Numerical Optimization

Lecture 13: Gradient Descent Method & Subgradient Method

王浩

信息科学与技术学院

Email: wanghao1@shanghaitech.edu.cn

Gradient Descent

Consider the following iteration from x_0:

1. Let $d_k = -\nabla f(x_k)$.
2. Compute α_k satisfying the Wolfe conditions.
3. Update $x_{k+1} \leftarrow x_k + \alpha_k d_k$, return to step 1.

- Obviously, the angle between d_k and $-\nabla f(x_k)$ is always $0°$, so

$$\cos\theta_k = 1 > 0 \text{ for all } k.$$

- Thus, we have global convergence due to Zoutendijk's theorem.

Interpretation of Gradient Descent

General nonlinear problem (may be nonconvex)

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

Gradient descent:

$$x^{k+1} \leftarrow x^k - \alpha \nabla f(x^k)$$

At each iteration, consider the expansion

$$f(y) \approx f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2\alpha} \|y - x\|^2$$

Quadratic approximation, replacing usual $\nabla^2 f(x)$ by $\frac{1}{\alpha} I$

$f(x) + \nabla f(x)^\top (y - x)$ linear approximation to f

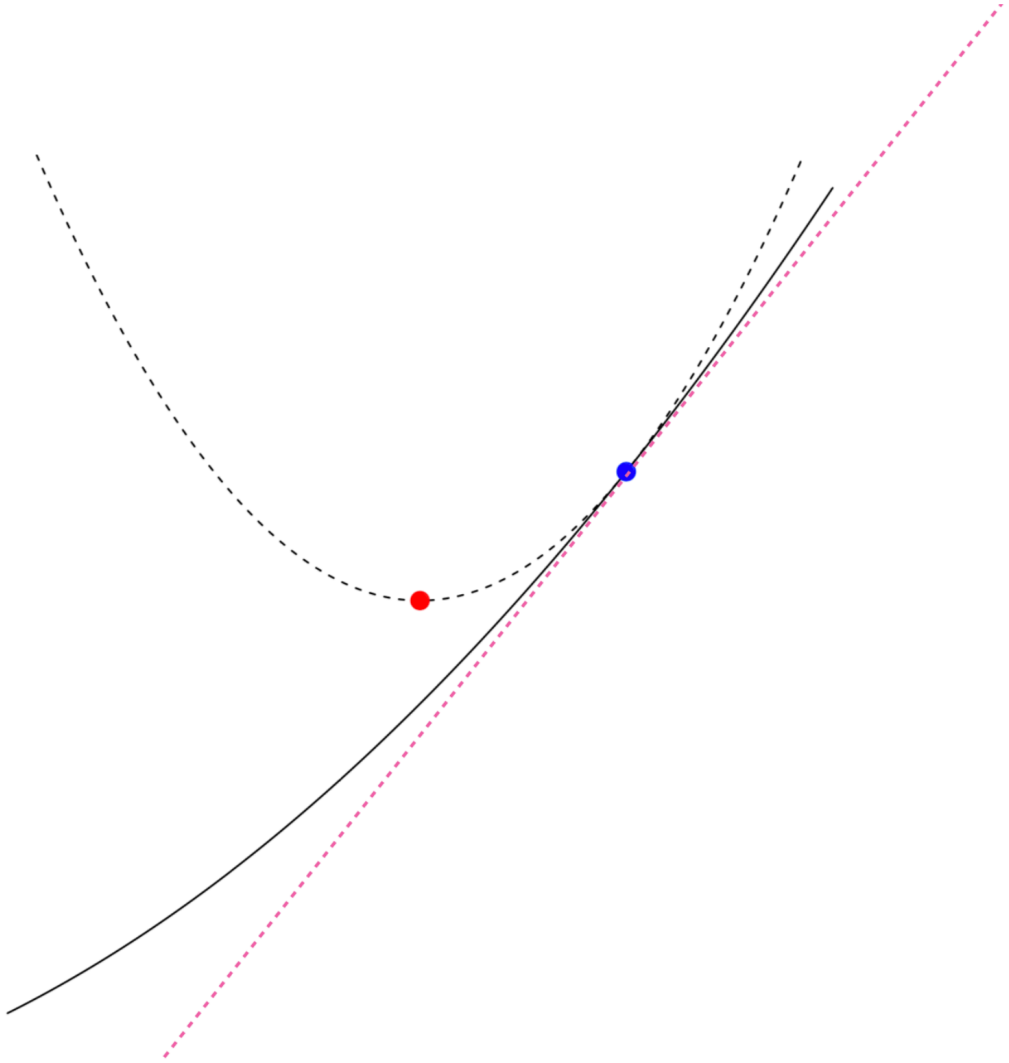
$\frac{1}{2t} \|y - x\|^2$ proximity term to x, with weight $1/(2\alpha)$

Choose next point $y = x^+$ to minimize quadratic approximation

$$x^+ = x - \alpha \nabla f(x)$$

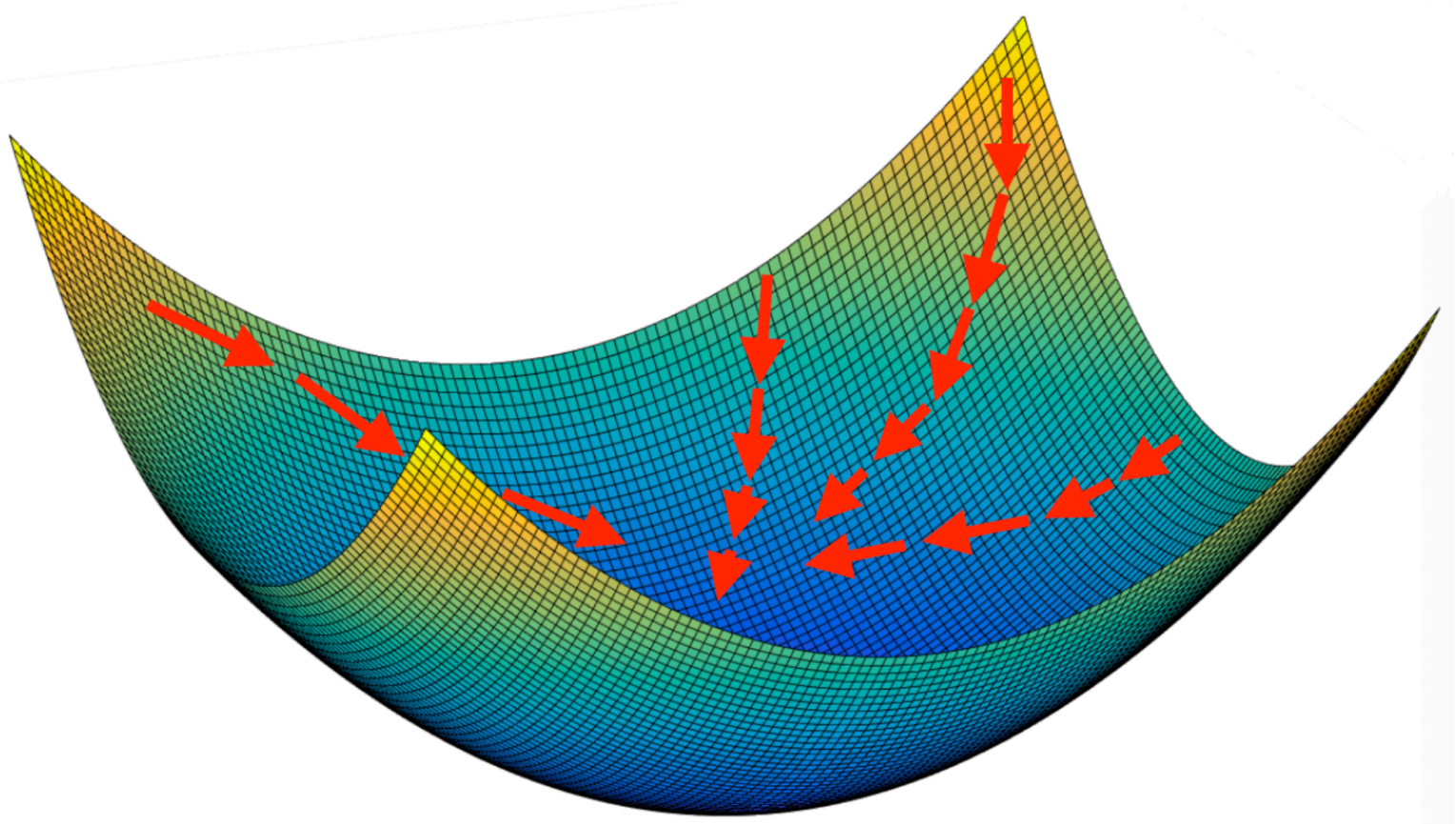
Gradient descent revisited Geoff Gordon & Ryan Tibshirani Optimization 10-725 / 36-725

Interpretation of Gradient Descent



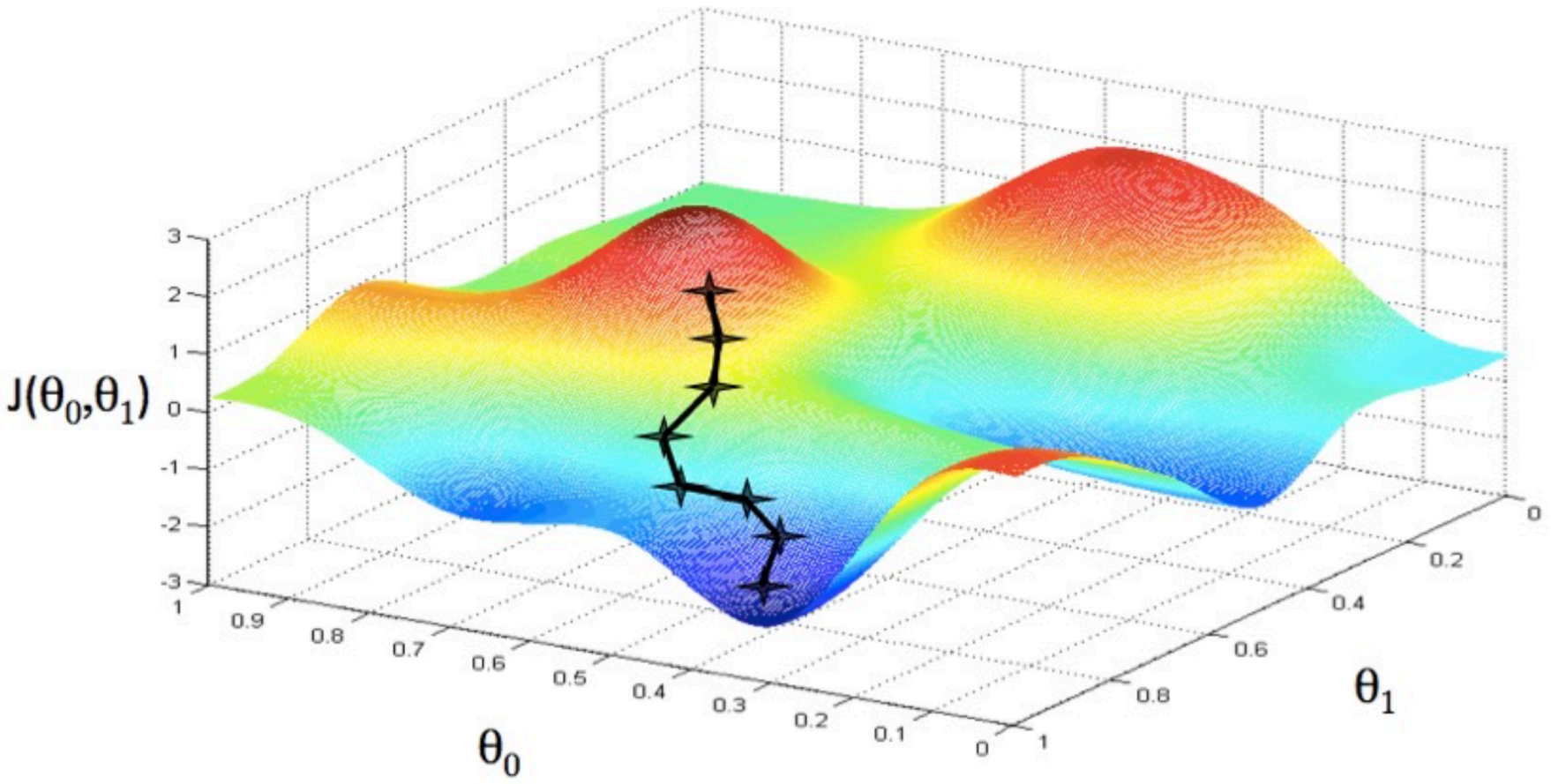
Blue point is x, red point is x^+

Starting point (global convergence)



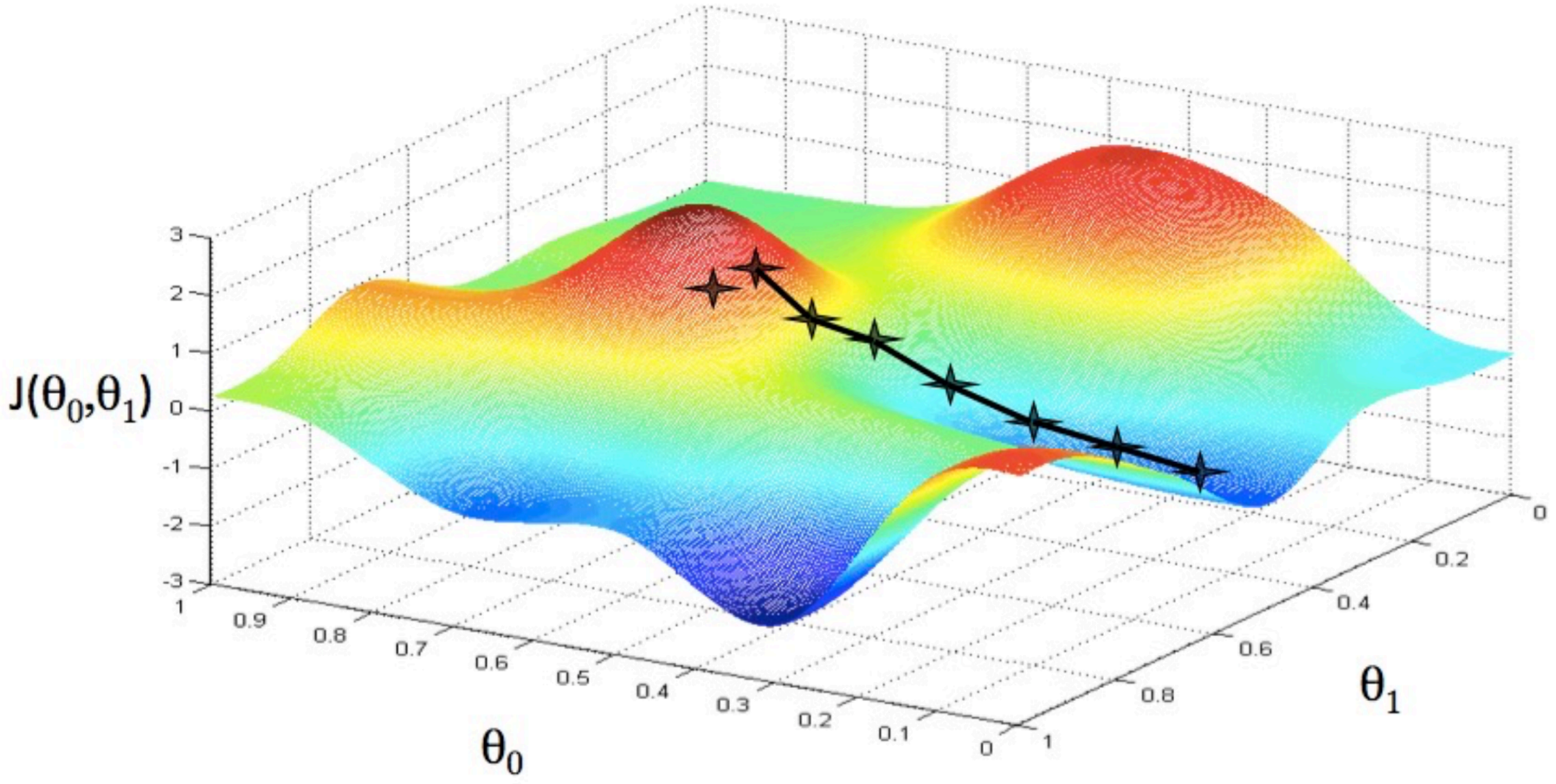
Starting point (global convergence)

For nonconvex cases, the algorithm converges starting with any initial point. But may lead to different optimal solution.



Starting point (global convergence)

For nonconvex cases, the algorithm converges starting with any initial point. But may lead to different optimal solution.



Stepsize/Learning Rate

Fixed stepsize α

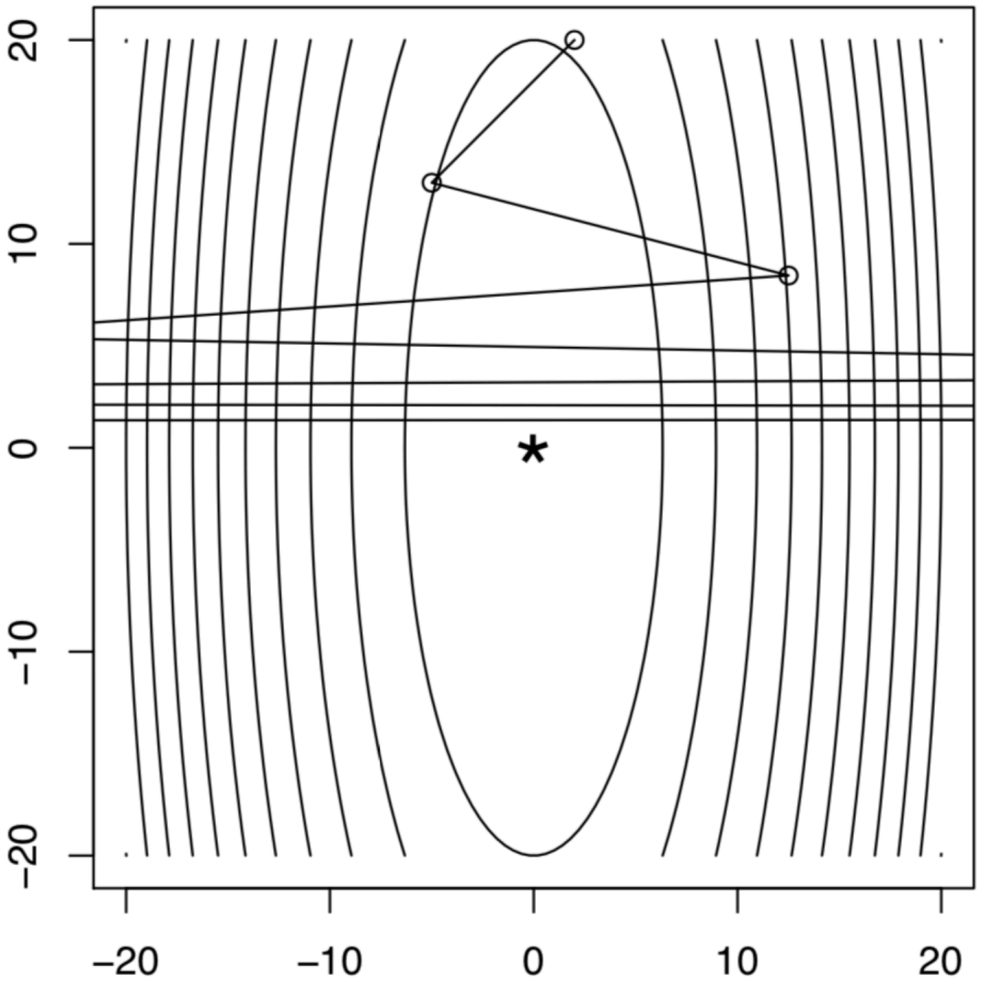
- Too small, slow progress, slow convergence
- Too large, oscillation, slow convergence, or even diverge

Dynamically changing stepsize α^k

- Too small, slow progress, slow convergence, or fail to converge to minimizers
- Too large, oscillation, slow convergence, or even diverge

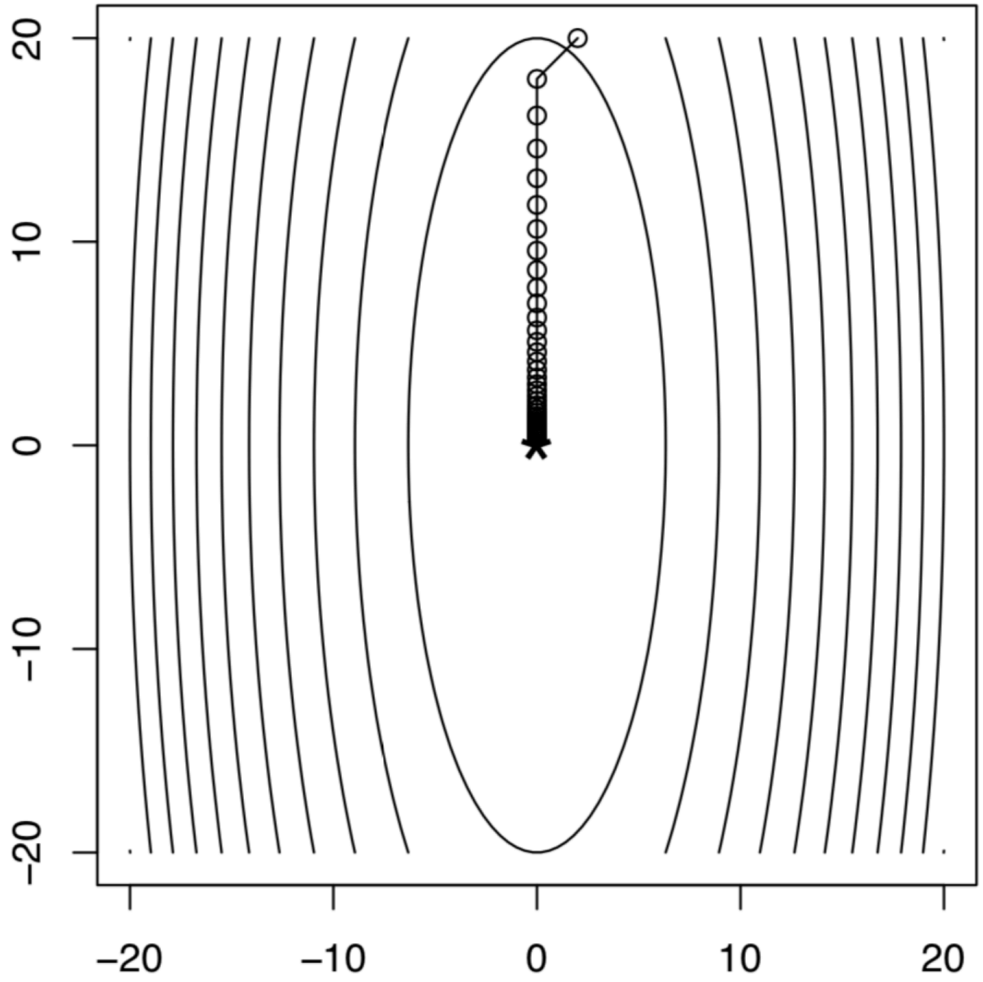
Fixed step size

Simply take $\alpha^k = \alpha$ for all $k = 1, 2, 3, \ldots$, can diverge if t is too big.
Consider $f(x) = (10x_1^2 + x_2^2)/2$, gradient descent after 8 steps:



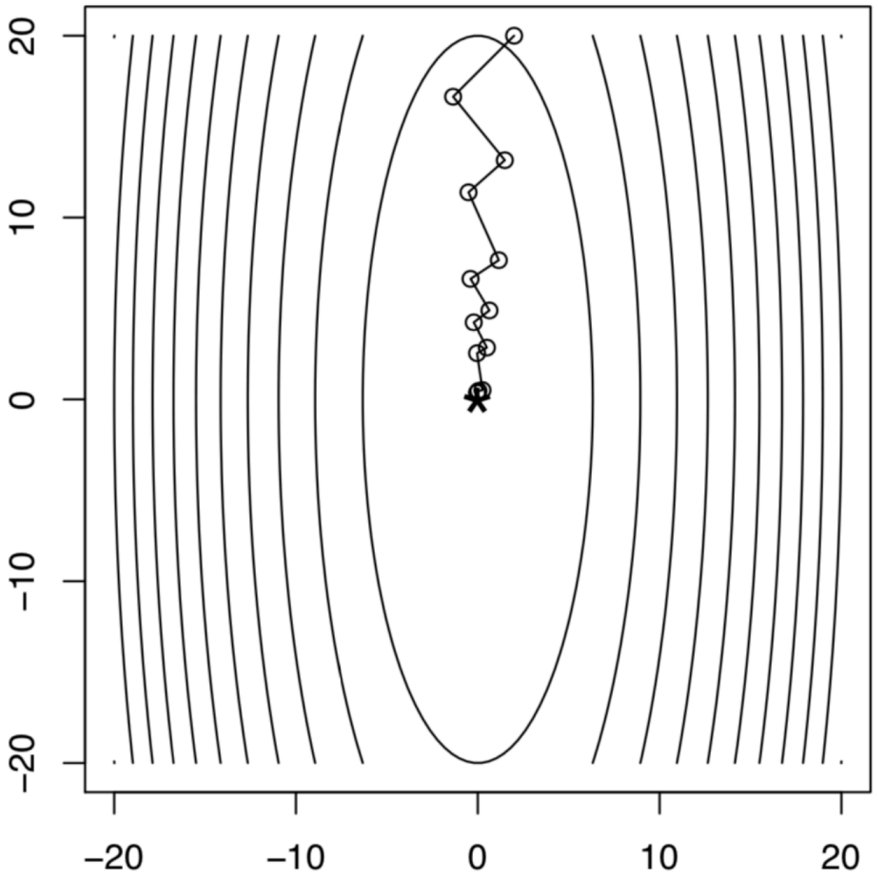
Fixed step size

Same example, gradient descent after 40 appropriately sized steps:



Interpretation of Backtracking line search

Backtracking picks up roughly the right step size (13 steps):



Here $\gamma = 0.8$ ($\gamma \in (0.1, 0.8)$ is recommended, and often chosen as 0.5).

Exact line search

At each iteration, do the best we can along the direction of the gradient,

$$t = \arg\min_{s \geq 0} f(x - s\nabla f(x))$$

Usually not possible to do this minimization exactly

Approximations to exact line search are often not much more efficient than backtracking, and it's not worth it

Theorem

If the steepest descent method with exact line searches is applied to

$$f(x) = c^T x + \tfrac{1}{2} x^T Q x, \quad Q \succ 0,$$

then the iterates satisfy

$$\|x_{k+1} - x_*\|_Q^2 \leq \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} \right)^2 \|x_k - x_*\|_Q^2,$$

where $0 < \lambda_1 \leq \cdots \leq \lambda_n$ *are the eigenvalues of* Q.

- Note: $\|x\|_Q^2 := x^T Q x$.
- If $\lambda_n = \lambda_1$, then convergence in one step.
- If $\lambda_n \neq \lambda_1$, then function values converge to $f(x_*)$ at a linear rate.
- We cannot expect anything better for more general $f(x)$!

Convergence analysis

Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, and additionally

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \text{for any } x, y.$$

i.e., ∇f is Lipschitz continuous with constant $L > 0$

Theorem

Gradient descent with fixed step size $\alpha \leq 1/L$ satisfies

$$f(x^k) - f(x^*) \leq \frac{\|x^0 - x^*\|^2}{2\alpha k}.$$

i.e., gradient descent has convergence rate $O(1/k)$

i.e., to get $f(x^k) - f(x^*) \leq \epsilon$, need $O(1/\epsilon)$ iterations

Proof

Key steps:

- ∇f Lipschitz with constant $L \implies$

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2 \quad \forall x, y$$

- Letting $x^+ = x - \alpha \nabla f(x)$ and taking $0 < \alpha \leq 1/L$, we then get

$$\begin{aligned} f(x^+) &\leq f(x) + \nabla f(x)^\top (x^+ - x) + \tfrac{L}{2}\|x^+ - x\|_2^2 \\ &= f(x) + \nabla f(x)^\top (x - \alpha \nabla f(x) - x) + \tfrac{L}{2}\|x - \alpha \nabla f(x) - x\|^2] \\ &= f(x) - \alpha \nabla f(x)^\top \nabla f(x) + \tfrac{L\alpha^2}{2}\|\nabla f(x)\|^2] \\ &= f(x) - (1 - \tfrac{1}{2}L\alpha)\alpha \|\nabla f(x)\|_2^2 \end{aligned}$$

- Notice that $-(1 - \frac{1}{2}L\alpha) = \frac{1}{2}L\alpha - 1 \leq \frac{1}{2}L(1/L) - 1 = \frac{1}{2} - 1 = -\frac{1}{2}$. Therefore,

$$f(x^+) \leq f(x) - \tfrac{1}{2}\alpha \|\nabla f(x)\|_2^2$$

$\implies$ objective monotonically decreases.

- By the convexity of f, we have

$$f(x) \leq f(x^*) + \nabla f(x)^\top (x - x^*)$$

- We can use this to further derive

$$\begin{aligned}
f(x^+) &\leq f(x) - \tfrac{1}{2}\alpha \|\nabla f(x)\|_2^2 \\
&\leq f(x^*) + \nabla f(x)^\top (x - x^*) - \frac{\alpha}{2}\|\nabla f(x)\|^2 \\
&= f(x^*) - \frac{1}{2\alpha}[-2(x - x^+)^\top (x - x^*) + \|x - x^+\|^2] \\
&= f(x^*) + \frac{1}{2\alpha}[2(x - x^+)^\top (x - x^*) - \|x - x^* - (x^+ - x^*)\|^2] \\
&= f(x^*) + \frac{1}{2\alpha}[2(x - x^+)^\top (x - x^*) + 2(x - x^*)^\top (x^+ - x^*) \\
&\quad - \|x - x^*\|^2 - \|x^+ - x^*\|^2] \\
&= f(x^*) + \frac{1}{2\alpha}(\|x - x^*\|^2 - \|x^+ - x^*\|^2)
\end{aligned}$$

算法定义

- Summing over iterations:

$$\sum_{i=1}^{k}(f(x^i) - f(x^*)) \leq \frac{1}{2\alpha}(\|x^0 - x^*\|^2 - \|x^k - x^*\|^2)$$
$$\leq \frac{1}{2\alpha}\|x^0 - x^*\|^2$$

- Since $f(x^k)$ is non-increasing

$$f(x^k) - f(x^*) \leq \frac{1}{k}\sum_{i=1}^{k}(f(x^i) - f(x^*)) \leq \frac{\|x^0 - x^*\|^2}{2\alpha k}$$

Convergence analysis for backtracking (Optional)

Same assumption, $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, and ∇f is Lipschitz continuous with constant $L > 0$

Same rate for a step size chosen by backtracking search

Theorem

Gradient descent with backtracking line search satis- fies

$$f(x^k) - f(x^*) \leq \frac{\|x^0 - x^*\|^2}{2\alpha_{\min} k}$$

with $\alpha_{\min} = \min\{1, \gamma/L\}$.

If γ is not too small, then we don't lose much compared to fixed step size (γ/L vs $1/L$)

$$f(x - \alpha\nabla f(x)) > f(x) - \frac{\alpha}{2}\|\nabla f(x)\|_2^2,$$

Only have to clarify the minimum stepsize is bounded by some constant. On the other hand, we have for any $\alpha \leq 1/L$,

$$f(x^+) \leq f(x) - \tfrac{1}{2}\alpha\|\nabla f(x)\|_2^2$$

So that the Line Search process terminate with $\alpha \geq \gamma\frac{1}{L}$ or no backtracking happens with $\alpha = 1$. Hence $\alpha_{\min} = \min\{1, \gamma/L\}$.

Linear Convergence for Strongly Convex Cases (Optional)

Strong convexity of f means for some $\mu > 0$,

$$\nabla^2 f(x) \succeq \mu I \quad \text{for any } x$$

Better lower bound than that from usual convexity

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|^2 \quad \forall x, y$$

Under Lipschitz assumption as before, and also strong convexity

Theorem

For L-smooth and μ-strongly convex f, Gradient descent with fixed step size $\alpha \leq 2/(\mu + L)$ satisfies

$$f(x^k) - f(x^*) \leq \left(\frac{L/\mu - 1}{L/\mu + 1}\right)^{2k} \frac{L}{2}\|x^0 - x^*\|^2.$$

Since $\exp(-x) \geq 1 - x$ for every x, we get:

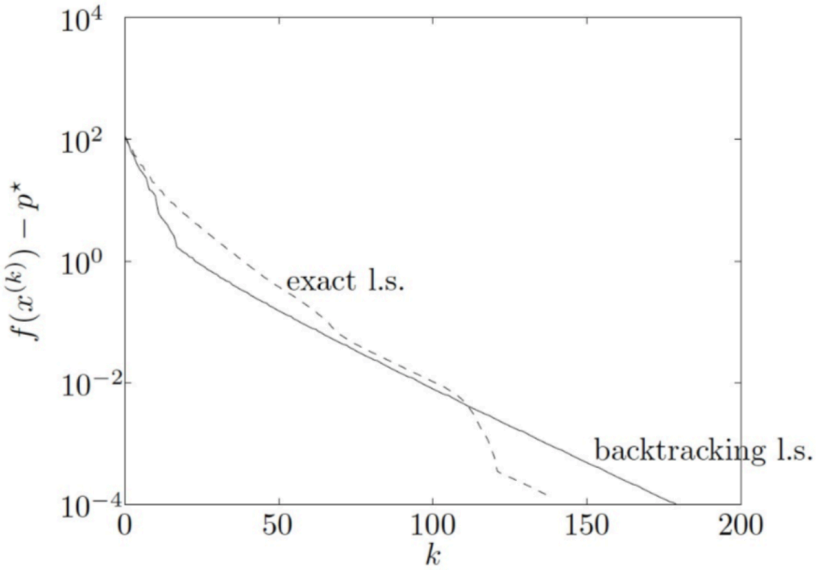
$$\left(\frac{L/\mu - 1}{L/\mu + 1}\right)^2 = \left(1 - \frac{4L\mu}{(L+\mu)^2}\right) \implies \left(\frac{L/\mu - 1}{L/\mu + 1}\right)^2 \leq \exp\left(-\frac{4L\mu}{(L+\mu)^2}\right)$$

$$\implies f(x^k) - f(x^*) \leq \frac{L}{2} \exp\left(-\frac{4L\mu k}{(L+\mu)^2}\right) \|x^0 - x^*\|^2$$

i.e., rate with strong convexity is exponentially fast!

i.e., to get $f(x^k) - f(x^*) \leq \epsilon$, need $O(\log(1/\epsilon))$ iterations

Called linear convergence, because looks linear on a semi-log plot:



Rate constant depends adversely on condition number L/μ (higher condition number) $\implies$ slower rate)

Key Steps of the Proof

Lemma

Let f be L-smooth and μ-strongly convex. Then for all $x, y \in \mathbb{R}^n$, one has

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \frac{\mu L}{\mu + L} \|x - y\|^2 + \frac{1}{\mu + L} \|\nabla f(x) - \nabla f(y)\|^2.$$

Proof.

Note that $\phi(x) := f(x) - \frac{\mu}{2}\|x\|^2$ is convex and $(L - \mu)$-smooth, so that (using a previous lemma)

$$[\nabla \phi(x) - \nabla \phi(y)]^\top (x - y) \geq \frac{1}{L - \mu} \|\nabla \phi(x) - \nabla \phi(y)\|^2 \quad \text{(课本6.2.4)}$$

implying

$$[\nabla f(x) - \mu x - (\nabla f(y) - \mu y)]^\top (x - y) \geq \frac{1}{L - \mu} \|\nabla f(x) - \mu x - (\nabla f(y) - \mu y)\|^2.$$

Rearranging we have the desired result.

□

Key Steps of the Proof

Now we are ready to prove the linear convergence of GD. Indeed recall that smoothness can be defined via the inequality:

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

$$\implies f(x^k) - f(x^*) \leq \nabla f(x^*)^\top (x^k - x^*) + \frac{L}{2}\|x^k - x^*\|^2$$

By strongly convexity,

$$\begin{aligned}
\|x^k - x^*\|^2 &= \|x^{k-1} - \alpha \nabla f(x^{k-1}) - x^*\|^2 \\
&= \|x^{k-1} - x^*\|^2 - 2\alpha \nabla f(x^{k-1})^\top (x^{k-1} - x^*) + \alpha^2 \|\nabla f(x^{k-1})\|^2 \\
&\leq \left(1 - 2\frac{\alpha\mu L}{L+\mu}\right) \|x^{k-1} - x^*\|^2 + \left(\alpha^2 - 2\frac{\alpha}{L+\mu}\right) \|\nabla f(x^{k-1}) - \nabla f(x^*)\|^2 \\
&\leq \left(1 - 2\frac{\alpha\mu L}{L+\mu}\right) \|x^{k-1} - x^*\|^2 + \left(\alpha^2 - 2\frac{\alpha}{L+\mu}\right) L^2 \|x^{k-1} - x^*\|^2 \\
&= \left(\frac{L + \mu - 2\alpha L\mu + \alpha^2 L^2 (L+\mu) - 2\alpha L^2}{L+\mu}\right) \|x^{k-1} - x^*\|^2 \\
&= \left(\frac{\alpha^2 L^2 (L+\mu) - 2\alpha L(L+\mu) + L + \mu}{L+\mu}\right) \|x^{k-1} - x^*\|^2 \\
&= \left(\alpha^2 L^2 - 2\alpha L + 1\right) \|x^{k-1} - x^*\|^2 \\
&= (\alpha L - 1)^2 \|x^{k-1} - x^*\|^2 = 0, \quad \text{if we set } \alpha = 1/L, \text{ so what is wrong??}
\end{aligned}$$

- If $0 \leq \alpha \leq \frac{2}{L+\mu}$,

$$\|x^k - x^*\|^2 \leq \left(1 - \frac{2\alpha\mu L}{L+\mu} + (\alpha^2 - \frac{2\alpha}{L+\mu})L^2\right)\|x^{k-1} - x^*\|^2 = (\alpha L - 1)^2\|x^k - x^*\|^2$$

$$\implies \alpha^* = \min(2/(L+\mu), 1/L) = 2/(L+\mu)$$

- If $\alpha \geq \frac{2}{L+\mu}$,

$$\|x^k - x^*\|^2 \leq \left(1 - \frac{2\alpha\mu L}{L+\mu}\right)\|x^{k-1} - x^*\|^2$$

$$\implies \alpha^* = 2/(L+\mu)$$

-

$$\implies \|x^k - x^*\|^2 \leq \left(\frac{L/\mu - 1}{L/\mu + 1}\right)^2 \|x^{k-1} - x^*\|^2 \quad \text{with } \alpha = 2/(L+\mu)$$

$$\implies \|x^k - x^*\|^2 \leq \left(\frac{L/\mu - 1}{L/\mu + 1}\right)^{2k} \|x^0 - x^*\|^2.$$

$$\implies f(x^k) - f(x^*) \leq \frac{L}{2}\left(\frac{L/\mu - 1}{L/\mu + 1}\right)^{2k} \|x^0 - x^*\|^2$$

- The number L/μ is called the condition number of f.

How about backtracking line search in strongly convex cases? (Optional)

Remember the backtracking line search terminates with $\alpha_{\min} = \min\{1, \gamma/L\}$.

Theorem

For L-smooth and μ-strongly convex f, Gradient descent with backtracking line search attains linear convergence

$$f(x^k) - f(x^*) \leq (1 - \mu\alpha_{\min})^k [f(w^0) - f^*] = (1 - \min\{\tfrac{\mu\gamma}{L}, \mu\})^k [f(w^0) - f^*].$$

Question: in case we have $\min\{\frac{\mu\gamma}{L}, \mu\} = \mu$, do we have $\mu \leq 1$?

$\min\{\frac{\mu\gamma}{L}, \mu\} = \mu \implies \mu \leq \gamma\frac{\mu}{L} \leq \gamma < 1$

How about backtracking line search in strongly convex cases? (Optional)

Alternatively, we have $\alpha_{\min} = \min\{\frac{\gamma}{L}, 1\} \leq 1/L \leq \frac{2}{L+\mu}$. Previously, we have shown that

If $0 \leq \alpha \leq \frac{2}{L+\mu}$,

$$\|x^k - x^*\|^2 \leq \left(1 - \frac{2\alpha\mu L}{L+\mu} + (\alpha^2 - \frac{2\alpha}{L+\mu})L^2\right)\|x^{k-1} - x^*\|^2 = (\alpha L - 1)^2\|x^k - x^*\|^2$$

$$\implies \|x^k - x^*\|^2 \leq (\alpha_{\min} L - 1)^2\|x^0 - x^*\|^2.$$

$$\implies f(x^k) - f(x^*) \leq \frac{L}{2}(\alpha_{\min} L - 1)^{2k}\|x^0 - x^*\|^2$$

If $\alpha_{\min} = 1$, then $\gamma/L \geq 1 \implies L \leq \gamma < 1$, and

$$\implies f(x^k) - f(x^*) \leq \frac{L}{2}(1 - L)^{2k}\|x^0 - x^*\|^2$$

If $\alpha_{\min} = \gamma/L$, then

$$\implies f(x^k) - f(x^*) \leq \frac{L}{2}(1 - \gamma)^{2k}\|x^0 - x^*\|^2$$

Overall, we have

$$\boxed{f(x^k) - f(x^*) \leq \frac{L}{2}\max\left[1 - L, 1 - \gamma\right]^{2k}\|x^0 - x^*\|^2}$$

Practicalities

Pros and cons:

- Pro: simple idea, and each iteration is cheap
- Pro: Very fast for well-conditioned, strongly convex problems
- Con: Often slow, because interesting problems aren't strongly convex or well-conditioned
- Con: can't handle nondifferentiable functions

Acceleration

- There are accelerated gradient methods for strongly-convex functions. They improve the rate to

$$f(w^k) - f^* \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k [f(w^0) - f^*]$$

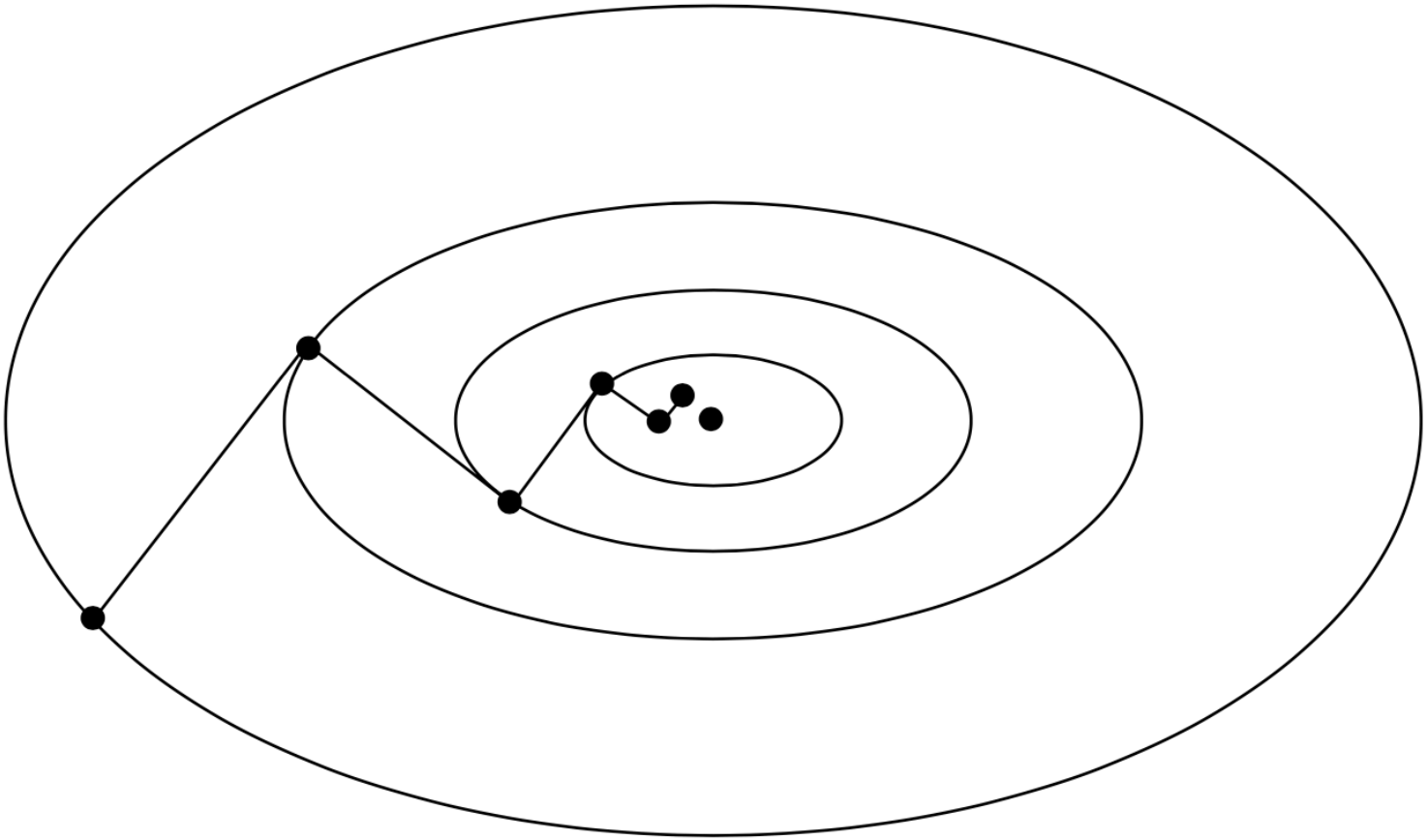
  which is a faster linear convergence rate.

- Alternately, Newton's method achieves superlinear convergence rate.
  - Under strong-convexity and using both ∇f and $\nabla^2 f$ being Lipschitz.
  - But unfortunately this gives a superlinear iteration cost.

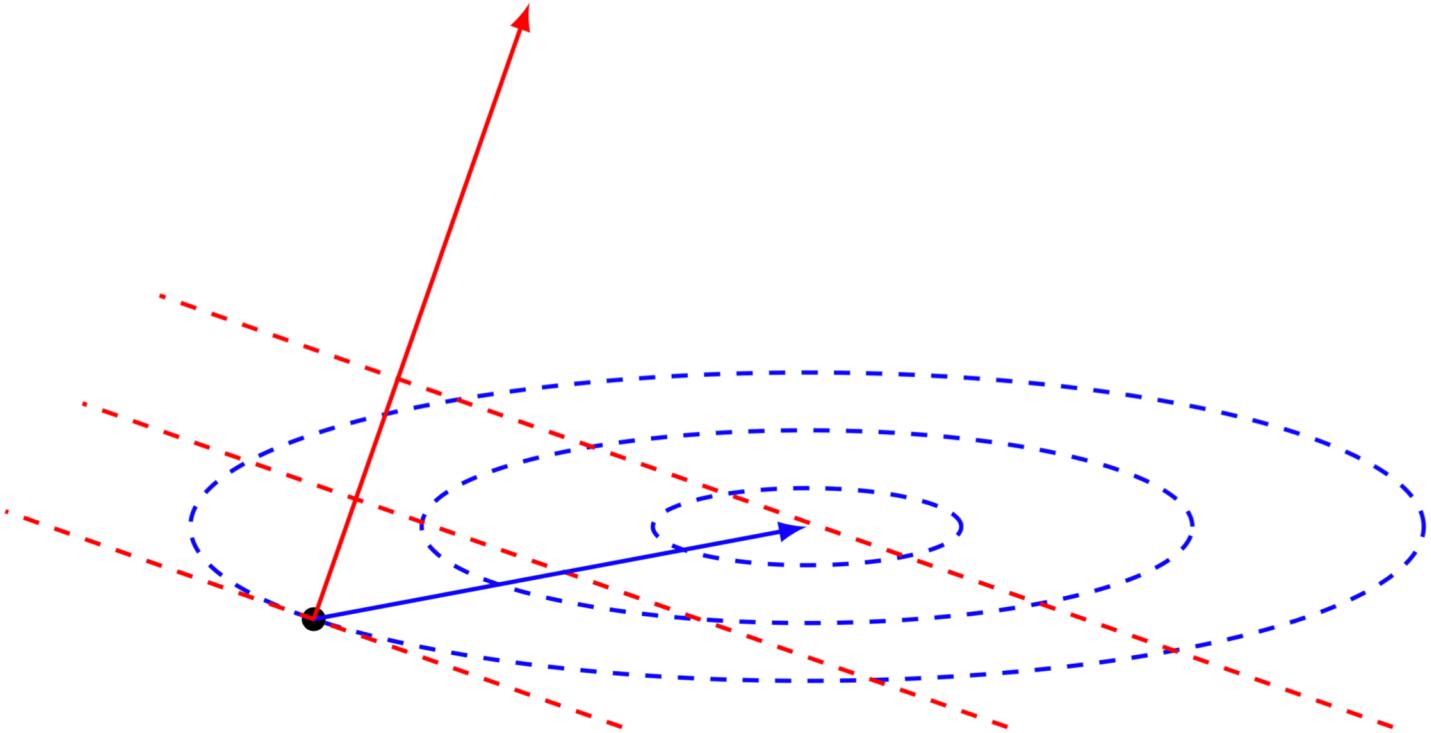
- There are also linear-time approximations to Newton:
  - Barzilai-Borwein step-size for gradient descent.
  - Limited-memory Quasi-Newton methods like L-BFGS.
  - Hessian-free Newton methods.

- Work amazing for many problems, but don't achieve superlinear convergence.

最速下降方向和步长并不是最好的选择



Keep in mind that steepest may not be wise



附录：

Definition (L-smoothness) L-Lipschitz 可微

f is L-smooth (locally L-smooth on $\mathcal{B} \subset \mathbb{R}^n$) in the sense that the gradient mapping ∇f is L-Lipschitz: for any $x, y \in \mathbb{R}^n$ ($\in \mathcal{B}$), one has $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$.

Important property:

Lemma

Let f be a L-smooth function on $\mathbb{R}^n$. Then for any $x, y \in \mathbb{R}^n$, one has

$$|f(x) - f(y) - \nabla f(y)^\top (x - y)| \leq \frac{L}{2}\|x - y\|^2.$$

Proof.

Applying smoothness, and Cauchy-Schwarz, which improves the constant and also get rid of the convexity condition, to represent the quantity $f(x) - f(y)$ as an integral:

$$
\begin{aligned}
& |f(x) - f(y) - \nabla f(y)^\top (x - y)| \\
= & \left| \int_0^1 \nabla f(y + t(x - y))^\top (x - y)\, dt - \nabla f(y)^\top (x - y) \right| \\
= & \left| \int_0^1 [\nabla f(y + t(x - y))^\top - \nabla f(y)]^\top (x - y)\, dt \right| \\
\leq & \left| \int_0^1 L \| t(x - y) \| \| x - y \|\, dt \right| \\
\leq & \int_0^1 L t \| x - y \|^2 dt = \frac{L}{2} \| x - y \|^2 .
\end{aligned}
$$

□

- For $f \in \mathcal{C}^2$, Lipschitz continuity $\iff$ $\| \nabla^2 f(x) \| \leq L$.

Strongly Convexity

Definition (Strongly Convex)

A function $f: \mathbb{R}^n \to \mathbb{R}$ is strongly convex with constant $\mu > 0$ if for all $\{x_1, x_2\} \in \mathcal{X}$ and $\alpha \in [0, 1]$, we have

$$f(\alpha x_1 + (1-\alpha)x_2) + \tfrac{1}{2}\mu\alpha(1-\alpha)\|x_1 - x_2\|_2^2 \leq \alpha f(x_1) + (1-\alpha)f(x_2).$$

- Equivalent definition if f is differentiable and for $\{x_1, x_2\} \in \mathcal{X}$, it holds

$$f(x_1) \geq f(x_2) + \nabla f(x_2)^\top (x_1 - x_2) + \tfrac{1}{2}\mu\|x_1 - x_2\|_2^2.$$

- Equivalent definition if f is differentiable and for $\{x_1, x_2\} \in \mathcal{X}$, it holds

$$(\nabla f(x_1) - \nabla f(x_2))^\top (x_1 - x_2) \geq \mu\|x_1 - x_2\|_2^2.$$

- Equivalent definition if f is twice differentiable and $\forall x \in \mathcal{X}$, it holds

$$\nabla^2 f(x) - \mu I \succeq 0.$$

Convex and L-Smooth f

From previous page, for convex L-smooth f, we have

$$0 \leq f(x) - f(y) - \nabla f(y)^\top (x - y) \leq \frac{L}{2}\|x - y\|^2.$$

Lemma

Let f be a convex and L-smooth function on $\mathbb{R}^n$. Then for any $x, y \in \mathbb{R}^n$, one has

$$f(x) - f(y) - \nabla f(x)^\top (x - y) \leq -\frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2.$$

Lemma

Let f be a convex and L-smooth function on $\mathbb{R}^n$. Then for any $x, y \in \mathbb{R}^n$, one has

$$[\nabla f(x) - \nabla f(x)]^\top (x - y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2.$$

Proof.

Let $\phi(y) = f(y) - \nabla f(x)^\top y$. Note that ϕ is convex and L-smooth. Remark also that x is the minimizer of ϕ since $\phi(y) - \phi(x) \geq \phi(x)^\top (y - x) = 0$. Thus:

L-smooth

x is optimal

$$f(x) - f(y) - \nabla f(x)^\top (x - y) = \phi(x) - \phi(y) \leq \phi\left(y - \frac{1}{L}\nabla\phi(y)\right) - \phi(y)$$

$$\leq \nabla\phi(y)^\top \left(y - \frac{1}{L}\nabla\phi(y) - y\right) + \frac{L}{2}\left\|y - \frac{1}{L}\nabla\phi(y) - y\right\|^2$$

$$= -\frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2,$$

which concludes the proof. □

Proof.

Summing up two inequalities

$$f(x) - f(y) - \nabla f(x)^\top (x - y) \leq -\frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2$$

$$f(y) - f(x) - \nabla f(y)^\top (y - x) \leq -\frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2.$$

we have

$$[\nabla f(x) - \nabla f(x)]^\top (x - y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2.$$

Subgradient Descent Method

"Go there" the Subgradient said.

Subgradient Method

$$x^{k+1} = x^k - \alpha_k g^k, \quad g^k \in \partial f(x^k)$$

其中 $\alpha_k > 0$ 为步长. 它通常有如下四种选择:

(1) 固定步长 $\alpha_k = \alpha$;

(2) 固定 $\|x^{k+1} - x^k\|$, 即 $\alpha_k \|g^k\|$ 为常数;

(3) 消失步长 $\alpha_k \to 0$ 且 $\sum_{k=0}^{\infty} \alpha_k = +\infty$;

(4) 选取 α_k 使其满足某种线搜索准则.

Subgradient method does NOT descend：方向

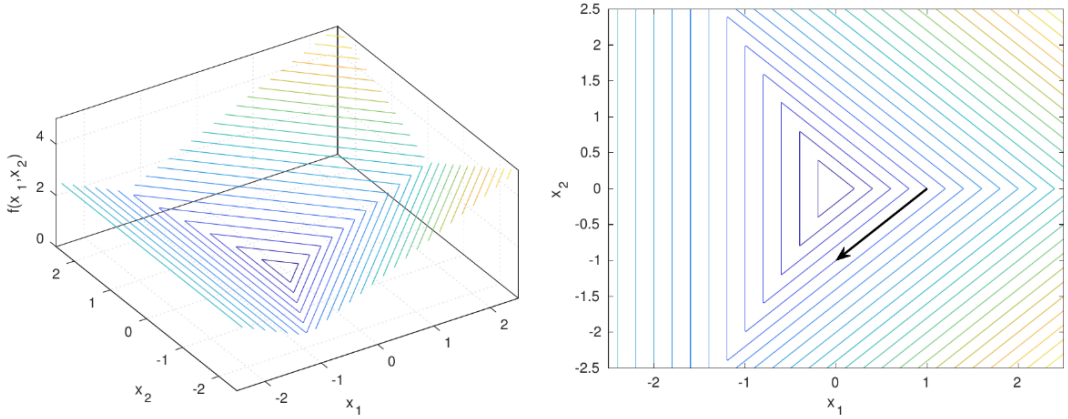


Figure 1. 3D plot (left) and level sets (right) of $f(\boldsymbol{x}) = \max[-x_1, x_1 - x_2, x_1 + x_2]$. The negative subgradient is indicated by the black arrow.

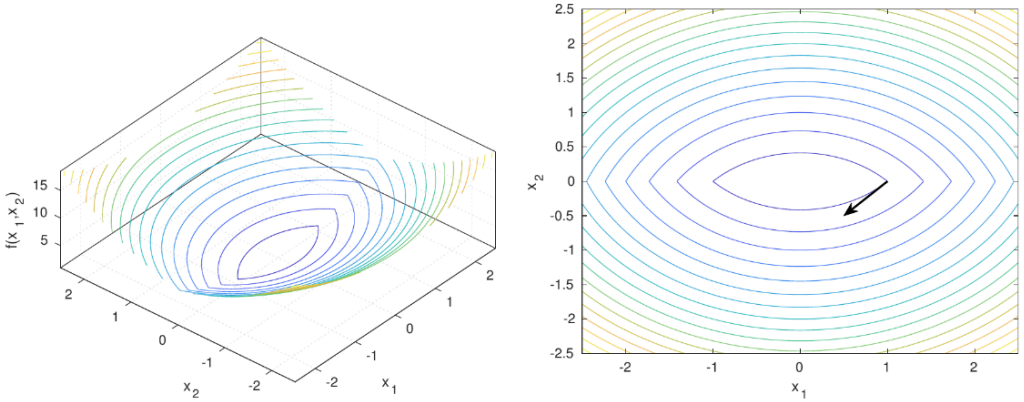


Figure 2. 3D plot (left) and level sets (right) of $f(\boldsymbol{x}) = \max[x_1^2 + (x_2+1)^2, x_1^2 + (x_2-1)^2]$. The negative subgradient is indicated by the black arrow.

Subgradient method does NOT descend：步长

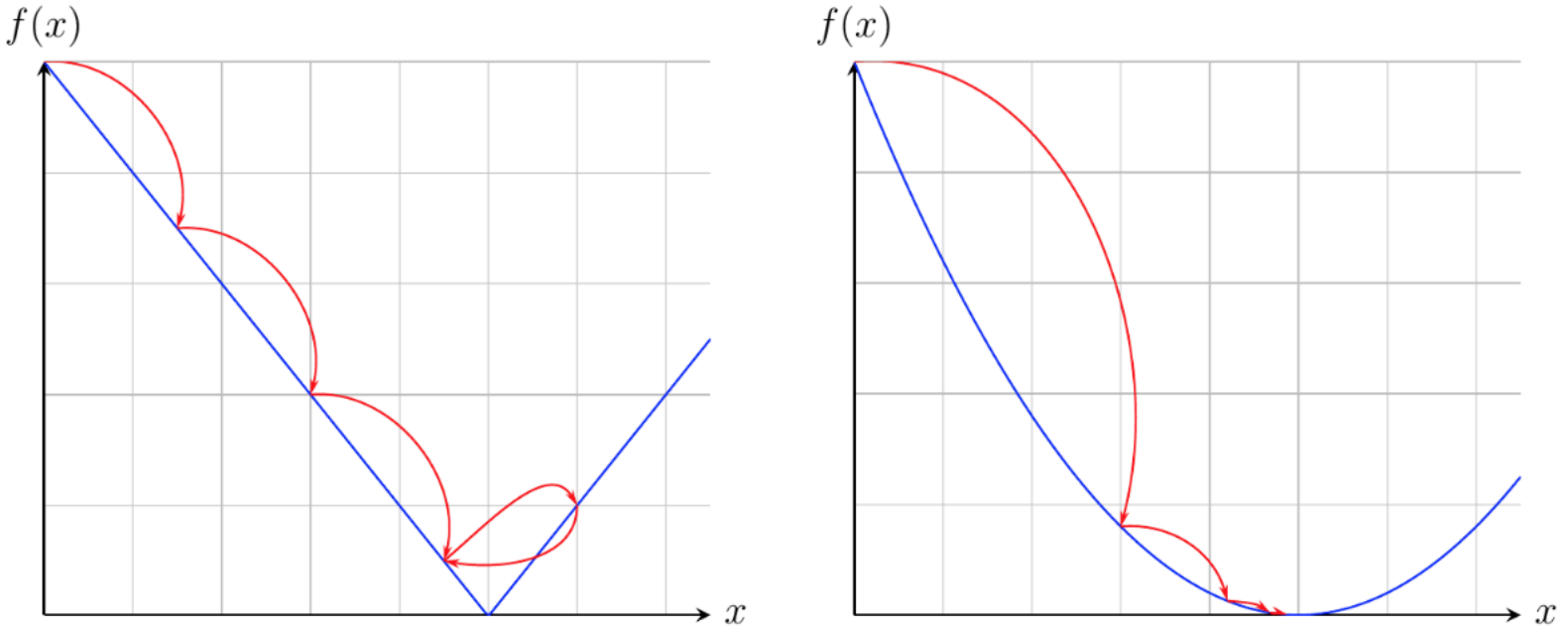


Figure 3. The effect of a constant stepsize on non-differentiable (left) and smooth (right) functions.

收敛性分析

假设 6.1 对无约束优化问题(6.0.1)，目标函数 $f(x)$ 满足：

(1) f 为凸函数；

(2) f 至少存在一个有限的极小值点 x^*，且 $f(x^*) > -\infty$；

(3) f 为利普希茨连续的，即

$$|f(x) - f(y)| \leqslant G\|x - y\|, \quad \forall\, x, y \in \mathbb{R}^n,$$

其中 $G > 0$ 为利普希茨常数.

引理 6.2 设 $f(x)$ 为凸函数，则 $f(x)$ 是 G-利普希茨连续的当且仅当 $f(x)$ 的次梯度是有界的，即

$$\|g\| \leqslant G, \quad \forall\, g \in \partial f(x), x \in \mathbb{R}^n.$$

收敛性分析

定理 6.5 (次梯度算法的收敛性) 在假设6.1的条件下，设 $\{\alpha_k > 0\}$ 为任意步长序列，$\{x^k\}$ 是由算法(6.3.1)产生的迭代序列，则对任意的 $k \geqslant 0$，有

$$2\left(\sum_{i=0}^{k}\alpha_i\right)(\hat{f}^k - f^*) \leqslant \|x^0 - x^*\|^2 + \sum_{i=0}^{k}\alpha_i^2 G^2, \tag{6.3.2}$$

其中 x^* 是 $f(x)$ 的一个全局极小值点，$f^* = f(x^*)$，$\hat{f}^k$ 为前 k 次迭代 $f(x)$ 的最小值，即

$$\hat{f}^k = \min_{0 \leqslant i \leqslant k} f(x^i).$$

收敛性分析

推论 6.2 在假设6.1的条件下，次梯度算法的收敛性满足（$\hat{f}^k$ 的定义和定理6.5中的定义相同）：

(1) 取 $\alpha_i = t$ 为固定步长，则

$$\hat{f}^k - f^* \leqslant \frac{\|x^0 - x^*\|^2}{2kt} + \frac{G^2 t}{2};$$

(2) 取 α_i 使得 $\|x^{i+1} - x^i\|$ 固定，即 $\alpha_i \|g^i\| = s$ 为常数，则

$$\hat{f}^k - f^* \leqslant \frac{G\|x^0 - x^*\|^2}{2ks} + \frac{Gs}{2};$$

(3) 取 α_i 为消失步长，即 $\alpha_i \to 0$ 且 $\sum_{i=0}^{\infty} \alpha_i = +\infty$，则

$$\hat{f}^k - f^* \leqslant \frac{\|x^0 - x^*\|^2 + G^2 \sum_{i=0}^{k} \alpha_i^2}{2\sum_{i=0}^{k} \alpha_i};$$

进一步可得 $\hat{f}^k$ 收敛到 f^*.