

数学最优化

○○○

随机梯度方法

○○○○○

随机梯度之外

○○○○○○

随机梯度方法

许坚

jian.xu@szihi.cn

2020 年 10 月 26 日

数学最优化

○○○

随机梯度方法

○○○○○

随机梯度之外

○○○○○

数学最优化

随机梯度方法

随机梯度之外

数学最优化与 AI

- 二十一世纪以来，随着计算机算力、大量数据和理论知识的不断发展，AI 技术开始复苏。
- 以监督学习、无监督学习、强化学习为代表的机器学习算法逐渐成为 AI 领域的重要组成部分。
- 通俗地说，机器学习算法基于样本数据建立数学模型，以达到进行预测或决策的目的。具体来说，数学模型将学习问题重新表述为最小化关于样本数据的损失函数，而数学最优化为其提供方法。

数学最优化

机器学习中的数学优化可以分为三类

- 有限终止 (适用特定结构的模型)
- 迭代方法 (一阶方法、二阶方法)
- 启发式 (模拟退火、遗传、蚁群)

数学最优化

○○●

随机梯度方法

○○○○○

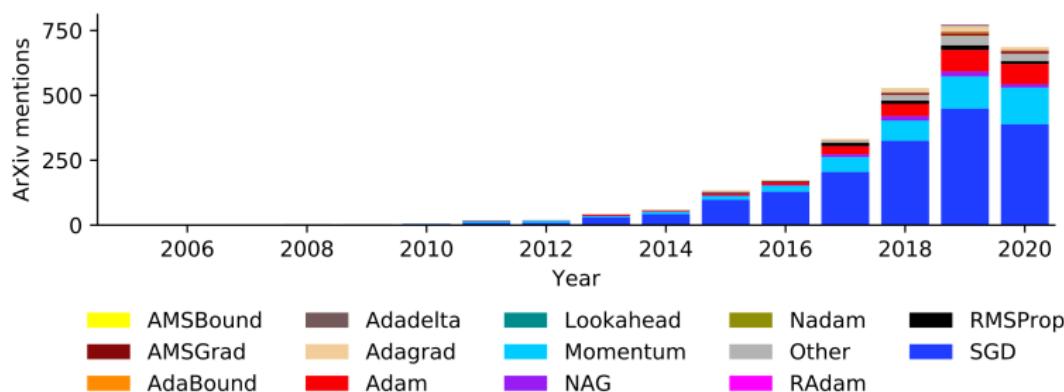
随机梯度之外

○○○○○

图解迭代方法

随机梯度方法

每年 ArXiv 标题和摘要提及随机梯度下降 (Stochastic Gradient Descent, SGD) 的次数。



随机梯度方法

随机梯度方法处理的问题一般有如下形式

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n f_i(w) = F(w)$$

其中 $f_i(w)$ 为可微函数。一般的，随机梯度算法的迭代格式为

$$w_{k+1} = w_k - t_k g(w_k, \xi_k)$$

这里 $g(w_k, \xi_k)$ 指第 k 步迭代中根据随机变量 ξ_k 随机抽取 n_k 个样本产生的方向， n_k 的值由不同的算法策略决定，当 n_k 恒为 1 时为最朴素的随机梯度方法。

收敛速率 (强凸)

- GD¹, 固定步长 $0 < \bar{\alpha} \leq 2/(c + L)$, $O(n \log(\frac{1}{\varepsilon}))$

$$F(w_k) - F_* \leq \frac{L}{2} \left(1 - \bar{\alpha} \frac{2cL}{c + L}\right)^{k-1} \|w_1 - w_*\|_2^2$$

- SGD², 固定步长 $0 < \bar{\alpha} \leq \frac{\mu}{LM_G}$

$$\mathbb{E}[F(w_k) - F_*] \leq \frac{\bar{\alpha}LM}{2c\mu} + (1 - \bar{\alpha}c\mu)^{k-1} \left(F(w_1) - F_* - \frac{\bar{\alpha}LM}{2c\mu} \right)$$

- SGD, $\alpha_k = \frac{\beta}{\gamma+k}$, $O(\frac{1}{\varepsilon})$

$$\mathbb{E}[F(w_k) - F_*] \leq \frac{\nu}{\gamma + k}$$

¹Beck, Amir. First-order methods in optimization. Society for Industrial and Applied Mathematics, 2017.

²Bottou, Léon, Frank E. Curtis, and Jorge Nocedal. "Optimization methods for large-scale machine learning." Siam Review 60.2 (2018): 223-311.

收敛速率 (非凸)

- GD, 固定步长 $0 < \bar{\alpha} \leq \frac{1}{L}$, $O(\frac{n}{\varepsilon^2})$

$$\frac{1}{K} \sum_{k=1}^K \|\nabla F(w_k)\|_2^2 \leq \frac{2(F(w_1) - F_{\inf})}{K\bar{\alpha}}$$

- SGD, 固定步长 $0 < \bar{\alpha} \leq \frac{\mu}{LM_G}$

$$\mathbb{E} \left[\frac{1}{K} \sum_{k=1}^K \|\nabla F(w_k)\|_2^2 \right] \leq \frac{\bar{\alpha} LM}{\mu} + \frac{2(F(w_1) - F_{\inf})}{K\mu\bar{\alpha}}$$

- SGD, $\sum_{k=1}^{\infty} \alpha_k = \infty$, $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$

$$\frac{\sum_{k=1}^K \alpha_k \mathbb{E} \left[\|\nabla F(w_k)\|_2^2 \right]}{\sum_{k=1}^K \alpha_k} \leq \frac{2(F(w_1) - F_{\inf})}{\mu \sum_{k=1}^K \alpha_k} + \frac{LM \sum_{k=1}^K \alpha_k^2}{\mu \sum_{k=1}^K \alpha_k}$$

收敛速率 (凸函数)

- GD, 固定步长 $0 < \bar{\alpha} \leq \frac{1}{L}$, $O(\frac{n}{\varepsilon})$

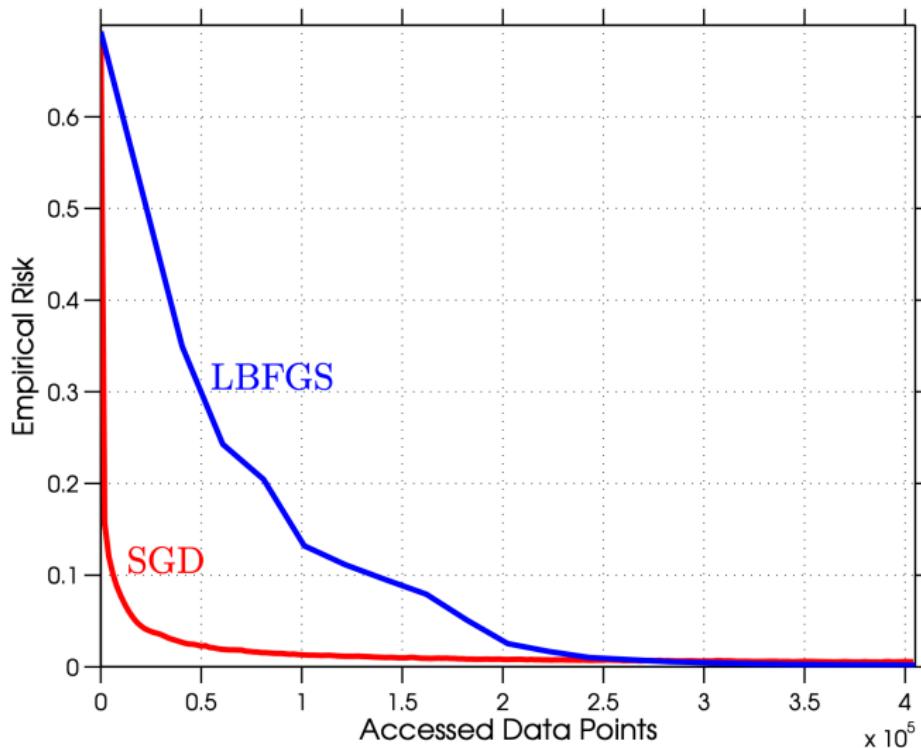
$$F(w_k) - F_* \leq \frac{\|w_1 - w_*\|_2^2}{2k\bar{\alpha}}$$

- GD, 任意步长策略, $O(\frac{n}{\sqrt{\varepsilon}})$

$$F(w_k) - F_* \geq \frac{3}{32} \frac{L \|w_1 - w_*\|_2^2}{k^2}$$

- SGD, ?

SGD 与 LBFGS



噪声减小类方法

这类方法尝试通过减小每一步迭代中由随机带来的方差由此来加快收敛速率，包括

- SVRG
- SAGA/SAG
- SARAH

且在强凸的假设下，我们有更小的收敛速率 $O(n' \log(\frac{1}{\varepsilon}))$, $n' < n$ 。

二阶方法

这类方法尝试利用二阶信息来加快收敛速率，包括

- Diagonal Scaling (RMSprop、**Adam**)
- quasi-Newton
- Gauss-Newton
- Hessian-free Newton
- Natural gradient(quasi natural gradient、batch normalization)

RMSprop、Adam 不能保证收敛。

Adam

$$m_k = \beta_1 m_{k-1} + (1 - \beta_1) g(\omega_k, \xi_k)$$

$$v_k = \beta_2 v_{k-1} + (1 - \beta_2) g(\omega_k, \xi_k) \odot g(\omega_k, \xi_k)$$

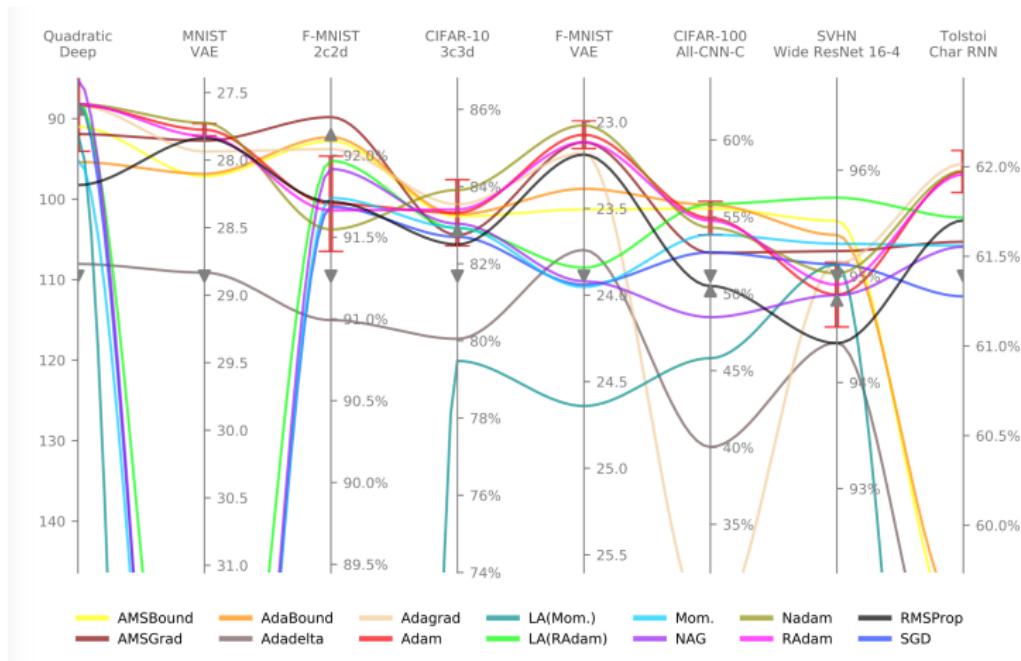
$$\hat{m}_k = m_k / (1 - \beta_1^k)$$

$$\hat{v}_k = v_k / (1 - \beta_2^k)$$

$$\omega_{k+1} = \omega_k - t_k \hat{m}_k / (\sqrt{\hat{v}_k} + \epsilon)$$

其中 $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$

优化算法对比



优化算法选择建议

优化算法的选择应当择优选择，图像领域的建议：

- RAdam(与 Adam 持平的表现、更智能的步长策略)
- Lookahead-RAdam(默认参数就能在残差网络表现很好)