矩阵论

2024年秋学期

第八讲

2024年10月9日

第4章 梯度分析与最优化

单变量函数

单变量函数的平稳点与极值点

平稳点： $f'(c)=0$

- 平稳点是函数图像上的一个点，在该点处，函数的导数为零。直观上，这意味着函数在这一点的切线是水平的。
- 平稳点是函数局部极值可能出现的地方，但不是所有平稳点都是极值点。

局部极小点: $f'(c)=0$ $\quad f''(c)=\left.\frac{\mathrm{d}^2 f(x)}{dx^2}\right|_{x=c}\geq 0$

局部极大点: $f'(c)=0$ $\quad f''(c)=\left.\frac{\mathrm{d}^2 f(x)}{dx^2}\right|_{x=c}\leq 0$

鞍点（saddle point）： $f'(c)=0$

$$f''(c+\Delta x)\leq 0 \qquad f''(c+\Delta x)\geq 0$$

多变量函数

多变量函数的平稳点与极值点

多变量函数无约束极小化问题

$$\min_{x \in S} f(\boldsymbol{x}) \quad f(\boldsymbol{x}): \mathbb{R}^n \to \mathbb{R}$$

开邻域 $B(\boldsymbol{c};r) = \left\{ \boldsymbol{x} \middle| \ \boldsymbol{x} \in \mathbb{R}^n, \| \boldsymbol{x} - \boldsymbol{c} \|_2 < r \right\}$

闭合邻域 $B(\boldsymbol{c};r) = \left\{ \boldsymbol{x} \middle| \ \boldsymbol{x} \in \mathbb{R}^n, \| \boldsymbol{x} - \boldsymbol{c} \|_2 \leq r \right\}$

多变量函数

多变量函数的平稳点与极值点

二阶泰勒级数逼近

$$
\begin{aligned}
f(\boldsymbol{c}+\Delta\boldsymbol{x}) &= f(\boldsymbol{c})+\left(\frac{\partial f(\boldsymbol{c})}{\partial \boldsymbol{c}}\right)^{\mathrm{T}}\Delta\boldsymbol{x}+\frac{1}{2}(\Delta\boldsymbol{x})^{\mathrm{T}}\cdot\frac{\partial^2 f(\boldsymbol{c})}{\partial\boldsymbol{c}\partial\boldsymbol{c}^{\mathrm{T}}}\Delta\boldsymbol{x} \\
&= f(\boldsymbol{c})+(\nabla f(\boldsymbol{c}))^{\mathrm{T}}\Delta\boldsymbol{x}+\frac{1}{2}(\Delta\boldsymbol{x})^{\mathrm{T}}\boldsymbol{H}(f(\boldsymbol{c}))\Delta\boldsymbol{x}
\end{aligned}
$$

梯度向量

$$
\nabla f(\boldsymbol{c})=\frac{\partial f(\boldsymbol{c})}{\partial \boldsymbol{c}}=\left.\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{c}}
$$

Hessian矩阵

$$
\boldsymbol{H}(f(c))=\frac{\partial^2 f(\boldsymbol{c})}{\partial\boldsymbol{c}\partial\boldsymbol{c}^{\mathrm{T}}}=\left.\frac{\partial^2 f(\boldsymbol{x})}{\partial\boldsymbol{x}\partial\boldsymbol{x}^{\mathrm{T}}}\right|_{\boldsymbol{x}=\boldsymbol{c}}
$$

多变量函数

多变量函数的平稳点与极值点

$$f(c) \leqslant f(c+\Delta x) \quad \forall\, 0 < \|\Delta x\|_2 \leqslant r$$

局部极小

$$\boldsymbol{H}(f(\boldsymbol{c})) = \left.\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}}\right|_{\boldsymbol{x}=\boldsymbol{c}} \succeq 0$$

局部极小

$$f(c) < f(c+\Delta x) \quad \forall\, 0 < \|\Delta x\|_2 \leqslant r$$

严格局部极小

$$f(\boldsymbol{c}) \leqslant f(\boldsymbol{x}) \quad \forall \boxed{\boldsymbol{x} \in S}$$

全局极小

$$f(\boldsymbol{c}) < f(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in S, \boldsymbol{x} \neq \boldsymbol{c}$$

严格全局极小

多变量函数

多变量函数 $f(X)$ 的平稳点与极值点

邻域

$$B(\boldsymbol{C};r)=\left\{\boldsymbol{X} \mid \boldsymbol{X}\in\mathbb{R}^{m\times n}, \|\mathrm{vec}(\boldsymbol{X})-\mathrm{vec}(\boldsymbol{C})\|_2<r\right\}$$

二阶泰勒级数逼近

$$f(\boldsymbol{C}+\Delta\boldsymbol{X})=f(\boldsymbol{C})+\left(\frac{\partial f(\boldsymbol{C})}{\partial\,\mathrm{vec}(\boldsymbol{C})}\right)^{\mathrm{T}}\mathrm{vec}(\Delta\boldsymbol{X})+\frac{1}{2}(\mathrm{vec}(\Delta\boldsymbol{X}))^{\mathrm{T}}\frac{\partial^2 f(\boldsymbol{C})}{\partial\,\mathrm{vec}(\boldsymbol{C})\partial(\mathrm{vec}\boldsymbol{C})^{\mathrm{T}}}\mathrm{vec}(\Delta\boldsymbol{X})$$

$$=f(\boldsymbol{C})+\left(\nabla_{\mathrm{vec}\boldsymbol{C}}\,f(\boldsymbol{C})\right)^{\mathrm{T}}\mathrm{vec}(\Delta\boldsymbol{X})+\frac{1}{2}(\mathrm{vec}(\Delta\boldsymbol{X}))^{\mathrm{T}}\boldsymbol{H}(f(\boldsymbol{C}))\mathrm{vec}(\Delta\boldsymbol{X})$$

$$\nabla_{\mathrm{vec}\boldsymbol{C}}\,f(\boldsymbol{C})=\left.\frac{\partial f(\boldsymbol{X})}{\partial\,\mathrm{vec}(\boldsymbol{X})}\right|_{\boldsymbol{X}=\boldsymbol{C}}\in\mathbb{R}^{mn}\qquad \boldsymbol{H}(f(\boldsymbol{C}))=\left.\frac{\partial^2 f(\boldsymbol{X})}{\partial\,\mathrm{vec}(\boldsymbol{X})\partial(\mathrm{vec}\,\boldsymbol{X})^{\mathrm{T}}}\right|_{\boldsymbol{X}=\boldsymbol{C}}\in\mathbb{R}^{mn\times mn}$$

总结

表 4.1.1 实变函数的平稳点和极值点的条件

实变函数	$f(x):\mathbb{R}\to\mathbb{R}$	$f(\boldsymbol{x}):\mathbb{R}^n\to\mathbb{R}$	$f(\boldsymbol{X}):\mathbb{R}^{m\times n}\to\mathbb{R}$
平稳点	$\left.\frac{\partial f(x)}{\partial x}\right\|_{x=c}=0$	$\left.\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}\right\|_{\boldsymbol{x}=\boldsymbol{c}}=\boldsymbol{0}$	$\left.\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}}\right\|_{\boldsymbol{X}=\boldsymbol{C}}=\boldsymbol{O}_{m\times n}$
局部极小点	$\left.\frac{\partial^2 f(x)}{\partial x\partial x}\right\|_{x=c}\geqslant 0$	$\left.\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x}\partial \boldsymbol{x}^T}\right\|_{\boldsymbol{x}=\boldsymbol{c}}\succeq 0$	$\left.\frac{\partial^2 f(\boldsymbol{X})}{\partial \mathrm{vec}(\boldsymbol{X})\partial(\mathrm{vec}\,\boldsymbol{X})^T}\right\|_{\boldsymbol{X}=\boldsymbol{C}}\succeq 0$
严格局部极小点	$\left.\frac{\partial^2 f(x)}{\partial x\partial x}\right\|_{x=c}>0$	$\left.\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x}\partial \boldsymbol{x}^T}\right\|_{\boldsymbol{x}=\boldsymbol{c}}\succ 0$	$\left.\frac{\partial^2 f(\boldsymbol{X})}{\partial \mathrm{vec}(\boldsymbol{X})\partial(\mathrm{vec}\,\boldsymbol{X})^T}\right\|_{\boldsymbol{X}=\boldsymbol{C}}\succ 0$
局部极大点	$\left.\frac{\partial^2 f(x)}{\partial x\partial x}\right\|_{x=c}\leqslant 0$	$\left.\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x}\partial \boldsymbol{x}^T}\right\|_{\boldsymbol{x}=\boldsymbol{c}}\preceq 0$	$\left.\frac{\partial^2 f(\boldsymbol{X})}{\partial \mathrm{vec}(\boldsymbol{X})\partial(\mathrm{vec}\,\boldsymbol{X})^T}\right\|_{\boldsymbol{X}=\boldsymbol{C}}\preceq 0$
严格局部极大点	$\left.\frac{\partial^2 f(x)}{\partial x\partial x}\right\|_{x=c}<0$	$\left.\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x}\partial \boldsymbol{x}^T}\right\|_{\boldsymbol{x}=\boldsymbol{c}}\prec 0$	$\left.\frac{\partial^2 f(\boldsymbol{X})}{\partial \mathrm{vec}(\boldsymbol{X})\partial(\mathrm{vec}\,\boldsymbol{X})^T}\right\|_{\boldsymbol{X}=\boldsymbol{C}}\prec 0$
鞍　点	$\left.\frac{\partial^2 f(x)}{\partial x\partial x}\right\|_{x=c}$ 不定	$\left.\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x}\partial \boldsymbol{x}^T}\right\|_{\boldsymbol{x}=\boldsymbol{c}}$ 不定	$\left.\frac{\partial^2 f(\boldsymbol{X})}{\partial \mathrm{vec}(\boldsymbol{X})\partial(\mathrm{vec}\,\boldsymbol{X})^T}\right\|_{\boldsymbol{X}=\boldsymbol{C}}$ 不定

4.1.4 一阶必要条件，二阶必要条件，二阶充分条件

复变函数的平稳点和极值点条件

一阶微分

$$\mathrm{d}f\left(\mathbf{z},\mathbf{z}^*\right)=\frac{\partial f\left(\mathbf{z},\mathbf{z}^*\right)}{\partial \mathbf{z}^{\mathrm{T}}}\mathrm{d}\mathbf{z}+\frac{\partial f\left(\mathbf{z},\mathbf{z}^*\right)}{\partial \mathbf{z}^{\mathrm{H}}}\mathrm{d}\mathbf{z}^*=\left[\frac{\partial f\left(\mathbf{z},\mathbf{z}^*\right)}{\partial \mathbf{z}^{\mathrm{T}}},\frac{\partial f\left(\mathbf{z},\mathbf{z}^*\right)}{\partial \mathbf{z}^{\mathrm{H}}}\right]\begin{bmatrix}\mathrm{d}\mathbf{z}\\ \mathrm{d}\mathbf{z}^*\end{bmatrix}$$

二阶微分

$$\mathrm{d}^2 f\left(\mathrm{z},\mathrm{z}^*\right)=\left(\frac{\partial f^2\left(\mathrm{z},\mathrm{z}^*\right)}{\partial \mathrm{z}\partial \mathrm{z}^{\mathrm{T}}}\mathrm{dz}+\frac{\partial f^2\left(\mathrm{z},\mathrm{z}^*\right)}{\partial \mathrm{z}^*\partial \mathrm{z}^{\mathrm{T}}}\mathrm{dz}^*\right)^{\mathrm{T}}\mathrm{dz}+\left(\frac{\partial f^2\left(\mathrm{z},\mathrm{z}^*\right)}{\partial \mathrm{z}\partial \mathrm{z}^{\mathrm{H}}}\mathrm{dz}+\frac{\partial f^2\left(\mathrm{z},\mathrm{z}^*\right)}{\partial \mathrm{z}^*\partial \mathrm{z}^{\mathrm{H}}}\mathrm{dz}^*\right)^{\mathrm{T}}\mathrm{dz}^*$$

$$=\left[\mathrm{dz}^{\mathrm{H}},\mathrm{dz}^{\mathrm{T}}\right]\begin{bmatrix}\dfrac{\partial^2 f\left(\mathrm{z},\mathrm{z}^*\right)}{\partial \mathrm{z}^*\partial \mathrm{z}^{\mathrm{T}}} & \dfrac{\partial^2 f\left(\mathrm{z},\mathrm{z}^*\right)}{\partial \mathrm{z}^*\partial \mathrm{z}^{\mathrm{H}}}\\ \dfrac{\partial^2 f\left(\mathrm{z},\mathrm{z}^*\right)}{\partial \mathrm{z}\partial \mathrm{z}^{\mathrm{T}}} & \dfrac{\partial^2 f\left(\mathrm{z},\mathrm{z}^*\right)}{\partial \mathrm{z}\partial \mathrm{z}^{\mathrm{H}}}\end{bmatrix}\begin{bmatrix}\mathrm{dz}\\ \mathrm{dz}^*\end{bmatrix}$$

梯度分析与最优化

复变函数的平稳点和极值点条件

二阶Taylor级数逼近

$$f\left(\boldsymbol{z},\boldsymbol{z}^*\right)\approx f\left(\boldsymbol{c},\boldsymbol{c}^*\right)+\left[\frac{\partial f\left(\boldsymbol{c},\boldsymbol{c}^*\right)}{\partial \boldsymbol{c}},\frac{\partial f\left(\boldsymbol{c},\boldsymbol{c}^*\right)}{\partial \boldsymbol{c}^*}\right]\begin{bmatrix}\Delta\boldsymbol{c}\\ \Delta\boldsymbol{c}^*\end{bmatrix}+\frac{1}{2}\left[\Delta\boldsymbol{c}^{\mathrm{H}},\Delta\boldsymbol{c}^{\mathrm{T}}\right]\begin{bmatrix}\dfrac{\partial^2 f\left(\boldsymbol{c},\boldsymbol{c}^*\right)}{\partial\boldsymbol{c}^*\partial\boldsymbol{c}^T} & \dfrac{\partial^2 f\left(\boldsymbol{c},\boldsymbol{c}^*\right)}{\partial\boldsymbol{c}^*\partial\boldsymbol{c}^{\mathrm{H}}}\\ \dfrac{\partial^2 f\left(\boldsymbol{c},\boldsymbol{c}^*\right)}{\partial\boldsymbol{c}\partial\boldsymbol{c}^{\mathrm{T}}} & \dfrac{\partial^2 f\left(\boldsymbol{c},\boldsymbol{c}^*\right)}{\partial\boldsymbol{c}\partial\boldsymbol{c}^{\mathrm{H}}}\end{bmatrix}\begin{bmatrix}\Delta\boldsymbol{c}\\ \Delta\boldsymbol{c}^*\end{bmatrix}$$

$$=f\left(\boldsymbol{c},\boldsymbol{c}^*\right)+\left(\nabla f\left(\boldsymbol{c},\boldsymbol{c}^*\right)\right)^{\mathrm{T}}\Delta\tilde{\boldsymbol{c}}+\frac{1}{2}(\Delta\tilde{\boldsymbol{c}})^{\mathrm{H}}\boldsymbol{H}\left(f\left(\boldsymbol{c},\boldsymbol{c}^*\right)\right)\Delta\tilde{\boldsymbol{c}}$$

梯度分析与最优化

极值点的辨识

必要条件

$$\left.\frac{\partial f\left(\mathbf{z},\mathbf{z}^*\right)}{\partial \mathbf{z}^*}\right|_{\mathbf{z}=\mathbf{z}_0}=\mathbf{0} \qquad \left.\begin{bmatrix} \boldsymbol{H}_{\mathbf{z}^*,\mathbf{z}} & \boldsymbol{H}_{\mathbf{z}^*,\mathbf{z}^*} \\ \boldsymbol{H}_{\mathbf{z},\mathbf{z}} & \boldsymbol{H}_{\mathbf{z},\mathbf{z}^*} \end{bmatrix}\right|_{\mathbf{z}=\mathbf{z}_0} \succeq 0$$

或

$$\left.\frac{\partial f\left(\boldsymbol{Z},\boldsymbol{Z}^*\right)}{\partial \boldsymbol{Z}^*}\right|_{\boldsymbol{Z}=\boldsymbol{Z}_0}=\boldsymbol{O} \qquad \left.\begin{bmatrix} \boldsymbol{H}_{\boldsymbol{Z}^*,\boldsymbol{Z}} & \boldsymbol{H}_{\boldsymbol{Z}^*,\boldsymbol{Z}^*} \\ \boldsymbol{H}_{\boldsymbol{Z},\boldsymbol{Z}} & \boldsymbol{H}_{\boldsymbol{Z},\boldsymbol{Z}^*} \end{bmatrix}\right|_{\boldsymbol{Z}=\boldsymbol{Z}_0} \succeq 0$$

充分条件

$$\left.\frac{\partial f\left(\mathbf{z},\mathbf{z}^*\right)}{\partial \mathbf{z}^*}\right|_{\mathbf{z}=\mathbf{z}_0}=\mathbf{0} \qquad \left.\begin{bmatrix} \boldsymbol{H}_{\mathbf{z}^*,\mathbf{z}} & \boldsymbol{H}_{\mathbf{z}^*,\mathbf{z}^*} \\ \boldsymbol{H}_{\mathbf{z},\mathbf{z}} & \boldsymbol{H}_{\mathbf{z},\mathbf{z}^*} \end{bmatrix}\right|_{\mathbf{z}=\mathbf{z}_0} \succ 0$$

或

$$\left.\frac{\partial f\left(\boldsymbol{Z},\boldsymbol{Z}^*\right)}{\partial \boldsymbol{Z}^*}\right|_{\boldsymbol{Z}=\boldsymbol{Z}_0}=\boldsymbol{O} \qquad \left.\begin{bmatrix} \boldsymbol{H}_{\boldsymbol{Z}^*,\boldsymbol{Z}} & \boldsymbol{H}_{\boldsymbol{Z}^*,\boldsymbol{Z}^*} \\ \boldsymbol{H}_{\boldsymbol{Z},\boldsymbol{Z}} & \boldsymbol{H}_{\boldsymbol{Z},\boldsymbol{Z}^*} \end{bmatrix}\right|_{\boldsymbol{Z}=\boldsymbol{Z}_0} \succ 0$$

则z_0为严格局部最小点

无约束最小化问题的梯度分析

给定一个实值目标函数 $f(\boldsymbol{w}, \boldsymbol{w}^*)$ 或 $f(\boldsymbol{W}, \boldsymbol{W}^*)$

无约束最小化问题的梯度分析

1. 共轭梯度矩阵决定最小化问题的闭式解。
2. 共轭梯度矩阵与 Hessian 矩阵给出局部极小点辨识的必要条件或充分条件。
3. 共轭梯度向量的负方向决定求解最小化问题的最速下降迭代算法。
4. Hessian 矩阵给出求解最小化问题的 Newton 算法。

无约束最小化问题的梯度分析——闭式解

例 4.2.1 考察求解超定矩阵方程 $\boldsymbol{Az}=\boldsymbol{b}$ 的最小二乘方法。定义误差平方和

$$J(\boldsymbol{z})=\|\boldsymbol{Az}-\boldsymbol{b}\|_2^2=(\boldsymbol{Az}-\boldsymbol{b})^{\mathrm{H}}(\boldsymbol{Az}-\boldsymbol{b})$$
$$=\boldsymbol{z}^{\mathrm{H}}\boldsymbol{A}^{\mathrm{H}}\boldsymbol{Az}-\boldsymbol{z}^{\mathrm{H}}\boldsymbol{A}^{\mathrm{H}}\boldsymbol{b}-\boldsymbol{b}^{\mathrm{H}}\boldsymbol{Az}+\boldsymbol{b}^{\mathrm{H}}\boldsymbol{b}$$

为准则函数。令其共轭梯度向量 $\nabla_{\boldsymbol{z}^*}J(\boldsymbol{z})=\boldsymbol{A}^{\mathrm{H}}\boldsymbol{Az}-\boldsymbol{A}^{\mathrm{H}}\boldsymbol{b}$ 等于零向量，易知：若 $\boldsymbol{A}^{\mathrm{H}}\boldsymbol{A}$ 非奇异，则

$$\boldsymbol{z}=(\boldsymbol{A}^{\mathrm{H}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{H}}\boldsymbol{b} \qquad (4.2.24)$$

极大值/极小值?

梯度向量

$$\nabla f(\boldsymbol{z},\boldsymbol{z}^*)=\begin{bmatrix}\dfrac{\partial f(\boldsymbol{z},\boldsymbol{z}^*)}{\partial \boldsymbol{z}}\\[2mm] \dfrac{\partial f(\boldsymbol{z},\boldsymbol{z}^*)}{\partial \boldsymbol{z}^*}\end{bmatrix}\in\mathbb{C}^{2n}$$

无约束最小化问题的梯度分析——闭式解

例 4.2.2 考察求解超定矩阵方程 $\boldsymbol{A}\boldsymbol{z}=\boldsymbol{b}$ 的最大似然方法。定义对数似然函数

$$l(\hat{\boldsymbol{z}})=C-\frac{1}{\sigma^2}\boldsymbol{e}^{\mathrm{H}}\boldsymbol{e}=C-\frac{1}{\sigma^2}(\boldsymbol{b}-\boldsymbol{A}\hat{\boldsymbol{z}})^{\mathrm{H}}(\boldsymbol{b}-\boldsymbol{A}\hat{\boldsymbol{z}}) \tag{4.2.25}$$

式中，C 为一实常数。求对数似然函数

$$l(\hat{\boldsymbol{z}})=C-\frac{1}{\sigma^2}\boldsymbol{b}^{\mathrm{H}}\boldsymbol{b}+\frac{1}{\sigma^2}\boldsymbol{b}^{\mathrm{H}}\boldsymbol{A}\hat{\boldsymbol{z}}+\frac{1}{\sigma^2}\hat{\boldsymbol{z}}^{\mathrm{H}}\boldsymbol{A}^{\mathrm{H}}\boldsymbol{b}-\frac{1}{\sigma^2}\hat{\boldsymbol{z}}^{\mathrm{H}}\boldsymbol{A}^{\mathrm{H}}\boldsymbol{A}\hat{\boldsymbol{z}} \tag{4.2.26}$$

相对于 $\boldsymbol{z}$ 的共轭梯度，得

$$\nabla_{\hat{\boldsymbol{z}}^*}l(\hat{\boldsymbol{z}})=\frac{1}{\sigma^2}\boldsymbol{A}^{\mathrm{H}}\boldsymbol{b}-\frac{1}{\sigma^2}\boldsymbol{A}^{\mathrm{H}}\boldsymbol{A}\hat{\boldsymbol{z}}$$

令其等于零，得 $\boldsymbol{A}^{\mathrm{H}}\boldsymbol{b}-\boldsymbol{A}^{\mathrm{H}}\boldsymbol{A}\boldsymbol{z}_{\mathrm{opt}}=\boldsymbol{0}$ 或 $\boldsymbol{A}^{\mathrm{H}}\boldsymbol{A}\boldsymbol{z}_{\mathrm{opt}}=\boldsymbol{A}^{\mathrm{H}}\boldsymbol{b}$，其中 $\boldsymbol{z}_{\mathrm{opt}}$ 是使对数似然函数 $l(\hat{\boldsymbol{z}})$ 极大化的 $\hat{\boldsymbol{z}}$ 值。于是，若 $\boldsymbol{A}^{\mathrm{H}}\boldsymbol{A}$ 非奇异，则

$$\boldsymbol{z}_{\mathrm{opt}}=(\boldsymbol{A}^{\mathrm{H}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{H}}\boldsymbol{b} \tag{4.2.27}$$

与最小二乘解具有等价性

无约束最小化问题的梯度分析——极值点的辨识

必要条件

$$\left.\frac{\partial f(z,z^*)}{\partial z^*}\right|_{z=z_0}=0,\quad \begin{bmatrix} H_{z^*,z} & H_{z^*,z^*} \\ H_{z,z} & H_{z,z^*} \end{bmatrix}_{z=z_0}\succeq 0$$

或

$$\left.\frac{\partial f(Z,Z^*)}{\partial Z^*}\right|_{Z=Z_0}=O,\quad \begin{bmatrix} H_{Z^*,Z} & H_{Z^*,Z^*} \\ H_{Z,Z} & H_{Z,Z^*} \end{bmatrix}_{Z=Z_0}\succeq 0$$

充分条件

$$\left.\frac{\partial f(z,z^*)}{\partial z^*}\right|_{z=z_0}=0,\quad \begin{bmatrix} H_{z^*,z} & H_{z^*,z^*} \\ H_{z,z} & H_{z,z^*} \end{bmatrix}_{z=z_0}\succ 0$$

或

$$\left.\frac{\partial f(Z,Z^*)}{\partial Z^*}\right|_{Z=Z_0}=O,\quad \begin{bmatrix} H_{Z^*,Z} & H_{Z^*,Z^*} \\ H_{Z,Z} & H_{Z,Z^*} \end{bmatrix}_{Z=Z_0}\succ 0$$

则z_0为严格局部最小点

无约束最小化问题的梯度分析—凸优化理论

标准的约束优化问题

$$\min_{\boldsymbol{x}} \ f_0(\boldsymbol{x}) \quad \text{subject to } f_i(\boldsymbol{x}) \leqslant 0,\ i=1,\cdots,m; \boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$$

约束优化问题很难求解，决策变量x很大，有很多局部解，收敛速度差，收敛的停止准则失败等。

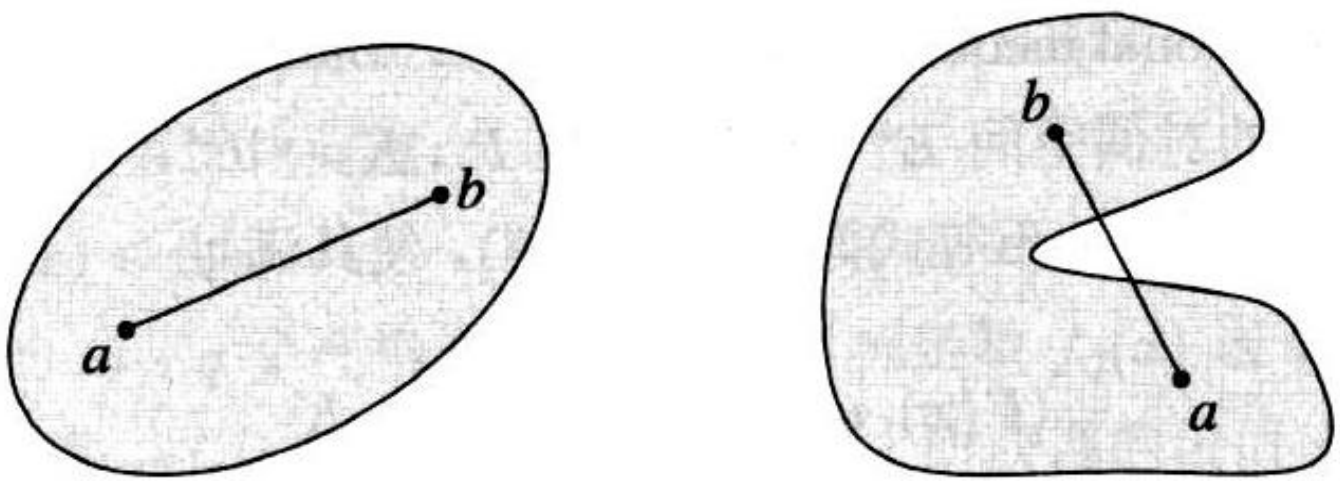
定义 4.3.1 一个集合 $S \in \mathbb{R}^n$ 称为凸集 (合)，若对任意两个点 $\boldsymbol{x}, \boldsymbol{y} \in S$，连接它们的线段也在集合 S 内，即

$$\boldsymbol{x}, \boldsymbol{y} \in S, \quad \theta \in [0,1] \quad \Longrightarrow \quad \theta\boldsymbol{x} + (1-\theta)\boldsymbol{y} \in S \tag{4.3.11}$$

无约束最小化问题的梯度分析—凸优化理论

定义 4.3.1 一个集合 $S \in \mathbb{R}^n$ 称为凸集 (合)，若对任意两个点 $\boldsymbol{x}, \boldsymbol{y} \in S$，连接它们的线段也在集合 S 内，即

$$\boldsymbol{x}, \boldsymbol{y} \in S, \quad \theta \in [0,1] \quad \Longrightarrow \quad \theta\boldsymbol{x} + (1-\theta)\boldsymbol{y} \in S \tag{4.3.11}$$



(a) 凸集 (b) 非凸集

凸函数

定义 4.3.4[363] 给定一个凸集 $S \in \mathbb{R}^n$ 和函数 $f : S \to \mathbb{R}$，则：

(1) 函数 $f : \mathbb{R}^n \to \mathbb{R}$ 称为凸函数 (convex function)，当且仅当 $S = \mathrm{dom}(f)$ 是凸集，并且对于所有 $\boldsymbol{x}, \boldsymbol{y} \in S$ 和每一个标量 $\alpha \in (0,1)$，函数满足 Jensen 不等式

$$f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \leqslant \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) \tag{4.3.18}$$

(2) 函数 $f(\boldsymbol{x})$ 称为严格凸函数 (strictly convex function)，当且仅当 $S = \mathrm{dom}(f)$ 是凸集，并且对于所有 $\boldsymbol{x}, \boldsymbol{y} \in S$ 和每一个标量 $\alpha \in (0,1)$，函数满足不等式

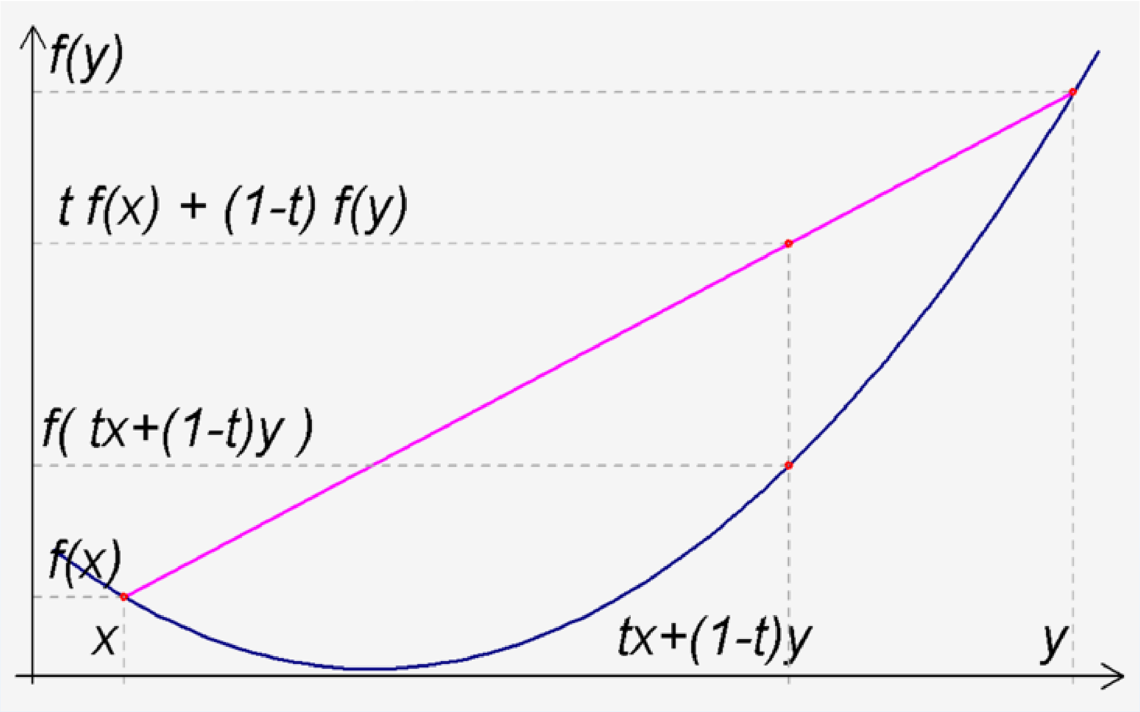
$$f(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) < \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) \tag{4.3.19}$$

无约束最小化问题的梯度分析—凸优化理论

凸函数

(2) 函数 $f(\boldsymbol{x})$ 称为严格凸函数 (strictly convex function)，当且仅当 $S = \mathrm{dom}(f)$ 是凸集，并且对于所有 $\boldsymbol{x}, \boldsymbol{y} \in S$ 和每一个标量 $\alpha \in (0, 1)$，函数满足不等式

$$f(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) < \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) \tag{4.3.19}$$



无约束最小化问题的梯度分析—凸优化理论

凸函数辨识的二阶充分必要条件

定理 4.3.3[328] 令 $f: S \to \mathbb{R}$ 是一个定义在 n 维向量空间 $\mathbb{R}^n$ 内的凸集 S 上的函数，并且可二次微分，则 $f(\boldsymbol{x})$ 是凸函数，当且仅当 Hessian 矩阵半正定

$$\boldsymbol{H}_{\boldsymbol{x}} f(\boldsymbol{x}) = \frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}} \succeq 0, \quad \forall\, \boldsymbol{x} \in S \tag{4.3.33}$$

注释 令 $f: S \to \mathbb{R}$ 是一个定义在 n 维向量空间 $\mathbb{R}^n$ 内的凸集 S 上的函数，并且可二次微分，则 $f(\boldsymbol{x})$ 是严格凸函数，当且仅当 Hessian 矩阵正定

$$\boldsymbol{H}_{\boldsymbol{x}} f(\boldsymbol{x}) = \frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^{\mathrm{T}}} \succ 0, \quad \forall\, \boldsymbol{x} \in S \tag{4.3.34}$$

与严格极小点的充分条件要求 Hessian 矩阵在 $\boldsymbol{c}$ 一点正定不同，这里要求 Hessian 矩阵在整个凸集 S 的所有点均正定。

无约束最小化问题的梯度分析—实值目标函数的最速下降方向

以复向量或者矩阵为变元的实值目标函数的平稳点存在两种选择

$$\left.\frac{\partial f(Z, Z^*)}{\partial Z}\right|_{Z=C} = O_{m\times n} \quad 或 \quad \left.\frac{\partial f(Z, Z^*)}{\partial Z^*}\right|_{Z=C} = O_{m\times n}$$

在设计优化迭代算法时，应该选哪一种梯度?

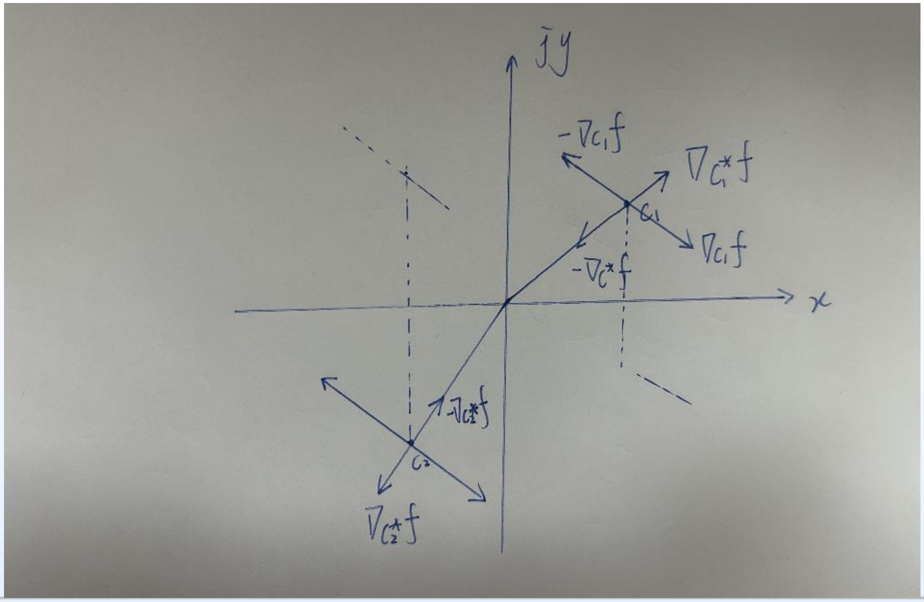
无约束最小化问题的梯度分析—实值目标函数的最速下降方向

定理 4.2.1[58] 令 $f(\boldsymbol{z})$ 是复向量 $\boldsymbol{z}$ 的实值函数。通过将 $\boldsymbol{z}$ 和 $\boldsymbol{z}^*$ 视为独立的变元，实目标函数 $f(\boldsymbol{z})$ 的曲率方向由共轭梯度向量 $\nabla_{\boldsymbol{z}^*} f(\boldsymbol{z})$ 给出。

曲率方向也就是函数的最大变化率方向。

(1) 共轭梯度向量 $\nabla_{\boldsymbol{z}^*} f(\boldsymbol{z}, \boldsymbol{z}^*)$ 或 $\nabla_{\mathrm{vec}(\boldsymbol{Z}^*)} f(\boldsymbol{Z}, \boldsymbol{Z}^*)$ 给出目标函数增长最快的方向；

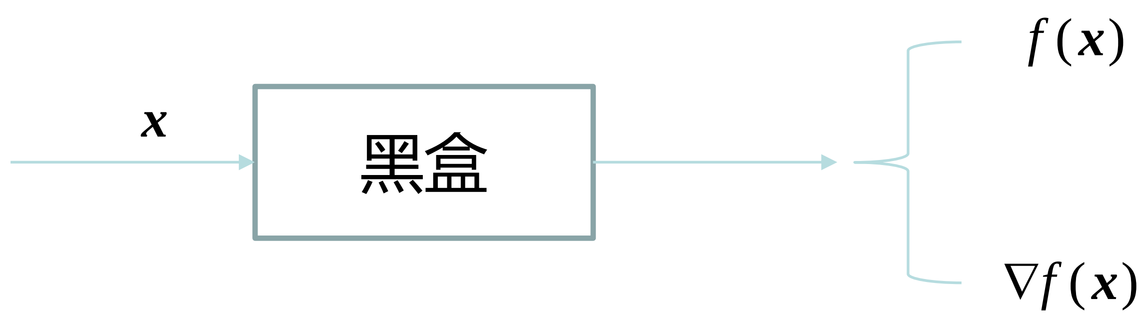
(2) 负共轭梯度向量 $-\nabla_{\boldsymbol{z}^*} f(\boldsymbol{z}, \boldsymbol{z}^*)$ 或 $-\nabla_{\mathrm{vec}(\boldsymbol{Z}^*)} f(\boldsymbol{Z}, \boldsymbol{Z}^*)$ 给出目标函数最陡减小的方向。



函数 $f(z)=|z|^2$ 梯度和共轭梯度方向示意图

平滑凸优化的一阶算法——梯度法/最陡下降方法

$$\min_{\boldsymbol{x}\in\mathcal{D}} f(\boldsymbol{x})$$



令 x_{opt} 表示 $\min f(x)$ 的最优解, 一阶黑盒优化 (first-order black-box optimization)就是只利用 $f(\boldsymbol{x})$ 和 $\nabla f(\boldsymbol{x})$ 求解向量 $\boldsymbol{y}\in Q$ 满足 $\boldsymbol{y}:\ f(\boldsymbol{y})-f\left(\boldsymbol{x}_{\text{opt}}\right)\leqslant\varepsilon$ ⟶ 给定的精度误差

$$\arg\min f(\boldsymbol{x})$$

↓

Argument，函数取最小值时 $\boldsymbol{x}$ 的取值

平滑凸优化的一阶算法——梯度法/最陡下降方法

梯度法（最陡下降法）是一种最简单的一阶优化算法

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \mu_k \Delta \boldsymbol{x}_k, \quad k = 1, 2, \cdots$$

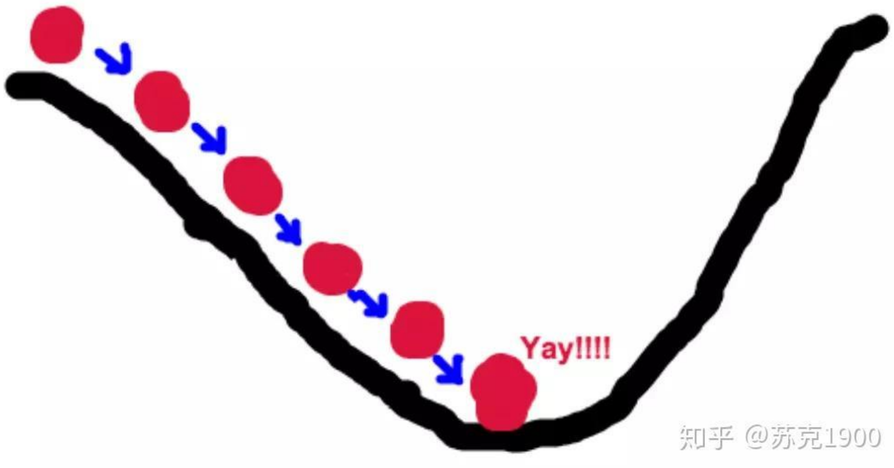
x_{opt}

搜索方向或者更新方向

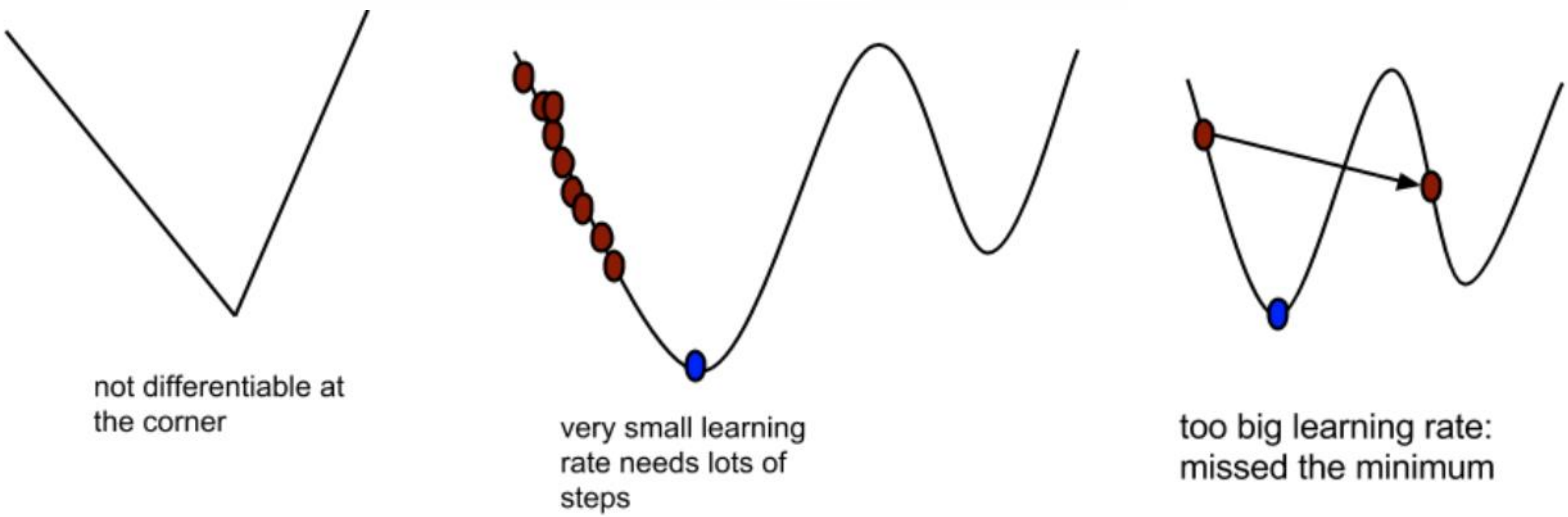
步长(学习率)
用于控制更新x寻优的步伐

$f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$ 迭代过程中目标函数下降

平滑凸优化的一阶算法——梯度法/最陡下降方法

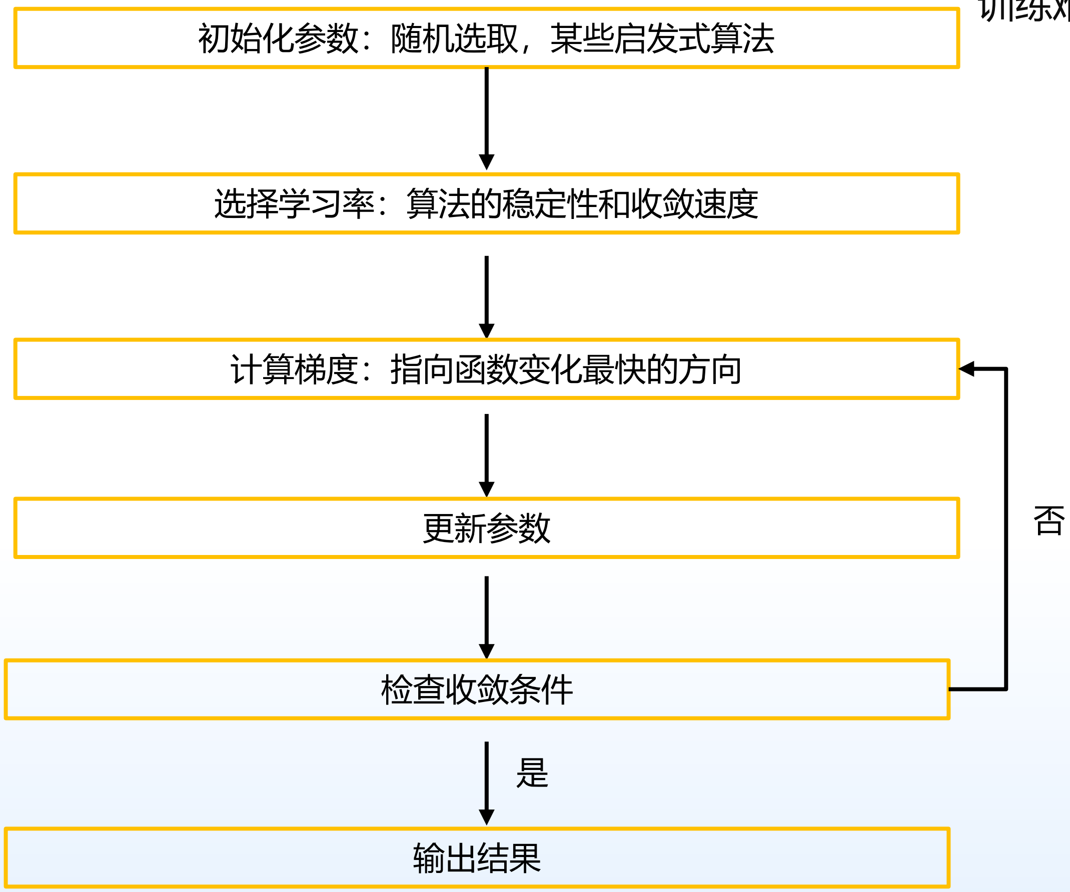


Gradient Descent lecture notes from UD262 Udacity Georgia Tech ML Course.



平滑凸优化的一阶算法——梯度法/最陡下降方法

梯度法的基本流程



梯度消失和梯度爆炸：在深度学习模型训练中，由于层数过多，梯度可能在传播过程中逐渐变小或变大，导致训练难以进行。需要采取特定策略来解决这些问题。

拓展——梯度下降的变体

- 批量梯度下降（Batch Gradient Descent）：每一步更新都使用所有训练样本来计算梯度。精度高但计算量大，对大数据集不够高效。
- 随机梯度下降（Stochastic Gradient Descent, SGD）：每一步更新只使用一个训练样本来计算梯度。计算速度快，但更新过程中有较多噪声。
- 小批量梯度下降（Mini-batch Gradient Descent）：每一步更新使用一小批训练样本来计算梯度。兼顾了批量梯度下降的精度和随机梯度下降的速度，是实际应用中的常用选择。

第三章习题

见学在浙大
作业版块
Homework2

10.13 交