大规模机器学习的一类自适应随机梯度方法

毕业论文答辩

答辩人：XX

October 24, 2025

研究背景

考虑大规模优化问题 $\min_{x\in D} f(x) := \frac{1}{m}\sum_{i=1}^{m} f_i(x)$.

随机梯度下降法（SGD）：用 ∇f 的一个无偏估计来代替 ∇f

对 $k \in \mathbb{N}_+$，选取梯度计算式 $g(x^k, a_k)$ s.t. $\mathbf{E}(g(x^k, a_k)|x^k) = \nabla f(x^k)$ ，执行

$$x^{k+1} = P_D(x^k - \eta_k g(x^k, a_k)) \tag{1}$$

其中 $\eta_k > 0$ 为每次迭代时的步长，P_D 表示到定义域的投影。

在后文中，我们主要考虑 $D = \mathbb{R}^n$, $g(x^k, a_k) = \nabla f_{a_k}\left(x^k\right)$, $a_k \sim U(\{1, \cdots, m\})$ 的情况。

- SGD 的优点：运算效率高、内存占用小；
- SGD 的缺陷：对步长选取敏感、在病态问题中收敛缓慢、在每个维度都采取相同步长，没有考虑不同维度上 f 的不同性质。

考虑大规模优化问题 $\min_{x\in\mathbb{R}^n} f(x) := \frac{1}{m}\sum_{i=1}^m f_i(x)$.

AdaGrad-Norm 算法：通过在分母累积历史梯度范数的平方，实现步长的自适应

对 $k \in \mathbb{N}_+$，随机等可能地选取 $a_k \in \{1,2,\cdots,m\}$，执行

$$G^k = G^{k-1} + \left\|\nabla f_{a_k}(x^k)\right\|_2^2, \quad x^{k+1} = x^k - \frac{\eta \nabla f_{a_k}(x^k)}{\sqrt{G^k + \varepsilon}}. \tag{2}$$

其中 $G^0 = 0$, $\eta > 0$ 为初始步长，$\varepsilon \approx 10^{-7}$ 为数值稳定性小量。

AdaGrad 算法 (John Duchi et al. 2011)：逐分量的 AdaGrad-Norm

对 $k \in \mathbb{N}_+$，随机等可能地选取 $a_k \in \{1,2,\cdots,m\}$，执行

$$G^k = G^{k-1} + \nabla f_{a_k}(x^k) \odot \nabla f_{a_k}(x^k), \quad x^{k+1} = x^k - \frac{\eta}{\sqrt{G^k + \varepsilon}} \odot \nabla f_{a_k}(x^k). \tag{3}$$

其中 $G^0 = \mathbf{0}$, $\eta > 0$, $\varepsilon \approx 10^{-7}$，除法、根号、加法均视为逐分量运算。

AdaGrad 在病态问题和稀疏数据上表现较好，逐分量自适应步长是关键。

AdaGrad 的理论性质（随机优化情境）

定理 1 (光滑非凸情境中的收敛速度)

用 *AdaGrad* 算法 *(3)* 求解大规模优化问题 $\min_{x\in\mathbb{R}^n} f(x) := \frac{1}{m}\sum_{i=1}^{m} f_i(x)$，如果

假设 *1*：目标函数下有界：$\forall\ x \in \mathbb{R}^n,\ f(x) \geq f_*$；

假设 *2*：随机梯度一致有界：$\exists\ M,\ \ \text{s.t.}\ \forall\ 1 \leq i \leq m,\ \forall\ x \in \mathbb{R}^n,\ \|\nabla f_i(x)\|_\infty \leq M$；

假设 *3*：目标函数 *L*-光滑：$\forall\ x, y \in \mathbb{R}^n,\ \|\nabla f(x) - \nabla f(y)\|_2 \leq L\,\|x - y\|_2$，

则在 $T-1$ 次迭代后，从迭代点列 $\{x^k\}_{k=1}^{T}$ 中随机选取一个点作为输出，该点处目标函数的梯度值的期望至多与 $\ln T/T$ 同阶：

$$\mathbf{E}_t\left(\left\|\nabla f(x^t)\right\|_2^2\right) := \mathbf{E}\left(\frac{1}{T}\sum_{k=1}^{T}\left\|\nabla f(x^k)\right\|_2^2\right) = O\left(\frac{\ln T}{T}\right). \tag{4}$$

Proof (sketch).

设 $x^{k+1} = x^k - \eta u_k$，$u_k = \nabla f_{a_k}(x^k) \odot (1/\sqrt{G^k + \varepsilon})$，利用 L-光滑函数的二次上界可见

$$f(x^{k+1}) \leq f(x^k) - \eta u^k \cdot \nabla f(x^k) + \frac{L}{2}\left\|\eta u^k\right\|_2^2, \tag{5}$$

用 $\mathbf{E}_i(\cdot)$ 表示条件期望 $\mathbf{E}(\cdot \mid a_1, \cdots, a_i)$, $\mathbf{E}_0(\cdot) := \mathbf{E}(\cdot)$，上式两端求条件期望可见

$$\mathbf{E}_{k-1}\left(f(x^{k+1})\right) \leq f(x^k) - \eta \mathbf{E}_{k-1}\left(u^k \cdot \nabla f(x^k)\right) + \frac{L\eta^2}{2}\mathbf{E}_{k-1}\left(\left\|u^k\right\|_2^2\right). \tag{6}$$

$\mathbf{E}_{k-1}\left(u^k \cdot \nabla f(x^k)\right)$ 表示更新方向与负梯度方向的平均偏差，具有下界

$$\mathbf{E}_{k-1}\left(\nabla f(x^k) \cdot u^k\right) \geq \frac{\left(\nabla f(x^k)\right)^2}{2\sqrt{kM^2+\varepsilon}} - 2M\mathbf{E}_{k-1}\left(\left\|u^k\right\|_2^2\right), \tag{7}$$

将其代入 (6)，对 $k = 1, 2, \cdots, T$ 求和后，两端再求期望。由条件期望的性质，对 $k \in \mathbb{N}_+$ 有 $\mathbf{E}\left(\mathbf{E}_{k-1}(\cdot)\right) = \mathbf{E}(\cdot)$，可见

$$\mathbf{E}\left(f(x^{T+1})\right) \leq \mathbf{E}\left(f(x^1)\right) + \sum_{k=1}^{T}\left(-\frac{\eta\mathbf{E}\left(\left\|\nabla f(x^k)\right\|_2^2\right)}{2M_1} + C\mathbf{E}\left(\left\|u^k\right\|_2^2\right)\right). \tag{8}$$

其中 $C = 2M\eta + \frac{L\eta^2}{2}$, $M_1 = \sqrt{TM^2+\varepsilon}$. 最后再说明

$$\sum_{k=1}^{T}\mathbf{E}\left(\left\|u^k\right\|_2^2\right) \leq \sum_{j=1}^{n}\mathbf{E}\left(\ln\left(1 + \frac{\sum_{s=1}^{T}(g_j^s)^2}{\varepsilon}\right)\right) \leq n\ln\left(1 + \frac{TM^2}{\varepsilon}\right), \tag{9}$$

代入整理可得欲证结论。 □

AdaGrad 的理论性质（在线学习情境）

定理 2 (凸优化情境下的收敛率)

用 *AdaGrad* 算法 *(3)* 求解大规模优化问题 $\min_{x\in D} f(x) := \frac{1}{m}\sum_{i=1}^{m} f_i(x)$，如果

假设 *1*：D 为紧凸集，$f_1, \cdots, f_m$ 为凸函数，且 f 存在最小值点 $x^* \in D$；

假设 *2*：随机（次）梯度一致有界：$\forall\ x \in D,\ i \in \{1, \cdots, m\}$，有 $\|\nabla f_i(x)\|_\infty \leq M$，

则在 $T-1$ 次迭代后，采用迭代点列的平均值作为输出，将得到期望意义下 $O(1/\sqrt{T})$ 的收敛率。具体而言，记 $\bar{x}_T = \frac{1}{T}\sum_{i=1}^{T} x^t$，$d = \sup_{x,y\in D,\,1\leq j\leq n} |x_j - y_j|$，则

$$\mathbf{E}\left(f(\bar{x}_T) - f(x^*)\right) \leq \frac{n}{T}\sqrt{TM^2+\varepsilon}\left(\frac{d^2}{2\eta}+\eta\right). \tag{10}$$

Proof (sketch).

只需证明 AdaGrad 具有与随机梯度选取无关的遗憾界

$$R(T) \leq B(T) = \sum_{j=1}^{n}\sqrt{TM^2+\varepsilon}\left(\frac{d_j^2}{2\eta}+\eta\right), \tag{11}$$

其中 $d_j = \sup_{x,y\in D} |x_j - y_j|$. 再利用遗憾的性质证明 $\mathbf{E}(f(\bar{x}_T) - f(x^*)) \leq \frac{B(T)}{T}$ 即可。 □

考虑大规模优化问题 $\min_{x\in D} f(x) := \frac{1}{m}\sum_{i=1}^{m} f_i(x)$，其中 D 为紧凸集，诸 f_i 为凸函数。

定义 1 (遗憾)

设某随机梯度方法第 k 次迭代起始点为 x^k，所选取的随机（次）梯度为 $\nabla f_{a_i}(x^k)$，称

$$R(T) = \sum_{i=1}^{T} f_{a_i}(x^i) - \min_{x\in D}\sum_{i=1}^{T} f_{a_i}(x) \tag{12}$$

为该方法迭代 T 轮所产生的遗憾（*regret*）. 遗憾代表在线算法的总损失与离线算法的理论最小损失之间的差值。

当随机梯度一致有界时，我们有

- 取步长 $\eta_k = \eta/\sqrt{k}$ $(\eta > 0)$，此时 SGD 具有遗憾界 $R(T) \le B(T) = O(\sqrt{T})$；
- 对任意初始步长，AdaGrad 和 AdaGrad-Norm 具有遗憾界 $R(T) \le B(T) = O(\sqrt{T})$；
- AdaGrad 和 AdaGrad-Norm 的最优遗憾界为后见之明意义下最优遗憾界的 $\sqrt{2}$ 倍。

Ada-系列算法

受历史梯度累积影响，AdaGrad 步长递减趋于 0，导致迭代后期收敛缓慢，对新数据的学习能力较弱。

RMSProp：将记录梯度的二阶矩的 G^k 改为指数滑动平均的形式

$$\begin{aligned} G^k &= \rho G^{k-1} + (1-\rho)(\nabla f_{a_k}(x^k) \odot \nabla f_{a_k}(x^k)) \\ x^{k+1} &= x^k - \frac{\eta}{\sqrt{G^k + \varepsilon}} \odot \nabla f_{a_k}(x^k), \qquad k = 1, 2, \cdots, \end{aligned} \tag{13}$$

其中 $G^0 = \mathbf{0}$，$\rho \in (0,1)$ 为衰减参数，一般取 0.9.

Adam：在 RMSprop 的基础上引入动量，并进行偏差修正

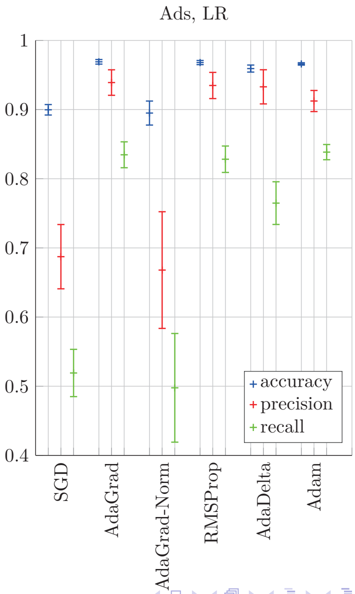
$$\begin{aligned} S^k &= \rho_1 S^{k-1} + (1-\rho_1)\nabla f_{a_k}(x^k), \\ G^k &= \rho_2 G^{k-1} + (1-\rho_2)\nabla f_{a_k}(x^k) \odot \nabla f_{a_k}(x^k), \\ \hat{S}^k &= \frac{S^k}{1-\rho_1^k}, \quad \hat{G}^k = \frac{G^k}{1-\rho_2^k}, \\ x^{k+1} &= x^k - \frac{\eta}{\sqrt{\hat{G}^k + \varepsilon}} \odot \hat{S}^k, \qquad k = 1, 2, \cdots. \end{aligned} \tag{14}$$

其中 $S^0 = G^0 = \mathbf{0}$，衰减参数 $\rho_1, \rho_2 \in (0,1)$，一般取 $\rho_1 = 0.9$，$\rho_2 = 0.999$.

数值实验

- 考虑对应了三类优化问题（光滑强凸、非光滑凸、光滑非凸）的三种模型（Logistic 回归（LR）、支持向量机（SVM）、多层感知机（MLP））；
- 数据集：Ads（稀疏）、Spambase（非稀疏）；
- 算法：SGD, AdaGrad, AdaGrad-Norm, RMSProp, AdaDelta 和 Adam；
- 主要结论：
  Ada-系列算法的整体表现较好，在稀疏数据集 Ads 上，差距更加明显。

右图是在稀疏数据集 Ads 上，用各种算法训练 LR 模型，迭代 100 轮后的结果。蓝色、红色、绿色线分别表示验证准确度、精确度、召回率。



总结与展望

SGD 与 Ada-系列算法各自的优缺点：

- SGD 超参数少、可解释性强、泛化能力好，但在病态问题中收敛慢、对步长敏感、难以逃脱鞍点；
- AdaGrad 适合稀疏数据、在病态问题下收敛快、无需手动调整步长，但其步长单减趋于 0，迭代后期收敛缓慢，对新数据的学习能力较弱；
- RMSProp, AdaDelta, Adam 等改进算法避免了步长过早衰减、对新数据学习能力较强、对不同模型和超参数的鲁棒性强，但迭代格式较复杂，缺乏收敛性的理论保证，并且在神经网络等模型中泛化能力可能较弱。

Ada-系列算法适合稀疏数据的关键因素：逐分量自适应步长。如果某个特征很少出现，那么在该特征对应的分量上，梯度的累积将很少，因此 Ada-系列算法能在特征出现时作出及时的反应（在该方向给予一个较大步长）。

论文关于 AdaGrad 收敛性刻画的不足之处：

- 在凸优化问题中，$O(1/\sqrt{T})$ 的收敛率与定义域直径有关；
- 在光滑非凸优化问题中，刻画收敛性的指标不如收敛率直接，并且 $O(\ln T/\sqrt{T})$ 的收敛速度与问题维数 n、数值稳定性小量 ε 有关。