梯度法II

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. 5

回顾：梯度下降法

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

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

- 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.

二次函数的梯度法

设二次函数$f(x,y)=x^2+10y^2$, 初始点(x^0,y^0) 取为$(10,1)$, 取固定步长$\alpha_k=0.085$. 我们使用梯度法$x^{k+1}=x^k-\alpha_k\nabla f(x^k)$ 进行15 次迭代.

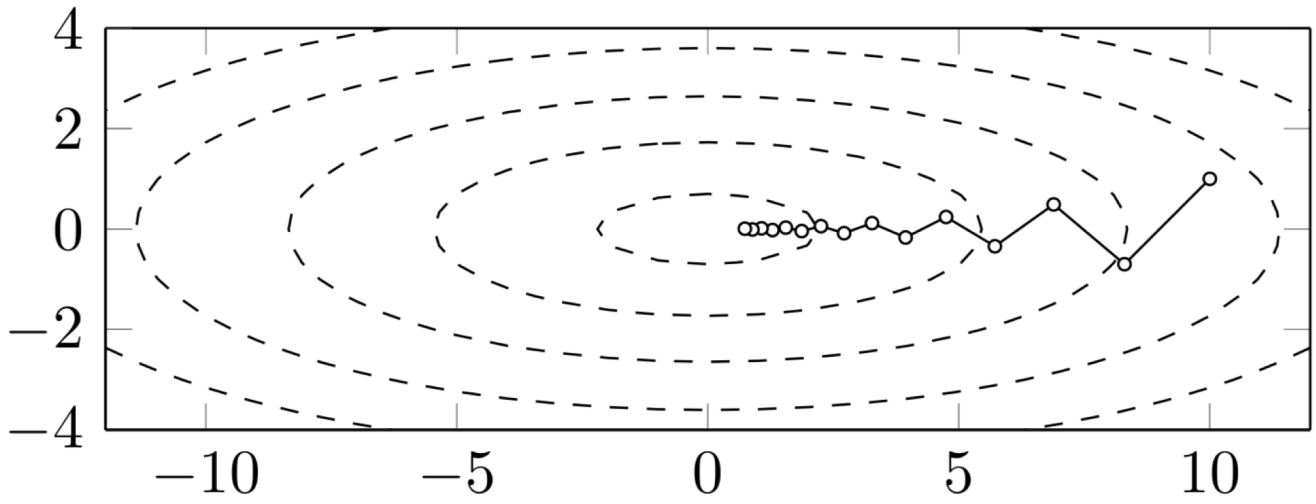


Figure: 梯度法的前15 次迭代

回顾：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$\mathcal{O}\left(\frac{1}{k}\right)$.

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

精确线性搜索

- 首先构造一元辅助函数

$$\phi(\alpha) = f(x^k + \alpha d^k),$$

其中d^k 是给定的下降方向, $\alpha > 0$ 是该辅助函数的自变量.

- 线搜索的目标是选取合适的α_k 使得$\phi(\alpha_k)$ 尽可能减小. 这要求:
  1. α_k 应该使得f 充分下降
  2. 不应在寻找α_k 上花费过多的计算量
- 一个自然的想法是寻找α_k 使得

$$\alpha_k = \underset{\alpha>0}{\operatorname{argmin}}\, \phi(\alpha),$$

即α_k 为最佳步长. 这种线搜索算法被称为**精确线搜索算法**

通常需要较大计算量, 在实际应用中并不常用

后退线性搜索

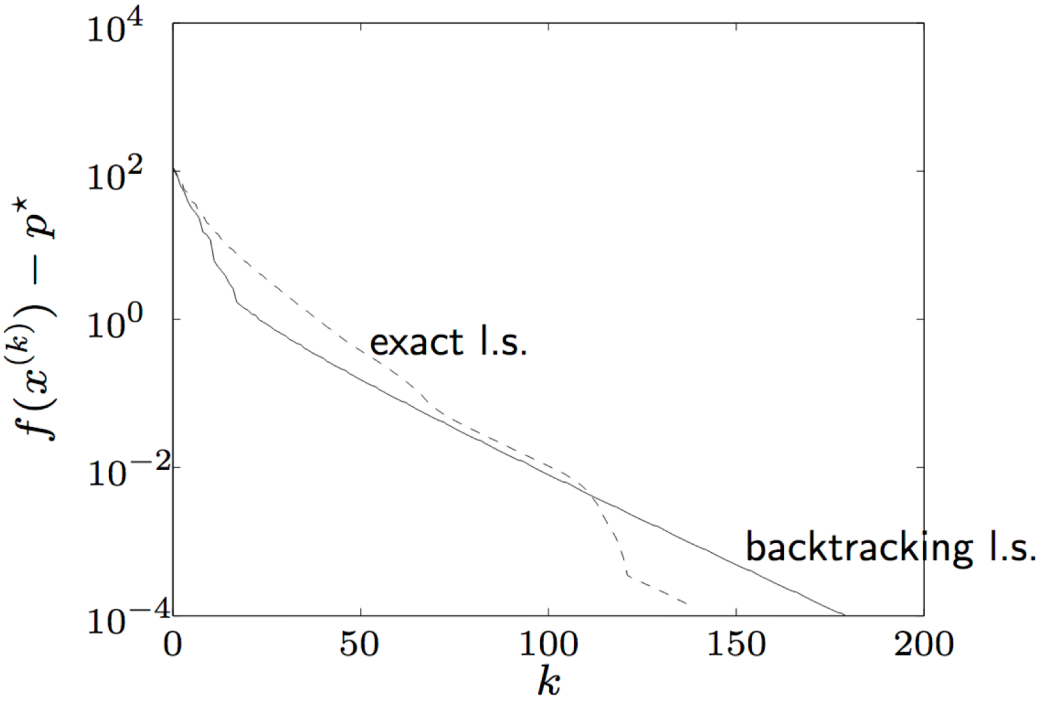
Algorithm 9.2 *Backtracking line search.*

given a descent direction Δx for f at $x \in \mathbf{dom}\, f$, $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$.
$t := 1$.
while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$, $\quad t := \beta t$.

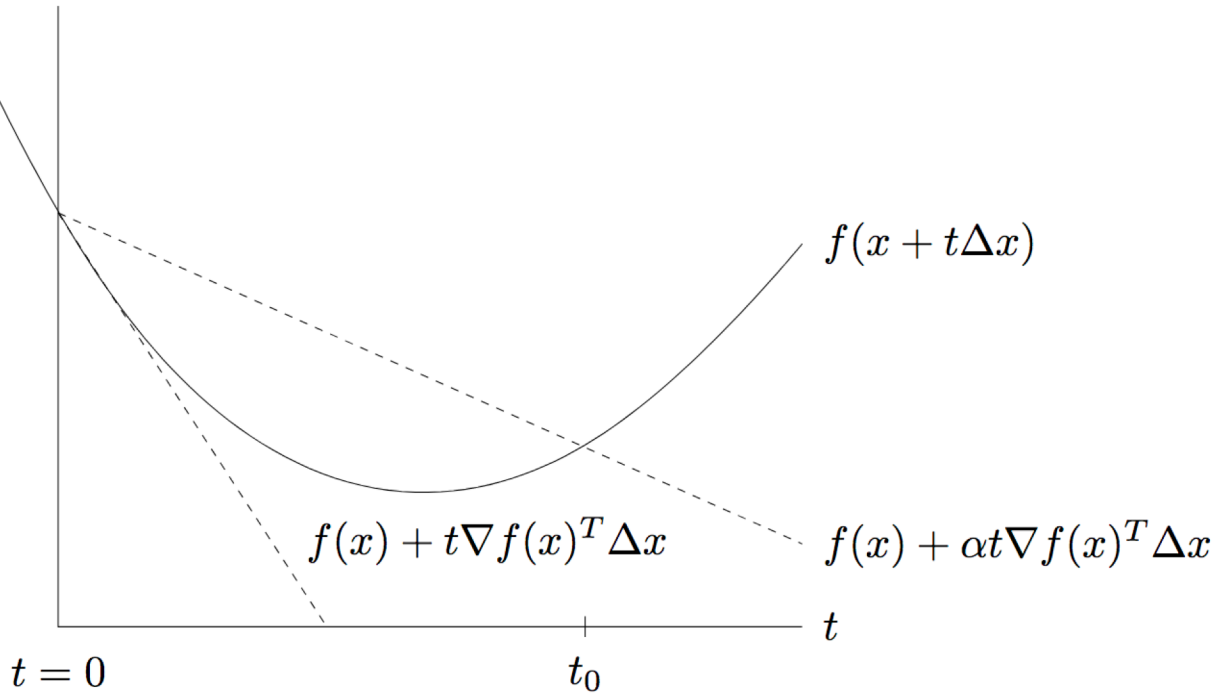
最常用的一类不精确搜索策略，更详细介绍见boyd版9.2与9.3节

a problem in $\mathbf{R}^{100}$

$$f(x) = c^T x - \sum_{i=1}^{500} \log(b_i - a_i^T x)$$



后退法的几何解释



For us $\Delta x = -\nabla f(x)$

对充分小的t, $f(x+t\Delta x) \approx f(x)+t\nabla f(x)^T\Delta x < f(x)+\alpha t\nabla f(x)^T\Delta x$

关于alpha取值的讨论请阅读Boyd版9.3节

强凸函数

f is strongly convex with parameter $m > 0$ if

$$g(x) = f(x) - \frac{m}{2}x^\top x \quad \text{is convex}$$

Jensen's inequality: Jensen's inequality for g is

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y) - \frac{m}{2}\theta(1-\theta)\|x-y\|_2^2$$

monotonicity: monotonicity of ∇g gives

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

this is called *strong monotonicity*(*covercivity*) of ∇f

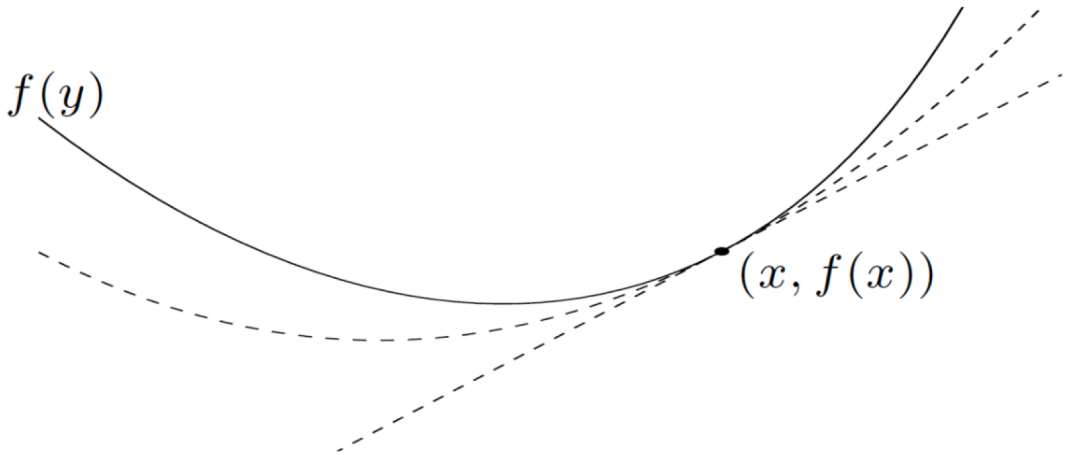
second-order condition: $\nabla^2 f(x) \succeq mI$ for all $x \in \mathbf{dom}\, f$

课后作业：完成上述强凸函数相互间等价性的证明

强凸函数的二次下界

form 1st order condition of convexity of g:

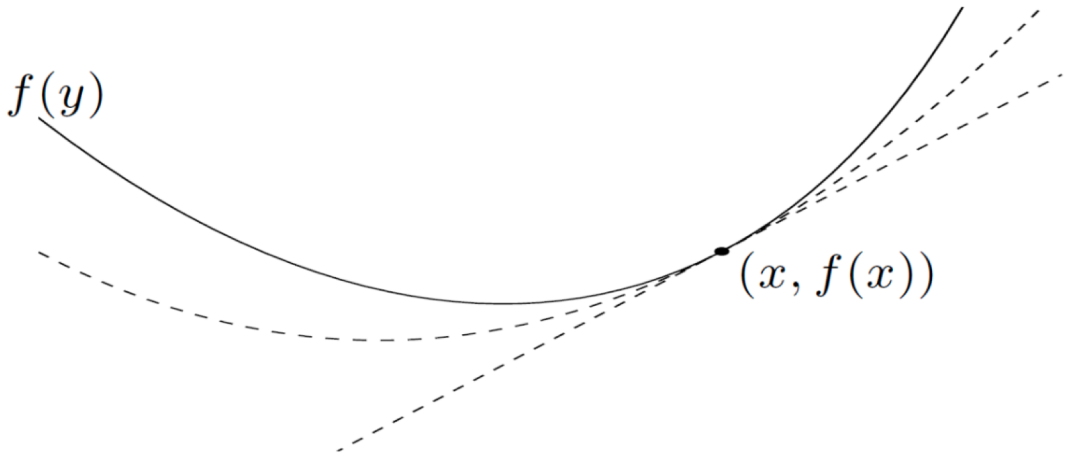
$$f(y) \geq f(x) + \nabla f(x)^{\top}(y-x) + \frac{m}{2}\|y-x\|_2^2 \quad \forall x, y \in \mathbf{dom}\, f$$



强凸函数的二次下界

form 1st order condition of convexity of g:

$$f(y) \geq f(x) + \nabla f(x)^{\top}(y-x) + \frac{m}{2}\|y-x\|_2^2 \quad \forall x, y \in \mathbf{dom}\, f$$



课堂作业：Let h(x)=f(x)+g(x) where f(x) is strongly convex function and g(x) is a convex function, then h(x) is strongly convex function.

二次下界的意义

$$
\begin{aligned}
f(y) &\geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2}\|y - x\|_2^2 \\
&\geq f(x) + \nabla f(x)^T (\tilde{y} - x) + \frac{m}{2}\|\tilde{y} - x\|_2^2 \\
&= f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2.
\end{aligned}
$$

这里： $\tilde{y} = x - (1/m)\nabla f(x)$

故 $p^\star \geq f(x) - \frac{1}{2m}\|\nabla f(x)\|_2^2$

If the gradient is small at a point, then the point is nearly optimal

回顾：梯度利普希茨连续

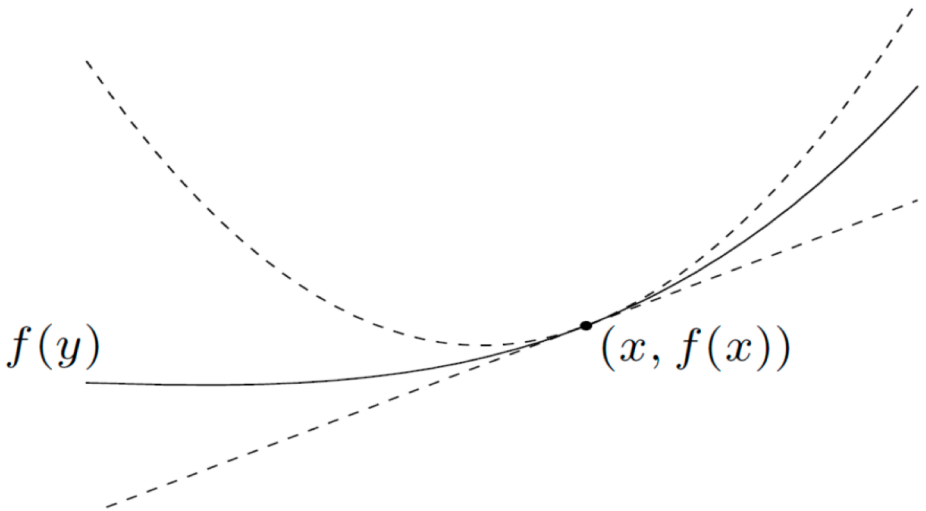
定义 (梯度利普希茨连续)

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

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \tag{3}$$

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**.

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



强凸情形的收敛性

Reminder: strong convexity of f means $f(x) - \frac{m}{2}\|x\|_2^2$ is convex for some $m > 0$

Assuming Lipschitz gradient as before, and also strong convexity:

Theorem: Gradient descent with fixed step size $t \leq 2/(m+L)$ or with backtracking line search search satisfies

$$f(x^{(k)}) - f^\star \leq \gamma^k \frac{L}{2}\|x^{(0)} - x^\star\|_2^2$$

where $0 < \gamma < 1$

Rate under strong convexity is $O(\gamma^k)$, exponentially fast! That is, it finds ϵ-suboptimal point in $O(\log(1/\epsilon))$ iterations

详细证明参见文再文版引理6.1与定理6.4

一些注记

Stopping rule: stop when $\|\nabla f(x)\|_2$ is small

- Recall $\nabla f(x^\star) = 0$ at solution $x^\star$

Pros and cons of gradient descent:

- Pro: simple idea, and each iteration is cheap (usually)
- Pro: fast for well-conditioned, strongly convex problems
- Con: can often be slow, because many interesting problems aren't strongly convex or well-conditioned
- Con: can't handle nondifferentiable functions

不可微的例子

- ℓ_1 regularized minimization

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) + \underbrace{\|\boldsymbol{x}\|_1}_{h(\boldsymbol{x}):\, \ell_1 \text{ norm}}$$

  - use ℓ_1 regularization to promote sparsity

- nuclear norm regularized minimization

$$\text{minimize}_{\boldsymbol{X}} \quad f(\boldsymbol{X}) + \underbrace{\|\boldsymbol{X}\|_*}_{h(\boldsymbol{X}):\, \text{nuclear norm}}$$

  - use nuclear norm regularization to promote low-rank structure

课程作业：多元线性回归

Least-squares

$$\min_x \frac{1}{2}\|Ax-b\|_2^2$$

数据集：

1. Advertising: https://github.com/rghan/ISLR/blob/master/Advertising.csv
2. Auto MPG: https://archive.ics.uci.edu/ml/datasets/Auto+MPG

作业要求：

1. 分别在上述两个数据集合上运用正规方程法与梯度下降法求解相应的最小二乘问题，并进行结果对比；
2. 梯度下降法需采用固定步长(注意步长的选择范围!)与后退法两类策略；
3. 用latex撰写详细的实验报告，以小组形式于5.20号之前提交，提交内容包括tex源文件与算法实现代码。

Thank you for your attentions !