Subject: Lecture 3, Descent Methods

Date: from July 14, 2025 to July 23, 2025

Contents

Descent Methods

Outline

本节主要介绍了基于梯度的优化算法及其理论分析，通过介绍下降方向所需要满足的条件，引出最速下降方向和最速梯度下降法.

针对定步长的最速梯度下降法，在 1. $L-$smooth; 2. $L-$smooth + convex; 3. $L-$smooth + strongly convex 三种情形下进行收敛性分析并比较了相关收敛速率.

针对一般的梯度下降法，在相邻函数下降量与梯度大小相关的假设(29)下对其进行了理论分析.

针对一般梯度下降法，介绍了如何使用 line search 对下降方向进行选择. 接着介绍了在得到下降方向后使用 line search 选择步长的三种方法：1. 精确线搜索；2. 近似线搜索；3. 回溯线搜索. 其中在近似线搜索中介绍了重要的条件：weak Wolfe conditions.

基于 Lecture 2 主要介绍的如何寻找满足一阶最有条件的方法，本节进一步介绍了如何寻找满足二阶必要条件点的基本方法并进行了收敛性分析.

在附录中，附加了几个较长证明. 同时介绍了最速梯度下降法的另一种理解：mirror descent 并进行了理论分析. 最后，简要介绍了 PL property 与 KL property，由于涉及后续内容，因此此处仅简单带过.

Introduction

Prelimilary

基于梯度的优化算法构成了本书的基本内容，其概括来说是一类使用梯度信息以获取每次算法迭代中目标函数下降的方法.

在本节中主要考虑如下光滑凸函数的无约束优化问题：

$$\min_{x\in\mathbb{R}^n} f(x) \tag{1}$$

其中 f 为光滑的凸函数. 并且假设其梯度 ∇f 可以准确获取.

Note 1. 光滑+凸的强假设是为方便进行基础的理论分析，在放宽假设后会有如下变体：

1. 目标函数 f 由一项光滑凸函数项和一项非凸正则化项*(regularization term)*相加
2. 在约束集合*(constraint sets)*上进行约束优化
3. f 的函数值或其梯度值 ∇f 不能被准确估计或很难直接获得，但二者的无偏估计获取较容易
4. 目标函数 f 光滑但非凸

Descent Directions

本书中的大多数方法都可以形式化为生成一个迭代序列 $\{x_k\}$，使得 $f(x_{k+1}) < f(x_k), k = 0, 1, 2, \ldots$ 使函数值变小的方向被称为下降方向(descent direction)，定义如下：

Definition 1 (Descent Direction)**.** 称 d 为 f 在点 x 处的下降方向*(descent direction)*若对于任意足够小的 t， $f(x + td) < f(x)$.

Proposition 1. 若 f 在 x 的某邻域中连续可微，那么任意使得 $d^T\nabla f(x) < 0$ 的方向 d 都是下降方向. 其中 $d = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$ 为最小化 $d^T\nabla f(x)$ 的标准化方向.

Proof. 假设在邻域 $\mathcal{N}(x)$ 内 f 连续可微，且 $d^T\nabla f(x) < 0$. 由 ∇f 在邻域内的连续性可知，存在 $\bar{t} > 0$，使得 $\nabla f(x + td)^T d < 0, \forall\, x + td \in \mathcal{N}(x),\ t \in [0, \bar{t}]$. 由 mean-value form 的 Taylor Theorem 可得存在 $\gamma \in (0, 1)$，使得

$$f(x + td) = f(x) + t\nabla f(x + \gamma td)^d < f(x) \tag{2}$$

因此 d 为下降方向. 由于 $d^T\nabla f(x) = \|\nabla f(x)\| d^T \frac{\nabla f(x)}{\|\nabla f(x)\|}$，其中 $d, \frac{\nabla f(x)}{\|\nabla f(x)\|}$ 均为标准化方向(向量)，显然二者方向相反时可以取得最小值为 -1. □

根据 Proposition 1，将 $-\nabla f(x)$ 称为最速下降方向(steepest-descent direction). 使用此方向进行函数值迭代更新可以形式化为：

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k), k = 0, 1, 2, \ldots \tag{3}$$

其中 $\alpha_k > 0$ 为步长(step length). 这种方法被称为 梯度下降法(gradient descent method)或最速梯度下降法(steepest-descent method). 后文将统一称这种方法为最速梯度下降法.

Steepest-Descent Method

How to choose the step length α_k?

如何选取步长 α_k 是一个重要问题，因为若 α_k 过大则有使函数值变大的风险，若 α_k 过小则可能需要过多次数的迭代才能找到解. 下面考虑最简单的情形： f 满足 $L-$smooth 且 α_k 固定为 α，即式(3)变为：

$$x^{k+1} = x^k - \alpha \nabla f(x^k), k = 0, 1, 2, \ldots \tag{4}$$

根据第一节中 Lemma 1 可知

$$f(x + \alpha d) \leqslant f(x) + \alpha \nabla f(x)^T d + \alpha^2 \frac{L}{2} \|d\|^2 \tag{5}$$

令 $d = -\nabla f(x)$，当 $\alpha = \frac{1}{L}$ 时不等号右端项取得最小值，此时满足

$$f(x^{k+1}) = f(x^k - \frac{1}{L}\nabla f(x^k)) \leqslant f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2. \tag{6}$$

Note 2. 由式*(6)*可知 f 函数值的下降量与当前迭代步的梯度范数 $\|\nabla f(x^k)\|$ 和梯度的 *Lipschitz* 系数 L 有关.

Convergence analysis – General Case

在一般情况(general case)下对于最速梯度下降方法有如下结论：

Theorem 1 (General Case)**.** 若 f 有下界的连续可微 $L-$*smooth* 函数，即存在 $\bar{f}$，使得

$$f(x) > \bar{f},\ \forall\, x \in dom(f) \tag{7}$$

那么在 T 步迭代后最速梯度下降方法可以找到点 x 满足：

$$\min_{0\leqslant k\leqslant T-1}\|\nabla f(x^k)\|^2\leqslant\frac{2L[f(x^0)-\bar{f}]}{T} \tag{8}$$

Proof. 将式(6)累和至 $k=T$ 可得

$$f(x^T)\leqslant f(x^0)-\frac{1}{2L}\sum_{T-1}^{k=0}\|\nabla f(x^k)\|^2\Rightarrow\sum_{T-1}^{k=0}\|\nabla f(x^k)\|^2\leqslant 2L[f(x^0)-f(x^T)]. \tag{9}$$

由于 $\bar{f}<f(x^T)$，得

$$\sum_{T-1}^{k=0}\|\nabla f(x^k)\|^2\leqslant 2L[f(x^0)-\bar{f}]. \tag{10}$$

由于式(10)右端有界，因此得到 $\lim_{T\to\infty}\|\nabla f(x^T)\|=0$. 进一步可以得到：

$$\min_{0\leqslant k\leqslant T-1}\|\nabla f(x^k)\|^2\leqslant\frac{1}{T}\sum_{k=1}^{T-1}\|\nabla f(x^k)\|^2\leqslant\frac{2L[f(x^0)-\bar{f}]}{T}, \tag{11}$$

即在 T 步迭代后最速梯度下降方法可以找到点 x 满足：

$$\min_{0\leqslant k\leqslant T-1}\|\nabla f(x^k)\|^2\leqslant\frac{2L[f(x^0)-\bar{f}]}{T} \tag{12}$$

□

Note 3. 式*(33)*中的收敛速度*(convergence rate)*很慢，并且只能告诉我们可以找打到一个近似一阶驻点 x^k. 若要得到更强的结论则需要更多更强的假设.

Convergence analysis – Convex Case

当 f 同时满足凸性(Convex Case)时，会有如下更强的结论：

Theorem 2 (Convex Case)**.** 若 f 为有下界的连续可微的凸 $L-$*smooth* 函数，且下界设为 $\bar{f}$. 假设问题*(1)*有最优解，设为 x^*，定义 $f^*\triangleq f(x^*)$. 那么 $\alpha_k=\frac{1}{L}$ 时定步长最速梯度下降方法产生的迭代序列 $\{x^k\}_{k=0}^{\infty}$ 满足

$$f(x^T)-f^*\leqslant\frac{L}{2T}\|x^0-x^*\|^2,\quad T=1,2,\ldots. \tag{13}$$

Proof. 由凸性可知

$$f(x^*)\geqslant f(x^k)+\nabla f(x^k)^T(x^*-x^k), \tag{14}$$

代入式(6)中可得

$$f(x^{k+1})\leqslant f(x^*)+\nabla f(x^k)^T(x^k-x^*)-\frac{1}{2L}\|\nabla f(x^k)\|^2 \tag{15}$$

由于 $\nabla f(x^k) = -L(x^{k+1} - x^k)$，因此

$$\begin{aligned}
&\nabla f(x^k)^T(x^k - x^*) - \frac{1}{2L}\|\nabla f(x^k)\|^2 \\
&= L(x^k - x^{k+1})^T(x^k - x^*) - \frac{L^2}{2L}(x^k - x^{k+1})^T(x^k - x^{k+1}) \\
&= (x^k - x^{k+1})^T(Lx^k - Lx^* - \frac{L}{2}x^k + \frac{L}{2}x^{k+1}) \\
&= \frac{L}{2}(x^k - x^{k+1})^T(x^k + x^{k+1} - 2x^*) \\
&= \frac{L}{2}[(x^k - x^*) - (x^{k+1} - x^*)]^T[(x^k - x^*) + (x^{k+1} - x^*)] \\
&= \frac{L}{2}[\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2]
\end{aligned} \tag{16}$$

代入式(15)中，得

$$\begin{aligned}
&f(x^{k+1}) - f(x^*) \leqslant \frac{L}{2}\left[\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right] \\
&\Rightarrow \sum_{k=0}^{T-1}\left(f(x^{k+1}) - f^*\right) \leqslant \frac{L}{2}\sum_{k=0}^{T-1}\left[\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right] \\
&= \frac{L}{2}\left[\|x^0 - x^*\|^2 - \|x^T - x^*\|^2\right] \leqslant \frac{L}{2}\|x^0 - x^*\|^2
\end{aligned} \tag{17}$$

由于序列 $\{f(x^k)\}$ 单调递减，因此有

$$f(x^T) - f^* \leqslant \frac{1}{T}\sum_{k=0}^{T-1}\left(f(x^{k+1}) - f(x^*)\right) \leqslant \frac{L}{2T}\|x^0 - x^*\|^2 \tag{18}$$

□

Convergence analysis – Strongly Convex Case

Lemma 1. 若 f 为连续可微的强凸函数，且凸模为 m，则

$$f(x) - f(x^*) \leqslant \frac{\|\nabla f(x)\|^2}{2m} \tag{19a}$$

$$\|x - x^*\|^2 \leqslant \frac{2}{m}\|\nabla f(x)\| \tag{19b}$$

Proof. 详见附录A. □

Note 4. 注意 *Lemma 1* 并不需要函数 f 满足 $L-smooth$.

将式(65b)代入式(6)中得

$$\begin{aligned}
&f(x^{k+1}) = f(x^k - \frac{1}{L}\nabla f(x^k)) \leqslant f(x^k) - \frac{1}{2L}\|f(x^k)\|^2 \leqslant f(x^k) - \frac{m}{L}(f(x^k) - f(x^*)) \\
&\Rightarrow f(x^{k+1}) - f(x^*) \leqslant (1 - \frac{m}{L})(f(x^k) - f(x^*))
\end{aligned} \tag{20}$$

从而得到如下线性收敛性：

$$f(x^T) - f(x^*) \leqslant (1 - \frac{m}{L})^T (f(x^0) - f(x^*)). \tag{21}$$

Comparison between Rates

下面分析对于足够小的 $\epsilon > 0$，找到近似满足一阶必要条件 $f(x^*) = 0$ 的点 x^k 使得 $\nabla f(x^k) \leqslant \epsilon$ 或 $f(x^k) = f(x^*)$ 对迭代次数 T 的要求.

General Case. 由式(33)可知，若满足

$$\frac{2L[f(x^0) - \bar{f}]}{T} \leqslant \epsilon^2 \Rightarrow T \geqslant \frac{2L[f(x^0) - \bar{f}]}{\epsilon^2} \tag{22}$$

则有

$$\min_{0 \leqslant k \leqslant T-1} \|\nabla f(x^k)\|^2 \leqslant \frac{2L[f(x^0) - \bar{f}]}{T} \leqslant \epsilon^2 \Rightarrow \|\nabla f(x^k)\| \leqslant \epsilon \tag{23}$$

Convex Case. 由式(13)可知

$$f(x^T) - f^* \leqslant \frac{L}{2T}\|x^0 - x^*\|^2 \leqslant \epsilon \Rightarrow T \geqslant \frac{L}{2\epsilon}\|x^0 - x^*\|^2. \tag{24}$$

Strongly Convex Case. 由式(21)可得

$$\begin{aligned} & f(x^T) - f(x^*) \leqslant (1 - \frac{m}{L})^T (f(x^0) - f(x^*)) \leqslant \epsilon \\ & \Rightarrow T \cdot \log(\frac{L-m}{L}) \leqslant \log \frac{\epsilon}{f(x^0) - f(x^*)} \\ & \Rightarrow T \geqslant \frac{\log \frac{\epsilon}{f(x^0)-f(x^*)}}{\log(1 - \frac{m}{L})} \geqslant \frac{\log \frac{\epsilon}{f(x^0)-f(x^*)}}{-\frac{m}{L}} = \frac{L}{m} \log \frac{f(x^0) - f(x^*)}{\epsilon} \end{aligned} \tag{25}$$

其中最后一个不等号使用了不等式 $\log(1 - x) \geqslant -x$.

Note 5. 值得强调和回忆的是，在计算过程中，由 *$L-$smooth* 的性质*(Lecture 1, Lemma 1)*，初识点与最优点函数值间的差异能够被二者距离控制：

$$f(x^0) - f^* \leqslant \frac{L}{2}\|x^0 - x^*\|^2 \tag{26}$$

Case(with $L-$smooth)	Iteration Threshold	Rate Type
General Case	$T \geqslant \frac{2L[f(x^0)-\bar{f}]}{\epsilon^2}$	
Convex Case	$T \geqslant \frac{L}{2\epsilon}\|x^0 - x^*\|^2$	linear rate
Strongly Convex Case	$T \geqslant \frac{L}{m} \log \frac{f(x^0)-f(x^*)}{\epsilon}$	sublinear rate

Table 1: Comparison between Rates

Convergence Analysis for General Descent Methods

Introduction

在前面的内容中，我们根据梯度的全局 Lipschitz 系数 L 选取了步长 $\frac{1}{L}$ 并分析了固定步长的最速梯度下降法的收敛性. 下面分析更一般的情况，即分析：

$$x^{k+1} = x^k + \alpha_k d^k, k = 1, 2, 3, \ldots, \tag{27}$$

其中 α_k, d^k 分别为第 k 步的步长($\alpha_k > 0$)和迭代方向.

注意，对最速梯度下降方法的分析都是基于式(6)，即

$$f(x^{k+1}) = f(x^k - \frac{1}{L}\nabla f(x^k)) \leqslant f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2. \tag{28}$$

在对更一般的梯度下降法式(27)的分子中，将证明如下类似的性质也会被满足：

$$f(x^{k+1}) \leqslant f(x^k) - C\|\nabla f(x^k)\|^2, \text{ for some } C > 0. \tag{29}$$

在证明性质(29)被一些梯度下降法满足前，先说明满足该性质的算法产生的序列的良好性质.

Theorem 3. 假设 f 有下界且满足*L−smooth*性，那么由某种满足性质*(29)*的模式产生的序列 $\{x_k\}$ 的聚点*(accumulation points)* $\bar{x}$ 是驻点*(stationary point)*，即满足 $\nabla f(\bar{x}) = 0$. 若 f 是凸的，则 $\bar{x}$ 是问题*(1)*的解.

Proof. 根据条件可知序列 $\{f(x_k)\}$ 单调递减有下界，因此必有下确界，故 $\lim\limits_{k\to\infty} f(x^k) - f(x^{k+1}) = 0$. 由式(29)可得

$$\|\nabla f(x^k)\|^2 \leqslant \frac{[f(x^k) - f(x^{k+1})]}{C}, \quad k = 0, 1, 2, \ldots, \tag{30}$$

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

由于 $\bar{x}$ 为聚点，根据定义可知存在 $\{1, 2, 3, \ldots\}$ 的子序列 $\mathcal{S}$ 使得 $\lim\limits_{k\in\mathcal{S}, k\to\infty} x^k = \bar{x}$. 根据 ∇f 的连续性，有

$$\nabla f(\bar{x}) = \lim_{k\in\mathcal{S}, k\to\infty} \nabla f(x^k) = 0 \tag{31}$$

□

Convergence Analysis under Property (29)

Corollary 1. 若 f 有下界的连续可微函数，即存在 $\bar{f}$，使得

$$f(x) > \bar{f}, \forall\, x \in dom(f) \tag{32}$$

那么在 T 步满足性质*(29)*的迭代后产生的序列 $\{x^k\}$ 满足：

$$\min_{0\leqslant k\leqslant T-1} \|\nabla f(x^k)\|^2 \leqslant \frac{f(x^0) - \bar{f}}{CT} \tag{33}$$

若 f 还满足强凸性，则

$$f(x^{k+1}) - f(x^*) \leqslant f(x^k) - f(x^*) - C\|\nabla f(x^k)\|^2 \leqslant (1 - 2mC)[f(x^k) - f(x^*)] \tag{34}$$

Proof. **Part 1.** 根据 Theorem 1 即得.

Part 2. 根据 Lemma 1 式(65a)可得

$$\|\nabla f(x^k)\|^2 \geqslant 2m(f(x^k) - f(x^*)) \tag{35}$$

因此

$$\begin{aligned} &f(x^{k+1}) - f(x^*) \leqslant f(x^k) - f(x^*) - C\|\nabla f(x^k)\|^2 \\ &\leqslant f(x^k) - f(x^*) - 2mC(f(x^k) - f(x^*)) = (1 - 2mC)[f(x^k) - f(x^*)] \end{aligned} \tag{36}$$

□

Theorem 4. 若 f 为光滑的凸函数，且其梯度 ∇f 满足 *Lipschitz* 性质*(*系数为 L*)*，设问题*(1)*有解 x^*. 如下定义的水平集*(level set)*有界，即 $R_0 < \infty$:

$$R_0 \triangleq \max\{\|x - x^*\| | f(x) \leqslant f(x^0)\}. \tag{37}$$

那么满足性质*(29)*的梯度下降法产生的序列 $\{x^k\}_0^\infty$ 满足:

$$f(x^T) - f(x^*) \leqslant \frac{R_0^2}{CT}, T = 1, 2, \ldots. \tag{38}$$

Proof. 详见附录B □

Note 6. 水平集*(level set)*定义了算法产生的点的与最优解 x^* 的最远距离.

Line-Search Methods: How to choose the direction

Line-Search Methods: Choosing the Direction

由于我们考虑的是一般意义的梯度下降法(27)，其中 α_k, d^k 都可以改变，也都可以试图寻找每一步的最佳步长 α_k 和方向 d^k，下面先考虑对前进方向 d^k 进行搜索.

在对 d^k 进行搜索时，常要求其满足其满足如下两条性质:

$$0 < \bar{\epsilon} \leqslant \frac{-(d^k)^T \nabla f(x^k)}{\|\nabla f(x^k)\| \|d^k\|}, \tag{39a}$$

$$0 < \gamma_1 \leqslant \frac{\|d^k\|}{\|\nabla f(x^k)\|} \leqslant \gamma_2, \tag{39b}$$

其中 $\bar{\epsilon}, \gamma_1, \gamma_2$ 均为正的常数.

Note 7. 条件*(39a)*要求 d^k 与 $-\nabla f(x^k)$ 之间的夹角是锐角，且 $\bar{\epsilon}$ 越靠近 *1* 则夹角越小. 条件*(39b)*要求 d^k 与 $-\nabla f(x^k)$ 的长度相差也不大. 注意若 $\nabla f(x^k) = 0$ 那么说明已达到驻点，算法将会停止，因此不考虑 $\|\nabla f(x^k)\| = 0$.

对于最速梯度下降法 $d^k = -\nabla f(x^k)$，此时 $\bar{\epsilon} = \gamma_1 = \gamma_2 = 1$.

在满足上述两条性质的前提下，由 Taylor's Theorem 可知

$$\begin{aligned} &f(x^{k+1}) = f(x^k + \alpha d^k) \leqslant f(x^k) + \alpha \nabla f(x^k)^T d^k + \alpha^2 \frac{L}{2} \|d^k\|^2 \\ &\leqslant f(x^k) - \alpha\bar{\epsilon}\|\nabla f(x^k)\| \|d^k\| + \alpha^2 \frac{L}{2} \|d^k\|^2 \leqslant f(x^k) - \alpha \left(\bar{\epsilon} - \alpha \frac{L}{2} \gamma_2\right) \|\nabla f(x^k)\| \|d^k\|. \end{aligned} \tag{40}$$

因此只需要 $\alpha \in (0, \frac{2\bar{\epsilon}}{L\gamma_2})$ 就可以保证 x^k 非驻点的前提下 $f(x^{k+1}) < f(x^k)$.

Possible choices of d^k

一些可供选择的前进方向列举如下：

1. 变化负梯度方向(transformed negative gradient direction) $d^k = -S^k \nabla f(x^k)$，其中 S^k 为特征值在 $[\gamma_1, \gamma_2]$ 范围内的对称正定矩阵，且 $\gamma_2 > \gamma_1 > 0$.

Note 8. 条件*(39b)*满足是由于 S^k 的特征值在 $[\gamma_1, \gamma_2]$ 范围内. 条件*(39a)*也满足，且 $\bar{\epsilon} = \frac{\gamma_1}{\gamma_2}$:

$$-(d^k)^T \nabla f(x^k) = \nabla f(x^k)^T S^k \nabla f(x^k) \geqslant \gamma_1 \|\nabla f(x^k)\|^2 \geqslant \frac{\gamma_1}{\gamma_2} \|\nabla f(x^k)\| \|d^k\|. \tag{41}$$

牛顿法*(Newton's method)*选取 $S^k = \nabla^2 f(x^k)^{-1}$. 若对于所有的 x，*Hessian* $\nabla^2 f(x)$ 具有一致有界在 $[\frac{1}{\gamma_2}, \frac{1}{\gamma_1}]$ 范围内的特征值.

2. 按坐标下降的 Gauss-Southwell 变体(Gauss–Southwell variant of coordinate descent) $d^k = -[\nabla f(x^k)]_{i_k} e_{i_k}$，其中 $i_k = \arg \max_{i=1,2,\ldots,n} |[\nabla f(x^k)]_i|$ 为 i_k 处为 1 的单位向量.

Line-Search Methods: How to choose the Step Length

Line-Search Methods: Choosing the Step Length

在根据原则(39a)(39b)得到下降方向 d^k 后，需要选择步长[a]. 步长选择有定步长和变步长两类，其中变步长介绍三种选择方式：1. 精确线搜索；2. 近似线搜索；3. 回溯线搜索.

Fixed Steplength. 将步长进行固定，即式(4)的更新模式. 这种方式可以得到相应的收敛性结果，但缺点是需要先验信息适当地选择步长，有如下两种方法：

1. **Trial and Error.** 试错法，通过多次实验寻找合适的步长，合理的想法是选择使算法不会发散的尽量大的 α. 这实际使用实验经验作为先验信息，通过反复试验来估计 ∇f 的 Lipschitz 系数.
2. **Based on Function Info.** 先验信息为函数的某些信息，如 ∇f 的 Lipschitz 系数 L，强凸模 m，或式(39a)和式(39b)中的系数.

Note 9. 例如给定式*(39a)*和式*(39b)*中的系数，令 $\alpha = \frac{\bar{\epsilon}}{L\gamma_2}$，由式*(40)*可得

$$f(x^{k+1}) \leqslant f(x^k) - \frac{\bar{\epsilon}}{2L\gamma_2} \|\nabla f(x^k)\| \|d^k\| \geqslant f(x^k) - \frac{\bar{\epsilon}}{2L\gamma_2} \|\nabla f(x^k)\|^2 \tag{42}$$

Exact Line Search. 对方向 d^k 使用一维精确线搜索(one-dimensional exact line search)：

$$\min_{\alpha > 0} \ f(x^k + \alpha d^k) \tag{43}$$

注意，这种方式想要 work 需要能够经济地计算函数值[b] $f(x^k + \alpha d^k)$. 许多情况下这都可以做到，例如当 f 为多元多项式时，问题(43)自然转化为寻根问题；当进行坐标方向梯度下降时，即 d^k 是坐标方向，对于某些函数计算函数值也较容易.

Approximate Line Search. 一般来说精确线搜索是昂贵且不必要的，为更好的实验性能常使用近似线搜索，即只需要近似地找到满足某种条件的步长即可. 这些条件使程序能经济地找到保证良好收敛性能的步长，其中两个经典的条件是 **weak Wolfe Conditions**：

$$f(x^k + \alpha d^k) \leqslant f(x^k) + c_1 \alpha \nabla f(x^k)^T d^k, \tag{44a}$$

$$\nabla f(x^k + \alpha d^k)^T d^k \geqslant c_2 \nabla f(x^k)^T d^k. \tag{44b}$$

这里的 $0 < c_1 < c_2 < 1$.

式(44a)被称为“充分减少条件”(**sufficient decrease condition**)，根据下降方向 d^k 满足的条件 (39a)可知：

$$f(x^k) - f(x^k + \alpha d^k) \geqslant -c_1 \alpha \nabla f(x^k)^T d^k > 0. \tag{45}$$

说明相邻迭代下降量至少是与一阶泰勒展开有关项的 c_1 倍.

式(44b)被称为“梯度条件”(**gradient condition**)，确保 α_k 不会太小，使得 f 在 d^k 的方向导数更大(更靠近 0 的负数或 0 或正数)，即在方向 d^k 移动了足够长的距离.

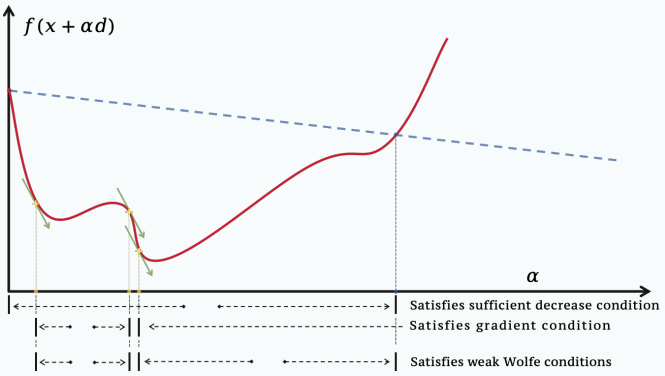

Figure 1: weak Wolfe Conditions

Theorem 5. 若 f 为连续可微的 *L−smooth* 函数，梯度下降法*(27)*中迭代方向满足条件*(39a)(39b)*，步长满足 *weak Wolfe Conditions (44a)(44b)*，则相邻迭代步函数值下降量满足性质*(29)*. 更详细的，为：

$$f(x^k) - f(x^{k+1}) \geqslant \frac{c_1(1-c_2)}{L} \bar{\epsilon}^2 \|\nabla f(x^k)\|^2. \tag{46}$$

Proof. 根据条件(44b)，可知

$$-(1-c_2)\nabla f(x^k)^T d^k \leqslant [\nabla f(x^k + \alpha_k d^k) - \nabla f(x^k)]^T d^k \tag{47}$$

根据 Cauchy Schwarz 不等式，同时由梯度的 Lipschtz 性可知

$$[\nabla f(x^k + \alpha_k d^k) - \nabla f(x^k)]^T d^k \leqslant \|\nabla f(x^k + \alpha_k d^k) - \nabla f(x^k)\| \|d^k\| \leqslant L\alpha_k \|d^k\|^2. \tag{48}$$

从而得到

$$\alpha_k \geqslant -\frac{(1-c_2)}{L} \frac{\nabla f(x^k)^T d^k}{\|d^k\|^2} \tag{49}$$

代入条件(44a)可得

$$\begin{aligned} f(x^k) - f(x^{k+1}) &= f(x^k) - f(x^k + \alpha_k d^k) \geqslant -c_1 \alpha_k \nabla f(x^k)^T d^k \\ &\geqslant \frac{c_1(1-c_2)}{L} \frac{\left(\nabla f(x^k)^T d^k\right)^2}{\|d^k\|^2} \geqslant \frac{c_1(1-c_2)}{L} \bar{\epsilon}^2 \|\nabla f(x^k)\|^2 \end{aligned} \tag{50}$$

□

(1)中提出了一种结合外推和二分法来找到满足 weak Wolfe Conditions 的方法，详见算法1.

Algorithm 1 Extrapolation-Bisection Line Search (EBLS)

Given $0 < c_1 < c_2 < 1$, set $L \leftarrow 0$, $U \leftarrow +\infty$, $\alpha \leftarrow 1$;
repeat
  if $f(x+\alpha d) > f(x) + c_1\alpha\nabla f(x)^T d$ **then**
    $U \leftarrow \alpha$, $\alpha \leftarrow \frac{U+L}{2}$
  else if $\nabla f(x+\alpha d)^T d < c_2 \nabla f(x)^T d$ **then**
    $L \leftarrow \alpha$;
    if $U = +\infty$ **then**
      $\alpha \leftarrow 2L$;
    else
      $\alpha \leftarrow \frac{L+U}{2}$;
    end if
  else
    Stop (Success!);
  end if
until Forever

Backtracking Line Search. 确定合适的 α_k 的另一种常用方法称为"回溯"(backtracking)，广泛应用于评估 f 较容易但是评估 ∇f 较困难的情况. 其优点在于：1. 易于部署（如不需要评估 Lipschitz 系数 L）；2. 仍能保证好的收敛性.

在其最简单的变体中，考虑 $\bar{\alpha} > 0$ 作为初始猜测，选取 $\beta \in (0,1)$ 构造序列 $\bar{\alpha}, \beta\bar{\alpha}, \beta^2\bar{\alpha}, \beta^3\bar{\alpha}, \ldots$ 使得 sufficient decrease condition (44a) 满足即可.

Note 10. 注意 *backtracking* 无需满足条件*(44b)*，因为该条件是为确保 α_k 不会太小，但是在 *backtracking* 中 α_k 是通过 $\bar{\alpha}$ 逐渐递减地"试探"得到的，所以不会特别小.

当初始猜测就满足条件(44a)，即无 backtracking 时，$\alpha_k = \bar{\alpha_k}$，根据下降方向条件(39a)(39b)有

$$f(x^{k+1}) \leqslant f(x^k) + c_1\bar{\alpha}\nabla f(x^k)^T d^k \leqslant f(x^k) - c_1\bar{\alpha}\bar{\epsilon}\gamma_1\|\nabla f(x^k)\|^2. \tag{51}$$

当需要 backtracking 时，考虑反面情况，即对于当前步长 α 只有使用了 backtracking 才能满足条件(44a)，此时 $\alpha = \beta\alpha_k$ 时不满足条件(44a)，有

$$f(x^k + \beta^{-1}\alpha_k d^k) > f(x^k) + c_1\beta^{-1}\alpha_k\nabla f(x^k)^T d^k. \tag{52}$$

根据 Lecture 2, Lemma 1 可知

$$f(x^k + \beta^{-1}\alpha_k d^k) \leqslant f(x^k) + \beta^{-1}\alpha_k\nabla f(x^k)^T d^k + \frac{L}{2}(\beta^{-1}\alpha_k)^2\|d^k\|^2. \tag{53}$$

从而得到

$$\alpha_k \geqslant -\frac{2}{L}\beta(1-c_1)\frac{\nabla f(x^k)^T d^k}{\|d^k\|^2}. \tag{54}$$

由于 α_k 满足条件(44a)，结合步长条件(39a)可得

$$\begin{aligned} f(x^{k+1}) &\leqslant f(x^k) + c_1\alpha_k\nabla f(x^k)^T d^k \leqslant f(x^k) - \frac{2}{L}\beta(1-c_1)c_1\frac{(\nabla f(x^k)^T d^k)^2}{\|d^k\|^2} \\ &\leqslant f(x^k) - \frac{2}{L}\beta c_1(1-c_1)\bar{\epsilon}^2\|\nabla f(x^k)\|^2. \end{aligned} \tag{55}$$

[a] 在机器学习(machine learning, ML)任务中常被称为学习率(learning rate, lr)
[b] 也可能需要计算梯度 $(d^k)^T\nabla f(x^k + \alpha d^k)$

Convergence to Approximate Second-Order Necessary Points

Convergence to Approximate Second-Order Necessary Points

在前面的内容中，我们介绍并分析了寻找满足一阶最优条件点的方法，下面介绍寻找满足二阶必要条件点的基本方法. 首先回顾一下二阶必要条件：

$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) \text{ positive semidefinite} \tag{56}$$

为进行理论分析，不仅需要对梯度 ∇f 进行 Lipschitz 连续性假设，还要对 Hessian $\nabla^2 f$ 进行 Lipschitz 连续性假设.

Definition 2 (Lipschitz continuity of the Hessian)**.** 称矩阵 $\nabla^2 f$ 满足 *Lipschitz* 连续性，若存在 $M > 0$，使得

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leqslant M\|x - y\|, \quad \textit{for all } x, y \in \operatorname{dom}(f). \tag{57}$$

Note 11. 有了 *Definition 2*，可以自然地将 *Lecture 2* 的 *Lemma 1* 做如下拓展：

$$f(x+p) \leqslant f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x) p + \frac{1}{6} M\|p\|^3. \tag{58}$$

下面介绍算法内容：

1. 假设：函数 f 光滑、可能非凸，有下界其梯度 ∇f 和 Hessian $\nabla^2 f$ 均有 Lipschitz 性，Lipschitz 常数分别为 L, M.
2. 算法目标：寻找点 x 满足二阶最优性条件：

   $$\|\nabla f(x)\| \leqslant \epsilon_g, \quad \lambda_{\min}(\nabla^2 f(x)) \geqslant -\epsilon_H, \tag{59}$$

   其中 ϵ_g, ϵ_H 为两很小的常数.
3. 算法流程：
   (a) 若 $\|\nabla f(x^k)\| > \epsilon_g$，则进行最速梯度下降法(3)，步长 $\alpha_k = \frac{1}{L}$.
   (b) 若 $\|\nabla f(x^k)\| \leqslant \epsilon_g$，定义 $\lambda_k \triangleq \lambda_{\min}(\nabla^2 f(x^k))$. 若 $\lambda_k < -\epsilon_H$，选取 λ_k 对应的特征向量中满足 $(p^k)^T \nabla f(x^k) \leqslant 0$ 的单位特征向量，记为 $p^k, \|p^k\| = 1$. 令

   $$x^{k+1} = x^k + \alpha_k p^k, \alpha_k = \frac{2|\lambda_k|}{M}. \tag{60}$$

在流程(a)中，根据最速下降法中式(6)可知

$$f(x^{k+1}) \leqslant f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2 \leqslant f(x^k) - \frac{\epsilon_g^2}{2L}. \tag{61}$$

在流程(b)中，根据式(58)可得

$$\begin{aligned} f(x^{k+1}) &\leqslant f(x^k) + \alpha_k \nabla f(x^k)^T p^k + \frac{1}{2}\alpha_k^2 (p^k)^T \nabla^2 f(x^k) p^k + \frac{1}{6} M \alpha_k^3 \|p^k\|^3 \\ &\leqslant f(x^k) + \frac{1}{2}\alpha_k^2 (p^k)^T \nabla^2 f(x^k) p^k + \frac{1}{6} M \alpha_k^3 \|p^k\|^3 = f(x^k) + \frac{1}{2}\alpha_k^2 \lambda_k + \frac{1}{6} M \alpha_k^3 \\ &= f(x^k) - \frac{1}{2}\left(\frac{2|\lambda_k|}{M}\right)^2 |\lambda_k| + \frac{1}{6} M \left(\frac{2|\lambda_k|}{M}\right)^3 = f(x^k) - \frac{2}{3}\frac{|\lambda_k|^3}{M^2} \leqslant f(x^k) - \frac{2}{3}\frac{\epsilon_H^3}{M^2} \end{aligned} \tag{62}$$

结合上述两式，若二阶最优条件未满足，则相邻迭代函数值下降量最少为

$$\min\left(\frac{\epsilon_g^2}{2L}, \frac{2}{3}\frac{\epsilon_H^3}{M^2}\right). \tag{63}$$

故需要的迭代步数 K 满足

$$K \cdot \min\left(\frac{\epsilon_g^2}{2L}, \frac{2}{3}\frac{\epsilon_H^3}{M^2}\right) \leqslant f(x^0) - \bar{f} \Rightarrow K \leqslant \max\left(2L\epsilon_g^{-2}, \frac{3}{2}M^2\epsilon_H^{-3}\right)\left(f(x^0) - \bar{f}\right). \tag{64}$$

A Proof of Lemma 1

Proof of Lemma 1

Proposition 2. 若 f 为连续可微的强凸函数，且凸模为 m，则

$$f(x) - f(x^*) \leqslant \frac{\|\nabla f(x)\|^2}{2m} \tag{65a}$$

$$\|x - x^*\|^2 \leqslant \frac{2}{m}\|\nabla f(x)\| \tag{65b}$$

Proof. 根据 Lecture 1 Lemma 5 可知

$$f(y) \geqslant f(x) + \nabla f(x)(y - x) + \frac{m}{2}\|y - x\|^2 \tag{66}$$

上述不等式左端当 $y = x^*$ 时取得最小值，右端当 $y = x - \frac{\nabla f(x)}{m}$(化为二次函数)时取得最小值，分别代入可得

$$\begin{aligned} &f(x^*) \geqslant f(x) - \nabla f(x)^T \frac{\nabla f(x)}{m} + \frac{m}{2}\|\frac{1}{m}\nabla f(x)\|^2 = f(x) - \frac{1}{2m}\|\nabla f(x)\|^2 \\ &\Rightarrow f(x) - f(x^*) \leqslant \frac{\|\nabla f(x)\|^2}{2m} \end{aligned} \tag{67}$$

根据式(66)，由 **Cauchy-Schwarz** 不等式可知

$$\begin{aligned} &f(x^*) \geqslant f(x) + \nabla f(x)^T(x^* - x) + \frac{m}{2}\|x - x^*\|^2 \geqslant f(x) - \|\nabla f(x)\|\|x^* - x\| + \frac{m}{2}\|x^* - x\|^2 \\ &\Rightarrow \frac{m}{2}\|x^* - x\|^2 - \|\nabla f(x)\| \cdot \|x^* - x\| + f(x) - f(x^*) \leqslant 0 \\ &\Rightarrow \|x^* - x\| \leqslant \frac{\|\nabla f(x)\| + \sqrt{\|\nabla f(x)\|^2 - 2m(f(x) - f(x^*))}}{m} \leqslant \frac{2}{m}\|\nabla f(x)\| \end{aligned} \tag{68}$$

□

B Proof of Theorem 4

Proof for Convergence of SGD

Proposition 3. 若 f 为光滑的凸函数，且其梯度 ∇f 满足 *Lipschitz* 性质*(*系数为 L*)*，设问题*(1)*有解 x^*. 如下定义的水平集*(level set)*有界，即 $R_0 < \infty$:

$$R_0 \triangleq \max\{\|x - x^*\| | f(x) \leqslant f(x^0)\}. \tag{69}$$

那么满足性质*(29)*的梯度下降法产生的序列 $\{x^k\}_0^\infty$ 满足:

$$f(x^T) - f(x^*) \leqslant \frac{R_0^2}{CT}, T = 1, 2, \ldots. \tag{70}$$

Proof. 定义 $\Delta_k \triangleq f(x^k) - f(x^*)$，那么有

$$\Delta_k = f(x^k) - f(x^*) \leqslant \nabla f(x^k)^T(x^k - x^*) \leqslant \|\nabla f(x^k)\| \cdot \|x^k - x^*\| \leqslant R_0\|\nabla f(x^k)\|. \tag{71}$$

代入式(29)中可得

$$f(x^{k+1}) \leqslant f(x^k) - \frac{C}{R_0^2}\Delta_k^2 \Rightarrow \Delta_{k+1} \leqslant \Delta_k - \frac{C}{R_0^2}\Delta_k^2 = \Delta_k\left(1 - \frac{C}{R_0^2}\Delta_k\right). \tag{72}$$

等价地得到

$$\frac{1}{\Delta_{k+1}} \geqslant \frac{1}{\Delta_k}\frac{1}{1 - \frac{C}{R_0^2}\Delta_k}. \tag{73}$$

根据不等式 $\frac{1}{1-\epsilon} \geqslant 1 + \epsilon$ for all $\epsilon \in [0, 1]$，且 $\frac{C}{R_0^2}\Delta_k \in [0, 1]$ 可得

$$\frac{1}{\Delta_{k+1}} \geqslant \frac{1}{\Delta_k}\left(1 + \frac{C}{R_0^2}\Delta_k\right) = \frac{1}{\Delta_k} + \frac{C}{R_0^2}. \tag{74}$$

对上式两边类和，可得

$$\frac{1}{\Delta_T} \geqslant \frac{1}{\Delta_0} + \frac{TC}{R_0^2} \geqslant \frac{TC}{R_0^2} \Rightarrow \Delta_T = f(x^T) - f(x^*) \leqslant \frac{R_0^2}{CT} \tag{75}$$

□

C The Property of $D_h(x, z)$

Proof for Convergence of SGD

Proposition 4. *Bregman divergences* $D_h(x, z)$ 对于 x 具有非负性与强凸性，其中 h 光滑且强凸，$D_h(x, z) \triangleq h(x) - h(z) - \nabla h(z)^T(x - z)$.

Proof. 由于 h 强凸，根据 Lecture 2, Lemma 5 可知

$$h(x) \geqslant h(z) + \nabla h(z)^T(x - z) + \frac{m}{2}\|x - z\|^2 \Rightarrow D_h(x, z) \geqslant \frac{m}{2}\|x - z\|^2 \geqslant 0 \tag{76}$$

从而非负性满足.

由于

$$\begin{aligned}
&D_h(y, z) + \nabla_y D_h(y, z)^T(x - y) + \frac{m}{2}\|x - y\|^2 \\
&= h(y) - h(z) - \nabla h(z)^T(y - z) + (\nabla h(y) - \nabla h(z))^T(x - y) + \frac{m}{2}\|x - y\|^2 \\
&= \left[h(y) + \nabla h(y)^T(x - y) + \frac{m}{2}\|x - y\|^2\right] - h(z) + \nabla h(z)^T z - \nabla h(z)^T x \\
&\leqslant h(x) - h(z) - \nabla h(z)^T(x - z) = D_h(x, z)
\end{aligned} \tag{77}$$

因此强凸性满足. □

D Proof of three points property

Proof of three points property

Proposition 5. 定义 $D_h(x, z) \triangleq h(x) - h(z) - \nabla h(z)^T(x - z)$，则 D_h 满足如下 *three points*

property:

$$D_h(x,y) = D_h(x,z) + D_h(z,y) - (x-z)^T(\nabla h(y) - \nabla h(z)). \tag{78}$$

Proof.

$$\begin{aligned} &D_h(x,z) + D_h(z,y) - (x-z)^T(\nabla h(y) - \nabla h(z)) \\ &= h(x) - h(z) - \nabla h(z)^T(x-z) + h(z) - h(y) - \nabla h(y)^T(z-y) - (x-z)^T(\nabla h(y) - \nabla h(z)) \\ &h(x) - h(y) + \nabla h(y)^T(y-x) = D_h(x,y) \end{aligned} \tag{79}$$

□

E Mirror Descent

Mirror Descent – Introduction

最速梯度下降法(3)实际可以从一阶泰勒展开角度来理解，实际上可以从如下简单二次问题得到:

$$x^{k+1} = \arg\min f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{1}{2\alpha_k}\|x - x^k\|^2. \tag{80}$$

因此新的迭代 x^{k+1} 可以看作从具有基于欧几里得范数的二次惩罚项的一阶泰勒级数模型中获得，惩罚项 $\frac{1}{2\alpha_k}\|x - x^k\|^2$ 惩罚远离当前迭代的移动. 因此可以看出，当 α_k 逐渐减小时，惩罚变大，因此相邻迭代的差异变小.

此处考虑更一般的模型：将惩罚项 $\frac{1}{2\alpha_k}\|x - x^k\|^2$ 换为 Bregman 散度(Bregman divergences) $D_h(\cdot,\cdot)$:

$$x^{k+1} = \arg\min f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{1}{\alpha_k}D_h(x, x^k). \tag{81}$$

Definition 3 (Bregman divergences)**.** 对于光滑且强凸(凸模为 m)的函数 f，称 $D_h(\cdot,\cdot)$ 为 h 生成的 ***Bregman divergences***，若

$$D_h(x,z) \triangleq h(x) - h(z) - \nabla h(z)^T(x-z), \tag{82}$$

或写为

$$D_h(x,z) \triangleq h(x) - h(z) - \langle \nabla h(z), x - z\rangle \tag{83}$$

其中 h 的强凸性中的 $\frac{m}{2}\|y-x\|^2$ 的范数可以为任意范数.

Note 12. 回忆 h 在点 z处的一阶泰勒近似为 $h(x) \approx h(z) + \nabla h(z)^T(x-z)$，因此 $D_h(x,z)$ 是 h 在 z 处的一阶泰勒近似式与 h 于 x 点处的差异. 另外由于 h 具有强凸性，$D_h(x,z)$ 对于 x 具有非负性和强凸性(证明见附录C).

对于一般范数 $\|\cdot\|$，其满足 "**three-point property**"，或称"**law of cosines**":

$$\begin{aligned} \|x-y\|^2 &= \|x-z\|^2 + \|z-y\|^2 - 2(x-z)(y-z) \\ &= \|x-z\|^2 + \|z-y\|^2 - 2\|x-z\|\|y-z\|\cos\gamma, \end{aligned} \tag{84}$$

其中 γ 为 $x-z$ 与 $y-z$ 在 z 处的夹角. 当 $\gamma = \frac{\pi}{2}$ 时此即勾股定理(Pythagorean theorem).

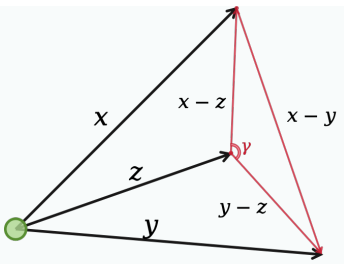

Figure 2: Law of cosines

Bregman divergences 也有"three-point property":

$$D_h(x,y) = D_h(x,z) + D_h(z,y) - (x-z)^T(\nabla h(y) - \nabla h(z)). \tag{85}$$

这一性质是分析 mirror descent 收敛性的基础. 证明见附录D.

Example 1 (Squared Euclidean Norm)**.** 当 $h(x) = \frac{1}{2}\|x\|^2$,

$$D_h(x,z) = \frac{1}{2}\|x\|^2 - \frac{1}{2}\|z\|^2 - z^T(x-z) = \frac{1}{2}\|x-z\|^2, \tag{86}$$

此即说明式*(80)*是式*(81)*的特殊情形.

Example 2 (Negative Entropy)**.** 考虑具有 n 个类别的离散概率分布*(*或称为 n-单纯形*, $n-simplex$)*，定义为:

$$\Delta_n \triangleq \left\{ p \in \mathbb{R}^n \,\middle|\, p \geqslant 0, \sum_{i=1}^n p_i = 1 \right\}. \tag{87}$$

定义 $h(p) = \sum_{i=1}^n p_i \log p_i$ 为分布 p 的负熵*(negative entropy)*．该函数是凸的，对任意$p, q \in \Delta_n$，有:

$$\nabla h(p) = [\log p_1 + 1, \log p_2 + 1, \ldots, \log p_n + 1]^T \Rightarrow \langle \nabla h(q), p-q \rangle = \sum_{i=1}^n (\log q_i + 1)(p_i - q_i) \tag{88}$$

$$\begin{aligned} D_h(p,q) &= \sum_{i=1}^n p_i \log p_i - \sum_{i=1}^n q_i \log q_i - \sum_{i=1}^n (\log q_i + 1)(p_i - q_i) \\ &= \sum_{i=1}^n [p_i \log p_i - q_i \log q_i - (\log q_i + 1)(p_i - q_i)] \\ &= \sum_{i=1}^n \left(p_i \log \frac{p_i}{q_i} - p_i + q_i \right) \overset{\sum_{i=1}^n p_i = 1}{\underset{\sum_{i=1}^n q_i = 1}{=}} \sum_{i=1}^n p_i \log \left(\frac{p_i}{q_i} \right). \end{aligned} \tag{89}$$

此时，$D_h(p,q)$ 又被称为 ***Kullback–Liebler divergence****(KL* 散度*)*.

Note 13. 定义 $h(p) = \sum_{i=1}^n p_i \log p_i$ 时，函数 h 对于 *1-norm* $\|\cdot\|_1$ 是强凸的且凸模为 *1*，因此有如下不等式:

$$h(p) \geqslant h(q) + \nabla h(q)^T(p-q) + \frac{1}{2}\|p-q\|_1^2, \quad \textit{for all } p, q \in \operatorname{int} \Delta_n. \tag{90}$$

此不等式被称为 ***Pinsker's inequality***.

Mirror Descent – Optimality Conditions

Mirror descent 由式(81)定义，由于

$$\nabla_x D_h(x,z) = \nabla h(x) - \nabla h(z), \tag{91}$$

因此式(81)的最优条件为：

$$\nabla f(x^k) + \frac{1}{\alpha_k}\nabla h(x^{k+1}) - \frac{1}{\alpha_k}\nabla h(x^k) = 0 \Rightarrow \nabla h(x^{k+1}) = \nabla h(x^k) - \alpha_k \nabla f(x^k). \tag{92}$$

从而得到 x^{k+1} 满足

$$x^{k+1} = (\nabla h)^{-1}\left\{\nabla h(x^k) - \alpha_k \nabla f(x^k)\right\}, \tag{93}$$

其中 $(\nabla h)^{-1}$ 表示 ∇h 的反函数. 除前面举的两个例子外，能够计算反函数的 ∇h 是很少的.

$$h(x) = \frac{1}{2}\|x\|^2 \Rightarrow (\nabla h)^{-1}(v) = v \tag{94a}$$

$$h(p) = \sum_{i=1}^n p_i \log p_i \Rightarrow (\nabla h)^{-1}(v)_i = \frac{e^{v_i}}{\sum_{j=1}^n e^{v_j}},\ i = 1,2,\ldots,n. \tag{94b}$$

Mirror Descent – Analysis

在式(81)中，最理想的情况是在全集 $\mathbb{R}^n$ 中寻找满足 $\arg\min$ 条件的 x^{k+1}，为分析方便我们将区域缩小至子集 $\mathcal{X} \subseteq \mathbb{R}^n$. 并且做出如下假设：

1. $\mathcal{X} \subseteq \mathbb{R}^n$ 是凸集
2. $h : \mathcal{X} \to \mathbb{R}$ 连续可微
3. 范数 $\|\cdot\|$ 可以是任意的，h 相对于此范数强凸，凸模为 m，即满足

$$h(x) \geqslant h(z) + \langle \nabla h(z), x - z\rangle + \frac{m}{2}\|x - z\|^2, \quad \text{for all } x, z \in \mathcal{X}. \tag{95}$$

这样我们分析的 mirror descent 就变为：

$$x^{k+1} = \arg\min_{x\in\mathcal{X}} f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{1}{\alpha_k} D_h(x, x^k), \quad k = 0,1,2,\ldots. \tag{96}$$

通过后面章节的结论，可以得到此问题的最优条件为

$$\left[\nabla f(x^k) + \frac{1}{\alpha_k}\nabla h(x^{k+1}) - \frac{1}{\alpha_k}\nabla h(x^k)\right]^T (x - x^{k+1}) \geqslant 0, \quad \text{for all } x \in \mathcal{X}. \tag{97}$$

此处的理论分析关注更一般的平均情况：迭代的加权平均(weighted average of the iterates)，即

$$\lambda_k = \sum_{j=0}^k \alpha_j, \quad \bar{x}^k = \lambda_k^{-1}\sum_{j=0}^k \alpha_j x^j. \tag{98}$$

Theorem 6. 设$\|\cdot\|$是定义在集合 $\mathcal{X}$ 上的任意范数，函数 h 关于该范数在 $\mathcal{X}$ 上是m-强凸的。假设函数 f 是凸的，并且关于$\|\cdot\|$ 是 *L-Lipschitz* 的，且问题 $\min_{x\in\mathcal{X}} f(x)$ 存在最优解 x^*, $f^* \triangleq$

$f(x^*)$。则对任意整数$T \geqslant 1$，有如下不等式成立：

$$f(\bar{x}^T) - f^* \leqslant \frac{D_h(x^*, x^0) + \frac{L^2}{2m}\sum_{t=0}^{T} \alpha_t^2}{\sum_{t=0}^{T} \alpha_t}, \tag{99}$$

其中，$\bar{x}^k = \lambda_k^{-1} \sum_{j=0}^{k} \alpha_j x^j$.

Proof. 证明暂略. □

F The KL and PL Properties

The KL and PL Properties

在对函数 f 使用梯度下降法得到的序列进行收敛性分析时，强凸性能够得到更好的收敛性，但有一些凸函数不满足强凸性而满足其他的性质也可以获得较好的收敛，例如 **Polyak–Łojasiewicz(PL) condition**.

Definition 4 (PL condition)**.** 若函数 f 存在最小值，设为 $f^* = f(x^*)$，那么称 f 满足 *Polyak–Łojasiewicz (PL) condition* 若存在 $m > 0$ 使得

$$\|\nabla f(x)\|^2 \geqslant 2, [f(x) - f(x^*)]. \tag{100}$$

Note 14. 事实上条件*(100)*弱于强凸，因为若 f 满足 $m-$强凸，则有

$$f(z) \geqslant f(x) + \nabla f(x)^T (z - x) + \frac{m}{2}\|z - x\|^2, \tag{101}$$

不等号左端取最小值当 $z = x^*$，右端取最小值当 $z = x - \frac{\nabla f(x)}{m}$，代入后整理即得*(100)*.

Example 3. 凸二次函数 $f(x) = \frac{1}{2}x^T Ax$ 当 $A \succeq 0$ 但 A 奇异*(singular)* 时不满足强凸性但满足 *PL condition*，系数 m 为 A 的最小非零特征值.

Proof. 暂略，需要使用后面结论. □

Note 15. *PL condition* 是 ***Kurdyka–Łojasiewicz (KL) condition*** 的特殊情形. 在 *KL condition* 下也可以证明迭代的局部收敛性.

First updated: July 23, 2025

References

[1] James V Burke and Abraham Engle. Line search methods for convex-composite optimization. *arXiv preprint arXiv:1806.05218*, 2018.