章节 6.2 梯度下降法

梯度下降法

梯度下降法取负梯度作为迭代算法的搜索方向，其迭代格式为

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

算法 梯度下降算法 GD

Require: 选取初始点 $\mathbf{x}^{(0)}$, 设置终止误差 $\varepsilon > 0$, 令 $k := 0$.

1: **while** $\|\nabla f(\mathbf{x}^{(k)})\| > \epsilon$ **do**

2: 令 $\mathbf{d}^{(k)} = -\nabla f(\mathbf{x}^{(k)})$, 并由一维搜索确定步长因子 α_k 使得 $f(\mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)})$ 满足 Backtracking linesearch 或者 Wolfe-Powell 条件

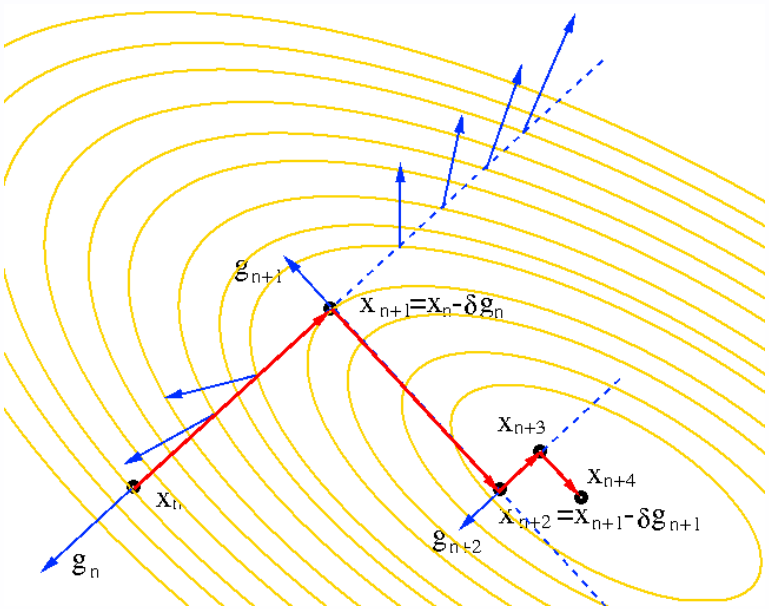
3: 迭代更新 $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{d}^{(k)}$, 置 $k := k + 1$ 。

4: **end while**

梯度下降法

梯度下降法全局收敛性定理：

设 $f(\mathbf{x}) \in C^1$，在梯度下降法中采用（精确或非精确）一维搜索，则产生的迭代点列 $\{\mathbf{x}^{(k)}\}$ 的每一个聚点都是驻点。



我们下面着重讲解，在非凸、凸函数、强凸三种情况下，梯度算法的收敛结果。

定理 1.1

判定凸函数的一阶条件：f 是一个凸集上的可微函数。f 是凸函数，当且仅当其满足

$$f(y) \geq f(x) + \nabla f(x)^T (y - x), \quad \forall x, y \in \mathrm{dom} f.$$

证明：

- **充分性：**对任意的 $x, y \in \mathrm{dom} f$ 以及任意的 $t \in (0, 1)$，定义 $z = tx + (1 - t)y$，应用两次一阶条件我们有

$$f(x) \geq f(z) + \nabla f(z)^T (x - z),$$

$$f(y) \geq f(z) + \nabla f(z)^T (y - z).$$

将上述第一个不等式两边同时乘 t，第二个不等式两边同时乘 $1 - t$，相加得

$$tf(x) + (1 - t)f(y) \geq f(z).$$

这正是凸函数的定义，因此充分性成立。

- **必要性：**设 f 是凸函数。对于任意可行域的 x, y, 以及 $t \in (0,1)$, 根据凸函数定义，可得

$$tf(y) + (1-t)f(x) \geq f(x + t(y-x)).$$

由上式，经过移项处理，两边同时除以 t 可得

$$f(y) - f(x) \geq \frac{f(x + t(y-x)) - f(x)}{t}.$$

令 $t \to 0$, 因为极限保号性，可得

$$f(y) - f(x) \geq \lim_{t \to 0} \frac{f(x + t(y-x)) - f(x)}{t} = \nabla f(x)^T (y-x).$$

证毕。

定理 1.2

判定凸函数二阶条件：f 是一个凸集上的二阶连续可微。f 是凸函数，当且仅当其满足

$$\nabla^2 f(x) \succcurlyeq 0.$$

证明.

- **必要性：** 反设 $f(x)$ 在点 x 处的 Hessian 矩阵 $\nabla^2 f(x) \not\succcurlyeq 0$，即存在非零向量 $v \in \mathbb{R}^n$ 使得 $v^T \nabla^2 f(x) v < 0$。根据二阶可微性，有泰勒展开，

$$f(x+tv) = f(x) + t\nabla f(x)^T v + \frac{t^2}{2} v^T \nabla^2 f(x) v + o(t^2).$$

移项后等式两边同时除以 t^2，

$$\frac{f(x+tv) - f(x) - t\nabla f(x)^T v}{t^2} = \frac{1}{2} v^T \nabla^2 f(x) v + o(1).$$

当 t 充分小时，

$$\frac{f(x+tv) - f(x) - t\nabla f(x)^T v}{t^2} < 0,$$

这显然和一阶条件矛盾，因此必有 $\nabla^2 f(x) \geq 0$ 成立。

- **充分性：**若 f 满足二阶条件。对于可行域的任意 x, y, 根据二阶可微性，

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x),$$

这里 z 是 x, y 连线上的一个点。由于 $\nabla^2 f(z) \succcurlyeq 0$, 可知一阶条件成立。故 f 是凸函数。

例 1.3

$f(x) = \frac{1}{2}\|Ax - b\|^2$. 则 $\nabla f(x) = A^T(Ax - b)$, $\nabla^2 f(x) = A^T A$ 是半正定的。故其为凸函数。

例: log-sum-exp 函数 $f(x) = \log \sum_{k=1}^{n} \exp x_k$ 是凸函数

$$\nabla^2 f(x) = \frac{1}{\sum_k z_k} \mathrm{diag}(z) - \frac{1}{(\sum_k z_k)^2} z z^T$$

这里 $z_k = \exp x_k$.

To prove $\nabla^2 f(x) \geq 0$, we only need to prove for any v, $v^T \nabla^2 f(x) v \geq 0$, i.e.,

$$v^T \nabla^2 f(x) v = \frac{\sum_k z_k v_k^2}{\sum_k z_k} - \frac{(\sum_k v_k z_k)^2}{(\sum_k z_k)^2} \geq 0$$

Using the Cauchy-Schwarz inequality, we have $(\sum_k v_k z_k)^2 \leq (\sum_k z_k v_k^2)(\sum_k z_k)$, thus f is a convex function.

作业 6.3: 证明:

- $f(x) = -(\sum_{k=1}^{n} x_k)^{\frac{1}{n}}$ (for $x \in \mathbb{R}_+^n$) 是凸函数. 即几何平均是凹函数。
- $f(x, y) = x^2/y$ 是定义域 $\{(x, y) \mid y > 0\}$ 上的凸函数。即二次函数的分式变换是凸函数。

定义 1.4

若给定函数 f 是可微函数，并且对于任意定义域的点 x, y, 梯度满足李氏连续性 (Lipschitz continuous), 即存在 $L > 0$, 使得

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|.$$

则称 f 是梯度李氏连续，或者李氏光滑（L-光滑）的。

引理 1.5

若 f 是李氏光滑的，则 f 有二次上界，即

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|^2.$$

证明： 由 f 可微，可得

$$\begin{aligned} f(y) &= f(x) + \int_0^1 \nabla f(x + t(y - x))^T(y - x)dt \\ &= f(x) + \nabla f(x)^T(y - x) + \int_0^1 (\nabla f(x + t(y - x)) - \nabla f(x))^T(y - x)dt \end{aligned}$$

证明 (续):

因此,

$$\begin{aligned} f(y) - f(x) - \nabla f(x)^T(y-x) &= \int_0^1 (\nabla f(x+t(y-x)) - \nabla f(x))^T(y-x)dt \\ &\le \int_0^1 \|\nabla f(x+t(y-x)) - \nabla f(x)\|\|y-x\|dt \\ &\le \int_0^1 Lt\|y-x\|^2 dt \\ &= \frac{L}{2}\|y-x\|^2. \end{aligned}$$

很多常用的函数满足李氏光滑性, 例如 $f(x) = \frac{1}{2}\|Ax-b\|^2$. 我们有

$$\|\nabla f(y) - \nabla f(x)\| \le \lambda_{\max}(A^TA)\|y-x\|,$$

这里 $\lambda_{\max}(A^TA)$ 是 A^TA 的最大特征根。因此，二次函数的 Lipschitz 常数是 $L = \lambda_{\max}(A^TA)$. 通常来说, $L = \max_x \lambda_{\max}(\nabla^2 f(x))$, 即定义域内的所有 Hessian 矩阵的最大特征根。

反例: $f(x) = e^x, f(x) = x^3$.

作业 6.5: 对于逻辑回归问题，$\min f(x) = \frac{1}{N}\sum_{i=1}^N \log(1+e^{-y_i a_i^T x})$, 这里 $y_i \ge 0, a_i$ 是已知的。估计 f 李氏常数 L.

梯度法的另一个理解

构造 $f(x)$ 的一个二次函数上界.

定义: $q_x(y)$ 是 f 的上界函数, 如果

- $q_x(y) = f(x)$
- $q_x(y) \geq f(y)$, for any y.

最大化-最小化方法:

$$x_{k+1} = \arg\min_y q_{x_k}(y)$$

我们有

$$f(x_{k+1}) \leq q_{x_k}(x_{k+1}) \leq q_{x_k}(x_k) = f(x_k)$$

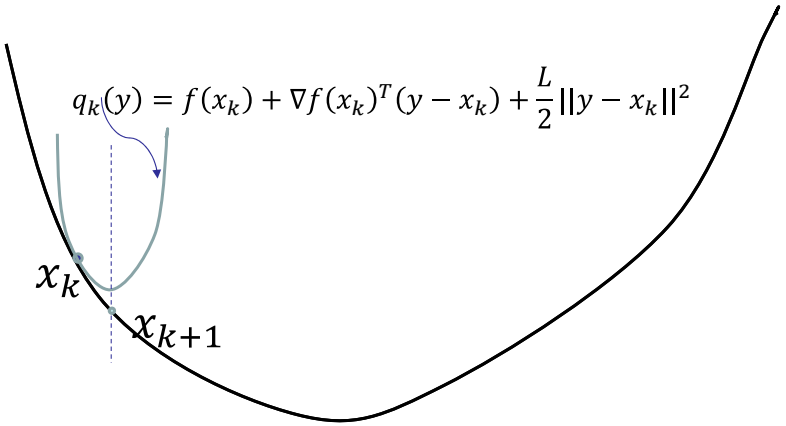


Figure: 最大化-最小化

梯度法对于非凸、李氏光滑的函数收敛

定理 1.6

若 f 是 $L-$ 光滑函数并且 f 有最小值 f^*，则选取步长 $\alpha_k = 1/L$, 我们有 $\lim\limits_{k\to\infty} \|\nabla f(x_k)\| = 0$ 并且并且对于任意正整数 T，有

$$\min_{k=0,1,\ldots,T} \|\nabla f(x_k)\|^2 \leq \frac{2L(f(x_0) - f^*)}{T}.$$

证明： 由李氏光滑性，$q_{x_k}(y) = f(x_k) + \nabla f(x_k)^T(y - x_k) + \frac{1}{2\alpha}\|y - x_k\|^2$ 为一个上界函数。梯度法迭代满足

$$x_{k+1} = \arg\min_y q_{x_k}(y) = x_k - 1/L\nabla f(x_k).$$

所以

$$f(x_{k+1}) \leq q_{x_k}(x_{k+1}) = q_{x_k}(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 = f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2. \tag{94}$$

因此

$$\sum_{k=0}^{\infty} \frac{1}{2L}\|\nabla f(x_k)\|^2 < \infty.$$

故

$$\lim_{k\to\infty} \|\nabla f(x_k)\| = 0.$$

并且对于任意正整数 T，有 $\min_{k=0,1,\ldots,T} \|\nabla f(x_k)\|^2 \leq \frac{2L(f(x_0)-f^*)}{T}$.

引理 1.7

设函数 $f(x)$ 是 $\mathbb{R}^n$ 上的凸可微函数, 则以下结论等价:

1. f 的梯度为 L -连续的;
2. $\nabla f(x)$ 有余强制性, 即对任意的 $x, y \in \mathbb{R}^n$, 有

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \tag{95}$$

我们只证明：(1)⇒ (3) 定义函数 $\phi(y) = f(y) - \nabla f(x)^T y$. 函数 $\phi(y)$ 是凸函数，并且也是 $L-$ 光滑的。因为 $\nabla \phi(x) = 0$, 故 x 是 ϕ 的最小值。根据 $L-$ 光滑，

$$\phi(x) \leq \phi(y - \frac{1}{L} \nabla \phi(y)) \leq \phi(y) - \frac{1}{2L} \|\nabla \phi(y)\|^2.$$

由 $\nabla \phi(y) = \nabla f(y) - \nabla f(x)$ 可得

$$\phi(y) - \phi(x) \geq \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2.$$

即

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2.$$

交换上面 x, y，得到的不等式与上述不等式相加，即可得到结论。

定理 1.8

若 f 是 $L-$ 光滑的凸函数，并且 f 有最小值 f^*，则选取步长 $\alpha_k = 1/L$, 对于任意 $T \geq 1$,

$$f(x_T) - f^* \leq \frac{L}{2T}\|x_0 - x^*\|^2.$$

证明： 由(94)和 f 是凸函数可得

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \\ &\leq f^* + \nabla f(x_k)^T(x_k - x^*) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \\ &= f^* + \frac{L}{2}\left(\|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L}\nabla f(x_k)\|^2\right) \\ &= f^* + \frac{L}{2}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right) \end{aligned} \tag{96}$$

(96)表明梯度法中，函数值和最小值的差是严格减小的。对上式 $k = 0, 1, ..., T$ 相加可得

证明（续）：

$$\begin{aligned}\sum_{k=1}^{T}(f(x_k)-f^*) &\le \frac{L}{2}\sum_{k=1}^{T}\left(\|x_k-x^*\|^2-\|x_{k+1}-x^*\|^2\right)\\ &= \frac{L}{2}\left(\|x_0-x^*\|^2-\|x_{T+1}-x^*\|^2\right)\\ &\le \frac{L}{2}\|x_0-x^*\|^2.\end{aligned}$$

因为 $f(x_k)$ 单调递减，所以

$$f(x_T)-f^* \le \frac{L}{2T}\|x_0-x^*\|^2.$$

结论： $f(x_k)-f^*$ **收敛的速度是次线性的。**收敛到 $f(x_k)-f^* \le \epsilon$ 的速度是 $\mathcal{O}(1/k)$.

梯度下降法: 一阶方法的收敛下界

一阶方法：任何选择 x_{k+1} 在集合中的迭代算法

$$x_0 + \text{span}\{\nabla f(x_0), \nabla f(x_1), \ldots, \nabla f(x_k)\}$$

问题类：满足 L 式光滑和凸假设的任何函数。
定理 (Nesterov)：对于每个整数 $k \leq \frac{n-1}{2}$ 和每个 x_0，存在在问题类中的函数，对于任何一阶方法

$$f(x_k) - f^* \geq \frac{3L\|x_0 - x^*\|^2}{32(k+1)^2}$$

- 表明梯度方法的 $\frac{1}{k}$ 速率不是最优的。
- Nesterov's 加速梯度方法有 $\frac{1}{k^2}$ 的收敛性。

该定理见 Yu. Nesterov, Lectures on Convex Optimization (2018), section 2.1. (Theorem 2.1.7 in the book.) Nesterov 加速梯度算法，感兴趣也参考此书 section 2.2.

梯度下降法: 强凸函数收敛性

定义 1.9

可微函数是 $\mu-$ 强凸函数，如果

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}\|y-x\|^2, \forall x, y \in \mathrm{dom} f.$$

假设中增加强凸性后，我们可以得到更好的结果。强凸性意味着最小值点唯一。

引理 1.10

设函数 $f(x)$ 是 $\mathbb{R}^n$ 上的 $\mu-$ 强凸可微函数, 则有如下不等式:

$$(\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{\mu L}{L+\mu}\|x-y\|^2 + \frac{1}{L+\mu}\|\nabla f(x) - \nabla f(y)\|^2, \forall x, y \in \mathrm{dom} f \quad (97)$$

证明: 记 $\phi(x) = f(x) - \frac{\mu}{2}\|x\|^2$. 则 $\phi(x)$ 是凸函数, 并且是 $L-\mu$ 李氏光滑。由余强制性(95), 可得

$$(\nabla \phi(x) - \nabla \phi(y))^T(x-y) \geq \frac{1}{L-\mu}\|\nabla \phi(x) - \nabla \phi(y)\|^2, \forall x, y \in\in \mathrm{dom} f.$$

带入 $\nabla \phi(x) = \nabla f(x) - \mu x$，可得(97).

梯度下降法: 强凸函数收敛性

如果 $x^+ = x - \alpha \nabla f(x)$ 且 $0 < \alpha \leq \frac{2}{\mu+L}$:

$$\begin{aligned}
\left\|x^+ - x^*\right\|^2 &= \|x - \alpha\nabla f(x) - x^*\|^2 \\
&= \|x - x^*\|^2 - 2\alpha\nabla f(x)^T(x - x^*) + \alpha^2 \|\nabla f(x)\|^2 \\
&\leq (1 - \alpha\frac{2\mu L}{\mu + L}) \|x - x^*\|^2 + \alpha(\alpha - \frac{2}{\mu + L}) \|\nabla f(x)\|^2 \\
&\leq (1 - \alpha\frac{2\mu L}{\mu + L}) \|x - x^*\|^2
\end{aligned}$$

$$\|x_k - x^*\|^2 \leq c^k \|x_0 - x^*\|^2$$

其中 $c = 1 - \alpha\frac{2mL}{\mu+L}$。

- 这意味着 x_k 线性收敛至最优值 x^*。
- 对于 $\alpha = \frac{2}{\mu+L}$，我们得到 $c = \left(\frac{\kappa-1}{\kappa+1}\right)^2$ 其中 $\kappa = \frac{L}{\mu}$ 被称为条件数。

$$f(x_k) - f^* \leq \frac{L}{2} \|x_k - x^*\|^2 \leq c^k \frac{L}{2} \|x_0 - x^*\|^2$$

结论：达到 $f(x_k) - f^* \leq \epsilon$ 所需的迭代次数是 $O(\log(1/\epsilon))$。

梯度下降法总结

问题类型	收敛描述	迭代复杂度
Nonconvex L-smooth	$\|\nabla f(x)\| \leq \epsilon$	$O\left(\frac{1}{\epsilon^2}\right)$
Convex L-smooth	$f(x_k) - f^* \leq \epsilon$	$O\left(\frac{1}{\epsilon}\right)$
Strongly convex μ-smooth	$\|x_k - x^*\|^2$	$O\left(\frac{L}{\mu}\log\frac{1}{\epsilon}\right)$

Table: Convergence for gradient method under function properties

章节 6.3 随机梯度算法

问题背景

我们以监督学习为例，作为随机梯度算法的重要应用。

什么是监督学习?

监督学习是一种机器学习范式，其中模型在带标签的数据上进行训练。训练数据包含输入-输出对。给定这些对，算法学习输入和输出之间的映射，然后可以用于预测新的、以前未见过的输入的输出。

它是如何工作的?

设想一个老师和一个学生。老师有一本习题集。当学生尝试这些问题时，老师会指出正确的解答来纠正任何错误。随着时间的推移，学生变得擅长独立解决类似的问题。

在监督学习中，算法扮演学生的角色，数据充当问题集，标签是正确的解决方案。算法对训练数据进行预测，并在出错时根据真实标签调整其理解。

关键组件:

1. **训练数据**：由输入特征和正确的输出组成，这是监督学习的基础。
2. **模型**：这是用来根据输入特征预测输出的算法或算法集。
3. **损失函数**：监督学习的核心概念是损失函数。这是一个数学函数，用于测量我们模型的预测与真实值之间的差距。监督学习的目标是最小化这种损失。

损失函数的例子：最常用于回归任务（输出是连续值）的损失函数是**均方误差 (MSE)**。如果我们有 N 个数据点，对于每个点 a_i，模型 ϕ 预测的值是 $\hat{y}_i$，真实值（标签）是 y_i，那么 MSE 为：

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 = \frac{1}{N}\sum_{i=1}^{N}(\phi(a_i) - y_i)^2$$

简单来说，这计算了预测值和真实值之间的平均平方差。我们记模型 ϕ 的参数为 x，则优化的目标为

$$\min_x \frac{1}{N}\sum_{i=1}^{N}(\phi(x; a_i) - y_i)^2$$

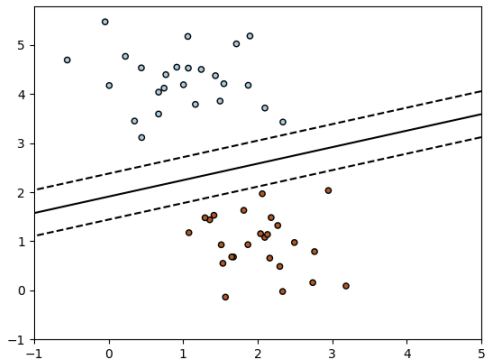


Figure: 分类任务示例。来源：https://scikit-learn.org/stable/modules/sgd.html

4. **优化技术**：这就是 SGD 等技术发挥作用的地方。它们的任务是调整模型的参数（如神经网络中的权重）以最小化损失。

深度神经网络 (DNN) 是一种前馈结构，通过堆叠多个神经元层来构建。每个神经元都会应用一个线性变换，然后再通过一个非线性激活函数进行激活。对于分类任务，最后一层通常后接一个 softmax 函数以产生类概率。然后，使用交叉熵损失来测量 DNN 的性能。神经网络的函数如下：

- 第 l 层, 给定前一层的激活值 $a^{(l-1)}$ ：

$$z^{(l)} = \phi(W^{(l)}, b^{(l)}; a^{(l-1)}) = W^{(l)} a^{(l-1)} + b^{(l)} \tag{98}$$

- ReLU作为激活函数，得到第 l 层的激活值

$$a^{(l)} = \text{ReLU}(z^{(l)}) \tag{99}$$

- 对于最后一层，我们使用 softmax 函数得到类概率：

$$\text{softmax}(z_i^{(L)}) = \frac{e^{z_i^{(L)}}}{\sum_{j=1}^{C} e^{z_j^{(L)}}} \tag{100}$$

  其中，C 是类别的数量。
- 测量网络预测与实际标签之间差异的损失函数是交叉熵损失：

$$\text{CE}(y, \hat{y}) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{c=1}^{C} y_{i,c} \log(\text{softmax}(z_{i,c}^{(L)})) \tag{101}$$

对于大规模神经网络训练，我们通常是求解下列形式的数学优化问题

$$\min_x f(x) = \frac{1}{N}\sum_{i=1}^{N} f(x, Y_i). \tag{102}$$

式子(102)中，未知量 x 表示网络参数，$Y_i, i = 1, 2, \ldots, N$ 是训练数据，它们组成总体训练集 $\{Y_1, \ldots, Y_N\}$。函数 $f(x)$ 表示的是模型的损失函数，例如在分类任务中，$f(x)$ 表示的是模型预测数据的分类与真实类别的误差。训练神经网络的目标即是寻找最优的参数 x 使得误差最小。一般来说，求解优化问题(102)关注的是最小化训练误差，而机器学习关注的是降低模型的泛化误差。在本课程中，我们更关注的是优化算法求解训练模型的最优解。

随机梯度算法

神经网络模型具有以下两个特征：（1）参数量非常大；（2）训练样本数量非常大。表29和表30 给出了一些经典的数据集和模型大小。模型规模和数据集大小制约了训练效率，大模型的训练有可能花费数天甚至数月时间。优化器的选择直接影响模型的训练效率。本节中，我们简单地探讨常用的模型训练优化器以及加速训练的并行分布式方法。下面，我们将会介绍常见的优化器的参数设定，如学习率，动量参数，批量大小等。

Table: 大数据集数据量

数据集	数据集大小
Cifar10, Cifar100	60000 张图像
ImageNet	约 1400 万张图像
MS Coco	约 33 万张图像
GPT-3	45TB

Table: 大模型参数量

模型名称	参数量
VGG16	1.4 亿
ResNet50	2500 万
DenseNet-201	2000 万
GPT-3	1750 亿
GPT-4	1.8 万亿

GPT-4 训练成本：一次的训练的成本为 6300 万美元 OpenAI 训练 GPT-4 的 FLOPS 约为 2.15e25，在大约 25000 个 A100 上训练了 90 到 100 天，利用率在 32% 到 36% 之间。

注意(102)的求和形式, 我们有

$$\nabla f(x) = \frac{1}{N}\sum_{i=1}^{N} \nabla f(x, Y_i).$$

因此当参数量非常大的时候，计算 $\nabla f(x, Y_i)$ 非常慢并且占用内存资源；当数据量 N 也非常大的时候，计算整体函数 $f(x)$ 的梯度是不可取的。所以，随机梯度算法应运而生。为了解决无法计算梯度的难点，SGD 的想法是随机抽取批量大小为 B 的子数据集 $\{Y_{i_1}, Y_{i_2}, \ldots, Y_{i_B}\}$，计算函数在子数据集上的梯度

$$\nabla f_B(x) = \frac{1}{B}\sum_{i=i_1,\ldots,i_B} \nabla f(x, Y_i).$$

在每一轮迭代（epoch）中，SGD 有两种方式选取子数据集（i）无重复（无放回）地选取子数据集进行批量梯度方向更新；（ii）可重复（有放回）地选取子数据集进行批量梯度方向更新。一般来说，实际中我们采用第（i）种方式，因为这种方式易于遍历所有数据集，使得训练得到的模型泛化能力更强。

随机梯度算法

算法 随机梯度算法 SGD

Require: 算法迭代轮数 epochs，步长 (学习率)γ_t，批数据量 B

1: **for** t=1,2,..., epochs **do**
2: **for** i=1,2,...,n **do**
3: 随机选取大小为 B 的批数据集, 计算梯度 $\nabla f_B(x_{t,i})$
4: $x_{t,i+1} = x_{t,i} - \gamma_t \nabla f_B(x_{t,i})$
5: **end for**
6: $x_{t+1,1} = x_{t,n}$
7: **end for**

算法2是 SGD 的迭代更新流程。其中，epochs 是总的更新轮数，t 是当前更新轮数。学习率 γ_t 是和 t 相关的参数。在每一轮更新中，我们采用无放回方式选取批数据，遍历全部数据集 $\{Y_1, Y_2, \ldots, Y_N\}$ 计算梯度进行更新。因此，共需要 $n = \lceil \frac{N}{B} \rceil$ 次内循环迭代 (即算法2中的第 2 到 5 行)，每个内层循环更新即为批量梯度更新。

由于我们不知道整体梯度信息，批次梯度 $-\nabla f_B(x_{t,i})$ 并非下降方向。所以线性搜索在随机梯度法中无法使用。

随机梯度算法

下面我们介绍 SGD 中常见的参数选取方式。

1. 迭代轮数设置。常见的 epochs 设定为 200 左右。如 Resnet 网络可在 100-300 次迭代得到很好的结果，Transformer 模型则需要更多的迭代，常用 300-500 次迭代。
2. 学习率设置。学习率 γ_t 和迭代轮数相关，随着 t 增大，γ_t 逐渐减小。一般有如下几种流行的方式。
   - γ_t 为一个恒定的常数。这种方式易于调参，但是最终训练误差会停在某个较大的值，所以不推荐此种方式。
   - γ_t 分段减小。例如，当 epochs 为 200 时，可以设定

$$\gamma_t = \begin{cases} 0.1, & t \leq 100 \\ 0.01, & 100 < t \leq 150 \\ 0.001, & 150 < t \leq 200 \end{cases} \tag{103}$$

   也就是说，在 $t = 100$ 和 $t = 150$ 时，我们缩小学习率 10 倍。这种分段缩小学习率的方法是最常见的设定方法。需要根据经验选择合适的初始学习率以及缩小阶段点。图12展示的是某神经网络，采用(103)学习率设定，函数值随着迭代轮数的变化情况。我们看到，在 100 和 150 次迭代时，由于学习率缩小 10 倍，函数值也有明显的快速下降。
3. 批量大小设置为了充分利用计算资源，我们尽可能选择最大的批量。因为大批量可以增加计算并行效率，提高训练速度。然而，有研究表明，批数据量大小直接影响泛化误差。因此，我们实际中也为设定一个上限，例如 $B = 128, 256$ 是最常见的选择。

随机梯度算法

- 余弦函数学习率。由分段减小学习率函数变化启发，当迭代点接近最优点时，应当选取很小的学习率。训练初期，学习率应缓慢变小，起到热启动的作用。中期则快速下降，达到快速训练的目标。而训练末期又缓慢减小，逐步逼近最优点。利用余弦函数

$$\gamma_t = \gamma_{\min} + \frac{1}{2}(\gamma_{\max} - \gamma_{\min})(1 + \cos(\frac{t}{\text{epochs}}))$$

有效模拟了该想法。其中 $\gamma_{\min}$ 和 $\gamma_{\max}$ 分别是学习率的最小值和最大值。图11所示为 cosine 学习率，以 $\gamma_{\min} = 10^{-5}$ 为最小，$\gamma_{\max} = 0.05$。此方法也是最常见的学习率设定方法。

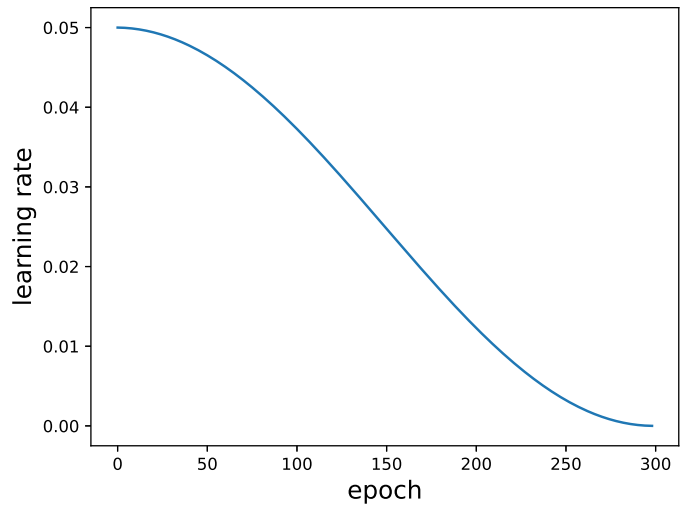


Figure: Cosine learning rate 示意图

随机梯度算法

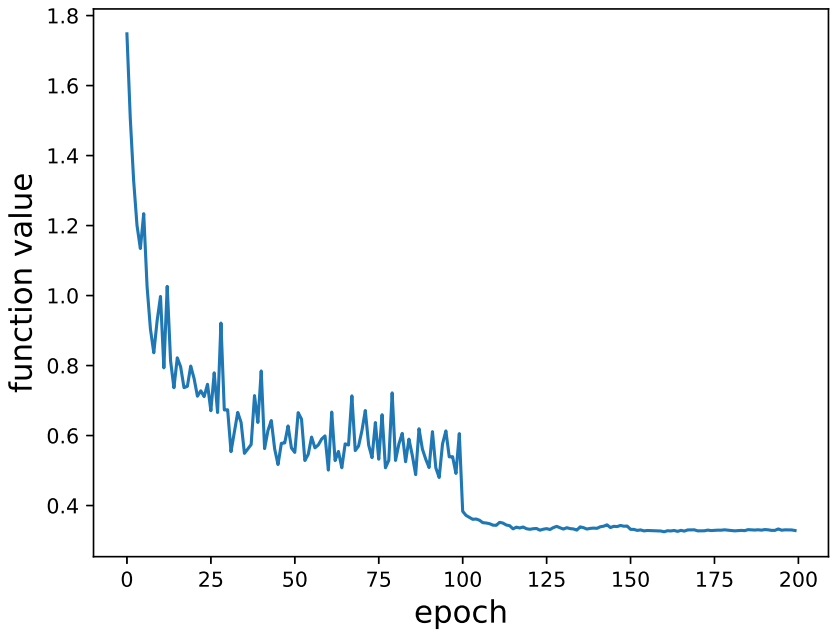


Figure: SGD 采用分段步长函数值变化示意图

动量随机梯度算法

(随机) 梯度法简单易用，但解决困难的问题（例如优化函数条件数非常大）时，常出现“之”字轨迹现象。如图13(a)所示的函数 $x_1^2 + 10x_2^2$ 等高线，由于函数在 x_2 方向变化更为陡峭，所以梯度更新轨迹在 x_2 方向过于“激进”，过于“信任”当前的梯度信息，而在 x_1 方向变化缓慢，造成了“之”字形轨迹，导致降低了算法收敛速度。动量随机梯度算法 (SGD with momentum) 可有效缓解这种情况，其描述见算法3。动量随机梯度法更新方向为动量方向 $v_{t,i+1}$，该算法也多一个动量参数 η。易知，$\eta = 0$ 时，其等价于随机梯度法。实际中，η 常用的值为 0.9。动量迭代 $v_{t,i+1} = \eta v_{t,i} + \gamma_t \nabla f_B(x_{t,i})$ 可理解为利用了“历史”梯度信息，纠正当前过于“激进”的梯度方向。

算法 动量随机梯度算法

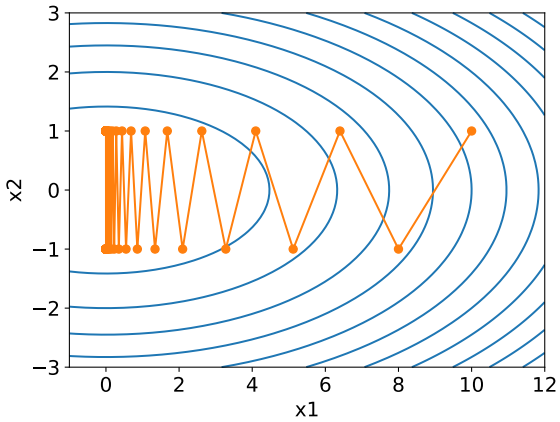
Require: 算法迭代轮数 epochs，学习率 γ_t，批数据量 B，初始动量 $v_{1,1} = 0$, 动量参数 $0 \leq \eta < 1$。

for t=1,2,..., epochs **do**
  for i=1,2,...,n **do**
    随机选取大小为 B 的批数据集, 计算梯度 $\nabla f_B(x_{t,i})$
    $v_{t,i+1} = \eta v_{t,i} + \gamma_t \nabla f_B(x_{t,i})$
    $x_{t,i+1} = x_{t,i} - v_{t,i+1}$
  end for
  $x_{t+1,1} = x_{t,n}, v_{t+1,1} = v_{t,n+1}$
end for

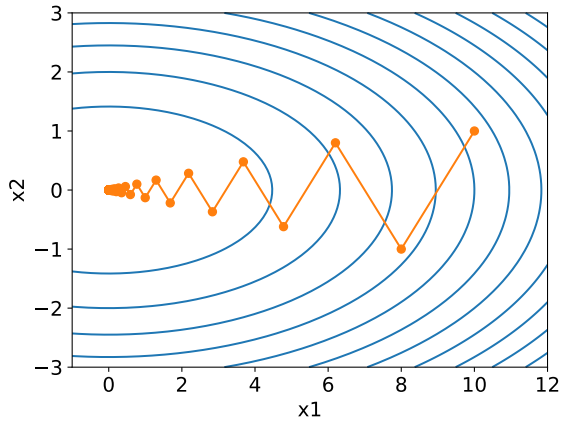
具体地，我们可以把迭代 $v_{t,i+1} = \eta v_{t,i} + \gamma_t \nabla f_B(x_{t,i})$ 展开为

$$\begin{aligned} v_{t,i+1} &= \gamma_t \nabla f_B(x_{t,i}) + \eta v_{t,i} \\ &= \gamma_t \nabla f_B(x_{t,i}) + \eta(\gamma_t \nabla f_B(x_{t,i-1}) + \eta v_{t,i-1}) \\ &= \dots \\ &= \gamma_t(\nabla f_B(x_{t,i}) + \eta \nabla f_B(x_{t,i-1}) + \eta^2 \nabla f_B(x_{t,i-2}) + \dots + \eta^j \nabla f_B(x_{t,i-j}) + \dots) \end{aligned}$$

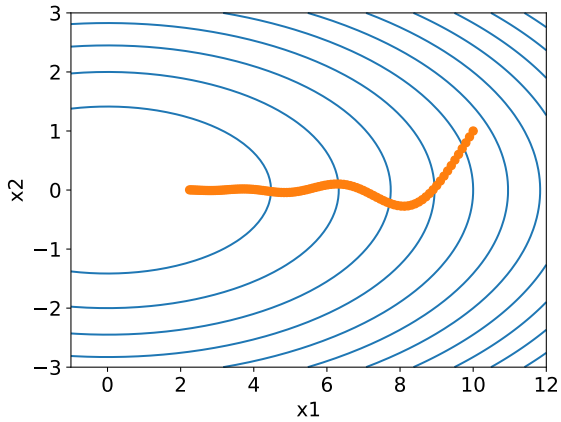
图13(b)展示的是动量梯度法求解函数 $x_1^2 + 10x_2^2$ 最小值的迭代轨迹。与图13(a)中的梯度法相对比，动量梯度法轨迹在 x_2 更加缓和，因为动量随机梯度法在历史相同的梯度方向累积起来形成动量，而梯度变化快的方向相互抵消，缓解了振荡现象。总体上来说，动量梯度法只需添加一个参数 η，其余参数设定方式可与 SGD 相似。因此，动量随机梯度法是训练神经网络非常流行的选择之一。



(a) 梯度法迭代轨迹示意图



(b) 动量梯度法迭代轨迹示意图



(c) Adam 迭代轨迹示意图

Figure: 三种优化器在函数 $x_1^2 + 10x_2^2$ 迭代轨迹

RMSProp 算法

前面我们知道，学习率很大程度上影响梯度算法的收敛表现。特别是对于如函数 $x_1^2 + 10x_2^2$，自变量 x_1, x_2 梯度方向大小不一致的情况，梯度法出现图13(a)中“之”字振荡情况。RMSProp 算法对于自变量 x 的不同坐标采用不同学习率，其算法描述见4。

算法 RMSProp 算法

Require: 算法迭代轮数 epochs，学习率 α，批数据量 B，初始向量 $s_{1,1} = 0$, 参数 $0 \leq \beta < 1$。

1: **for** t=1,2,..., epochs **do**
2: **for** i=1,2,...,n **do**
3: 随机选取大小为 B 的批数据集, 计算梯度 $\nabla f_B(x_{t,i})$
4: $s_{t,i+1} = \beta s_{t,i} + (1-\beta)\nabla f_B(x_{t,i}) \odot \nabla f_B(x_{t,i})$
5: $x_{t,i+1} = x_{t,i} - \frac{\alpha}{\sqrt{s_{t,i+1}+\epsilon}} \odot \nabla f_B(x_{t,i})$
6: **end for**
7: $x_{t+1,1} = x_{t,n}, s_{t+1,1} = s_{t,n+1}$
8: **end for**

初始 $s_{1,1}$ 是一个和变量 x 维度相同的全 0 向量。迭代 $s_{t,i+1} = \beta s_{t,i} + (1-\beta)\nabla f_B(x_{t,i}) \odot \nabla f_B(x_{t,i})$ 采用**指数平均**的方式更新。其中，符号 $\odot$ 表示向量之间按元素相乘。$s_{t,i+1}$ 追踪了函数梯度在不同坐标上的大小，指数平均的方式使得累加的梯度平方项变化更加平滑。具体地，我们考虑一般形式的指数平均迭代

$$y_t = \beta y_{t-1} + (1-\beta) b_t.$$

上式可以展开为

$$\begin{aligned} y_t &= \beta y_{t-1} + (1-\beta) b_t \\ &= (1-\beta) b_t + (1-\beta)\beta b_{t-1} + \beta^2 y_{t-2} \\ &= (1-\beta) b_t + (1-\beta)\beta b_{t-1} + (1-\beta)\beta^2 b_{t-2} + \beta^3 y_{t-3} \\ &= \dots \end{aligned}$$

参数 β 作为权重系数，一般选为 0.9。因此，指数平均赋予了距离当前时间 t 近的参数相对大的权重系数，而逐渐忽略很久之前的参数，从而达到平滑化 y_t 随着时间的变化率。

因此，RMSProp 有效追踪了梯度在不同坐标分量的情况，并且变化率比较平滑。若是累积梯度平方项 s 较大，通过 $\frac{\gamma_t}{\sqrt{s_{t,i+1}+\epsilon}}$ 处理后使用较小的学习率, 这里的 $\epsilon > 0$ 是为了防止分母太小，保证数值稳定，例如可设置为 $\epsilon = 10^{-6}$。反之，若是某项梯度较小，则使用更大的学习率平衡各个方向学习率。特别注意到，机器学习中部分参数是稀疏的。在这种情况下，使用 RMSProp 将放大稀疏部分的更新量，从而加速收敛。

Adam 算法

由指数平均启发，动量随机梯度法也可以写成以下指数平均的形式：

$$v_{t,i+1} = (1-\eta)\frac{\gamma_t}{1-\eta}\nabla f_B(x_{t,i}) + \eta v_{t,i}.$$

结合 RMSProp 中的指数平均，可以组合动量估计 v_t 和梯度平方量估计 s_t 设计算法。因此，Adam 算法应运而生，其完整描述见算法5。β_1, β_2 分别是累积梯度的权重系数和动量权重系数，Adam 算法提出者建议设置为 $\beta_1 = 0.999, \beta_2 = 0.9$。算法中第 6 行为偏差修正：$\hat{v}_{t,i+1} = \frac{v_{t,i+1}}{1-\beta_1^t}$, $\hat{s}_{t,i+1} = \frac{s_{t,i+1}}{1-\beta_2^t}$, 第 7 行更新使用修正后的 $\hat{v}_{t,i+1}$ 和 $\hat{s}_{t,i+1}$ 进行更新，其中 $\epsilon > 0$ 为保证数值稳定的常数，一般设置为 $\epsilon = 10^{-8}$。

算法 Adam 算法

Require: 算法迭代轮数 epochs，学习率 α，批数据量 B，初始向量 $s_{1,1} = 0$, 参数 $0 \leq \beta_1 < 1$，初始动量 $v_{1,1} = 0$, 动量参数 $0 \leq \beta_2 < 1$。

for t=1,2,..., epochs **do**
  for i=1,2,...,n **do**
    随机选取大小为 B 的批数据集, 计算梯度 $\nabla f_B(x_{t,i})$
    $s_{t,i+1} = \beta_1 s_{t,i} + (1-\beta_1)\nabla f_B(x_{t,i}) \odot \nabla f_B(x_{t,i})$
    $v_{t,i+1} = \beta_2 v_{t,i} + (1-\beta_2)\nabla f_B(x_{t,i})$
    偏差修正：$\hat{v}_{t,i+1} = \frac{v_{t,i+1}}{1-\beta_1^t}$, $\hat{s}_{t,i+1} = \frac{s_{t,i+1}}{1-\beta_2^t}$
    $x_{t,i+1} = x_{t,i} - \frac{\alpha}{\sqrt{\hat{s}_{t,i+1}+\epsilon}} \odot \hat{v}_{t,i+1}$
  end for
  $x_{t+1,1} = x_{t,n}, s_{t+1,1} = s_{t,n+1}, v_{t+1,1} = v_{t,n+1}$
end for

Adam 算法因结合了动量法和 RMSProp 算法，其优势在于学习率 α 更易于调节，算法在各种任务上表现更为稳定。实际中 α 对于不同任务可以采用大致相同量级的初始设定。另外，α 也可像 SGD 中的学习率 γ_t 一样随着时间改变，例如设为 cosine 学习率。图13(c)也展示了 Adam 在求解 $x_1^2 + 10x_2^2$ 最小值的收敛轨迹。