优化理论与算法

第八章

随机梯度下降法

郭加熠|助理教授

目录

随机梯度

收敛性分析

方差缩减方法

自适应步长方法

讲　　员

郭加熠，江波，刘慧康

最小化求和

$$\min \quad \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

- 最小二乘法： $f_i(x) = (a_i^T x - b_i)^2$
- 逻辑回归： $f_i(x) = -\log(1 + \exp(-b_i a_i^T x))$
- 最大似然估计： $f_i(x)$ 是给定参数 x 时观测 i 的对数似然的负值
- 机器学习： f_i 是模型 x 在样本 i 上的误差

最小化求和

求解

$$\min \quad \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

困境：

- 直观上，数据越多越容易解决该问题（有效求解）
- 但算法的复杂性随 n 的增加而增加！

目标：找到在给定更多数据时表现更好（或至少不更差）的算法

最小化期望

求解

$$\min \quad \mathbf{E}f(x) = \mathbf{E}_\omega f(x;\omega)$$

随机损失函数为 f，也可以表示为函数 $f(\cdot;\omega)$ 是关于随机变量 ω 的。

示例：观测值 $\omega = (a, b)$ 是随机的

- 最小二乘法：$f(x;\omega) = (a^T x - b)^2$
- 逻辑回归：$f(x;\omega) = -\log(1+\exp(-ba^T x))$
- 最大似然估计：$f(x;\omega)$ 是给定参数 x 的观测值 ω 的负对数似然
- 机器学习：$f(x;\omega)$ 是模型 x 在示例 ω 上的误差

随机梯度下降法

考虑函数之和：

$$\min_x \frac{1}{n}\sum_{i=1}^n f_i(x)$$

将梯度下降应用于该问题会重复以下步骤：

$$x^{(k)} = x^{(k-1)} - t_k \cdot \frac{1}{n}\sum_{i=1}^n \nabla f_i(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

相比之下，**随机梯度下降**（或增量梯度下降）重复以下步骤：

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f_{i_k}(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

其中 $i_k \in \{1, \ldots, n\}$ 是在迭代 k 时选择的某个索引

如何选取随机梯度

- 通常，我们随机（均匀）选择 $i_k \in \{1, \ldots, n\}$
- 还有另一种常见情况：小批量 (min-batch) 随机梯度下降，其中我们选择一个随机子集 $I_k \subset \{1, \ldots, n\}$，大小为 $b \ll n$，并根据以下方式更新：

$$x^{(k)} = x^{(k-1)} - t_k \cdot \frac{1}{b} \sum_{i \in I_k} \nabla f_i(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

- 在这两种情况下，我们通过一个有噪声的估计来近似完整梯度，并且我们的有噪声估计是**无偏**的：

$$\mathbb{E}[\nabla f_{i_k}(x)] = \nabla f(x)$$

$$\mathbb{E}\left[\frac{1}{b} \sum_{i \in I_k} \nabla f_i(x)\right] = \nabla f(x)$$

- 批量方法可以将方差降低 $\frac{1}{b}$ 倍，但计算成本也高出 b 倍！

示例：正则化逻辑回归

给定标签 $y_i \in \{0,1\}$，特征 $x_i \in \mathbb{R}^p$，$i = 1,\ldots,n$。考虑带有岭正则化的逻辑回归：

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}\left(-y_i x_i^T\beta + \log(1+e^{x_i^T\beta})\right) + \frac{\lambda}{2}\|\beta\|_2^2$$

将准则写成：

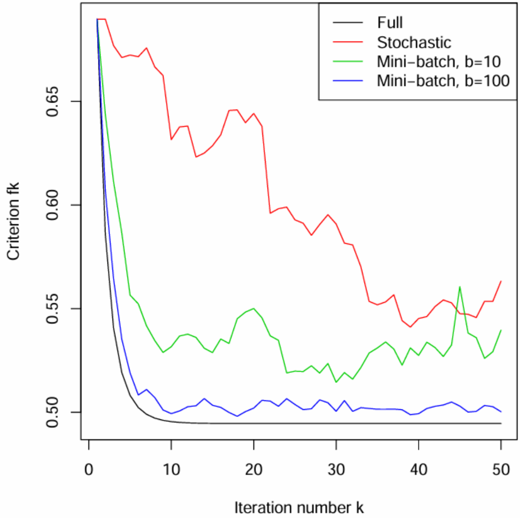
$$f(\beta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\beta), \quad f_i(\beta) = -y_i x_i^T\beta + \log(1+e^{x_i^T\beta}) + \frac{\lambda}{2}\|\beta\|_2^2$$

梯度计算 $\nabla f(\beta)$ 在 n 较小时是可以处理的，但在 n 极大时则不然：

- 一次（完整）批量更新的成本为 $O(np)$
- 一次随机更新的成本为 $O(p)$
- 一次小批量更新的成本为 $O(bp)$

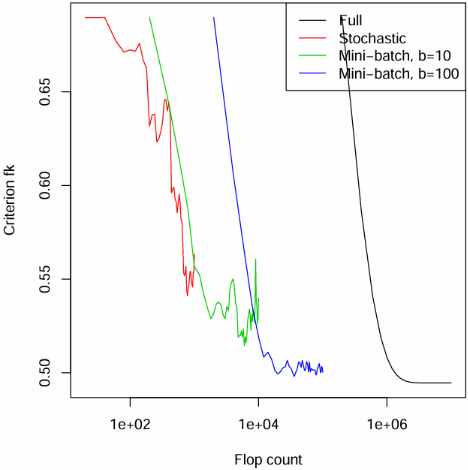
示例：正则化逻辑回归

示例中 $n = 10,000$，$p = 20$，所有方法均采用固定步长（递减步长给出的结果大致相似）：



发生了什么？

迭代随着批量大小 b 的增加而取得更好的进展。但现在让我们以浮点运算次数（flops）为基准：



收敛速率

回顾一下，在合适的步长下，当 f 是凸函数并且具有 Lipschitz 梯度时，全梯度（FG）下降满足：

$$f(x^{(k)}) - f^* = O(1/k)$$

对于随机梯度（SG）下降呢？在递减的步长下，当 f 是凸函数（加上其他条件）时：

$$\mathbb{E}[f(x^{(k)})] - f^* = O(1/\sqrt{k})$$

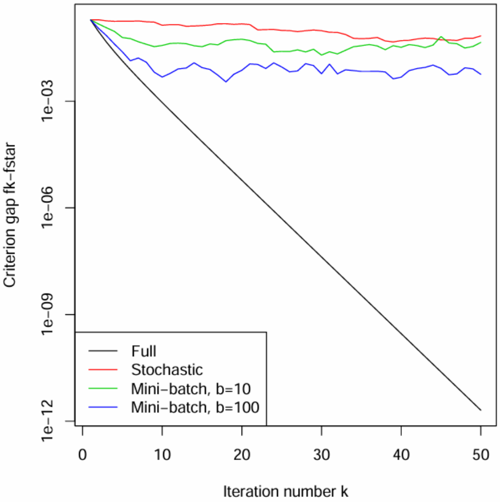
最后，小批量随机梯度下降呢？同样，在递减的步长下，对于凸函数 f (加上其他条件)：

$$\mathbb{E}[f(x^{(k)})] - f^* = O(1/\sqrt{bk} + 1/k)$$

但这里的每次迭代成本是 b 倍的。而且对于小的 b，在浮点运算方面，速率是相同的

回到岭逻辑回归示例

通过查看次优性差距（在对数尺度上），我们观察如下：



目录

随机梯度

收敛性分析

方差缩减方法

自适应步长方法

讲　　　　员

郭加熠，江波，刘慧康

随机次梯度

随机向量 $\tilde{g} \in \mathbf{R}^n$ 是 $f : \mathbf{R}^n \to \mathbf{R}^n$ 在 $x \in \mathbf{R}^n$ 处的**随机次梯度**，如果：

$$\mathbb{E}[\tilde{g}] \in \partial f(x)$$

即，对于所有 z：

$$f(z) \geq f(x) + (\mathbb{E}[\tilde{g}])^T (z - x)$$

- 等价地，$\tilde{g} = g + v$，其中 $g \in \partial f(x)$，$\mathbb{E}[v] = 0$
- v 可以表示计算 g 时的误差、测量噪声、蒙特卡罗抽样误差等。

随机次梯度方法

随机次梯度方法是使用随机次梯度的次梯度方法：

$$x^{(k+1)} = x^{(k)} - t_k \tilde{g}^{(k)}$$

- $x^{(k)}$ 是第 k 次迭代点
- $\tilde{g}^{(k)}$ 是（凸函数）f 在 $x^{(k)}$ 处的任何无偏次梯度，即：

$$\mathbb{E}[\tilde{g}^{(k)} | x^{(k)}] = g^{(k)} \in \partial f(x^{(k)})$$

- $t_k > 0$ 是第 k 次步长

假设条件

- $f^* = \inf_x f(x) > -\infty$，且存在 $f(x^*) = f^*$
- $\mathbb{E}[\|g^{(k)}\|_2^2] \leq G^2$ 对所有 k
- $\mathbb{E}[\|x^{(1)} - x^*\|_2^2] \leq R^2$（这里可以取等号）
- 步长是平方可求和的，但不是可求和的：

$$t_k \geq 0, \quad \sum_{k=1}^{\infty} t_k^2 = \|t\|_2^2 < \infty, \quad \sum_{k=1}^{\infty} t_k = \infty$$

这些假设条件比实际需要的更强，只是为了简化证明。

符号表示

与之前一样，定义：

- 最优目标值 $f_{\text{best}}^{(k)} := \min\{f(x^{(1)}), \ldots, f(x^{(k)})\}$
- 平均迭代 $\bar{x}^{(k)} = \frac{1}{k}\sum_{i=1}^{k} x^{(i)}$

根据最优性：

$$f_{\text{best}}^{(k)} = \min\{f(x^{(1)}), \ldots, f(x^{(k)})\} \leq \frac{1}{k}\sum_{i=1}^{k} f(x^{(i)})$$

并且根据凸性：

$$f(\bar{x}^{(k)}) \leq \frac{1}{k}\sum_{i=1}^{k} f(x^{(i)})$$

收敛结果

- 期望收敛：
$$\lim_{k\to\infty} \mathbf{E} f_{\text{best}}^{(k)} = f^*$$
- 概率收敛：对于任何 $\epsilon > 0$，
$$\lim_{k\to\infty} \text{prob}(f_{\text{best}}^{(k)} \geq f^* + \epsilon) = 0$$
- 几乎必然收敛：
$$\lim_{k\to\infty} f_{\text{best}}^{(k)} = f^*$$
（我们不会展示这个）

收敛证明

关键误差测度：到最优集的期望平方欧几里得距离

$$\mathbf{E}\left(\left\|x^{(k+1)}-x^{*}\right\|_{2}^{2} \mid x^{(k)}\right)=\mathbf{E}\left(\left\|x^{(k)}-t_{k}\tilde{g}^{(k)}-x^{*}\right\|_{2}^{2} \mid x^{(k)}\right)$$

$$=\left\|x^{(k)}-x^{*}\right\|_{2}^{2}-2t_{k}\mathbf{E}\left(\tilde{g}^{(k)T}\left(x^{(k)}-x^{*}\right) \mid x^{(k)}\right)+t_{k}^{2}\mathbf{E}\left(\left\|\tilde{g}^{(k)}\right\|_{2}^{2} \mid x^{(k)}\right)$$

$$=\left\|x^{(k)}-x^{*}\right\|_{2}^{2}-2t_{k}\mathbf{E}(\tilde{g}^{(k)} \mid x^{(k)})^{T}\left(x^{(k)}-x^{*}\right)+t_{k}^{2}\mathbf{E}\left(\left\|\tilde{g}^{(k)}\right\|_{2}^{2} \mid x^{(k)}\right)$$

$$\leq\left\|x^{(k)}-x^{*}\right\|_{2}^{2}-2t_{k}(f(x^{(k)})-f^{*})+t_{k}^{2}\mathbf{E}\left(\left\|\tilde{g}^{(k)}\right\|_{2}^{2} \mid x^{(k)}\right)$$

使用 $\mathbf{E}(\tilde{g}^{(k)} \mid x^{(k)}) \in \partial f(x^{(k)})$

收敛证明（续）

现在取全期望:

$$\mathbf{E}\left\|x^{(k+1)}-x^*\right\|_2^2 \leq \mathbf{E}\left\|x^{(k)}-x^*\right\|_2^2 - 2t_k(\mathbf{E}f(x^{(k)})-f^*) + t_k^2\mathbf{E}\left\|\tilde{g}^{(k)}\right\|_2^2$$

递归应用，并使用 $\mathbf{E}\left\|\tilde{g}^{(k)}\right\|_2^2 \leq G^2$ 得到:

$$\mathbf{E}\left\|x^{(k+1)}-x^*\right\|_2^2 \leq \mathbf{E}\left\|x^{(1)}-x^*\right\|_2^2 - 2\sum_{i=1}^{k} t_i(\mathbf{E}f(x^{(i)})-f^*) + G^2\sum_{i=1}^{k} t_i^2$$

因此:

$$\min_{i=1,\ldots,k}\left(\mathbf{E}f(x^{(i)})-f^*\right) \leq \frac{R^2+G^2\|t\|_2^2}{2\sum_{i=1}^k t_i}$$

如果 $t_i = t$ 为常数，也得到: $\mathbf{E}f(\bar{x}^{(i)})-f^* \leq \frac{R^2+G^2\|t\|_2^2}{2tk}$

收敛结果（续）

- 我们已经得出 $\min_{i=1,\ldots,k} \mathbf{E} f(x^{(i)}) \to f^*$
- 根据 Jensen 不等式和最小值的凹性：

$$\mathbf{E} f_{\text{best}}^{(k)} = \mathbf{E} \min_{i=1,\ldots,k} f(x^{(i)}) \leq \min_{i=1,\ldots,k} \mathbf{E} f(x^{(i)})$$

  因此 $\mathbf{E} f_{\text{best}}^{(k)} \to f^*$（期望收敛）
- 马尔可夫不等式：对于 $\epsilon > 0$

$$\text{prob}(f_{\text{best}}^{(k)} - f^* \geq \epsilon) \leq \frac{\mathbf{E}(f_{\text{best}}^{(k)} - f^*)}{\epsilon}$$

  右边趋近于零，因此我们得到概率收敛

收敛速率（续）

回顾一下，在合适的步长下，当 f 是强凸函数并且具有 Lipschitz 梯度时，梯度下降满足：

$$f(x^{(k)}) - f^* = O(\rho^k)$$

其中 $\rho < 1$。但在递减的步长下，当 f 是强凸函数（加上其他条件）时，随机梯度下降给出：

$$\mathbb{E}[f(x^{(k)})] - f^* = O(1/k)$$

因此，随机方法在强凸性下不享有梯度下降的线性收敛速率

- 有一段时间，这被认为是不可避免的，因为 Nemirovski 和其他人已经建立了匹配的下界。
- 实际上，这些下界讨论适用于随机优化问题:

$$\min f(x) = \int F(x, \xi) d\xi$$

- 对于有限和形式（$\min \frac{1}{n} \sum_{i=1}^{n} f_i(x)$），我们能否做得更好吗?

目录

讲　　　员

郭加熠，江波，刘慧康

随机梯度：偏差

对于求解 $\min_x f(x)$，随机梯度实际上是一类使用以下迭代的算法：

$$x^{(k)} = x^{(k-1)} - \eta_k g(x^{(k-1)}; \xi_k),$$

其中 $g(x^{(k-1)}, \xi_k)$ 是目标函数 $f(x)$ 在 $x^{(k-1)}$ 处的**随机梯度**。

偏差： 随机梯度的偏差定义为：

$$\mathrm{bias}(g(x^{(k-1)}; \xi_k)) := \mathbb{E}_{\xi_k}\left(g(x^{(k-1)}; \xi_k)\right) - \nabla f(x^{(k-1)}).$$

无偏差： 当 $\mathbb{E}_{\xi_k}\left(g(x^{(k-1)}; \xi_k)\right) = \nabla f(x^{(k-1)})$ 时，随机梯度被认为是无偏差的。（例如，到目前为止讨论的随机梯度方案）

有偏差： 我们可能也对有偏差的估计器感兴趣，但偏差很小，使得 $\mathbb{E}_{\xi_k}\left(g(x^{(k-1)}; \xi_k)\right) \approx \nabla f(x^{(k-1)})$。

随机梯度：方差

方差：除了小（或零）偏差外，我们还希望估计量的方差较小：

$$\begin{aligned}\text{variance}(g(x^{(k-1)},\xi_k)) &:= \mathbb{E}_{\xi_k}\left(g(x^{(k-1)},\xi_k)-\mathbb{E}_{\xi_k}\left(g(x^{(k-1)},\xi_k)\right)\right)^2\\ &\le \mathbb{E}_{\xi_k}\left(g(x^{(k-1)},\xi_k)\right)^2.\end{aligned}$$

我们目前看到的随机梯度方案的缺点是其方差很大，并且特别是不会随着迭代次数的增加而衰减到零。

大致来说：由于上述原因，我们不得不将步长 η_k 衰减到零，这意味着我们不能采取“大”步长，因此收敛速率很慢。

我们能否使方差很小，并且随着迭代次数的增加而衰减到零？

方差缩减

考虑参数 θ 的估计量 X。对于无偏差的估计量，有 $\mathbb{E}(X) = \theta$。

现在考虑以下修改后的估计量：$Z := X - Y$，且 $\mathbb{E}(Y) \approx 0$。然后 Z 的偏差也接近零，因为：

$$\mathbb{E}(Z) = \mathbb{E}(X) - \mathbb{E}(Y) \approx \theta.$$

估计量 X 的方差情况如何呢?

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X, Y).$$

如果 Y 与 X 高度相关，Var$(X - Y)$ 可能比 Var(X) 小得多。

因此，给定任何估计器 X，如果我们能构造一个 Y，它（a）期望接近零，且（b）与 X 高度相关，我们就可以降低其方差。

这是 SAG、SAGA、SVRG、SDCA 等方差缩小方法所遵循的思想。

随机平均梯度

随机平均梯度（Stochastic Average Gradient 或 SAG）（Schmidt, Le Roux, Bach 2013）是随机优化中的突破性方法。其思想相当简单：

- 维护一个表格，包含 f_i 的梯度 g_i，$i = 1, \ldots, n$
- 初始化 $x^{(0)}$，设 $g_i^{(0)} = x^{(0)}$，$i = 1, \ldots, n$
- 在迭代 $k = 1, 2, 3, \ldots$ 过程中，随机选择 $i_k \in \{1, \ldots, n\}$，然后令

  $$g_{i_k}^{(k)} = \nabla f_i(x^{(k-1)}) \quad \text{（}f_i\text{ 的最新梯度）}$$

  将所有其他 $g_i^{(k)} = g_i^{(k-1)}$，$i \neq i_k$，即这些保持不变
- 更新

  $$x^{(k)} = x^{(k-1)} - t_k \cdot \frac{1}{n} \sum_{i=1}^{n} g_i^{(k)}$$

分析

- SAG 的关键是允许每个 f_i 在每一步中更新梯度估计的一部分
- 这个基本思想可以追溯到增量聚合梯度（Blatt, Hero, Gauchman, 2006）
- SAG 梯度估计**不再是无偏差的，但它们的方差大大降低**
- 平均所有这些梯度是否昂贵?（特别是如果 n 很大?）
  实际上，SAG 基本上与随机梯度下降**一样高效**!

$$x^{(k)} = x^{(k-1)} - t_k \cdot \left(\frac{g_{i_k}^{(k)}}{n} - \frac{g_{i_k}^{(k-1)}}{n} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} g_i^{(k-1)}}_{\text{旧表格平均}} \right)$$

SAG 方差缩减

SAG 中的随机梯度：

$$\underbrace{g_{i_k}^{(k)}}_{X} - \underbrace{(g_{i_k}^{(k-1)} - \sum_{i=1}^{n} g_i^{(k-1)})}_{Y}.$$

可以看出 $\mathbb{E}(X) = \nabla f(x^{(k)})$，但 $\mathbb{E}(Y) \neq 0$，这是个**有偏差**的估计量。

但我们确实有 Y 好像与 X 相关（符合方差缩减思想）。

特别是，当 $k \to \infty$ 时，$X - Y \to 0$。这是因为 $x^{(k-1)}$ 和 $x^{(k)}$ 收敛到 $\bar{x}$，而 $X - Y$ 前两个项之间的差收敛到零，最后一项收敛到最优解的梯度，即也收敛到零。

因此，总体估计量的 ℓ_2 范数（及其方差）衰减到零。

SAG 收敛分析

假设 $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$，其中每个 f_i 是可微的，且 ∇f_i 是 Lipschitz 连续的，常数为 L。

记 $\bar{x}^{(k)} = \frac{1}{k}\sum_{\ell=0}^{k-1} x^{(\ell)}$，即经过 $k-1$ 步后的平均迭代。

定理（Schmidt，Le Roux，Bach） SAG 算法使用固定步长 $t = 1/(16L)$，以及初始化

$$g_i^{(0)} = \nabla f_i(x^{(0)}) - \nabla f(x^{(0)}), \quad i = 1, \ldots, n,$$

满足

$$\mathbb{E}[f(\bar{x}^{(k)})] - f^* \leq \frac{48n}{k}\left(f(x^{(0)}) - f^*\right) + \frac{128L}{k}\|x^{(0)} - x^*\|_2^2,$$

其中，期望是对每次迭代中随机选择索引取的。

分析

- 结果以平均迭代 $\bar{x}^{(k)}$ 表示，但也可以证明对目前为止看到的最佳迭代 $x_{\text{best}}^{(k)}$ 也成立
- SAG 是 $O(1/k)$ 收敛速率。对比 FG 的 $O(1/k)$ 和 SG 的 $O(1/\sqrt{k})$
- 但，**常数项不同**！经过 k 步后：

$$\begin{aligned}
\text{SAG}: &\quad \frac{48n}{k}\left(f(x^{(0)}) - f^*\right) + \frac{128L}{k}\|x^{(0)} - x^*\|_2^2 \\
\text{FG}: &\quad \frac{L}{2k}\|x^{(0)} - x^*\|_2^2 \\
\text{SG*}: &\quad \frac{L\sqrt{5}}{\sqrt{2k}}\|x^{(0)} - x^*\|_2 \quad \text{(进行了适当松弛)}
\end{aligned}$$

- 可以看出，SAG 的第一项受到 n 的影响；
- 建议使用更好的初始化使 $f(x^{(0)}) - f^*$ 很小（例如，使用 n 次 SG 迭代后的结果）

强凸性下的收敛分析

进一步假设每个 f_i 是强凸的，参数为 m。

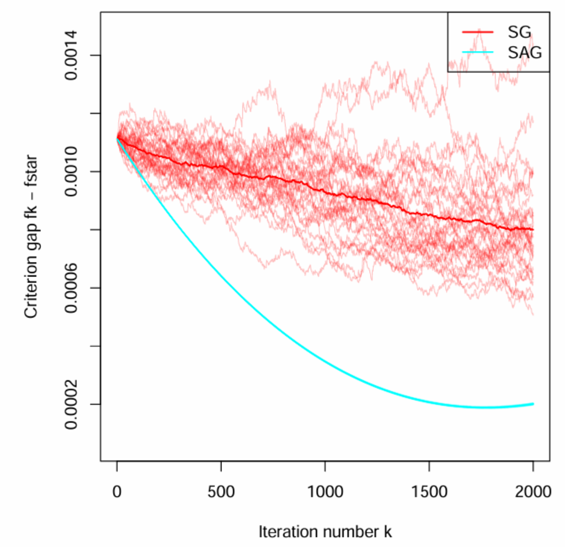
定理（Schmidt, Le Roux, Bach） SAG，使用步长 $t = 1/(16L)$ 和与之前相同的初始化，有如下式子成立：

$$\mathbb{E}[f(x^{(k)})] - f^* \leq \left(1 - \min\left\{\frac{m}{16L}, \frac{1}{8n}\right\}\right)^k \cdot \left(\frac{3}{2}\left(f(x^{(0)}) - f^*\right) + \frac{4L}{n}\|x^{(0)} - x^*\|_2^2\right)$$

- 这是 SAG 的**线性收敛速率** $O(\rho^k)$。对比 FG 的 $O(\rho^k)$ 和 SG 的 $O(1/k)$
- 这些结果的证明并不简单：15 页！

回到岭逻辑回归示例：SG 与 SAG

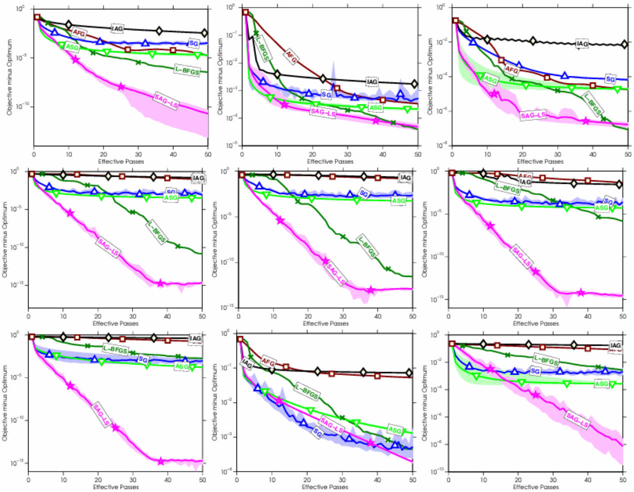
在对这些随机算法进行 30 次重新运行后，SG 与 SAG 的表现如下：



注释

- SAG 表现良好，但并非开箱即用；需要特定的设置
- 使用一次完整的 SG 循环（一次数据遍历）来获取 $\beta^{(0)}$，然后从 $\beta^{(0)}$ 开始同时启动 SG 和 SAG。这种热启动帮助很大
- SAG 初始化为 $g_i^{(0)} = \nabla f_i(\beta^{(0)})$，$i = 1, \ldots, n$，在初始 SG 循环期间计算。
- 调整 SAG 的固定步长非常棘手；现在手工调整到在发散前尽可能大的值

每个图表都是不同的问题设置



SAGA

SAGA（Defazio, Bach, Lacoste-Julien, 2014）是另一种最近的随机方法，与 SAG 精神相似。其思想同样简单：

- 维护一个表格，包含 f_i 的梯度 g_i，$i = 1, \ldots, n$
- 初始化 $x^{(0)}$，并 $g_i^{(0)} = x^{(0)}$，$i = 1, \ldots, n$
- 在步骤 $k = 1, 2, 3, \ldots$ 中，随机选择 $i_k \in \{1, \ldots, n\}$，然后令

  $$g_{i_k}^{(k)} = \nabla f_i(x^{(k-1)}) \quad (f_i \text{ 的最新梯度})$$

  将所有其他 $g_i^{(k)} = g_i^{(k-1)}$，$i \neq i_k$，即这些保持不变
- 更新

  $$x^{(k)} = x^{(k-1)} - t_k \cdot \left(g_{i_k}^{(k)} - g_{i_k}^{(k-1)} + \frac{1}{n} \sum_{i=1}^{n} g_i^{(k-1)} \right)$$

注释

- SAGA 梯度估计 $g_{i_k}^{(k)} - g_{i_k}^{(k-1)} + \frac{1}{n}\sum_{i=1}^{n} g_i^{(k-1)}$
- 对比 SAG 梯度估计 $\frac{1}{n}g_{i_k}^{(k)} - \frac{1}{n}g_{i_k}^{(k-1)} + \frac{1}{n}\sum_{i=1}^{n} g_i^{(k-1)}$
- 回顾一下，SAG 估计是有偏差的；值得注意的是，SAGA 估计是无偏差的！

SAGA 方差缩减

SAGA 中的随机梯度:

$$\underbrace{g_{i_k}^{(k)}}_{X} - \underbrace{\left(g_{i_k}^{(k-1)} - \frac{1}{n}\sum_{i=1}^{n} g_i^{(k-1)}\right)}_{Y}.$$

可以看出 $\mathbb{E}(X) = \nabla f(x^{(k)})$,并且 $\mathbb{E}(Y) = 0$,因此我们有一个无偏差的估计器。

此外,我们有 Y 似乎与 X 相关(符合方差缩减思想)

同时,$x^{(k-1)}$ 和 $x^{(k)}$ 收敛到 $\bar{x}$ 得:当 $k \to \infty$ 时,有 $X - Y \to 0$ 成立(前两个项之间的差收敛到零,最后一项收敛到最优解的梯度,即也收敛到零)。

因此,总体估计器的 ℓ_2 范数(及其方差)衰减到零。

SAGA 方差缩减（续）

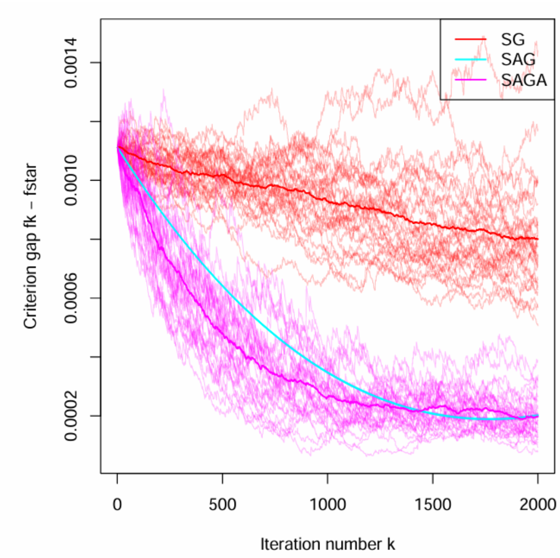
- SAGA 基本上匹配了 SAG 的强收敛速率（对于 Lipschitz 梯度和强凸情况），但这里的证明要简单得多
- SAGA 的另一个优势是它可以扩展到复合问题，形式如下

$$\min_x \frac{1}{n}\sum_{i=1}^{n} f_i(x) + h(x)$$

其中每个 f_i 是光滑且凸的，而 h 是凸的但非光滑的，但具有已知的 prox。更新现在变为

$$x^{(k)} = \mathrm{prox}_{h,t_k}\left(x^{(k-1)} - t_k \cdot \left(g_{i_k}^{(k)} - g_{i_k}^{(k-1)} + \frac{1}{n}\sum_{i=1}^{n} g_i^{(k-1)}\right)\right)$$

回到岭逻辑回归示例，现在加入 SAGA：



注释

- SAGA 表现良好，但同样需要特定的设置
- 与之前一样，使用一次完整的 SG 循环（一次数据遍历）来获取 $\beta^{(0)}$，然后从 $\beta^{(0)}$ 开始同时启动 SG、SAG 和 SAGA。这种热启动帮助很大
- SAGA 初始化为 $g_i^{(0)} = \nabla f_i(\beta^{(0)})$，$i = 1, \ldots, n$，在初始 SG 循环期间计算。对这些梯度进行中心化要差得多（将它们初始化为 0 也同样差）
- 调整 SAGA 的固定步长很好；似乎与调整 SG 的步长相当，并且比调整 SAG 的步长更稳健
- 有趣的是，SAGA 的准则曲线看起来像 SG 曲线（实现是锯齿形且高度可变的）；SAG 看起来非常不同，这确实佐证了其更新的方差要小得多

目录

讲　　　员

郭加熠，江波，刘慧康

自适应步长

其他自适应步长的想法:

- AdaGrad
- AdaDelta (= AdaGrad with stepsize $\nrightarrow 0$)
- RMSProp (= AdaGrad with stepsize $\nrightarrow 0$)
- Adam (= AdaGrad + momentum)
- ...

比较: http://imgur.com/a/Hqolp
这些主要用于深度学习，因此网上有很多信息!

AdaGrad

求解（f_i 不一定是光滑的）

$$\min \quad f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$$

算法 AdaGrad ([Duchi Singer Hazan, 2010])

输入：起始点 $x^{(0)} \in \mathrm{dom}\ f$，步长 t，小常数 $\delta > 0$

for $k = 0, 1, 2, \dots$ **do**

计算随机次梯度 $g^{(k)} \in \partial f_i(x^{(k)})$，

- 设 $H_k = \frac{1}{t}\mathrm{diag}(\left[\delta + \sqrt{\sum_{j=1}^{k}[g_1^{(j)}]^2}, \delta + \sqrt{\sum_{j=1}^{k}[g_2^{(j)}]^2}, ..., \delta + \sqrt{\sum_{j=1}^{k}[g_n^{(j)}]^2}\right])$

更新迭代 $x^{(k+1)} := x^{(k)} - H_k^{-1} g^{(k)}$

end for

AdaGrad - 动机

- 对于固定的 $H_k = H$，我们有估计：

$$f_{\text{best}}^{(k)} - f^* \leq \frac{R^2 + \sum_{i=1}^k \|g^{(i)}\|_{H^{-1}}^2}{2k}$$

- 思想：选择 $H_k \succ 0$ 以最小化这一估计：

$$H_k = \arg\min_H \sum_{i=1}^k \|g^{(i)}\|_{H^{-1}}^2$$

  约束 $\text{trace}(H)$ 为确定常数。

- 最优解 $H_k = \frac{1}{t}\text{diag}\left(\sqrt{\sum_{j=1}^k [g_1^{(j)}]^2}, \ldots, \sqrt{\sum_{j=1}^k [g_n^{(j)}]^2}\right)$
- 直观理解：根据历史步长自适应调整步长

AdaGrad - 收敛性

- 定义 $H_i = \frac{1}{t}\text{diag}(h_i)$，由于 $h_i \geq h_{i-1}$ ，AdaGrad 收敛为：

$$f_{\text{best}}^k - f^* \leq \frac{\sum_{i=1}^k \|g^{(i)}\|_{H_i^{-1}}^2}{2k} + \frac{R^2 \frac{1}{t}\|h_k\|_1}{2k}$$

- 可以证明：

$$\sum_{i=1}^k \|g^{(i)}\|_{H_i^{-1}}^2 \leq 2t\|h_k\|_1$$

从而推出：

$$f_{\text{best}}^k - f^* \leq \frac{(2t + \frac{R^2}{t})\|h_k\|_1}{2k}$$

- 如果 $\|g\|_\infty \leq G$，可以证明：

$$\|h_k\|_1 \leq n(\delta + G\sqrt{k})$$

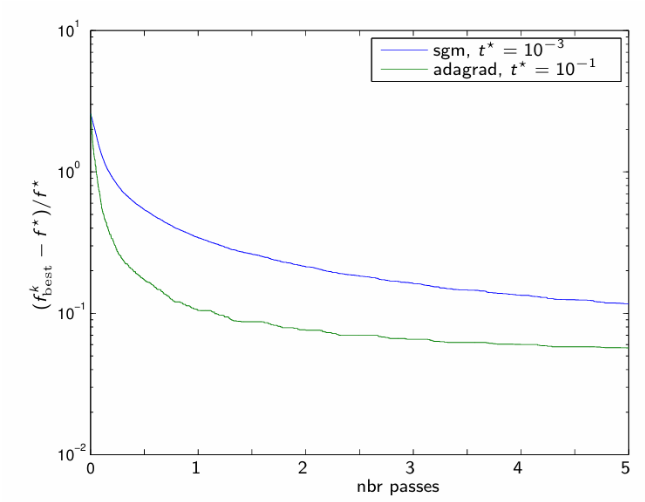
对于 $R < \infty$，这意味着 $f_{\text{best}}^k - f^* \to 0$ 对于任何 $t > 0$

示例

分类问题：

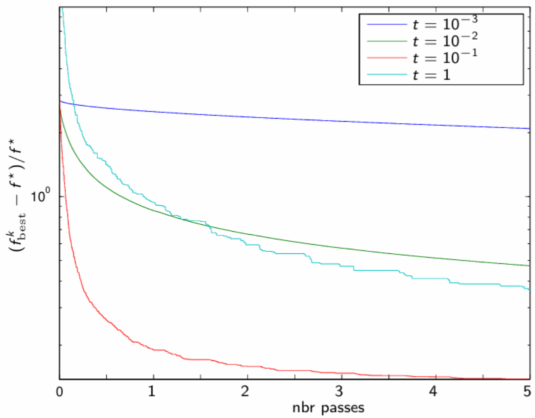
- 数据：$\{a_i, b_i\}$, $i = 1, \ldots, 50000$
  - $a_i \in \mathbf{R}^{1000}$
  - $b \in \{-1, 1\}$
  - 数据创建时相对于 $w = \mathbf{1}$, $v = 0$ 的误分类率为 5
- 目标：找到分类器 $w \in \mathbf{R}^{1000}$ 和 $v \in \mathbf{R}$ 使得：
  - $a_i^T w + v > 1$ 如果 $b = 1$
  - $a_i^T w + v < 1$ 如果 $b = -1$
- 优化方法：
  - 最小化铰链损失：$\sum_i \max(0, 1 - b_i(a_i^T w + v))$
  - 随机均匀选择示例，针对该示例进行次梯度步长更新

最佳次梯度方法与最佳 AdaGrad



通常最佳 AdaGrad 表现优于最佳次梯度方法

不同步长 t 的 AdaGrad：



对步长选择敏感（如同标准次梯度方法）

改进：AdaDelta, RMSProp, ADAM, ...

算法配方

组合算法配方：

- 随机（如 SGD）
- 方差缩减（如 SAGA）
- 自适应（如 AdaGrad）
- 加速（如 Nesterov/动量方法）
- 在线（如 SGD）
- 对偶（如 dual proximal gradient）
- 坐标（如 coordinate descent)）
- 异步（如 Hogwild）
- 分布式（如 Chambolle-Pock）