x^\star + b_i) a_i = 0,$$

which in general has no analytical solution, so here we must resort to an iterative algorithm. For this problem, $\mathrm{dom} f = \mathbb{R}^n$, so any point can be chosen as the initial point $x^{(0)}$.

Definition 1: Strong convexity

Analysis: The objective function is strongly convex on S, which means that there exists an $m > 0$ such that

$$\nabla^2 f(x) \succeq mI \quad (2) \tag{2}$$

for all $x \in S$. Strong convexity has several interesting consequences. For $x, y \in S$ we have

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x) \tag{3}$$

for some z on the line segment $[x, y]$. By the strong convexity assumption (2), the last term on the right-hand side is at least $(m/2)||x - y||_2^2$, so we have the inequality

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2}|y - x|_2^2 \tag{4}$$

for all x and y in S. When $m = 0$, we recover the basic inequality characterizing convexity; for $m > 0$ we obtain a better lower bound on $f(y)$ than follows from convexity alone.

We will first show that the inequality (4) can be used to bound $f(x) - p^*$, which is the suboptimality of the point x, in terms of $||\nabla f(x)||_2$. The right-hand side of (4) is a convex quadratic function of y (for fixed x). Setting the gradient with respect to y equal to zero, we find that $\tilde{y} = x - (1/m)\nabla f(x)$ minimizes the right-hand side. Therefore we have

$$\begin{aligned} f(y) &\geq f(x) + \nabla f(x)^T(y - x) + \frac{m}{2}|y - x|2^2 \\ &\geq f(x) + \nabla f(x)^T(\tilde{y} - x) + \frac{m}{2}|\tilde{y} - x|2^2 \\ &= f(x) - \frac{1}{2m}|\nabla f(x)|2^2. \end{aligned} \tag{5}$$

Since this holds for any $y \in S$, we have

$$p^\star \geq f(x) - \frac{1}{2m}|\nabla f(x)|2^2. \tag{6}$$

This inequality shows that if the gradient is small at a point, then the point is nearly optimal. The inequality can also be interpreted as a condition for suboptimality which generalizes the optimality condition :

$$|\nabla f(x)|2 \leq (2m\epsilon)^{\frac{1}{2}} \Rightarrow f(x) - p^\star \leq \epsilon. \quad (9.10) \tag{7}$$

We can also derive a bound on $||x - x^*||_2$, the distance between x and any optimal point x^*, in terms of $||\nabla f(x)||_2$:

$$|x - x^\star|2 \leq \frac{2}{m}|\nabla f(x)|2. \quad (9.11) \tag{8}$$

To see this, we apply (4) with $y = x^*$ to obtain

$$p^\star = f(x^\star) \geq f(x) + \nabla f(x)^T(x^\star - x) + \frac{m}{2}|x^\star - x|2^2 \quad \geq f(x) - |\nabla f(x)|2|x^\star - x|2 + \frac{m}{2}|x^\star - x|_2^2, \tag{9}$$

where we use the Cauchy-Schwarz inequality in the second inequality. Since $p^* \leq f(x)$, we must have

$$-||\nabla f(x)||_2||x^\star - x||_2 + \frac{m}{2}||x^\star - x||_2^2 \leq 0 \tag{10}$$

from which follows. One consequence of is that the optimal point x^* is unique.

Definition 2: Smoothness

Analysis: The inequality (4) implies that the sublevel sets contained in S are bounded, so in particular, S is bounded. Therefore the maximum eigenvalue of $\nabla^2 f(x)$, which is a continuous function of x on S, is bounded above on S, i.e., there exists a constant N such that

$$\nabla^2 f(x) \preceq MI \tag{11}$$

for all $x \in S$. This upper bound on the Hessian implies for any $x, y \in S$,

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{M}{2}||y - x||_2^2, \quad (9.13) \tag{12}$$

which is analogous to (4). Minimizing each side over y yields

$$p^\star \leq f(x) - \frac{1}{2M}||\nabla f(x)||_2^2, \quad (9.14) \tag{13}$$

the counterpart of (5).

Theorem 1: The gradient descent method with exact line search and backtracking has linear convergence rate for unconstrained convex problems with strong convexity and smoothness.

Proof: Using the lighter notation $x^+ = x + t\Delta x$ for $x(k+1) = x(k) + t(k)\Delta x(k)$, where $\Delta x = -\nabla f(x)$. We assume f is strongly convex on S, so there are positive constants $\tilde{m}$ and M such that

$$mI \preceq \nabla^2 f(x) \preceq MI, \quad \forall x \in S.$$

Define the function $\tilde{f} : \mathbb{R} \to \mathbb{R}$ by

$$\tilde{f}(t) = f(x - t\nabla f(x)),$$

i.e., $\tilde{f}$ as a function of the step length t in the negative gradient direction. In the following discussion, we will only consider t for which $x - t\nabla f(x) \in S$.

From the inequality (9.13), with $y = x - t\nabla f(x)$, we obtain a quadratic upper bound on $\tilde{f}$:

$$\tilde{f}(t) \le f(x) - t\|\nabla f(x)\|^2 + \frac{Mt^2}{2}\|\nabla f(x)\|^2.$$

Analysis for Exact Line Search We now assume that an exact line search is used, and minimize over t both sides of the inequality (9.17). On the left-hand side, we get

$$\tilde{f}(t_{\text{exact}}),$$

where t_{exact} is the step length that minimizes $\tilde{f}(t)$.

The step length that minimizes $\tilde{f}$ is on the right-hand side, which is a simple quadratic, minimized by $t = \frac{1}{M}$, and has the minimum value

$$f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2.$$

Therefore, we have

$$f(x^+) = \tilde{f}(t_{\text{exact}}) \le f(x) - \frac{1}{2M}\|\nabla f(x)\|_2^2.$$

Subtracting $p^\star$ from both sides, we get

$$f(x^+) - p^\star \le f(x) - p^\star - \frac{1}{2M}\|\nabla f(x)\|_2^2.$$

We combine this with

$$\|\nabla f(x)\|_2^2 \ge 2m(f(x) - p^\star)$$

(which follows from (5)) to conclude

$$f(x^+) - p^\star \le (1 - \frac{m}{M})(f(x) - p^\star).$$

Applying this inequality recursively, we find that

$$f(x^{(k)}) - p^\star \le c^k(f(x^{(0)}) - p^\star),$$

where $c = 1 - \frac{m}{M} < 1$, which shows that $f(x^{(k)})$ converges to $p^\star$ as $k \to \infty$. In particular, we must have

$$f(x^{(k)}) - p^\star \le \epsilon$$

after at most

$$\frac{\log((f(x^{(0)}) - p^\star)/\epsilon)}{\log(1/c)}$$

iterations of the gradient method with exact line search.

This bound on the number of iterations required, even though crude, can give some insight into the gradient method. The numerator,

$$\log((f(x^{(0)}) - p^\star)/\epsilon),$$

can be interpreted as the logarithm of the ratio of the initial suboptimality (i.e., the gap between $f(x^{(0)})$ and $p^\star$) to the final suboptimality (i.e., less than ϵ). This term suggests that the number of iterations depends on how good the initial point is and what the final required accuracy is.

The denominator $\log(1/c)$, is a function of M/m, which we have seen is a bound on the condition number of $\nabla^2 f(x)$ over S, or the condition number of the sublevel sets $\{z \mid f(z) \leq \alpha\}$. For large condition number bound M/m, we have

$$\log(1/c) = -\log(1 - m/M) \approx m/M,$$

so our bound on the number of iterations required increases approximately linearly with increasing M/m.

We will see that the gradient method does in fact require a large number of iterations when the Hessian of f, near $x^\star$, has a large condition number. Conversely, when the sublevel sets of f are relatively isotropic, so that the condition number bound M/m can be chosen to be relatively small, the bound shows that convergence is rapid, since c is small, or at least not too close to one.

The bound shows that the error $f(x^{(k)}) - p^\star$ converges to zero at least as fast as a geometric series. In the context of iterative numerical methods, this is called *linear convergence*, since the error lies below a line on a log-linear plot of error versus iteration number.

Analysis for Backtracking Line Search

Now we consider the case where a backtracking line search is used in the gradient descent method. We will show that the backtracking exit condition,

$$\tilde{f}(t) \leq f(x) - \alpha t \|\nabla f(x)\|_2^2,$$

is satisfied whenever $0 \leq t \leq 1/M$. First, note that

$$0 \leq t \leq 1/M \implies -t + \frac{Mt^2}{2} \leq -t/2$$

(which follows from the convexity of $-t + \frac{Mt^2}{2}$). Using this result and the bound (9.17), we have, for $0 \leq t \leq 1/M$,

$$\tilde{f}(t) \leq f(x) - t\|\nabla f(x)\|_2^2 + \frac{Mt^2}{2}\|\nabla f(x)\|_2^2 \leq f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2 \leq f(x) - \alpha t\|\nabla f(x)\|_2^2,$$

since $\alpha < 1/2$. Therefore, the backtracking line search terminates either with $t = 1$ or with a value $t \geq \beta/M$. This provides a lower bound on the decrease in the objective function.

In the first case, we have

$$f(x^+) \leq f(x) - \alpha\|\nabla f(x)\|_2^2,$$

and in the second case, we have

$$f(x^+) \leq f(x) - \frac{\beta\alpha}{M}\|\nabla f(x)\|_2^2.$$

Putting these together, we always have

$$f(x^+) \leq f(x) - \min\{\alpha, \beta\alpha/M\}\|\nabla f(x)\|_2^2.$$

Now we can proceed exactly as in the case of exact line search. Subtracting $p^\star$ from both sides, we get

$$f(x^+) - p^\star \leq f(x) - p^\star - \min\{\alpha, \beta\alpha/M\}\|\nabla f(x)\|_2^2,$$

and combine this with

$$\|\nabla f(x)\|_2^2 \geq 2m(f(x) - p^\star)$$

to obtain

$$f(x^+) - p^\star \leq \big(1 - \min\{2m\alpha, 2\beta\alpha m/M\}\big)(f(x) - p^\star).$$

From this, we conclude

$$f(x^{(k)}) - p^\star \leq c^k(f(x^{(0)}) - p^\star),$$

where

$$c = 1 - \min\{2m\alpha, 2\beta\alpha m/M\} < 1.$$

In particular, $f(x^{(k)})$ converges to $p^\star$ at least as fast as a geometric series with an exponent that depends (at least in part) on the condition number bound M/m. In the terminology of iterative methods, the convergence is *at least linear*.

Problem 3: For the function $f(x) = -\log x + x$, starting with the initial point $x^{(0)} = 3$ and using a fixed step size of $\alpha = 0.1$, calculate the value of $x^{(1)}$ after one step of gradient descent.

Solution: Given the function:

$$f(x) = -\log x + x$$

The derivative of $f(x)$ is:

$$f'(x) = -\frac{1}{x} + 1$$

With the initial point $x^{(0)} = 3$ and a step size of $\alpha = 0.1$, we compute the gradient at $x^{(0)}$:

$$f'(3) = -\frac{1}{3} + 1 = \frac{2}{3}$$

The update rule for gradient descent is:

$$x^{(1)} = x^{(0)} - \alpha f'(x^{(0)})$$

Thus, the updated value after one step is:

$$x^{(1)} \approx 2.9333$$

Problem 4: For the function $f(x) = x_1^2 + x_2$, starting with the initial point $x^{(0)} = (2, 1)$, calculate the value of $x^{(1)}$ after one step of gradient descent with an exact step size.

Solution: Given the function:

$$f(x) = x_1^2 + x_2$$

The gradient of $f(x)$ is:

$$\nabla f(x) = \begin{bmatrix} 2x_1 \\ 1 \end{bmatrix}$$

With the initial point $x^{(0)} = (2, 1)$, the gradient at $x^{(0)}$ is:

$$\nabla f(x^{(0)}) = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$$

The update rule for gradient descent is:

$$x^{(1)} = x^{(0)} - \alpha \nabla f(x^{(0)})$$

Using an exact step size α that minimizes $f(x)$, we solve:

$$f(x) = 16\alpha^2 - 17\alpha + 5$$

$$\frac{df}{d\alpha} = 32\alpha - 17$$

$$\alpha = \frac{17}{32}$$

Substituting $\alpha = \frac{17}{32}$ into the update rule:

$$x^{(1)} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} - \frac{17}{32} \begin{bmatrix} 4 \\ 1 \end{bmatrix}$$

$$x^{(1)} = \begin{bmatrix} -\frac{1}{8} \\ \frac{15}{32} \end{bmatrix}$$