



梯度法 I

王尧

西安交通大学智能决策与机器学习中心
(Email: yao.s.wang@gmail.com)

2022. 4

无约束优化问题

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \quad i = 1, \dots, m \\ & Ax = b \end{aligned}$$

can be rewritten as

$$\min_x f(x) \quad \text{subject to} \quad x \in C$$

where $C = \{x : g_i(x) \leq 0, i = 1, \dots, m, Ax = b\}$, the feasible set. Hence the latter formulation is **completely general**

Important special case: if $C = \mathbb{R}^n$ (unconstrained optimization)

无约束可微问题

无约束可微优化问题通常表示为如下形式：

$$\min_{x \in \mathbb{R}^n} f(x),$$

其中 f 是连续可微函数.

- 给定一个点 \bar{x} , 我们想要知道这个点是否是函数 f 的一个局部极小解或者全局极小解.

一阶必要条件

定理 (一阶必要条件)

假设 f 在全空间 \mathbb{R}^n 可微. 如果 x^* 是一个局部极小点, 那么

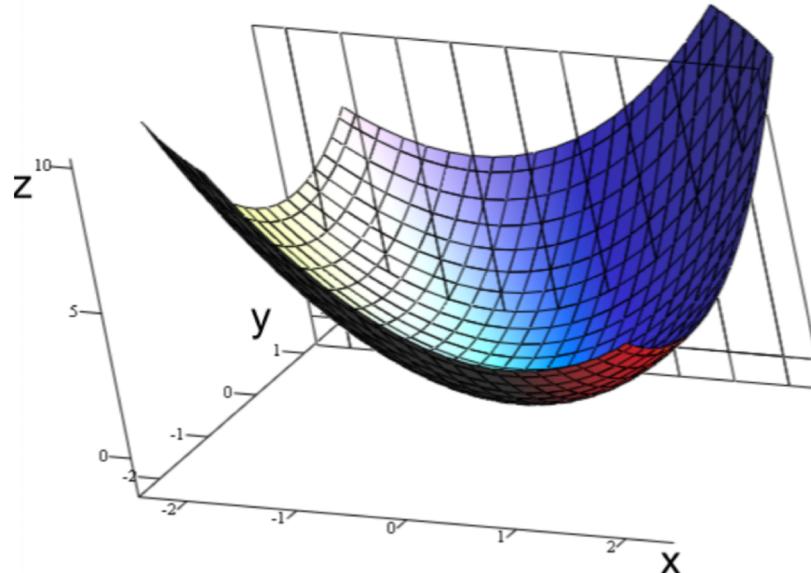
$$\nabla f(x^*) = 0.$$

- 对于 $f(x) = x^2, x \in \mathbb{R}$, 我们知道满足 $f'(x) = 0$ 的点为 $x^* = 0$, 并且其也是全局最优解.
- 对于 $f(x) = x^3, x \in \mathbb{R}$, 满足 $f'(x) = 0$ 的点为 $x^* = 0$, 但其不是一个局部最优解.
- 称满足 $\nabla f(x) = 0$ 的点 x 为 f 的稳定点(有时也称为驻点或临界点).
- 除了一阶必要条件, 还需要对函数加一些额外的限制条件, 才能保证最优解的充分性.

注: 若目标函数为凸, 则上述必要条件为充要条件

课后作业: 请完成上述一阶必要性定理的证明

全局和局部最优解



For convex optimization problems, **local minima are global minima**

Formally, if x is feasible— $x \in D$, and satisfies all constraints—and minimizes f in a local neighborhood,

$$f(x) \leq f(y) \text{ for all feasible } y, \|x - y\|_2 \leq \rho,$$

then

$$f(x) \leq f(y) \text{ for all feasible } y$$

迭代下降算法

定义 (下降方向)

对于可微函数 f 和点 $x \in \mathbb{R}^n$, 如果存在向量 d 满足

$$\nabla f(x)^T d < 0,$$

那么称 d 为 f 在点 x 处的一个下降方向.

General descent method.

given a starting point $x \in \text{dom } f$.

repeat

1. Determine a descent direction Δx .
2. *Line search.* Choose a step size $t > 0$.
3. *Update.* $x := x + t\Delta x$.

until stopping criterion is satisfied.

梯度下降法

- 注意到 $\phi(\alpha) = f(x^k + \alpha d^k)$ 有泰勒展开

$$\phi(\alpha) = f(x^k) + \alpha \nabla f(x^k)^T d^k + \mathcal{O}(\alpha^2 \|d^k\|^2).$$

- 由柯西不等式, 当 α 足够小时取 $d^k = -\nabla f(x^k)$ 会使函数下降最快.
- 因此梯度法就是选取 $d^k = -\nabla f(x^k)$ 的算法, 它的迭代格式为

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

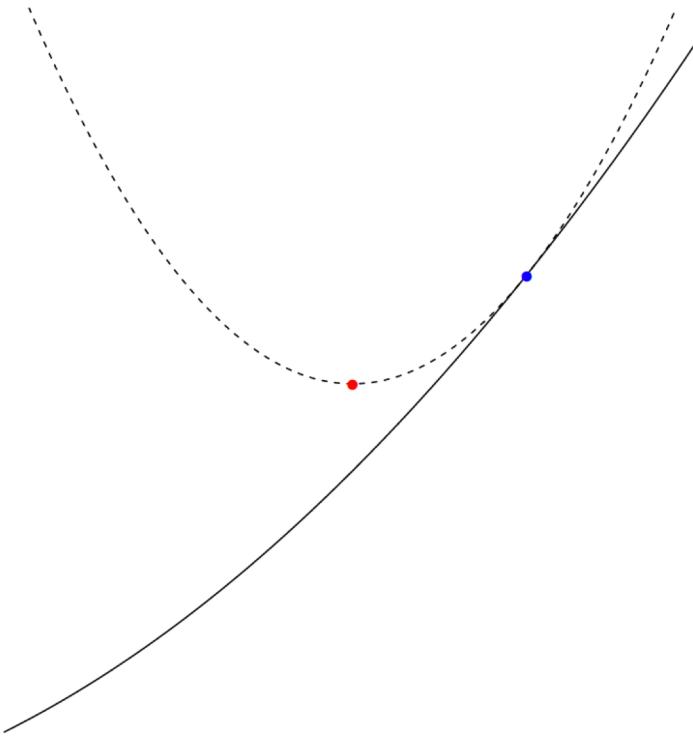
步长 α_k 的选取可依赖于线搜索算法, 也可直接选取固定的 α_k .

梯度下降的另一种理解

$$\begin{aligned}x^{k+1} &= \arg \min_x f(x^k) + \nabla f(x^k)^\top (x - x^k) + \frac{1}{\alpha_k} \|x - x^k\|_2^2 \\&= \arg \min_x \|x - (x^k - \alpha_k \nabla f(x^k))\|_2^2 \\&= x^k - \alpha_k \nabla f(x^k)\end{aligned}$$

核心思想为在目标函数的二阶Tylor展开中，用二次项替换Hessian

梯度下降的另一种理解

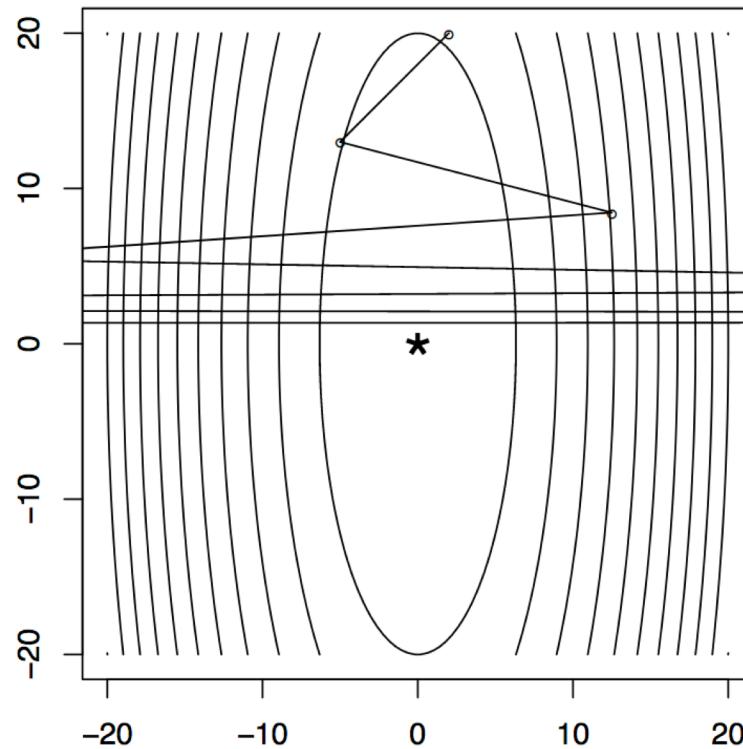


Blue point is x , red point is

$$x^+ = \underset{y}{\operatorname{argmin}} \ f(x) + \nabla f(x)^T(y - x) + \frac{1}{2t} \|y - x\|_2^2$$

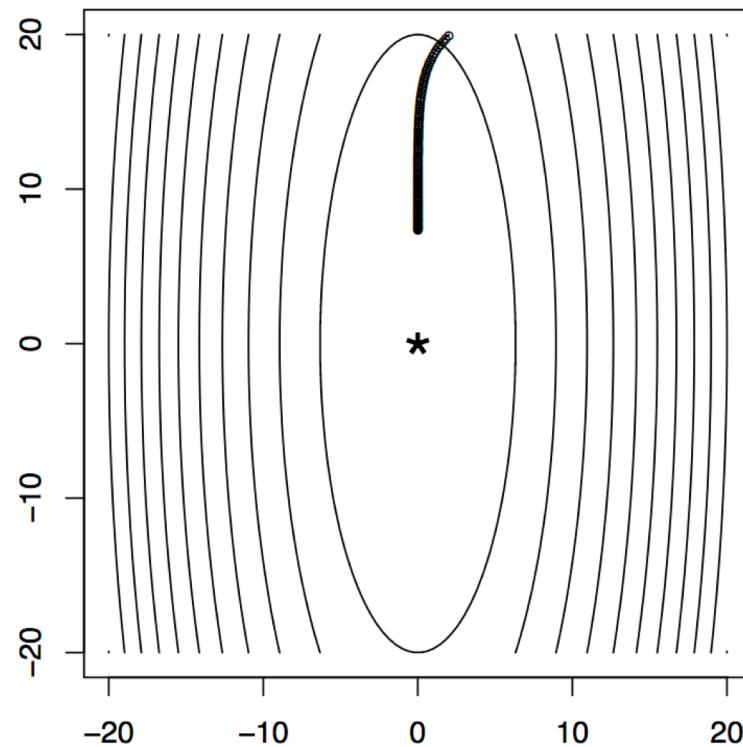
固定步长

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



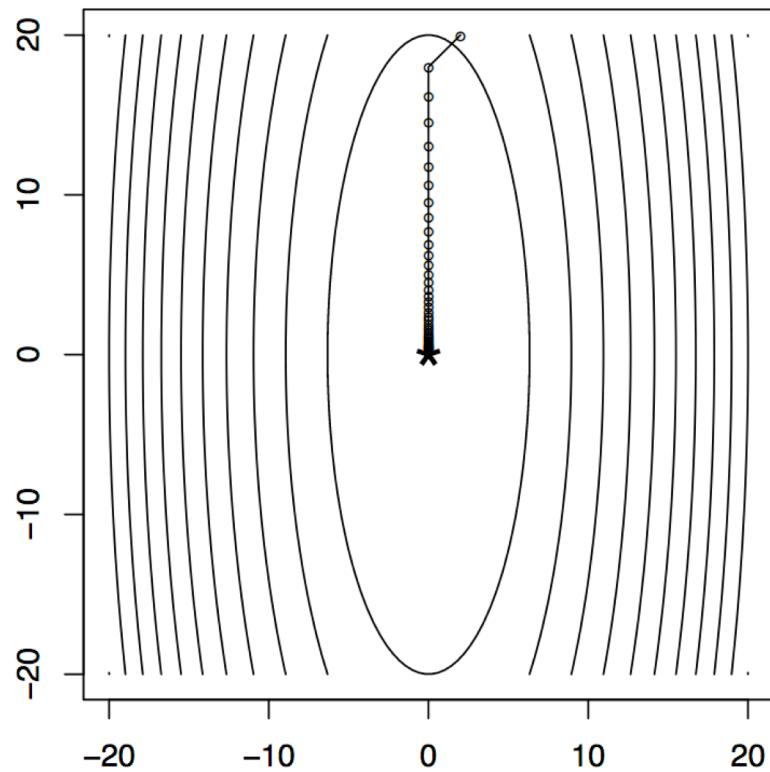
固定步长

Can be **slow** if t is too small. Same example, gradient descent after 100 steps:



固定步长

Converges nicely when t is “just right”. Same example, 40 steps:



梯度利普希茨连续

定义 (梯度利普希茨连续)

给定可微函数 f , 若存在 $L > 0$, 对任意的 $x, y \in \text{dom}f$ 有

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad (3)$$

则称 f 是梯度利普希茨连续的, 相应利普希茨常数为 L . 有时也简记为梯度 L -利普希茨连续或 L -光滑.

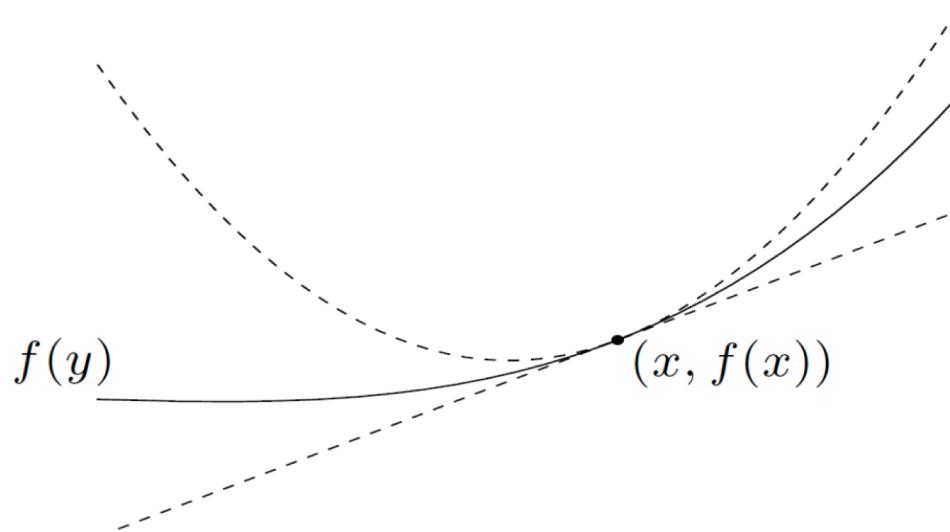
- functions f with this property are also called *L -smooth*
- the definition does not assume convexity of f (and holds for $-f$ if it holds for f)

二次上界

suppose ∇f is Lipschitz continuous with parameter L and $\text{dom } f$ is convex

- Then $g(x) = (L/2)x^\top x - f(x)$, with $\text{dom } g$, is convex
- convexity of g is equivalent to a quadratic upper bound on f :

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \|y - x\|_2^2 \quad \forall x, y \in \text{dom } f$$



二次上界的证明

Proof.

- Lipschitz continuity of ∇f and Cauchy-Schwarz inequality imply

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \leq L\|x - y\|_2^2 \quad \forall x, y \in \mathbf{dom} f$$

this is monotonicity of the gradient $\nabla g(x) = Lx - \nabla f(x)$

- hence, g is a convex function if its domain $\mathbf{dom} g = \mathbf{dom} f$
- the quadratic upper bound is the first-order condition for the convexity of g

$$g(y) \geq g(x) + \nabla g(x)^\top (y - x) \quad \forall x, y \in \mathbf{dom} g$$



在凸函数上的收敛性

考虑梯度法

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k)$$

假设：

- 设函数 $f(x)$ 为凸的梯度 L -利普希茨连续函数
- 极小值 $f^* = f(x^*) = \inf_x f(x)$ 存在且可达.
- 如果步长 α_k 取为常数 α 且满足 $0 < \alpha < \frac{1}{L}$

结论：点列 $\{x^k\}$ 的函数值收敛到最优值，且在函数值的意义下收敛速度为 $\mathcal{O}\left(\frac{1}{k}\right)$.

证明

- 因为函数 f 是利普希茨可微函数, 对任意的 x , 根据二次上界引理,

$$f(x - \alpha \nabla f(x)) \leq f(x) - \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla f(x)\|^2.$$

- 记 $\tilde{x} = x - \alpha \nabla f(x)$ 并限制 $0 < \alpha < \frac{1}{L}$, 我们有

$$\begin{aligned} f(\tilde{x}) &\leq f(x) - \frac{\alpha}{2} \|\nabla f(x)\|^2 \\ &\leq f^* + \nabla f(x)^T (x - x^*) - \frac{\alpha}{2} \|\nabla f(x)\|^2 \\ &= f^* + \frac{1}{2\alpha} (\|x - x^*\|^2 - \|x - x^* - \alpha \nabla f(x)\|^2) \\ &= f^* + \frac{1}{2\alpha} (\|x - x^*\|^2 - \|\tilde{x} - x^*\|^2), \end{aligned}$$

其中第一个不等式是因为 $0 < \alpha < \frac{1}{L}$, 第二个不等式为 f 的凸性.

证明

- 在上式中取 $x = x^{i-1}$, $\tilde{x} = x^i$ 并将不等式对 $i = 1, 2, \dots, k$ 求和得到

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

- 由于 $f(x^i)$ 是非增的, 所以

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

*Thank you for your
attentions !*