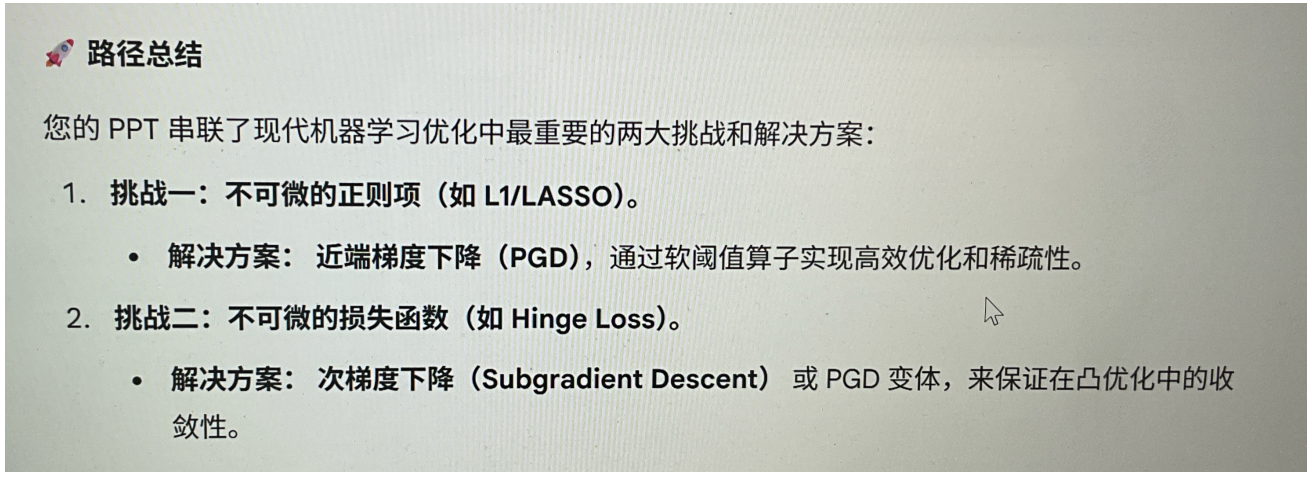
Stochastic Gradient Descent



🚀 **路径总结**

您的 PPT 串联了现代机器学习优化中最重要的两大挑战和解决方案：

1. **挑战一：不可微的正则项（如 L1/LASSO）。**
   - **解决方案： 近端梯度下降（PGD）**，通过软阈值算子实现高效优化和稀疏性。
2. **挑战二：不可微的损失函数（如 Hinge Loss）。**
   - **解决方案： 次梯度下降（Subgradient Descent）** 或 PGD 变体，来保证在凸优化中的收敛性。

→ Ulit 了的一个总结（两个clip的梳理）

Objectives

- Simplify gradient descent update
- Common methods for cycling through data
- Benefits
- Examples

Stochastic gradient descent updates weights using part of the data

$$f(\underline{w}) = \ell(\underline{w}) + \lambda r(\underline{w})$$

cost function

"loss" "regularize"

$$\underline{w}^{(k+1)} = \underline{w}^{(k)} - \frac{\tau}{2}\nabla_{\underline{w}} f(\underline{w})$$

沿的方向下坡

gradient

Gradient-descent through iterations converge to global min due to convex property of function

squared error

$$\ell(\underline{w}) = \sum_{i=1}^{N} (d_i - \underline{x}_i^T \underline{w})^2$$

$$\nabla_{\underline{w}} \ell(\underline{w}) = -2 \sum_{i=1}^{N} (d_i - \underline{x}_i^T \underline{w})\, \underline{x}_i$$

hinge loss

$$\ell(\underline{w}) = \sum_{i=1}^{N} (1 - d_i \underline{x}_i^T \underline{w})_+$$

$$\nabla_{\underline{w}} \ell(\underline{w}) = -\sum_{i=1}^{N} I_{\{d_i \underline{x}_i^T \underline{w} < 1\}}\, \underline{x}_i$$

$(d_i, \underline{x}_i), i = 1, \dots, N$

labels features

depends on all the data

SGD:

$$f(\underline{w}) = \sum_{i=1}^{N} f_i(\underline{w})$$

Define i_k, $k = 1, 2, \dots$

$$\underline{w}^{(k+1)} = \underline{w}^{(k)} - \frac{\tau}{2} \nabla_{\underline{w}} f_{i_k}(\underline{w}^{(k)})$$

depends on one sample $(d_{i_k}, \underline{x}_{i_k})$

SGD cycles through training data

N个样本，每个N epoch to converge

1) Cyclical (按顺序遍历) (incremental gradient descent)

$i_k = k \bmod N$ e.g. $i_k = 1,2,3,4,1,2,3,4,1,2,3\cdots$

2) Random permutation (先打乱一次，遍历，再打乱→遍历…) (reshuffle every N roundrs)

$i_k = 2,4,1,3,\boxed{2,1,4,3},4,3,1,2\ \cdots$

常用的

3) Stochastic gradient descent (uniformly at random)

$i_k = \text{uniform}\{1,2,\cdots,N\}$ $i_k = 2,1,3,1,4,4,2,3,1,3\ \cdots$

Update by $-\tau_2 \nabla_{\underline{w}} f_{i_k}(\underline{w})$ at each iteration

On average gives gradient $E\{\nabla_{\underline{w}} f_{i_k}(\underline{w})\} \simeq \frac{\nabla_{\underline{w}} f(\underline{w})}{N}$

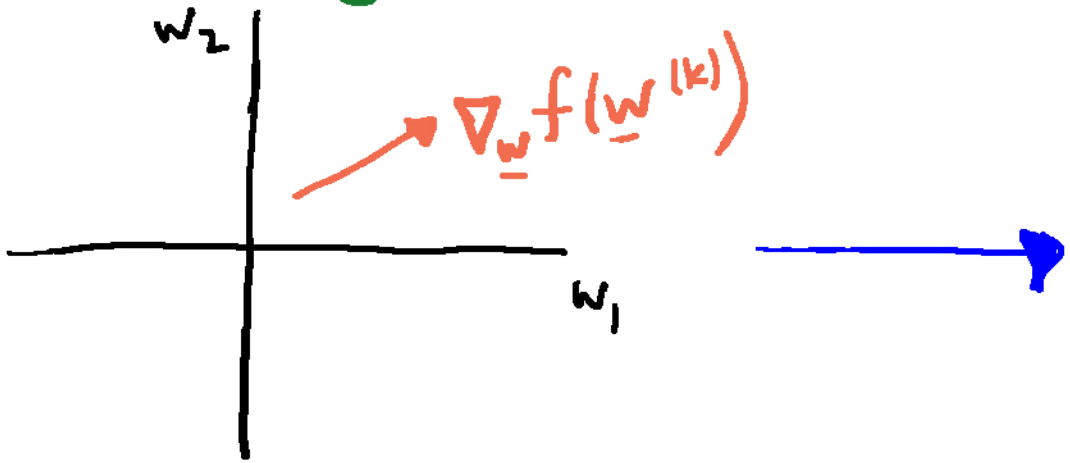
平均意义上等同于总梯度叠加

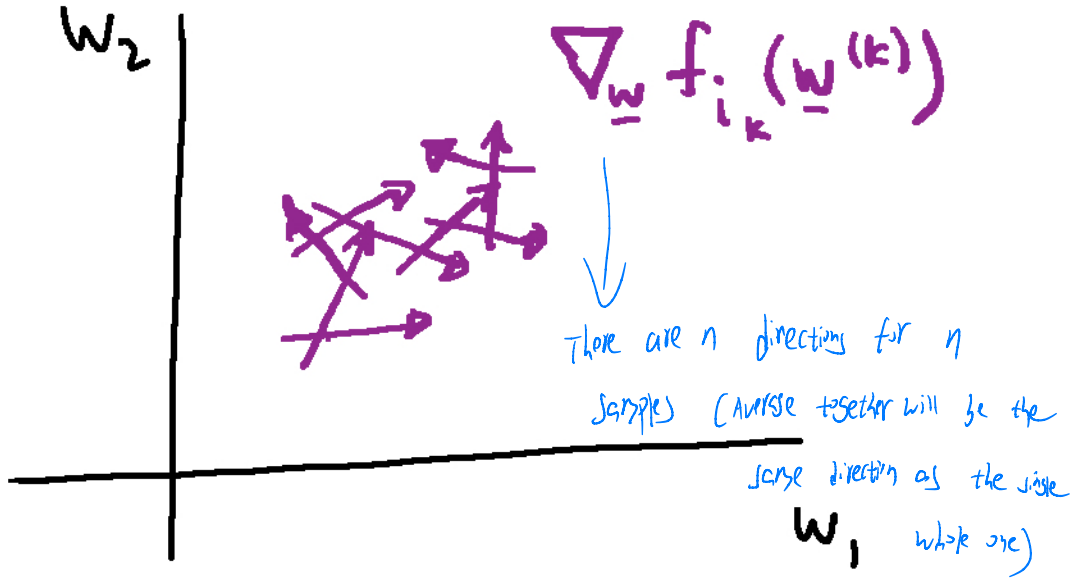
SGD has computational benefits

1) Computing $\nabla_{\underline{w}} f_{i_k}(\underline{w}^{(k)})$ is easier/faster than $\nabla_{\underline{w}} f(\underline{w}^{(k)})$

2) May not be able to store $\underline{x}_i$, $i=1,\dots,N$ in memory

3) Noisy gradient $\nabla_{\underline{w}} f_{i_k}(\underline{w}^{(k)})$ introduces added regularization





Example: Ridge Regression

$$f(\underline{w}) = \sum_{i=1}^{N} (d_i - \underline{x}_i^T \underline{w})^2 + \lambda \|\underline{w}\|_2^2 = \sum_{i=1}^{N} \underbrace{\left\{ (d_i - \underline{x}_i^T \underline{w})^2 + \frac{\lambda}{N} \|\underline{w}\|_2^2 \right\}}_{f_i(\underline{w})}$$

$f_i(\underline{w})$: 算一个 sample 的 error

$$\nabla_{\underline{w}} f_i(\underline{w}) = \nabla_{\underline{w}} \left[(d_i - \underline{x}_i^T \underline{w})^2 + \frac{\lambda}{N} \underline{w}^T \underline{w} \right]$$

$$= -2 (d_i - \underline{x}_i^T \underline{w}) \underline{x}_i + \frac{2\lambda}{N} \underline{w}$$

$$\underline{w}^{(k+1)} = \underline{w}^{(k)} - \frac{\tau}{2} \nabla_{\underline{w}^{(k)}} f_{i_k}(\underline{w}^{(k)})$$

(k here means the k^{th} iteration)

$$= \underline{w}^{(k)} + \tau (d_{i_k} - \underline{x}_{i_k}^T \underline{w}^{(k)}) \underline{x}_{i_k} - \frac{\tau\lambda}{N} \underline{w}^{(k)}$$

VS. $\underline{w}^{(k+1)} = \underline{w}^{(k)} + \tau \underline{A}^T (\underline{A} \underline{w}^{(k)} - \underline{d}) - \lambda\tau \underline{w}^{(k)}$

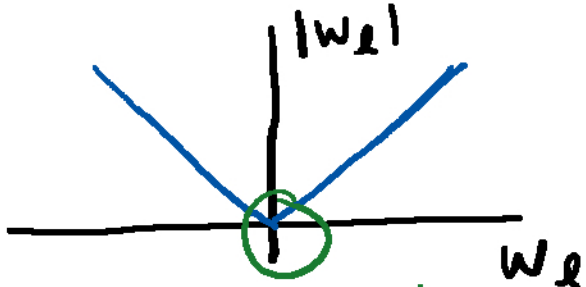
$\underline{A}: N \times M$

Example: Gradient descent for LASSO

$$f(\underline{w}) = \sum_{i=1}^{N} (d_i - \underline{x}_i^T \underline{w})^2 + \lambda \|\underline{w}\|_1 = \sum_{i=1}^{N} \underbrace{\left\{ (d_i - \underline{x}_i^T \underline{w})^2 + \frac{\lambda}{N} \|\underline{w}\|_1 \right\}}_{f_i(\underline{w})}$$

这里的 L1 正则化也是导致 function 不可导的原因

Consider $\nabla_{\underline{w}} \sum_{\ell=1}^{M} |w_\ell|$



Use subgradient

Write $\nabla_{\underline{w}} \|\underline{w}\|_1 = \text{sign}(\underline{w})$

$$\frac{d}{dw_\ell} |w_\ell| = \begin{cases} \text{sign}(w_\ell) & w_\ell \neq 0 \\ [-1, 1] & w_\ell = 0 \end{cases}$$

"0" popular

$$\nabla_w f_i(\underline{w}) = -2(d_i - \underline{x}_i^T \underline{w}) \underline{x}_i + \frac{\lambda}{N} \text{sign}(\underline{w})$$

$$\underline{w}^{(k+1)} = \underline{w}^{(k)} + \tau (d_{i_k} - \underline{x}_{i_k}^T \underline{w}^{(k)}) \underline{x}_{i_k} - \frac{\lambda \tau}{2N} \text{sign}(\underline{w}^{(k)})$$

↓ Each sample