DCS440 最优化理论

第三章：无约束优化算法

杨 磊

yanglei39@mail.sysu.edu.cn

计算机学院，2023 秋

致谢：本课件由朱嘉懿、谷曜东协助准备

- 线搜索
- 梯度类算法
- 梯度下降法的扩展
- 二阶算法

引言

在本章，我们将从连续可微无约束优化问题出发，认识并学习优化算法。

考虑连续可微无约束优化问题

$$\min_x \quad f_0(x),$$

其中 $f_0(x) : \mathbb{R}^n \to \mathbb{R}$ 在 $\operatorname{dom} f_0 = \mathbb{R}^n$ 上连续可微，且 $f_0(x)$ 的极小值 p^* 和最优解 x^* 存在。

我们将求解 $f_0(x)$ 的最小值点 x^* 的过程想象为“旅者下山”问题：

- 假设旅者当前处于某点 x 处，$f_0(x)$ 表示此地的高度；
- 为了到达山底，旅者在当前点 x 处需思考两件事：

  1. 下一步该向哪个方向走可以下山？（搜索方向）

  2. 沿该方向走多远后停下，以便选取下一个下山方向？（迭代步长）
- 沿上述确定方向行走，并重复以上思考，直至到达 $f(x)$ 的最小值点。

回顾（大多数）迭代算法的基本框架：

- **步 1.** 取初始点 $x^0 \in \mathbb{R}^n$ 及其他有关参数，令 $k = 0$。
- **步 2.** 验证停机准则。
- **步 3.** 求 x^k 点处的搜索方向 $d^k \in \mathbb{R}^n$。通常，我们要求 d^k 是下降方向，即 $\left(\nabla f_0(x^k)\right)^\top d^k < 0$。
- **步 4.** 计算迭代步长 $\alpha_k \geq 0$，通常要求 $f_0\left(x^k + \alpha_k d^k\right) < f_0\left(x^k\right)$。
- **步 5.** 产生下一迭代点：令 $x^{k+1} := x^k + \alpha_k d^k$，$k := k + 1$，转**步 2**。

上述基本框架的思想：先确定搜索方向 $\longrightarrow$ 后搜寻迭代步长

- 搜索方向 d^k 有着不同选取方法：梯度类方向、牛顿类方向等。有时还可能没有显示形式，需根据实际情况而定。
- 迭代步长 α_k 的计算本质上归结为一个一维问题，因此其寻找过程往往称作线搜索 (line search)。具体如下：

精确线搜索

考虑辅助函数

$$\phi(\alpha) := f_0\left(x^k + \alpha d^k\right).$$

容易观察到，寻找合适的迭代步长使目标函数值下降实际就是将目标函数 $f_0(x)$ 限制在射线 $\left\{x^k + \alpha d^k \mid \alpha > 0\right\}$ 上，以寻找 α_k 使 $\phi(\alpha_k) < \phi(0)$。

线搜索的目标：选取合适的 α_k，使得 $\phi(\alpha_k)$ 尽可能小！

一个自然的想法是考虑“最优”策略：

$$\alpha^k = \arg\min_{\alpha \geq 0} \ \phi(\alpha) \qquad \longleftarrow \qquad \begin{pmatrix} \text{含有非负约束} \\ \text{若 } f_0 \text{ 凸，则该问题为凸} \end{pmatrix}$$

此时 α_k 为最优步长，该策略也被称作**精确线搜索** (exact line search)。

注意到：

- 当 $\phi(\alpha)$ 较简单时，可有效地计算最优步长 α_k；
- 但实际中，$\phi(\alpha)$ 往往较复杂，导致精确线搜索的计算代价高。

*精确线搜索：二分法

假设 $\phi(\alpha)$ 在区间 $[0,\infty)$ 上是下单峰函数，即在 $(0,\infty)$ 内有唯一极小点 α^*，在 α^* 的左侧 $\phi(\alpha)$ 严格下降，在 α^* 的右侧 $\phi(\alpha)$ 严格上升。

当 $\phi(\alpha)$ 连续可微时，可通过二分法求解方程 $\phi'(\alpha)=0$，得到最优步长 α^k：

初始化：令 $a=0$，并任取 $b>0$ 得到区间 $[a,b]$，使得 $\phi'(a)$ 和 $\phi'(b)$ 正负符号相异；重复以下步骤：

- **步 1.** 取区间 $[a,b]$ 的中点，计算 $\phi'\left(\frac{a+b}{2}\right)$。若 $\left|\phi'\left(\frac{a+b}{2}\right)\right|$ 满足精度要求，则 $\alpha^*=\frac{a+b}{2}$，终止迭代；否则，转**步 2**；
- **步 2.** 考虑如下情形：
  - 若 $\phi'\left(\frac{a+b}{2}\right)$ 和 $\phi'(a)$ 同号，令 $a:=\frac{a+b}{2}$（左端点右移）；
  - 若 $\phi'\left(\frac{a+b}{2}\right)$ 和 $\phi'(b)$ 同号，令 $b:=\frac{a+b}{2}$（右端点左移）；

  得到新的区间 $[a,b]$，转**步 1**；

*精确线搜索：黄金分割法

假设 $\phi(\alpha)$ 在区间 $[0,\infty)$ 上是下单峰函数，即在 $(0,\infty)$ 内有唯一极小点 α^*，在 α^* 的左侧 $\phi(\alpha)$ 严格下降，在 α^* 的右侧 $\phi(\alpha)$ 严格上升。

当 $\phi(\alpha)$ 不易求导时，可采用**黄金分割法（或 0.618 法）**搜索最优步长 α^k：

初始化：令 $a=0$，并任取 $b>0$ 得到区间 $[a,b]$，使得最优解包含于其中；
重复以下步骤：

- **步 1.** 令 $x_2=a+0.618(b-a)$，$\phi_2=\phi(x_2)$；
- **步 2.** 令 $x_1=a+0.382(b-a)$，$\phi_1=\phi(x_1)$；
- **步 3.** 若 $|b-a|$ 满足精度要求，则 $\alpha^*=\frac{a+b}{2}$，终止迭代；否则，转**步 4**；
- **步 4.** 考虑如下情形：
  - 若 $\phi_1<\phi_2$，则令 $b=x_2$，$x_2=x_1$，$\phi_2=\phi_1$，转**步 2**；（右端点左移）
  - 若 $\phi_1=\phi_2$，则令 $a=x_1$，$b=x_2$，转**步 1**；
  - 若 $\phi_1>\phi_2$，则令 $a=x_1$，$x_1=x_2$，$\phi_1=\phi_2$，转**步 5**；（左端点右移）
- **步 5.** 令 $x_2=a+0.618(b-a)$，$\phi_2=\phi(x_2)$，转**步 3**；

非精确线搜索

实际中，精确线搜索一般计算代价高，因此人们往往考虑非精确地求解 $\min\limits_{\alpha\geq 0}\left\{f_0\left(x^k+\alpha d^k\right)\right\}$ 来寻找迭代步长 α_k，使之满足某些准则即可，这一过程称为**非精确线搜索**。

下面介绍一些常用的线搜索准则。

Armijo 准则 (Armijo rule)

Armijo rule 是最常用的线搜索准则之一，它要求 α_k 满足：

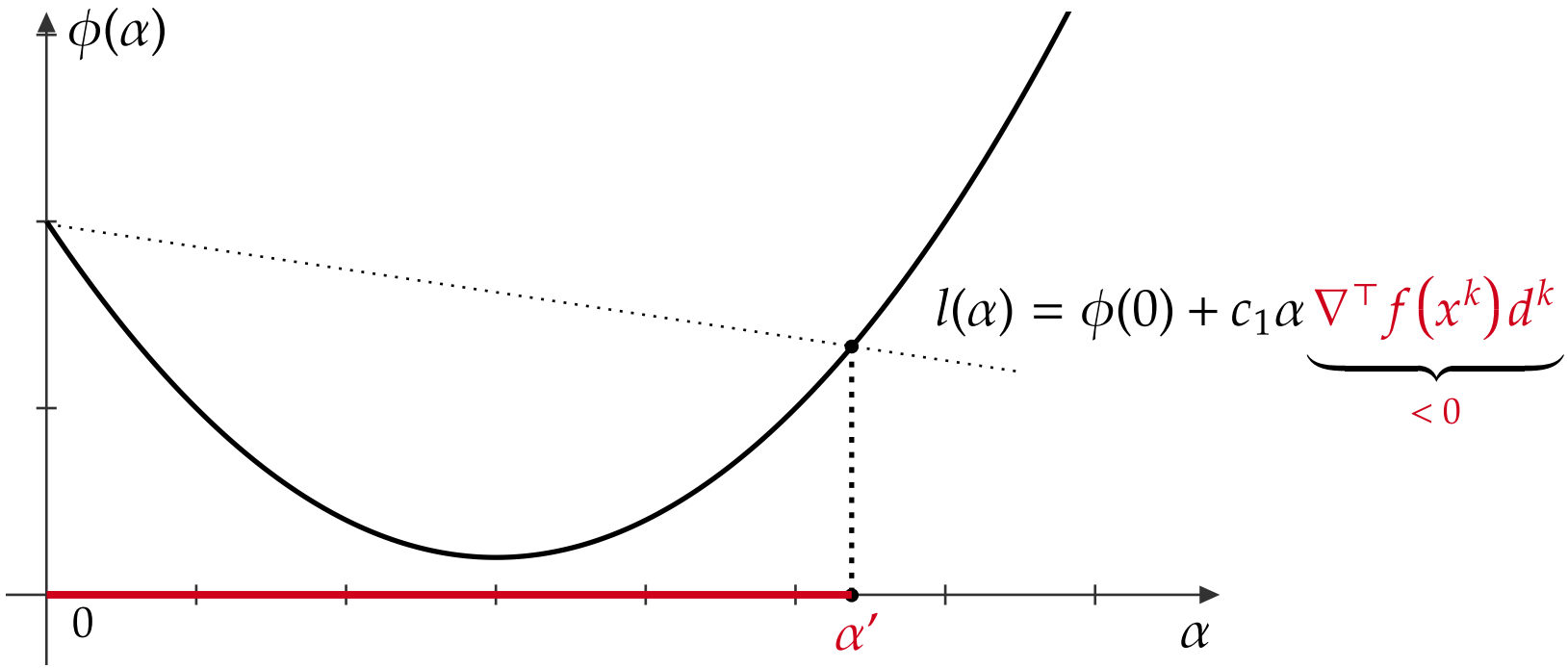
$$f_0(x^k+\alpha_k d^k)\leqslant f_0\left(x^k\right)+c_1\alpha_k\left(\nabla f_0(x^k)\right)^\top d^k,$$

其中 d^k 是点 x^k 处的下降方向，$c_1\in(0,1)$ 是一个常数（通常取值较小，如 $c_1=10^{-3}$）。

由于 d^k 是下降方向，因此 $\left(\nabla f_0(x^k)\right)^\top d^k<0$。于是，目标函数值在新的迭代点 $x^k+\alpha_k d^k$ 处减小，即 $f_0(x^k+\alpha_k d^k)<f_0(x^k)$。

非精确线搜索：Armijo rule

几何解释：要求在直线 $l(\alpha) = \phi(0) + c_1\alpha\left(\nabla f_0(x^k)\right)^\top d^k$ 的下方寻找合适的迭代步长 $\alpha_k > 0$。



非精确线搜索：Armijo rule

思考：在第 k 步，给定点 x^k 和该点处的下降方向 d^k，如何寻找满足 Armijo rule 的迭代步长？

一个常用的寻找方法是回退法 (back tracking)。

Algorithm 1: Back tracking with Armijo rule

输入： 初始步长 $\alpha_{\max}$，缩减比例 $\gamma \in (0,1)$，参数 $c_1 \in (0,1)$.

输出： $\alpha_k = \alpha$.

初始化 $\alpha \leftarrow \alpha_{\max}$;

while $f_0(x^k + \alpha d^k) > f_0(x^k) + c_1 \alpha \nabla f_0(x^k)^\top d^k$ **do**

  令 $\alpha \leftarrow \gamma\alpha$;

end

/* 若当前步长 α 使得 $f_0(x^k + \alpha d^k)$ 有足够的下降，则停止；否则继续减小 α */

非精确线搜索：Armijo rule

讨论：

1. **Armijo rule 的设计思路**：希望 $f_0(x)$ 在每步迭代相较于前一步都能得到“充分下降” (sufficient descent)；

2. 参数 c_1 控制着 Armijo rule 的“充分下降”程度：
若 c_1 过大，则 Armijo rule 不易满足；
反之，若 c_1 过小，则 Armijo rule 较容易满足，但 $f_0(x)$ 的下降程度也随之减弱；

3. 当 d^k 是下降方向时，算法 1（回退法）不会无限迭代，即总存在充分小的 α 使得 Armijo rule 成立，说明这一线搜索准则是良好定义的 (well-defined)。实际使用时也可以考虑给 α 设置一个下界，防止步长过小；

4. 值得注意的是，当 $\alpha = 0$ 时，Armijo rule 显然也满足，但此时迭代序列中的点固定不变。因此 Armijo rule 需要结合其它线搜索准则或加强问题自身的性质以保证线搜索算法的收敛性！

Goldstein 准则

Goldstein 准则 (Goldstein rule)

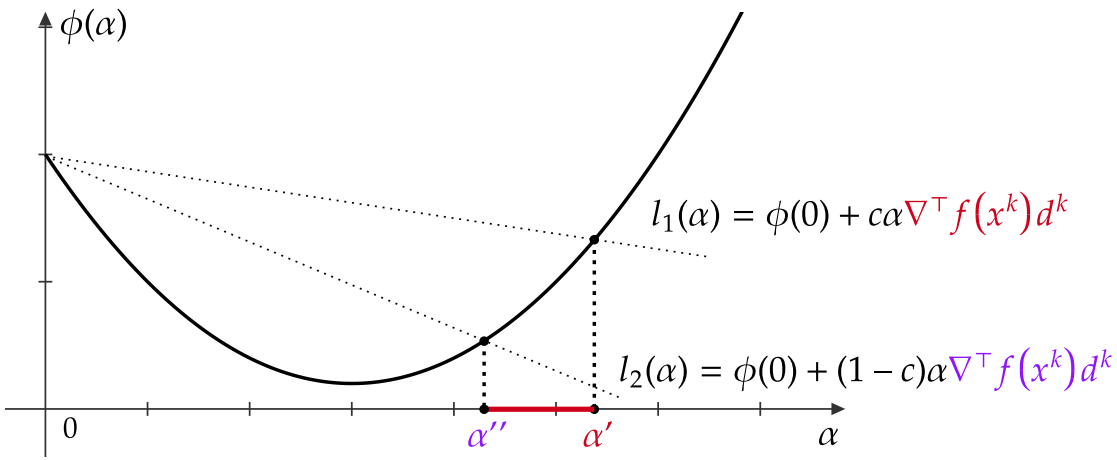
Goldstein rule 用于克服 Armoji rule 的一些缺陷，它要求 α_k 满足：

$$f_0(x^k + \alpha_k d^k) \leqslant f_0(x^k) + c\,\alpha_k \left(\nabla f_0(x^k)\right)^\top d^k,$$

$$f_0(x^k + \alpha_k d^k) \geqslant f_0(x^k) + (1-c)\,\alpha_k \left(\nabla f_0(x^k)\right)^\top d^k,$$

其中 d^k 是点 x^k 处的下降方向，$c \in (0, \frac{1}{2})$。

几何解释：要求在直线 $l_1(\alpha) = \phi(0) + c\alpha \left(\nabla f_0(x^k)\right)^\top d^k$ 和 $l_2(\alpha) = \phi(0) + (1-c)\alpha \left(\nabla f_0(x^k)\right)^\top d^k$ 之间寻找合适的迭代步长。



区间 $[\alpha'', \alpha']$ 之间的点均满足 Goldstein rule。该准则可以避免了过小的 α。

Wolfe 准则

Wolfe 准则 (Wolfe rule)

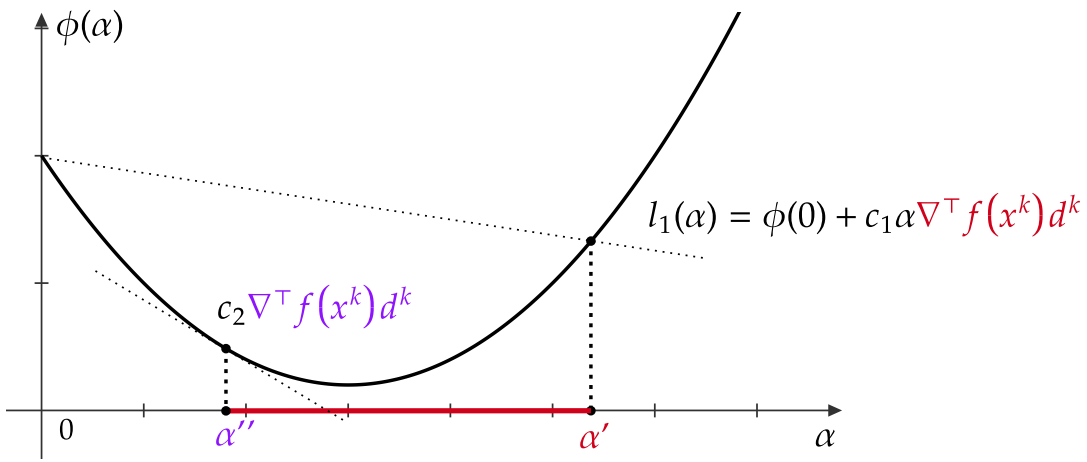
Goldstein rule 可能避开函数 $\phi(\alpha)$ 的最优点，为此引入 **Wolfe rule**，它要求 α_k 满足：

$$f_0(x^k+\alpha_k d^k) \leqslant f_0\left(x^k\right)+c_1\alpha_k\left(\nabla f_0(x^k)\right)^{\top} d^k,$$

$$\left(\nabla f(x^k+\alpha_k d^k)\right)^{\top} d^k \geqslant c_2\left(\nabla f_0(x^k)\right)^{\top} d^k,$$

其中 d^k 是点 x^k 处的下降方向，$c_1, c_2 \in (0,1)$，且 $c_1 < c_2$。

几何解释：$\left(\nabla f(x^k+\alpha d^k)\right)^{\top} d^k$ 是 $\phi(\alpha)$ 的导数，因此 Wolfe rule 实际要求 $\phi(\alpha)$ 在点 α_k 处的切线斜率不能小于 $\phi'(0)$ 的 c_2 倍。



Grippo 准则

以上所介绍的均为单调线搜索准则，即要求目标函数在每一步都有充分下降。实际中，人们也会采用非单调线搜索准则，以达到更好的计算效果。

Grippo 准则

设 d^k 是点 x^k 处的下降方向。以下线搜索准则要求 α_k 满足：

$$f_0(x^k+\alpha_k d^k) \leq \max_{0\leqslant j\leqslant \min\{k,\, M\}} f_0\left(x^{k-j}\right) + c_1\alpha_k\left(\nabla f_0(x^k)\right)^\top d^k,$$

其中 $M>0$ 为给定的正整数，$c_1\in(0,1)$ 为给定的常数。

相比于 Armijo rule:

$$f_0(x^k+\alpha_k d^k) \leqslant f_0(x^k) + c_1\alpha_k\nabla f_0(x^k)^\top d^k,$$

Grippo rule 只需 $f_0(x^k+\alpha_k d^k)$ 相比前面至多 M 步以内迭代的函数值有下降即可。显然，该准则比 Armijo rule 更宽泛，不要求 $f_0(x^k)$ 的单调性。

基于线搜索的算法收敛性

下面，我们给出基于线搜索的算法框架的一般收敛性结果，虽然结论较弱，但可以帮助我们理解线搜索类算法收敛的一些要求。

定理 1 (Zoutendijk 定理)

考虑利用迭代格式 $x^{k+1} = x^k + \alpha_k d^k$ 求解可微无约束优化问题

$$\min_x \quad f_0(x),$$

其中 d^k 是下降方向, α_k 是满足 Wolfe rule 的步长。假设目标函数 f_0 有下界，且 f_0 是梯度L-利普希茨连续 (L-Lipschitz continuous) 的，即 $\exists L > 0$ 使得

$$\|\nabla f_0(x) - \nabla f_0(y)\| \leqslant L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

那么，我们有：

$$\sum_{k=0}^{\infty} \cos^2 \theta_k \|\nabla f_0(x^k)\|^2 < +\infty,$$

其中 $\cos\theta_k$ 为负梯度 $-\nabla f_0(x^k)$ 和下降方向 d^k 夹角的余弦，即

$$\cos\theta_k = \frac{-\nabla f_0(x^k)^\top d^k}{\|\nabla f_0(x^k)\| \, \|d^k\|}.$$

基于线搜索的算法收敛性

证明： 回顾 Wolfe rule：

$$f_0(x^k + \alpha_k d^k) \leqslant f_0(x^k) + c_1 \alpha_k \nabla f_0(x^k)^\top d^k, \tag{1}$$

$$\nabla f(x^k + \alpha_k d^k)^\top d^k \geqslant c_2 \nabla f_0(x^k)^\top d^k. \tag{2}$$

1. 根据迭代格式可知 $\nabla \phi(\alpha_k) = \nabla f_0(x^{k+1})^\top d^k$。根据式 (2)，在不等式两侧同时减去 $\nabla f_0(x^k)^\top d^k$，可得

$$\left(\nabla f_0(x^{k+1}) - \nabla f_0(x^k)\right)^\top d^k \geqslant (c_2 - 1)\nabla f_0(x^k)^\top d^k.$$

2. 根据柯西不等式、f_0 的梯度L-利普希茨连续性和迭代格式，可得

$$\begin{aligned}\left(\nabla f_0(x^{k+1}) - \nabla f_0(x^k)\right)^\top d^k &\leqslant \|\nabla f_0(x^{k+1}) - \nabla f_0(x^k)\| \|d^k\| \\ &\leqslant L\|x^{k+1} - x^k\| \|d^k\| = \alpha_k L \|d^k\|^2.\end{aligned}$$

结合上述两式，可得到步长 α_k 的下界：

$$\alpha_k \geqslant \frac{c_2 - 1}{L} \frac{\nabla f_0(x^k)^\top d^k}{\|d^k\|^2}.$$

基于线搜索的算法收敛性

3. 由于 d^k 是下降方向，即 $\nabla f_0(x^k)^\top d^k < 0$，于是将上式代入式(1)，可得

$$f_0(x^{k+1}) \leqslant f_0(x^k) + c_1 \frac{c_2 - 1}{L} \frac{\left(\nabla f_0(x^k)^\top d^k\right)^2}{\|d^k\|^2}.$$

根据夹角 θ_k 的定义，上述不等式可等价表述为：

$$\begin{aligned} f_0(x^{k+1}) &\leqslant f_0(x^k) + c_1 \frac{c_2 - 1}{L} \frac{\left(\nabla f_0(x^k)^\top d^k\right)^2}{\|d^k\|^2} \\ &= f_0(x^k) + c_1 \frac{c_2 - 1}{L} \frac{\left(\nabla f_0(x^k)^\top d^k\right)^2 \|\nabla f_0(x^k)\|^2}{\|\nabla f_0(x^k)\|^2 \|d^k\|^2} \\ &= f_0(x^k) + c_1 \frac{c_2 - 1}{L} \cos^2 \theta_k \|\nabla f_0(x^k)\|^2. \end{aligned}$$

基于线搜索的算法收敛性

将上述不等式关于 k 求和，可得

$$f_0(x^{k+1}) \leqslant f_0(x^0) - c_1 \frac{1-c_2}{L} \sum_{j=0}^{k} \cos^2 \theta_j \|\nabla f_0(x^j)\|^2.$$

又因为函数 f_0 是下有界的，且由 $0 < c_1 < c_2 < 1$ 可知 $c_1(1-c_2) > 0$，因此当 $k \to \infty$ 时，

$$\sum_{j=0}^{\infty} \cos^2 \theta_j \|\nabla f_0(x^j)\|^2 < +\infty.$$

证毕!

基于线搜索的算法收敛性

Zoutendijk 定理表明：采用合适的线搜索准则，配合下降方向 d^k 的选取方式，我们可以得到最基本的收敛性。

推论 1 (基于线搜索的算法收敛性)

记 θ_k 为每一步负梯度 $-\nabla f(x^k)$ 与下降方向 d^k 的夹角，并假设

1. **定理 1** 中的条件成立；
2. 存在常数 $\gamma > 0$，使得对任意的 k 有

$$\theta_k < \frac{\pi}{2} - \gamma.$$

于是，我们有

$$\lim_{k\to\infty} \nabla f(x^k) = 0.$$

基于线搜索的算法收敛性

证明：（反证）假设结论不成立，即存在子列 $\{k_l\}$ 和正常数 $\delta > 0$，使得

$$\|\nabla f(x^{k_l})\| \geqslant \delta, \quad l = 1, 2, \cdots.$$

根据 θ_k 的假设可知，对任意的 k，

$$\cos\theta_k > \sin\gamma > 0.$$

于是，我们有

$$\sum_{k=0}^{\infty} \cos^2\theta_k \|\nabla f(x^k)\|^2 \geqslant \sum_{l=1}^{\infty} \underbrace{\cos^2\theta_{k_l}}_{>\sin^2\gamma} \underbrace{\|\nabla f(x^{k_l})\|^2}_{\geq\delta} \quad \to \quad +\infty.$$

显然，这和 **Zoutendijk 定理**矛盾，因此必有

$$\lim_{k\to\infty} \nabla f\left(x^k\right) = 0.$$

证毕。

- 线搜索
- 梯度类算法
- 梯度下降法的扩展
- 二阶算法

梯度类算法

考虑连续可微无约束优化问题

$$\min_x \quad f_0(x),$$

其中 $f_0(x):\mathbb{R}^n \to \mathbb{R}$ 在 $\mathrm{dom}\, f_0 = \mathbb{R}^n$ 上连续可微，且 $f_0(x)$ 的极小值 p^* 和最优解 x^* 存在。

- **求解方法**：$x^{k+1} = x^k + \alpha_k d^k$，配合下降方向 d^k 和迭代步长 α_k 的选取
- **梯度类算法**：仅仅使用函数的一阶“梯度”信息选取下降方向 d^k

梯度下降法：在 x^k 处选择负梯度 $-\nabla f_0(x^k)$ 作为下降方向 d^k

- $d^k = -\nabla f_0(x^k)$;
- $x^{k+1} = x^k + \alpha_k d^k$;
- 步长 α_k 的选取有多种策略：（非）精确线搜索、固定步长 α、递减步长（如 $\alpha_k = \frac{1}{k+1}$，$\frac{1}{\sqrt{k+1}}$）、自适应步长等 ...

梯度下降法的分析

梯度下降法以及大多数优化算法的分析主要关注如下几点：

- 能否收敛？ $\longrightarrow$ 稳
- 收敛到哪？ $\longrightarrow$ 准
- 收敛速度如何？ $\longrightarrow$ 快

接下来，我们简单介绍当 $f_0(x)$ 是梯度利普希茨连续的凸函数时，固定步长的梯度下降法的收敛性质。

梯度利普希茨连续

假设 1 (梯度利普希茨连续 (Lipschitz continuous gradient))

存在常数 $L > 0$，使得

$$\|\nabla f_0(x) - \nabla f_0(y)\| \leqslant L\|x - y\|, \quad \forall\, x, y \in \mathbb{R}^n.$$

其中 L 称为利普希茨常数。以后简记为，∇f_0 是 L-利普希茨的。

例： $f_0(x) = \mathbf{1}^\top x$， $f_0(x) = \frac{1}{2}\|x\|^2$

等价定义：

1. 若 f_0 二阶可微，则 $\nabla^2 f_0(x) \preceq LI$, $\forall\, x \in \mathbb{R}^n$;
2. $\langle \nabla f_0(x) - \nabla f_0(y), x - y\rangle \geq \frac{1}{L}\|\nabla f_0(x) - \nabla f_0(y)\|^2$, $\forall\, x, y \in \mathbb{R}^n$;
3. $\langle \nabla f_0(x) - \nabla f_0(y), x - y\rangle \leq L\|x - y\|^2$, $\forall\, x, y \in \mathbb{R}^n$;
4. $f_0(y) \leq f_0(x) + \langle \nabla f_0(x), y - x\rangle + \frac{L}{2}\|x - y\|^2$, $\forall\, x, y \in \mathbb{R}^n$;
5. $\frac{L}{2}\|x\|^2 - f(x)$ 是凸函数;

分析固定步长的梯度下降法

定理 2 (固定步长的梯度下降法的收敛性)

设 f_0 是定义在 $\mathbb{R}^n$ 上的连续可微凸函数，$x^* \in \arg\min\limits_{x\in\mathbb{R}^n} f_0(x)$ 存在，且 ∇f_0 是 L- 利普希茨的。若迭代步长 α_k 取为常数且满足 $0 < \alpha < \frac{2}{L}$，那么由**梯度下降法**得到的点列 $\{x^k\}$ 的函数值序列 $\{f_0(x^k)\}$ 收敛到最优值 $f(x^*)$，且

$$f_0(x^k) - f_0(x^*) \leqslant \frac{2\left(f_0(x^0) - f_0(x^*)\right)\|x^0 - x^*\|^2}{2\|x^0 - x^*\|^2 + k\,\alpha(2 - L\alpha)\left(f_0(x^0) - f_0(x^*)\right)}, \quad \forall\, x^*.$$

可以观察到：

- 当 $k \to \infty$ 时，$f_0(x^k)$ 收敛到最优值 $f_0(x^*)$，且收敛速率为 $\mathcal{O}\left(\frac{1}{k}\right)$
- 在此结论中，“最优”固定步长为 $\alpha = \frac{1}{L}$ $\longrightarrow$ 极大化分母中的 $\alpha(2 - L\alpha)$
- 收敛速率 $\mathcal{O}\left(\frac{1}{k}\right)$ 称为**次线性收敛速率**

分析固定步长的梯度下降法

证明： **Step 1.** 点的单调性（与任意最优解的距离在不断缩小）。事实上

$$\begin{aligned}\|x^{k+1}-x^*\|^2 &= \|x^k-\alpha\nabla f_0(x^k)-x^*\|^2\\ &= \|x^k-x^*\|^2-2\alpha\langle x^k-x^*,\nabla f_0(x^k)\rangle+\alpha^2\|\nabla f_0(x^k)\|^2\\ &= \|x^k-x^*\|^2-2\alpha\langle x^k-x^*,\nabla f_0(x^k)-\nabla f_0(x^*)\rangle+\alpha^2\|\nabla f_0(x^k)\|^2,\end{aligned}$$

最后等式由最优性条件 $\nabla f_0(x^*)=0$ 得到。又由 ∇f_0 是 L-利普希茨的，故

$$\left\langle x^k-x^*,\nabla f_0\left(x^k\right)-\nabla f_0\left(x^*\right)\right\rangle \geq \frac{1}{L}\|\nabla f_0(x^k)-\nabla f_0(x^*)\|^2=\frac{1}{L}\|\nabla f_0(x^k)\|^2.$$

回代后，结合 $\alpha\in\left(0,\frac{2}{L}\right)$，可得：

$$\begin{aligned}\|x^{k+1}-x^*\|^2 &\leq \|x^k-x^*\|^2-\frac{2\alpha}{L}\|\nabla f_0(x^k)\|^2+\alpha^2\|\nabla f_0(x^k)\|^2\\ &\leq \|x^k-x^*\|^2+\alpha\left(\alpha-\frac{2}{L}\right)\|\nabla f_0(x^k)\|^2\leq\|x^k-x^*\|^2.\end{aligned}$$

故对 $\forall x^*$ 均有 $\|x^{k+1}-x^*\|^2\leq\|x^k-x^*\|^2$，进而 $\|x^k-x^*\|^2\leq\|x^0-x^*\|^2$ 对任意的 k 成立。

分析固定步长的梯度下降法

证明续： **Step 2.** 目标函数值 $\{f_0(x)^k\}_{k=0}^{\infty}$ 的单调性。根据 ∇f_0 的 L-利普希茨性，有：

$$
\begin{aligned}
f_0(x^{k+1}) &\le f_0(x^k) + \langle \nabla f_0(x^k),\, x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&= f_0(x^k) + \langle \nabla f_0(x^k),\, -\alpha \nabla f_0(x^k) \rangle + \frac{L}{2}\alpha^2 \|\nabla f_0(x^k)\|^2 \\
&= f_0(x^k) - \alpha \|\nabla f_0(x^k)\|^2 + \frac{L}{2}\alpha^2 \|\nabla f_0(x^k)\|^2 \\
&= f_0(x^k) - \underbrace{\alpha\left(1 - \frac{L\alpha}{2}\right)}_{>0,\ \text{因为}\ \alpha \in \left(0,\, \frac{2}{L}\right)} \|\nabla f_0(x^k)\|^2 \\
&\le f_0(x^k).
\end{aligned}
$$

故对 $\forall k$ 均有 $f_0(x^{k+1}) \le f_0(x^k)$。

分析固定步长的梯度下降法

证明续：Step 3. 目标函数值的充分下降。事实上，在 **Step 2** 中，已得

$$f_0(x^{k+1}) \le f_0(x^k) - w\|\nabla f_0(x^k)\|^2, \quad \text{其中 } w = \alpha\left(1 - \frac{L\alpha}{2}\right).$$

将等式两侧同时减去 $f_0(x^*)$：

$$f_0(x^{k+1}) - f_0(x^*) \le f_0(x^k) - f_0(x^*) - w\|\nabla f_0(x^k)\|^2. \qquad (\star)$$

进一步，根据 f_0 的凸性，对 $\forall k$ 有：$f_0(x^*) \ge f_0(x^k) + \langle \nabla f_0(x^k), x^* - x^k\rangle$，

$$\Longrightarrow \quad f_0(x^k) - f_0(x^*) \le \underbrace{\langle \nabla f_0(x^k), x^k - x^*\rangle \le \|\nabla f_0(x^k)\|\|x^k - x^*\|}_{\text{根据柯西-施瓦茨不等式}}$$

$$\le \|\nabla f_0(x^k)\|\|x^0 - x^*\| \quad \text{(根据点列的单调性)}$$

$$\Longrightarrow \quad \|\nabla f_0(x^k)\| \ge \frac{f_0(x^k) - f_0(x^*)}{\|x^0 - x^*\|} \ge 0.$$

分析固定步长的梯度下降法

证明续：将上式代入式 $(\star)$ 可得

$$\underbrace{f_0(x^{k+1}) - f_0(x^*)}_{\triangleq \Delta_{k+1}} \leq \underbrace{f_0(x^k) - f_0(x^*)}_{\triangleq \Delta_k} - \frac{w}{\|x^0 - x^*\|^2} \underbrace{\left(f_0(x^k) - f_0(x^*)\right)^2}_{\triangleq \Delta_k^2}.$$

于是，

$$\Rightarrow \Delta_{k+1} \leq \Delta_k - \frac{w}{\|x^0 - x^*\|^2} \Delta_k^2$$

$$\Rightarrow \frac{1}{\Delta_k} \leq \frac{1}{\Delta_{k+1}} - \frac{w}{\|x^0 - x^*\|^2} \frac{\Delta_k}{\Delta_{k+1}} \quad (\text{等式两侧同时除以 } \Delta_k \Delta_{k+1})$$

$$\Rightarrow \frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{w}{\|x^0 - x^*\|^2} \frac{\Delta_k}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{w}{\|x^0 - x^*\|^2} \quad (\text{根据 } \frac{\Delta_k}{\Delta_{k+1}} \geq 1)$$

分析固定步长的梯度下降法

证明续：对不等式 $\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{w}{\|x^0 - x^*\|^2}$ 关于 k 求和，可得

$$\frac{1}{\Delta_k} \geq \frac{1}{\Delta_0} + \frac{k\,w}{\|x^0 - x^*\|^2}.$$

重新整理不等式，可得

$$\frac{1}{\Delta_k} \geq \frac{1}{\Delta_0} + \frac{kw}{\|x^0 - x^*\|^2} \quad \Longleftrightarrow \quad \Delta_k \leq \underbrace{\frac{\Delta_0 \|x^0 - x^*\|^2}{\|x^0 - x^*\|^2 + kw\Delta_0}}_{\text{注意 } w = \alpha\left(1 - \frac{L\alpha}{2}\right)}$$

$$\Longrightarrow \quad f_0(x^k) - f_0(x^*) \leqslant \frac{2\left(f_0(x^0) - f_0(x^*)\right)\|x^0 - x^*\|^2}{2\|x^0 - x^*\|^2 + k\alpha(2 - L\alpha)\left(f_0(x^0) - f_0(x^*)\right)}.$$

证毕！

强凸性

为得到更强收敛性结果，我们需要对 f_0 作进一步假设。

假设 2 (强凸性 (strongly convexity))

存在常数 $\mu > 0$，使得

$$f_0(y) \geqslant f_0(x) + \langle \nabla f_0(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2, \quad \forall\, x, y \in \mathbb{R}^n,$$

其中 μ 称为强凸系数。简记为，f_0 是 μ-强凸的。

例： $f_0(x) = \frac{1}{2}\|x\|^2$

等价定义：

1. 若 f_0 二阶可微，则 $\nabla^2 f_0(x) \succeq \mu I$, $\forall\, x \in \mathbb{R}^n$;
2. $\langle \nabla f_0(x) - \nabla f_0(y), x - y \rangle \geq \mu\|x - y\|^2$;
3. $f_0(y) - f_0(x) - \langle \nabla f_0(x), y - x \rangle \geq \frac{\mu}{2}\|x - y\|^2$;
4. $f(x) - \frac{\mu}{2}\|x\|^2$ 是凸函数;

重要性质： 强凸 $\Rightarrow$ 严格凸 $\Rightarrow$ f_0 的最优解 x^* 是唯一的

分析固定步长的梯度下降法

定理 3 (固定步长的梯度下降法的收敛性)

设 f_0 是定义在 $\mathbb{R}^n$ 上的连续可微μ-强凸函数，$x^* = \arg\min_{x\in\mathbb{R}^n} f_0(x)$ 存在，且 ∇f_0 是 L-利普希茨的。若迭代步长 α_k 取为常数且满足 $\alpha \in \left(0, \frac{2}{\mu+L}\right)$，那么由**梯度下降法**得到的点列 $\{x^k\}$ 收敛到最优解 x^*，且

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

- 易知 $\frac{2\alpha\mu L}{\mu+L} \in (0,1)$，故 $1 - \frac{2\alpha\mu L}{\mu+L} \in (0,1)$；
- 当 $k \to \infty$ 时，x^k 收敛到最优解 x^*，且收敛速率为 $\mathcal{O}\left(\left(1 - \frac{2\alpha\mu L}{\mu+L}\right)^k\right)$；
- 记 $c = 1 - \frac{2\alpha\mu L}{\mu+L}$，这样的收敛速率 $\mathcal{O}(c^k)$ 通常称为**线性收敛速率**。

分析固定步长的梯度下降法

证明：

Step 1. 点的单调性：

$$\begin{aligned}&\|x^{k+1}-x^*\|^2\\ =&\|x^k-x^*\|^2-2\alpha\langle x^k-x^*,\nabla f_0(x^k)-\nabla f_0(x^*)\rangle+\alpha^2\|\nabla f_0(x^k)\|^2.\end{aligned}$$

Step 2. 在强凸假设下，可得 $2\alpha\langle x^k-x^*,\nabla f_0(x^k)-\nabla f_0(x^*)\rangle$ 更紧的下界。

事实上，根据 f_0 的 μ-强凸性和 ∇f_0 的 L-利普希茨性，有：

$$g(x)=f_0(x)-\frac{\mu}{2}\|x\|^2,\text{ 为凸函数 (根据 } f_0 \text{ 的 } \mu\text{-强凸性)}$$

$$\frac{L-\mu}{2}\|x\|^2-g(x)=\frac{L}{2}\|x\|^2-f_0(x),\text{ 为凸函数 (根据 } \nabla f_0 \text{ 的 } L\text{-利普希茨性)}$$

因此，$\nabla g(x)$ 是 $(L-\mu)$-利普希茨的，故对 $\forall\, x,y$ 有：

$$\langle\nabla g(x)-\nabla g(y),x-y\rangle\geqslant\frac{1}{L-\mu}\|\nabla g(x)-\nabla g(y)\|^2.$$

分析固定步长的梯度下降法

将 $g(x)=f_0(x)-\frac{\mu}{2}\|x\|^2$ 带入上述不等式，有

$$\begin{aligned}&\langle\nabla f_0(x)-\nabla f_0(y)-\mu(x-y),x-y\rangle\\&\geqslant\frac{1}{L-\mu}\|\nabla f_0(x)-\nabla f_0(y)-\mu(x-y)\|^2.\end{aligned}\tag{$\star$}$$

将不等式 ($\star$) 左侧展开：

$$\langle\nabla f_0(x)-\nabla f_0(y),x-y\rangle-\mu\|x-y\|^2$$

将不等式 ($\star$) 右侧展开：

$$\frac{1}{L-\mu}\left(\|\nabla f_0(x)-\nabla f_0(y)\|^2+\mu^2\|x-y\|^2-2\mu\langle\nabla f_0(x)-\nabla f_0(y),x-y\rangle\right)$$

然后，整理不等式可得：

$$\langle\nabla f_0(x)-\nabla f_0(y),x-y\rangle\geqslant\frac{\mu L}{\mu+L}\|x-y\|^2+\frac{1}{\mu+L}\|\nabla f_0(x)-\nabla f_0(y)\|^2.$$

分析固定步长的梯度下降法

由 x, y 的任意性，用 x^k 替换 x，用 x^* 替换 y，可得：

$$\langle x^k - x^*, \nabla f_0(x^k) - \nabla f_0(x^*) \rangle \geqslant \frac{\mu L}{\mu + L} \|x^k - x^*\|^2 + \frac{1}{\mu + L} \|\nabla f_0(x^k) - \nabla f_0(x^*)\|^2.$$

于是，结合 **Step 1** 和 $\nabla f_0(x^*) = 0$，有

$$\begin{aligned}
&\|x^{k+1} - x^*\|^2 \\
&= \|x^k - x^*\|^2 - 2\alpha \langle x^k - x^*, \nabla f_0(x^k) - \nabla f_0(x^*) \rangle + \alpha^2 \|\nabla f_0(x^k)\|^2 \\
&\leqslant \|x^k - x^*\|^2 - \frac{2\alpha\mu L}{\mu + L} \|x^k - x^*\|^2 - \frac{2\alpha}{\mu + L} \|\nabla f_0(x^k)\|^2 + \alpha^2 \|\nabla f_0(x^k)\|^2 \\
&= \left(1 - \frac{2\alpha\mu L}{\mu + L}\right) \|x^k - x^*\|^2 + \left(\alpha^2 - \frac{2\alpha}{\mu + L}\right) \|\nabla f_0\left(x^k\right)\|^2 \\
&\leqslant \left(1 - \frac{2\alpha\mu L}{\mu + L}\right) \|x^k - x^*\|^2. \quad \left(\text{根据 } \alpha^2 - \frac{2\alpha}{\mu + L} < 0\right)
\end{aligned}$$

由此可得想要的结论。**证毕!**

对步长的思考

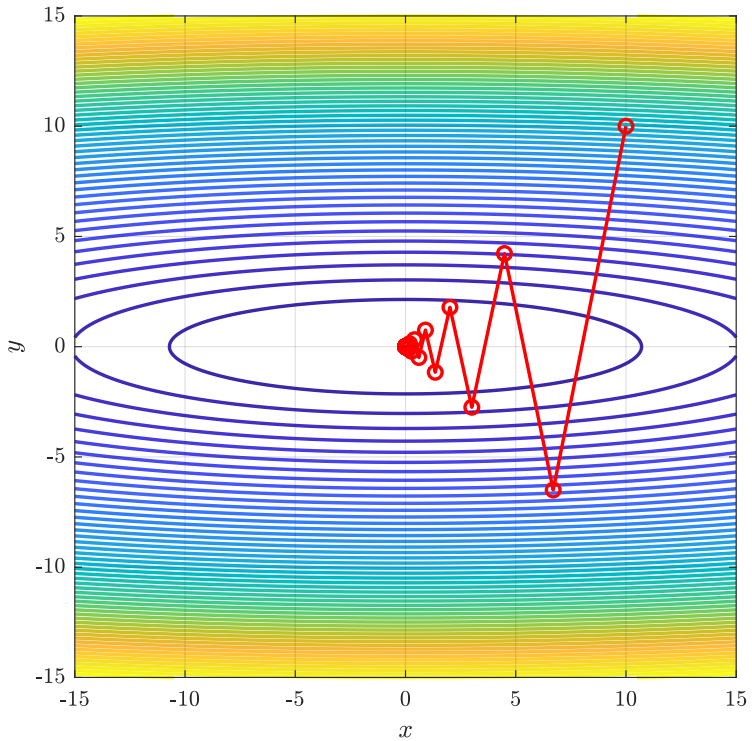
最优步长 α 应使得 $c = 1 - \frac{2\alpha\mu L}{\mu+L}$ 尽可能小，故 α 应尽可能大。由于 $\alpha \in \left(0, \frac{2}{\mu+L}\right)$，易知 $\left(\frac{L-\mu}{L+\mu}\right)^2$ 是 c 的下确界，故 $\frac{L-\mu}{L+\mu}$ 的大小决定 c 可能达到多小。

注意到 $\frac{L-\mu}{L+\mu} = \frac{\frac{L}{\mu}-1}{\frac{L}{\mu}+1}$，记 $\kappa := \frac{L}{\mu} \geq 1$，它正是 $\nabla^2 f_0(x)$ 的**条件数**[1]! 易知，κ 越小时，$\frac{L-\mu}{L+\mu}$ 越小。

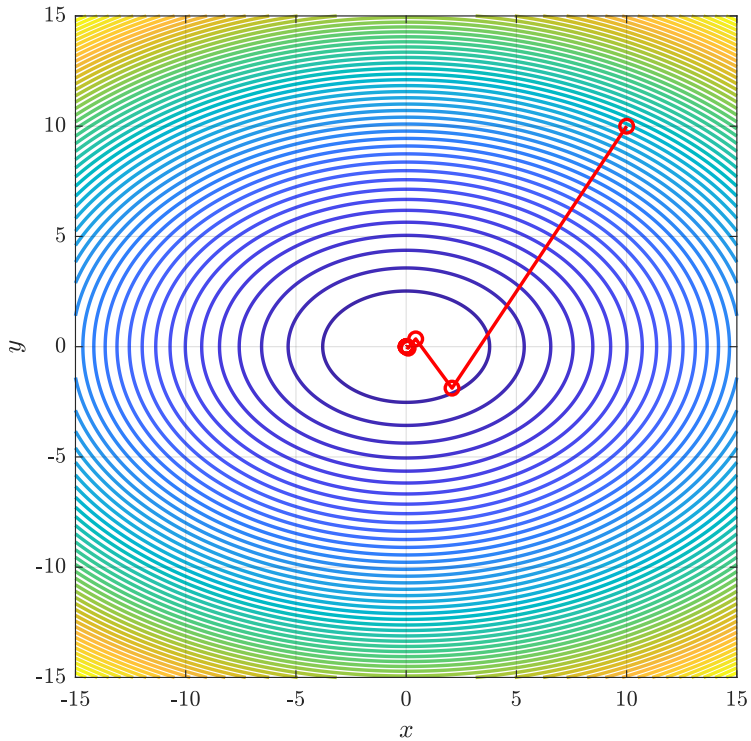
- 若 $\kappa = 1$，可选择合适的方向与步长实现 1 步收敛！例如，$f_0(x) = \frac{1}{2}x^2$，$L = \mu = 1$，选择下降方向 $d^k = -f_0'(x^k) = -x^k$ 和迭代步长 $\alpha = 1$，可知从任意初始点 $x^0 \neq 0$ 出发，有 $x^1 = x^0 + d^0 = 0$。
- 若 $\kappa \gg 1$，则收敛慢!

[1]对称正定矩阵 A 的条件数是最大特征值与最小特征值的比值 $\left|\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}\right|$

对步长的思考



$\kappa = 5$



$\kappa = 1.5$

Fig. 条件数 κ 对梯度下降法的影响，其中步长固定为 $\alpha = 0.99 \times \dfrac{2}{\mu + L}$.

梯度下降法的一种解释

梯度下降法的另一种解释

对连续可微的凸函数 f_0 在 x^k 附近作二阶近似：

$$f_0(x) \approx f_0(x^k) + \left\langle \nabla f_0(x^k),\, x - x^k \right\rangle + \frac{1}{2\alpha} \left\| x - x^k \right\|^2. \qquad (\triangle)$$

然后，通过求解一系列二阶近似函数 $(\triangle)$ 达到求解原问题的目的，其迭代格式如下：

$$\begin{aligned} x^{k+1} &= \arg\min_x \left\{ f_0(x^k) + \left\langle \nabla f_0(x^k),\, x - x^k \right\rangle + \frac{1}{2\alpha} \left\| x - x^k \right\|^2 \right\} \\ &= x^k - \alpha \nabla f_0\left(x^k\right) \qquad \Leftarrow \qquad \text{梯度下降法的迭代格式!} \end{aligned}$$

注意： 二阶近似的好坏影响着收敛的速度，进而依赖于 α 的选取!

思考： 若 $f_0(x)$ 不可微，该如何选择搜索方向？ ⋆ 采用负次梯度方向!

- 线搜索
- 梯度类算法
- 梯度下降法的扩展
- 二阶算法

次梯度

例： $\min\limits_{x} f_0(x) = |x|$，显然 f_0 在 $x = 0$ 处不可微。

定义 1 (次梯度 (subgradient) 和次微分 (subdifferential))

设 $f_0 : \mathbb{R}^n \to (-\infty, +\infty]$ 为凸函数，x 为定义域 $\operatorname{dom} f_0$ 中的某一点。若向量 $\nu \in \mathbb{R}^n$ 满足

$$f_0(y) \geqslant f_0(x) + \langle \nu, y - x \rangle, \quad \forall\, y \in \operatorname{dom} f_0,$$

则称 ν 为函数 f_0 在点 x 处的一个次梯度。进一步地，称集合

$$\partial f_0(x) = \{\nu \in \mathbb{R}^n \mid f_0(y) \geqslant f_0(x) + \langle \nu, y - x \rangle,\ \forall\, y \in \operatorname{dom} f_0\}$$

为 f_0 在点 x 处的次微分。

次梯度和次微分有如下性质：

- 若 f_0 在点 x 处可微，则 $\partial f_0(x) = \{\nabla f_0(x)\}$ $\longleftarrow$ 次微分中仅有一个元素
- 点 x^* 是 $\min\limits_{x} \{f_0(x)\}$ 的最优点 $\iff$ $0 \in \partial f_0(x^*)$

次梯度方法

例： $f_0(x)=|x|$ 的次微分 $\partial|x|=\begin{cases}1, & x>0,\\ [-1,1], & x=0,\\ -1, & x<0.\end{cases}$

★ **次梯度方法：** 在 x^k 处选择一个负的次梯度 $\nu^k\in\partial f_0(x^k)$ 作为下降方向：

- $d^k=-\nu^k,\quad \nu^k\in\partial f_0(x^k)$;
- $x^{k+1}=x^k+\alpha_k d^k$;

定理 4 (次梯度方法的收敛性)

假设 (i) f_0 是定义在 $\mathbb{R}^n$ 上的凸函数；(ii) f_0 至少存在一个有限且可达的极小值点 x^*，即 $f_0(x^*)>-\infty$；(iii) f_0 自身是 G-利普希茨的。设次梯度方法经过了 k 轮迭代，记 $\hat{f}_0^k=\min\left\{f_0(x^0),f_0(x^1),\cdots,f_0(x^k)\right\}=\min\limits_{0\leq t\leq k}\left\{f_0(x^t)\right\}$，为历史迭代的最优值，则 $\hat{f}_0^k$ 满足不等式：

$$\hat{f}_0^k-f(x^*)\leqslant\frac{\|x^0-x^*\|^2+G^2\sum_{i=0}^k\alpha_i^2}{2\sum_{i=0}^k\alpha_i}.$$

次梯度方法

由上述结论可知，不同的步长选取会带来不同的收敛性效果：

- 取固定步长（即 $\alpha_k = \alpha$），则 $\lim\limits_{k\to\infty} \hat{f}_0^k - f(x^*) \leq \frac{G^2\alpha}{2}$；
- 取 α_k 满足 $\sum_{k=0}^{\infty} \alpha_k = +\infty$ 但 $\sum_{k=0}^{\infty} \alpha_k^2 < +\infty$，则 $\lim\limits_{k\to\infty} \hat{f}_0^k - f(x^*) = 0$；

  例： 当 $\alpha_k = \frac{1}{k+1}$ 时，$\hat{f}_0^k - f(x^*)$ 的收敛速率为 $\mathcal{O}\left(\frac{1}{\ln k}\right)$。

- 取 α_k 为递减步长，如 $\alpha_k = \frac{1}{\sqrt{k+1}}$，其满足 $\sum_{k=0}^{\infty} \alpha_k = +\infty$ 且 $\sum_{k=0}^{\infty} \alpha_k^2 = +\infty$，则 $\lim\limits_{k\to\infty} \hat{f}_0^k - f(x^*) = 0$，且收敛速率为 $\mathcal{O}\left(\frac{\ln k}{\sqrt{k}}\right)$.

总结： 次梯度方法能够处理一般的不可微凸优化问题，但收敛速率往往很缓慢！

坐标下降法

思考： 当自变量 x 的维度较高或 $f_0(x)$ 较复杂时，（次）梯度往往难以计算，能否不借助梯度信息求解问题 $\min\limits_{x} \ f_0(x)$？

★ 考虑一种“分而治之”的策略：**坐标下降法** (Coordinate Descent Method)

- 在每次迭代中，选择第 i 个坐标轴方向；
- 在当前点 x^k 处沿第 i 个坐标轴方向进行极小化：

$$x_i^{k+1} \leftarrow \arg\min_{x_i} \ f_0(x_1^{k+1}, \cdots, x_{i-1}^{k+1}, x_i, x_{i+1}^{k}, \cdots, x_n^{k});$$

- 整个算法流程中循环遍历所有的坐标方向。

定义 2 (Coordinate-wise Minimizer)

对任意 $f_0 : \mathbb{R}^n \to (-\infty, +\infty]$，如果存在 $\bar{x} \in \operatorname{dom} f_0$ 满足：

$$f_0(\bar{x} + \alpha e_i) \geq f_0(\bar{x}), \quad \forall i \in [n], \ \forall \alpha \in (-\infty, +\infty),$$

则称 $\bar{x}$ 为 f_0 的一个 **coordinate-wise minimizer**，其中 $e_i = [0, \cdots, 1, \cdots, 0]^\top \in \mathbb{R}^n$ 表示第 i 个坐标轴的标准基向量。

含义： 在 $\bar{x}$ 处沿任意坐标轴极小化，都无法使 f_0 进一步减小！

坐标下降法

Algorithm 2: 坐标下降法

输入： 初始点 x^0;

输出： $\bar{x} = x^k$;

初始化：$k \leftarrow 0$;

repeat

$$x_1^{k+1} \leftarrow \arg\min_{x_1} f_0\left(x_1, x_2^k, x_3^k, \cdots, x_n^k\right);$$

$$x_2^{k+1} \leftarrow \arg\min_{x_2} f_0\left(x_1^{k+1}, x_2, x_3^k, \cdots, x_n^k\right);$$

$$\vdots$$

$$x_n^{k+1} \leftarrow \arg\min_{x_n} f_0\left(x_1^{k+1}, x_2^{k+1}, x_3^{k+1}, \cdots, x_n\right);$$

$k \leftarrow k + 1$

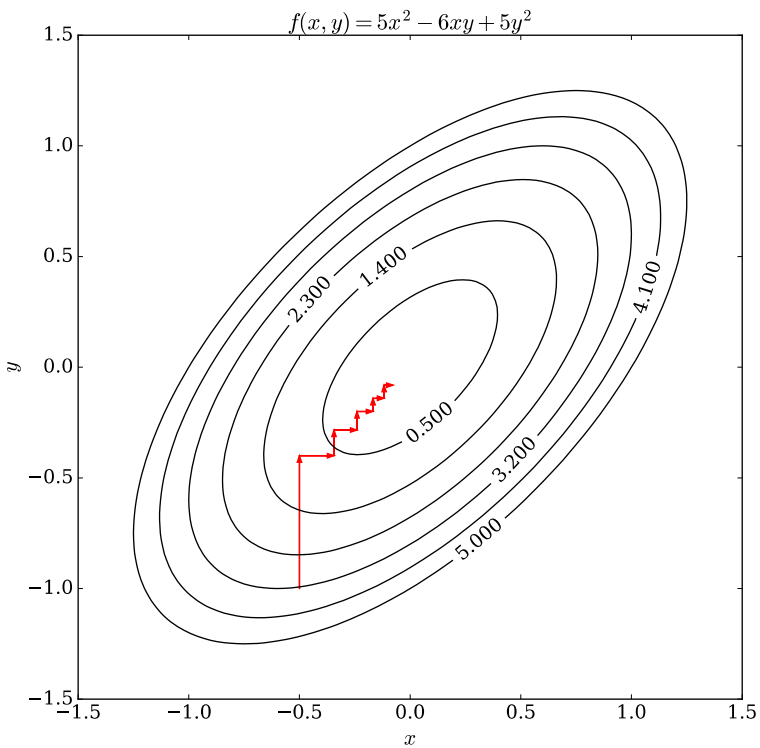
until 满足收敛准则;

思考： 若坐标下降法收敛，那么 $\bar{x}$ 是否为 f_0 的一个最优点?

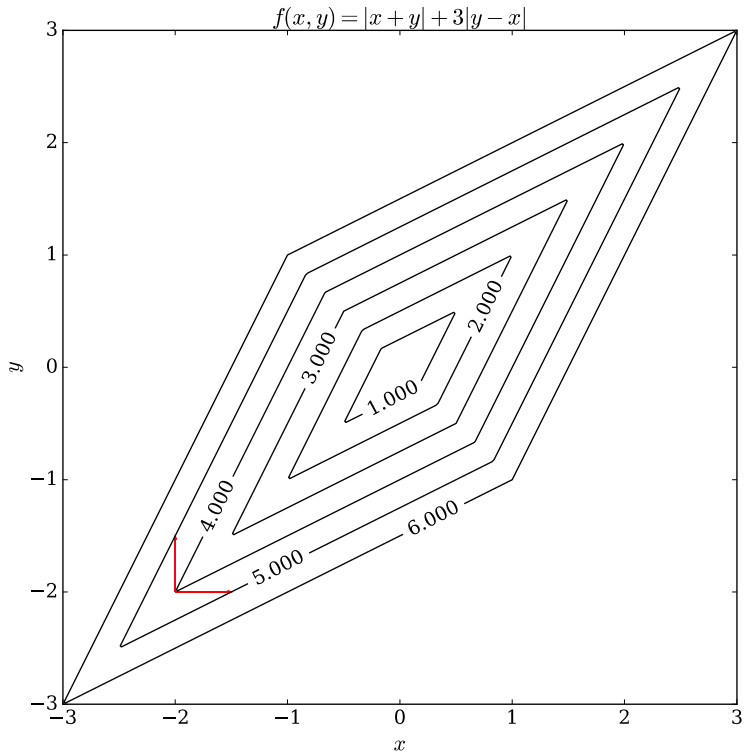
坐标下降法

结论

- 若 f_0 是连续可微的凸函数，则 $\bar{x}$ 是 f_0 的一个最优点！
- 若 f_0 是凸函数，但不连续可微，则 $\bar{x}$ 不一定是 f_0 的最优点！



连续可微凸函数



不可微且不可分凸函数

复合型优化问题

若优化问题不可微，但具有一些特殊结构，我们则仍有可能为其设计有效的**梯度类求解算法!**

考虑如下具有结构但不可微的复合型优化问题：

$$\min_x \ f_0(x) := g(x) + r(x),$$

其中

- $g(x)$ 表示损失函数，一般为光滑函数，如 $\frac{1}{2}\|Ax-b\|_2^2$；
- $r(x)$ 表示正则函数，一般为非光滑函数，用来诱导/限制解的特殊结构，如稀疏性，低秩性等。

例： LASSO 问题 $\min_x \ f_0(x) = \underbrace{\frac{1}{2}\|Ax-b\|_2^2}_{g(x)} + \underbrace{\lambda\|x\|_1}_{r(x)}$

求解方法：次梯度方法？局限性： 1. 收敛速度缓慢；2. 不能充分利用 $g(x)$ 的光滑性；3. 难以在迭代中保证 $r(x)$ 想要诱导的解的结构。

邻近算子

★ 另一种思路： 利用邻近算子来处理非光滑项 $r(x)$!

定义 3 (邻近算子 (proximal operator/mapping))

对于一个适当闭凸函数 r，以及任意常数 $\alpha > 0$，其邻近算子定义为：

$$\mathrm{prox}_{\alpha r}(x) = \arg\min_u \left\{ r(u) + \frac{1}{2\alpha}\|u - x\|_2^2 \right\}.$$

邻近算子的重要性质：

- **良定义性**：若 r 是适当闭凸函数，则 $\forall\, x \in \mathbb{R}^n$， $\mathrm{prox}_{\alpha r}(x)$ 存在且唯一
- **与次梯度的关系**：若 r 是适当闭凸函数，则

$$u^* = \mathrm{prox}_r(x) \iff x - u^* \in \partial r(u^*)$$

邻近算子的案例

- **示性函数：** $\delta_C(x) = \begin{cases} 0, & x \in C, \\ +\infty, & x \notin C. \end{cases}$，$C = \{x \in \mathbb{R}^n \mid x_i \in [a_i, b_i]\,,\ \forall i\}$

$$u = \mathrm{prox}_{\delta_C}(x) \quad \to \quad u_i = \begin{cases} a_i, & x_i < a_i, \\ x_i, & x_i \in [a_i, b_i], \\ b_i, & x_i > b_i. \end{cases}$$

- **ℓ_1-范数：** $\|x\|_1$，对任意 $\lambda > 0$，有

$$u = \mathrm{prox}_{\lambda\|\cdot\|_1}(x) \quad \to \quad u_i = \begin{cases} x_i - \lambda, & x_i > \lambda, \\ 0, & x_i \in [-\lambda,\, \lambda], \\ x_i + \lambda, & x_i < -\lambda. \end{cases}$$

邻近梯度法

对于复合型凸优化问题 (convex composite optimization problem)：

$$\min_x \quad f_0(x) := g(x) + r(x),$$

其中

- $g(x) : \mathbb{R}^n \to \mathbb{R}$ 凸、连续可微
- $r(x) : \mathbb{R}^n \to \mathbb{R}$ 凸、**邻近算子容易计算**

★ **邻近梯度法 (Proximal Gradient Method)**：

- $x^{k+\frac{1}{2}} = x^k - \alpha_k \nabla g(x^k)$ $\longleftarrow$ 对光滑部分 g 做梯度下降
- $x^{k+1} = \mathrm{prox}_{\alpha_k r}(x^{k+\frac{1}{2}})$ $\longleftarrow$ 对非光滑部分 r 使用邻近算子
- 结合起来得到简洁迭代格式：$x^{k+1} = \mathrm{prox}_{\alpha_k r}\left(x^k - \alpha_k \nabla g(x^k)\right)$

其中步长 $\alpha_k > 0$ 可以固定，也可通过线搜索寻找以提高实际收敛速度。

邻近梯度法的收敛性

定理 5 (固定步长的邻近梯度法的收敛性)

设 $g(x)$ 是定义在 $\mathbb{R}^n$ 上的连续可微凸函数，且 $\nabla g(x)$ 是 L-利普希茨的；$r(x)$ 是适当闭凸函数，且 $\mathrm{prox}_{\alpha r}$ 容易计算；$f_0(x) := g(x) + r(x)$ 的最优解 x^* 存在。若步长 $\alpha_k = \alpha$ 为常数且满足 $0 < \alpha \leq \frac{1}{L}$，那么由**邻近梯度法**得到的点列 $\{x^k\}$ 的函数值序列 $\{f_0(x^k)\}$ 收敛到最优值 $f(x^*)$，且

$$f_0(x^k) - f_0(x^*) \leqslant \frac{1}{2\alpha k}\|x^0 - x^*\|^2, \quad \forall\, x^*.$$

可以观察到：

- 当 $k \to \infty$ 时，$f_0(x^k)$ 收敛到最优值 $f_0(x^*)$，且收敛速率为 $\mathcal{O}\left(\frac{1}{k}\right)$
- 在此结论中，“最优”固定步长：$\alpha = \frac{1}{L}$
- 与次梯度法相比，邻近梯度法可采用固定步长，选取更便利，且收敛速率更优
- 实际计算中，全局利普希茨常数 L 往往难以估计，或者值很大，故可在局部通过线搜索来确定，进而确定步长的选取

邻近梯度法的一种解释

根据邻近算子的定义，展开迭代格式：

$$
\begin{aligned}
x^{k+1} &= \operatorname{prox}_{\alpha_k r}\left(x^k - \alpha_k \nabla g(x^k)\right) \\
&= \arg\min_u \left\{ r(u) + \frac{1}{2\alpha_k}\left\| u - x^k + \alpha_k \nabla g(x^k)\right\|^2 \right\} \\
&= \arg\min_u \left\{ r(u) + \frac{1}{2\alpha_k}\|u - x^k\|^2 + \nabla g(x^k)^\top (u - x^k) + \frac{\alpha_k}{2}\|\nabla g(x^k)\|^2 \right\} \\
&= \arg\min_u \left\{ r(u) + g(x^k) + \nabla g(x^k)^\top (u - x^k) + \frac{1}{2\alpha_k}\|u - x^k\|^2 \right\}.
\end{aligned}
$$

可以观察到：

- 对 f_0 的光滑部分 g 进行二阶近似，并保留非光滑部分 r；
- 然后极小化该近似问题，得到最优点 x^{k+1} 作为每轮迭代的更新；
- 当非光滑项 $r(x) \equiv 0$ 时，**邻近梯度法**就退化为**梯度下降法**：

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

加速邻近梯度法

回顾：

- 邻近梯度法适用于求解如下不可微复合型凸优化问题：

$$\min_x \ f_0(x) := g(x) + r(x)$$

  - $g(x) : \mathbb{R}^n \to \mathbb{R}$ 凸、连续可微
  - $r(x) : \mathbb{R}^n \to \mathbb{R}$ 凸、邻近算子容易计算（不要求光滑）

- 产生的函数值序列 $\left\{f_0(x^k)\right\}$ 单调非增
- 算法实现简单，但 $\mathcal{O}\left(\frac{1}{k}\right)$ 的收敛速率仍然不够令人满意！

思考：如果仅利用问题的一阶信息，能否得到收敛速度更快的算法？

⋆ Nesterov 加速方法！

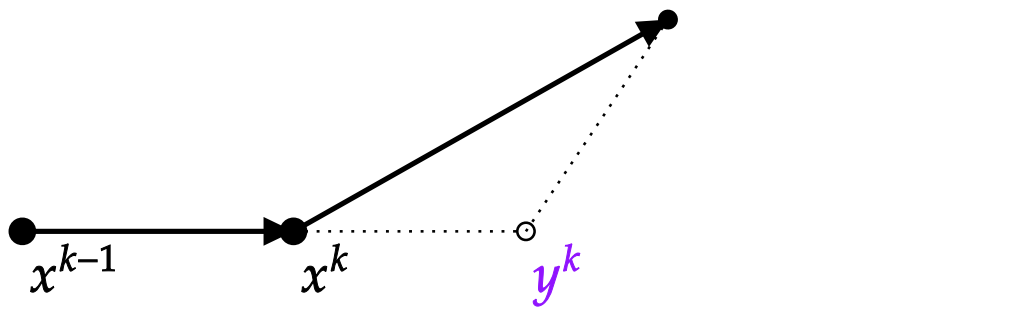
加速邻近梯度法

Nesterov 加速方法简介

- Nesterov 分别在 1983 年、1988 年和 2005 年提出了三种类型的加速梯度算法，它们的收敛速度能达到 $\mathcal{O}\left(\frac{1}{k^2}\right)$。这三种加速技术都可以应用于加速邻近梯度法。

- **加速思想：** 利用迭代历史信息！ 简单来说，我们认为点从 x^{k-1} 移动到 x^k 时会有“惯性” (momentum)，从而“外推”出一个新的点 y^k，然后在 y^k 处执行一步（邻近）梯度下降：

$$x^{k+1} = \mathrm{prox}_{\alpha_k r}\left(y^k - \alpha_k \nabla g\left(y^k\right)\right)$$



- 随着机器学习、数据科学等应用领域的发展，Nesterov 加速技术作为简单有效的加速方法被重新挖掘并迅速流行起来。

Nesterov 第一加速方法

考虑复合型凸优化问题：

$$\min_x \quad f_0(x) := g(x) + r(x),$$

其中 $r(x)$ 是适当闭凸函数；$g(x)$ 是凸且连续可微的函数，且 ∇g 是 L-利普希茨的。

Nesterov's 1st acceleration：

$$y^k = x^k + \theta_k(\theta_{k-1}^{-1} - 1)(x^k - x^{k-1}),$$

$$x^{k+1} = \underbrace{\arg\min_u \left\{ r(u) + g(y^k) + \langle \nabla g(y^k),\, u - y^k \rangle + \frac{1}{2\alpha_k}\|u - y^k\|^2 \right\}}_{\mathrm{prox}_{\alpha_k r}(y^k - \alpha_k \nabla g(y^k))},$$

$$\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}.$$

Nesterov 第一加速方法

- Nesterov 在 1983 首次发表了加速梯度算法（即上式中 $r \triangleq 0$ 的情形）。
- Beck 和 Teboulle 在 2007 年基于 Nesterov 第一加速技术给出了一种**加速邻近梯度法**，称作 **FISTA** (Fast Iterative Shrinkage-Thresholding Algorithm)[1]。
- **FISTA** 的迭代格式可简化为：给定初始点 x^0，记 $x^{-1} = x^0$，对 $k \geq 1$，

$$y^k = x^k + \frac{k-1}{k+2}(x^k - x^{k-1}),$$
$$x^{k+1} = \mathrm{prox}_{\alpha_k r}(y^k - \alpha_k \nabla g(y^k)).$$

- 可选择固定步长 $\alpha_k = \alpha$，或通过线搜索方法寻找步长。
- FISTA 生成的点列使得 $f(x^k) - f(x^*)$ 的收敛速度为 $O\left(\frac{1}{k^2}\right)$！

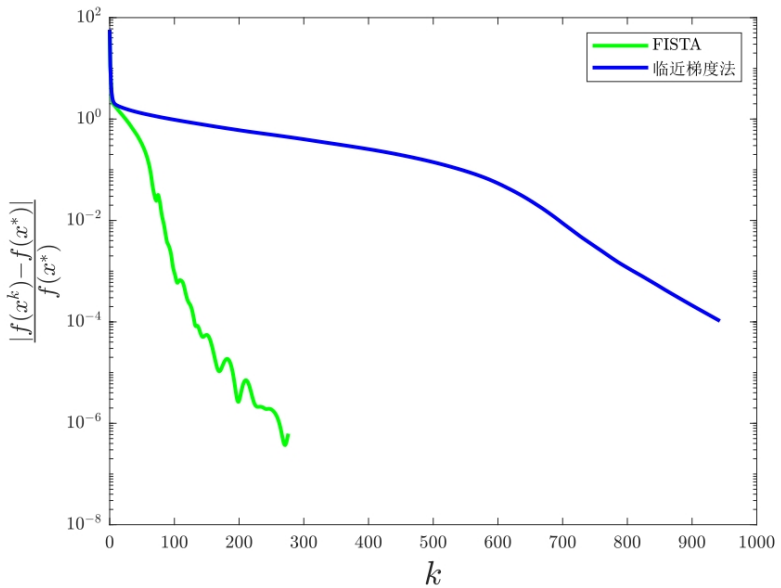
[1] Beck Amir and Marc Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems." SIAM Journal on Imaging Sciences, 2(1): 183-202, 2009.

Nesterov 第一加速方法

例. 求解 LASSO 问题：

$$\min_x \quad \frac{1}{2}\|Ax-b\|_2^2+\lambda\|x\|_1.$$

令 $m=100$，$n=500$，随机生成数据 $A\in\mathbb{R}^{m\times n}$，$b\in\mathbb{R}^m$，设置 $\lambda=1$，选取固定步长（由利普希茨常数决定）。



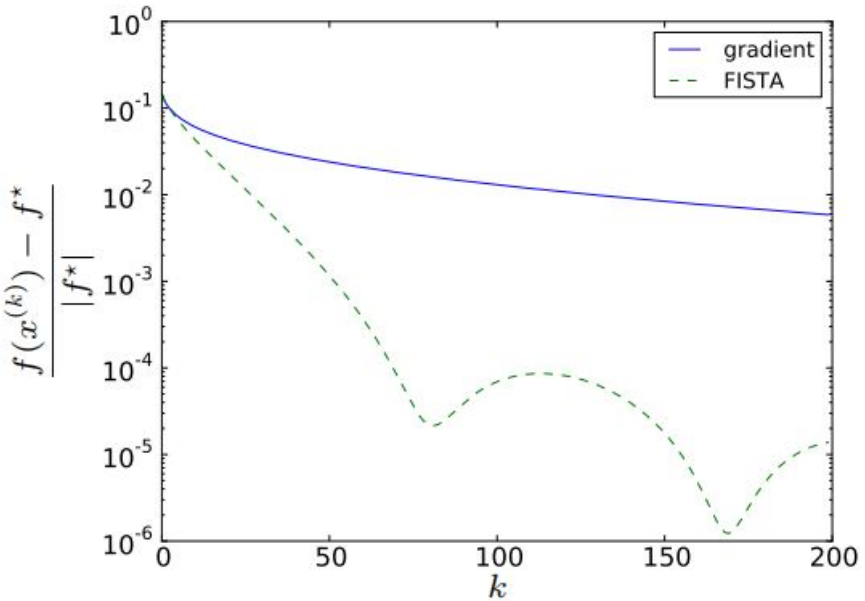
请留意 FISTA 产生的函数值序列 $\{f(x^k)\}$ 的非单调性!

Nesterov 第一加速方法

例. 最小化 log-sum-exp 函数：

$$\min_x \quad \log \sum_{i=1}^{m} \exp(a_i^\top x + b_i)$$

令 $m = 2000$，$n = 1000$，随机生成数据 $a_i \in \mathbb{R}^n$，$b_i \in \mathbb{R}$，$i = 1, 2, \cdots, m$，选取固定步长（由利普希茨常数决定），并记 $f^* := f(x^*)$。



*Nesterov 第二加速方法

Nesterov's 2nd acceleration:（只作简单了解）

$$
\begin{aligned}
y^k &= (1-\theta_k)x^k + \theta_k z^k, \\
z^{k+1} &= \arg\min_x \left\{ r(u) + g(y^k) + \langle \nabla g(y^k), u - y^k \rangle + \theta_k L \mathcal{D}_\phi(x, z^k) \right\}, \\
x^{k+1} &= (1-\theta_k)x^k + \theta_k z^{k+1}, \quad \leftarrow \text{更灵活!} \\
\theta_{k+1} &= (\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2)/2.
\end{aligned}
$$

- Nesterov 在 1988 年提出第二类加速梯度方法 (即上式中 $r \triangleq 0$ 的情形)[1]
- Auslender 和 Teboulle 在 2006 年将 Nesterov 第二加速技术推广，用于求解带有约束的凸优化问题。
- 许多学者针对 x^{k+1} 的更新提出了改进。
- 该框架能使用 Bregman 散度来刻画两点之间的“距离”：

$$\mathcal{D}_\phi(x, z^k) := \phi(x) - \phi(z^k) - \nabla\phi(z^k)^\top (x - z^k)$$

例： 当 $\phi(x) = \|x\|^2$ 时， $\mathcal{D}_\phi(x, z^k) = \|x - z^k\|^2$。

[1] See Section 2 in “Nesterov, Y.: Introductory Lectures on Convex Optimization, 2004” for more details.

*Nesterov 第三加速方法

Nesterov's 3rd acceleration:（只作简单了解）

$$
\begin{aligned}
y^k &= (1-\theta_k)x^k + \theta_k z^k, \\
z^{k+1} &= \arg\min_x \left\{ \sum_{i=0}^{k} \frac{\ell_{f_0}(x; y^i)}{\theta_i} + L\phi(x) \right\}, \\
x^{k+1} &= (1-\theta_k)x^k + \theta_k z^{k+1}, \quad \leftarrow \text{可选择其他更新形式!} \\
\theta_{k+1} &= \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}.
\end{aligned}
$$

其中 $\ell_{f_0}(x; y^i) := r(x) + g(y^i) + \langle \nabla g(y^i), x - y^i \rangle$。

- Nesterov 在 2005 年提出第三类加速梯度方法。
- 在此基础上，Nesterov 又在 2007 年提出了算法的改进版本。
- Paul Tseng 于 2008 年扩展并统一了几种 Nesterov 加速梯度方法。

*邻近点算法

回顾：

- 邻近梯度法适用于求解如下复合型凸优化问题：

$$\min_x \ f_0(x) := g(x) + r(x)$$

  - $g(x) : \mathbb{R}^n \to \mathbb{R}$ 凸、连续可微
  - $r(x) : \mathbb{R}^n \to \mathbb{R}$ 凸、邻近算子容易计算（不要求光滑）
- 迭代格式：$x^{k+1} = \mathrm{prox}_{\alpha_k r}\left((x^k - \alpha_k \nabla g(x^k)\right)$
- 当非光滑项 $r(x) \equiv 0$ 时， 邻近梯度法就退化为梯度下降法

思考：当光滑项 $g(x) \equiv 0$ 时，我们可以得到一个什么样的算法呢？

⋆ 邻近点算法 **(proximal point algorithm, PPA)**

$$x^{k+1} = \mathrm{prox}_{\alpha_k r}(x^k) \left(:= \arg\min_u \left\{ r(u) + \frac{1}{2\alpha_k}\|u - x^k\|^2 \right\} \right)$$

*邻近点算法

考虑一般形式的凸优化问题

$$\min_x \quad f_0(x),$$

其中 $f_0(x):\mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ 是适当闭凸函数，它的极小值 p^* 和最优解 x^* 存在。注意：这里并不要求 f_0 是连续可微的。

- **邻近点算法**的迭代格式：

$$x^{k+1} = \mathrm{prox}_{\alpha_k f_0}(x^k) = \arg\min_u \left\{ f_0(u) + \frac{1}{2\alpha_k}\left\|u - x^k\right\|^2 \right\},$$

其中步长 $\alpha_k > 0$ 可以固定，也可以自适应调整以提高实际收敛速度。

- 但值得注意的是，实际中 f_0 的邻近算子 $\mathrm{prox}_{\alpha_k f_0}(x^k)$ 往往没有显式解，因此需调用其它迭代算法对子问题进行（不精确地）求解！

*邻近点算法的收敛性

定理 6 (邻近点算法的收敛性)

设 $f_0(x)$ 是适当闭凸函数，它的极小值 p^* 和最优解 x^* 存在。则由**邻近点算法**得到的点列 $\{x^k\}$ 的函数值序列 $\{f_0(x^k)\}$ 收敛到最优值 $f_0(x^*)$，且

$$f_0(x^k) - f_0(x^*) \leqslant \frac{\|x^0 - x^*\|_2^2}{2\sum_{t=1}^k \alpha_t}, \quad \forall\, k \geqslant 1.$$

可以观察到：

- 若 $\lim\limits_{k\to\infty} \sum_{t=1}^k \alpha_t = +\infty$，则邻近点算法收敛。
- 显然，不同的步长会带来不同的收敛效果：
  - 若选取固定步长 $\alpha_k = \alpha$，则 $\sum_{t=1}^k \alpha_t \sim \mathcal{O}(k)$，此时算法具有次线性收敛率 $f_0(x^k) - f_0(x^*) \sim \mathcal{O}\left(\frac{1}{k}\right)$
  - 若 $\alpha_k = k+1$，则 $f_0(x^k) - f_0(x^*) \sim \mathcal{O}\left(\frac{1}{k^2}\right)$
  - 若 $\alpha_k = \alpha_0 c^k$，其中 $c > 1$，则 $f_0(x^k) - f_0(x^*) \sim \mathcal{O}\left(c^{-k}\right)$

*邻近点算法

思考：根据上述结论，是否可以令 α_k 极快地增长，从而使算法可以极快地收敛？

答：理论上可行，但实际中并不可行！

- 过大的 α_k 往往会导致子问题变地很病态，使之难以求解；
- 而当 α_k 非常大时，子问题与原问题几乎相当了。

邻近点算法 vs 邻近梯度法

- 邻近点算法可视作邻近梯度法在光滑项 $g(x) \equiv 0$ 时的特例；
- 事实上，邻近梯度法也可视作是一种**非精确的**邻近点算法！

- 线搜索
- 梯度类算法
- 梯度下降法的扩展
- 二阶算法

牛顿法

首先，回顾梯度下降法：

对连续可微的凸函数 f_0，在 x^k 附近作二阶近似：

$$f_0(x) \approx f_0(x^k) + \langle \nabla f_0(x^k),\, x - x^k \rangle + \frac{1}{2\alpha_k}\|x - x^k\|^2. \qquad (\triangle)$$

然后，求解二阶近似函数 $(\triangle)$ 的极小值点 x^{k+1} 作为每轮迭代的更新：

$$\begin{aligned} x^{k+1} &= \arg\min_x \left\{ f_0(x^k) + \langle \nabla f_0(x^k),\, x - x^k \rangle + \frac{1}{2\alpha_k}\|x - x^k\|^2 \right\} \\ &= x^k - \alpha_k \nabla f_0(x^k) \qquad \Leftarrow \text{ 梯度下降的迭代格式!} \end{aligned}$$

- 次线性收敛率（f_0 凸 + ∇f_0 利普西茨连续）；
- 线性收敛率（f_0 强凸 + ∇f_0 利普西茨连续）；

思考：若函数 $f_0(x)$ 充分光滑，能否利用问题的二阶信息实现对 f_0 更精确的近似，从而获得收敛速率更佳的算法？

牛顿法

考虑二次连续可微的无约束凸优化问题

$$\min_x \quad f_0(x),$$

其中 $f_0(x):\mathbb{R}^n\to\mathbb{R}$ 在 $\mathrm{dom}\, f_0=\mathbb{R}^n$ 上二次连续可微、凸，且 $f_0(x)$ 的极小值 p^* 和最优解 x^* 存在。

基于 Hessian 信息，对 $f_0(x)$ 在 x^k 附近作如下二阶近似：

$$f_0(x^k+d)\approx f_0(x^k)+\nabla f_0(x^k)^\top d+\frac{1}{2}d^\top\nabla^2 f_0(x^k)d. \qquad (\blacktriangle)$$

然后，极小化该二次近似函数得到的极小值点 d^k 作为新的迭代方向。

事实上，由最优性条件可知，极小值点 d^k 满足

$$\nabla f_0(x^k)+\nabla^2 f_0(x^k)d^k=0. \qquad \text{(牛顿方程)}$$

当 $\nabla^2 f_0(x^k)\succ 0$ 时，计算可得 $d^k=-\nabla^2 f_0(x^k)^{-1}\nabla f_0(x^k)$，这一方向称为牛顿方向! 易分析，牛顿方向是下降方向。

牛顿法

★ **牛顿法**：在 x^k 处选择牛顿方向 $-\nabla^2 f_0(x^k)^{-1}\nabla f_0(x^k)$ 作为下降方向 d^k

- $x^{k+1} = x^k - \nabla^2 f_0(x^k)^{-1}\nabla f_0(x^k)$;
- **注意**：迭代格式隐含了 $\alpha_k = 1$；称 $\alpha_k = 1$ 的牛顿法为**经典牛顿法**。

定理 7 (经典牛顿法的局部收敛性)

设 f_0 是定义在 $\mathbb{R}^n$ 上的二阶连续可微的凸函数，$x^* \in \arg\min_x f_0(x)$ 存在，且 Hessian 矩阵 在最优值点 x^* 的一个邻域 $N_\delta(x^*)$ 内是 L_H-利普希茨的，即存在常数 $L_H > 0$ 使得

$$\left\|\nabla^2 f_0(x) - \nabla^2 f_0(y)\right\| \leqslant L_H\|x-y\|, \quad \forall x, y \in N_\delta(x^*).$$

若 $f_0(x)$ 在 x^* 处满足 $\nabla f_0(x^*) = 0$，$\nabla^2 f_0(x^*) \succ 0$，则

1. （**局部收敛性**） 若初始点 x^0 距离 x^* 足够近，则牛顿法产生的点列 $\{x^k\}$ 收敛至 x^*;
2. （**局部二次收敛率**） 序列 $\{x^k\}$ 收敛到 x^* 的速度是 Q-二次的;
3. （**局部二次收敛率**） 序列 $\{\|\nabla f(x^k)\|\}$ 收敛到 0 的速度是 Q- 二次的。

牛顿法

经典牛顿法的局限性：

- 只有局部收敛性，故对初始点 x^0 的选取要求很高；
- 要求 $\nabla^2 f_0(x^*) \succ 0$，否则无法保证局部二次收敛率，甚至无法保证在 x^* 的附近满足牛顿方程的方向是下降方向；
- 计算代价高，每轮迭代需计算 $\nabla f_0(x^k)$ 和 $\nabla^2 f_0(x^k)$，并且须要高精确地求解牛顿方程。

针对上述缺陷，学者们提出了很多改进方法：

- 因为牛顿方向是下降方向，故可引入线搜索准则寻找步长 α_k 以实现全局收敛，得到**阻尼牛顿法** (Damped Newton method)；
- 当满足牛顿方程的方向不是下降方向时，可改用负梯度方向，保证搜索方向的下降性，得到梯度法与牛顿法相结合的一种“混合算法”；
- 为减少计算量，可考虑近似计算牛顿方向，基于该思路设计的方法称为**拟牛顿法** (Quasi-Newton method)。

拟牛顿法

拟牛顿法的思想：构造 $\nabla^2 f_0(x^k)$ 的近似矩阵 B_k 或 $\nabla^2 f_0(x^k)^{-1}$ 的近似矩阵 H_k，使得方程更易求解以减少计算量。

★ **拟牛顿法**的通用迭代格式：

- 求解如下模型得到拟牛顿方向 d^k：

$$\min_d \quad f_0(x^k) + \nabla f_0(x^k)^\top d + \frac{1}{2} d^\top B_k d,$$

  其中 B_k 是 $\nabla^2 f_0(x^k)$ 的一个近似矩阵。当 $B_k \succ 0$ 时，可直接计算得到拟牛顿方向 $d^k = -B_k^{-1} \nabla f_0(x^k)$；
- 由线搜索寻找合适的步长 α_k；
- 更新 $x^{k+1} = x^k - \alpha_k d^k = x^k - \alpha_k B_k^{-1} \nabla f_0(x^k)$；
- 更新近似矩阵 B_{k+1}；

拟牛顿法的核心问题：如何找到一个简单方法构建矩阵 B_k 以近似 $\nabla^2 f_0(x^k)$，并能保证拟牛顿方向 d^k 容易计算？

由此衍生出了 DFP，BFGS 等著名的拟牛顿法。本课程将不再介绍。