Lecture 11. Line Search

11.1 Exact line search

We have analysed the convergence (and the rate) of the gradient descent with a fixed step size, where we should choose the step size $\eta \le 1/L$ if the objective function is L-smooth. However, if the smoothness is difficult to determine, how can we set the step size?

Note that a convex function restricted to any line is also convex. A naive idea is to improve the length of step so that the restricted function achieve its minimum value. Specifically, since $x_{k+1} = x_k - \eta_k \nabla f(x_k)$ where the step size η_k is to be determined, we consider the function $g(s) = f(x_k - s\nabla f(x_k))$. It is also a convex function of s, and $f(x_{k+1}) = g(\eta_k)$. We can greedily make $f(x_{k+1})$ as small as possible, which is to set

$$\eta_k = \arg\min_s g(s) = \arg\min_s f\left(x_k - s\nabla f(x_k)\right).$$

This method is called the *exact line search*.

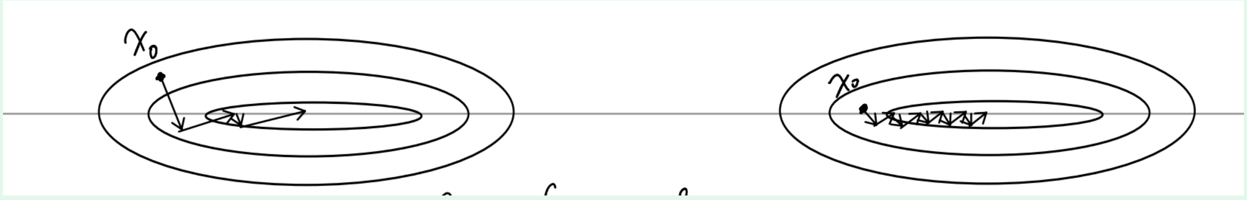
Example

Consider $f(x) = x^\top Qx + w^\top x, Q \succ \mathbf{0}$. Let $d_k = \nabla f(x_k) = 2Qx_k + w$. Then, by definition,

$$\begin{aligned}
\eta_k &= \arg\min_s f(x_k - sd_k) \\
&= \arg\min_s \left(f(x_k) - 2sd_k^\top Qx_k + s^2 d_k^\top Qd_k - sw^\top d_k\right) \\
&= \arg\min_s \left(-sd_k^\top (2Qx_k + w) + s^2 d_k^\top Qd_k\right) \\
&= \arg\min_s \left(-sd_k^\top d_k + s^2 d_k^\top Qd_k\right) \\
&= \frac{d_k^\top d_k}{2d_k^\top Qd_k}
\end{aligned}$$

Proposition

Applying the method of exact line search, successive gradient directions are always orthogonal.



Proof

Let $g(s) = f\left(x_k - s\nabla f(x_k)\right)$, then $\eta_k = \arg\min_s g(s)$. By the first-order necessary condition, we have $g'(\eta_k) = 0$, which means

$$0 = g'(\eta_k) = \langle \nabla f(x_k - \eta_k \nabla f(x_k)), -\nabla f(x_k)\rangle = -\langle \nabla f(x_{k+1}), -\nabla f(x_k)\rangle.$$

Therefore, $-\nabla f(x_{k+1})$ and $-\nabla f(x_k)$ are orthogonal.

Newton's method to find zero points

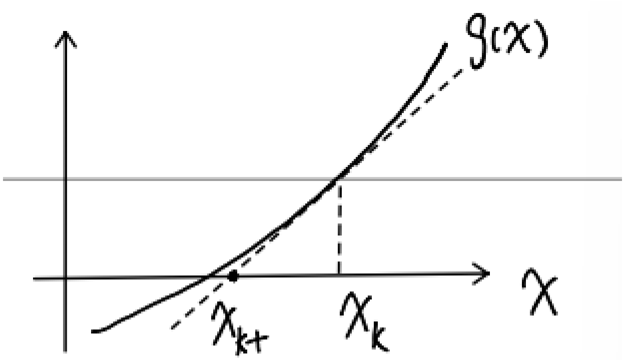
In the method of exact line search, we need to find the minimum point of a $\mathbb{R} \to \mathbb{R}$ convex function. This is equivalent to find the zero points of $g(s)$ if we let $g(s) = f'\left(x_k - s\nabla f(x_k)\right)$.

Now we introduce some methods to find the zero points. Note that $g(s)$ is an increasing function, since $g(s)$ is convex. Thus, there is a smart method using *binary search*. If g is continuous and we know that there exists $s_1 < s_2$ such that $g(s_1) < 0$ and $g(s_2) > 0$, then we can check whether $g((s_1 + s_2)/2)$ is zero and continue partitioning the interval into two parts and updating the left and right points until finding the zero point.

A better method is to apply the so-called *Newton's method*. The idea is that given a value of $s = \hat{s}$, we can approximate $g(s)$ locally (near $\hat{s}$):

$$g(s) \approx g(\hat{s}) + g'(\hat{s})(s - \hat{s})\,.$$

The zero point of the right hand side is $\hat{s} - \frac{g(\hat{s})}{g'(\hat{s})}$, which should be a good approximation of the zero point of $g(s)$ (if the zero point is near $\hat{s}$).



The Newton method is to do the approximation iteratively. Namely, we choose an arbitrary $s_0 = \hat{s}$ and let

$$s_{k+1} = s_k - \frac{g(s_k)}{g'(s_k)}\,.$$

Example (*Fast inverse square root*)

Given a positive real number x, how to calculate $\frac{1}{\sqrt{x}}$?
Let $g(y) = \frac{1}{y^2} - x$. Then $g(y) = 0$ iff $y = \frac{1}{\sqrt{x}}$. Applying the Newton's method, we choose a *magic value* of y_0 and let $y_1 = y_0\left(\frac{3}{2} - \frac{x}{2}y_0^2\right)$. It converges to $\frac{1}{\sqrt{x}}$ rapidly.
The algorithm is best known for its implementation in 1999 in *Quake III Arena*.

Convergence rate of exact line search

Theorem

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex and L-smooth. Let $\{x_k\}$ be the sequence given by the gradient descent with exact line search. Then

$$f(x_T) - f^* \le \left(1 - \frac{\mu}{L}\right)^T (f(x_0) - f^*)\,.$$

Proof

Let $g(s) = f\left(x_k - s\nabla f(x_k)\right)$. Since f is L-smooth, we have

$$g(s) \le f(x_k) - \langle \nabla f(x_k), s\nabla f(x_k)\rangle + \frac{L}{2}\|s\nabla f(x_k)\|^2 = f(x_k) - s\|\nabla f(x_k)\|^2 + \frac{Ls}{2}$$

Denote the right hand side by $\tilde{g}(s)$. The minimizer of $\tilde{g}(s)$ is $s = 1/L$. Since

$\eta_k = \arg\min_s g(s)$, we have

$$f(x_{k+1}) = \min_s g(s) \leq \min_s \tilde{g}(s) = \tilde{g}\Big(\frac{1}{L}\Big) = f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 . \tag{1}$$

Moreover, f is μ-strongly convexity, so it holds that

$$f(x) \geq f(x_k) + \langle \nabla f(x_k),\, x - x_k \rangle + \frac{\mu}{2}\|x_k - x\|^2 .$$

Denote the right hand side by $\tilde{f}(x)$. The minimizer of $\tilde{f}(x)$ is $x = x_k - \frac{1}{\mu}\nabla f(x_k)$, since $\nabla \tilde{f}(x) = \mu(x - x_k) + \nabla f(x_k)$. Thus,

$$\begin{aligned} f(x^*) \geq \tilde{f}(x) \geq \min_x \tilde{f}(x) &= \tilde{f}\Big(x_k - \frac{1}{\mu}\nabla f(x_k)\Big) \\ &= f(x_k) - \frac{1}{\mu}\|\nabla f(x_k)\|^2 + \frac{1}{2\mu}\|\nabla f(x_k)\|^2 \\ &= f(x_k) - \frac{1}{2\mu}\|\nabla f(x_k)\|^2 . \end{aligned} \tag{2}$$

Combining (1) and (2), we have $f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)(f(x_k) - f(x^*))$.

If we know that f is L-smooth and set the fixed step size $\eta = 1/L$, it is easy to see that the result of convergence rate is the same as that with exact line search. But the advantage of the exact line search method is that we do not need to know the smoothness of f in advance.

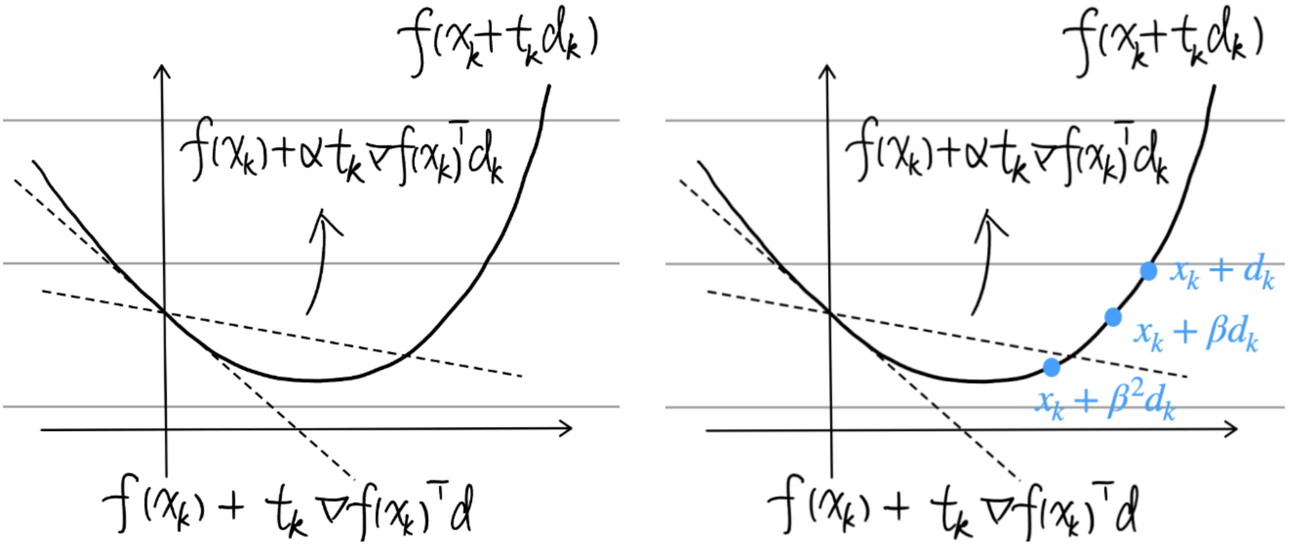
11.2 Backtracking line search

In general, it is expensive to calculate the step size in exact line search, and we usually do not need to know the *exact minimizer*. It is sufficient to show that the value of the objective function decreases sufficiently at each step. So we consider a so-called *backtracking line search method*.

Definition (*Armijo's rule*)

Armijo's rule is a well-known and widely applied backtracking rule to update the step size. Given a descending direction d_k, and $\alpha, \beta < 1$, we first initialize $\eta = 1$ and then update η to $\beta\eta$ as long as
$f(x_k + \eta d_k) > f(x_k) + \alpha\eta\langle \nabla f(x_k),\, d_k \rangle$.
In particular, if we set $d_k = -\nabla f(x_k)$, then $\langle \nabla f(x_k),\, d_k \rangle = -\|\nabla f(x_k)\|^2$.

The intuition is that, we know $f(x_k+\eta d_k) \geq f(x_k)+\eta\langle\nabla f(x_k),\, d_k\rangle$ by convexity. If $f(x_k+\eta d_k) \leq f(x_k)+\alpha\eta\langle\nabla f(x_k),\, d_k\rangle$ for some $\alpha<1$, we think $f(x_k+\eta d_k)$ decrease by a sufficient amount (then we can get an inequality similar to the descent lemma $f(x_{k+1}) \leq f(x_k)-\frac{\eta}{2}\|\nabla f(x_k)\|^2$ for gradient descent for L-smooth functions). Moreover, this requirement is always true if η is sufficiently close to 0, since $f(x_k+\eta d_k) \approx f(x_k)+\eta\langle\nabla f(x_k),\, d_k\rangle$ if $\eta \to 0$. Therefore, the update process stops after a finite number of iterations.


$f(x_k+t_k d_k)$
$f(x_k)+\alpha t_k \nabla f(x_k)^T d_k$
$f(x_k) + t_k \nabla f(x_k)^T d$
$f(x_k+t_k d_k)$
$f(x_k)+\alpha t_k \nabla f(x_k)^T d_k$
$x_k + d_k$
$x_k + \beta d_k$
$x_k + \beta^2 d_k$
$f(x_k) + t_k \nabla f(x_k)^T d$

Armijo choose $\alpha=\beta=\frac{1}{2}$. In the textbook, it suggests $\alpha\in[0.01,0.3]$ and $\beta\in[0.1,0.8]$.

Convergence rate of backtracking line search

We first give a lower bound of the step size. Initially set $\eta=1$ and $\alpha\leq\frac{1}{2}$. Since f is L-smooth, we have

$$\begin{aligned} f(x_k-\eta\nabla f(x_k)) &\leq f(x_k)-\eta\left\langle\nabla f(x_k),\nabla f(x_k)\right\rangle+\frac{L\eta^2}{2}\|\nabla f(x_k)\|^2 \\ &\leq f(x_k)-\frac{\eta}{2}\|\nabla f(x_k)\|^2 \\ &\leq f(x_k)-\alpha\eta\,\|\nabla f(x_k)\|^2 \end{aligned}$$

if $\eta\leq 1/L$. Thus the update process terminates once we have $\eta\leq 1/L$, which further implies that $\eta\geq\beta/L$.

Then, applying the lower bound of η, we give the following result of the convergence rate.

Theorem

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex and L-smooth. Let $\{x_k\}$ be the sequence given by the gradient descent with Armijo's backtracking line search. Then

$$f(x_T) - f^* \leq \Big(1 - \min\big\{2\mu\alpha, 4\mu\alpha(1-\alpha)\beta/L\big\}\Big)^T \big(f(x_0) - f^*\big)\,.$$