Стохастические задачи оптимизации

Мальковский Н. В.

Санкт-Петербургский академический университет

Проблема традиционных задач

Математическая постановка задач оптимизации подразумевает возможность **точно** измерять

- Значения целевой функции f в любых точках множества $\mathcal{D}$.
- Значения градиента целевой функции f.
- Значения гессиана целевой функции f.

Что делать, если можно померить только значения функции f?

Что делать, если можно померить только неточные значения функции f?

Стохастический градиент

Определение

Стохастическим градиентом выпуклой функции $f : \mathcal{D} \subset \mathbb{R}^n \to \mathbb{R}$ *в точке* $x \in \mathcal{D}$ *называется случайная величина* $g : \Omega \to \mathbb{R}^n$ *такая, что*

$$Eg = \int_\Omega g \; dp(g) = \nabla f(x)$$

Стохастический градиентный спуск

Пусть $\alpha_k > 0$, g_k – стохастический градиент f в точке x_k,
стохастическим градиентным спуском называется следующая процедура

$$x_{k+1} = x_k - \alpha_k g_k \tag{1}$$

Общий анализ сходимости

Теорема

Пусть f сильно выпукла с константой m, $\alpha_k = \frac{\beta}{k+1}$, $\beta > \frac{1}{2m}$, g_k – стохастический градиент f в точке x_k, $||g_k|| \leq M$, тогда последовательность (1) сходится в среднеквадратичном смысле, и при этом выполняется оценка

$$E||x_k - x^*||^2 \leq \max\left\{\frac{\beta^2 M^2}{2\beta m - 1}, ||x_0 - x^*||^2\right\}\frac{1}{k+1}$$

Док-во.

$$\begin{aligned} E\{||x_{k+1} - x^*||^2 \mid x_k\} &= ||x_k - x^*||^2 - 2\alpha_k E\{g_k\}^T(x_k - x^*) + \alpha_k^2 E||g_k||^2 \\ &= ||x_k - x^*||^2 - 2\frac{\beta}{k+1}\nabla f(x_k)^T(x_k - x^*) + \frac{E\{||g_k||^2\}\beta^2}{(k+1)^2} \\ &\leq \left(1 - \frac{2m\beta}{k+1}\right)||x_k - x^*||^2 + \frac{M^2\beta^2}{(k+1)^2} \end{aligned}$$

Используя индукцию это неравенство дает утверждение теоремы (см. лекцию о градиентном спуске). ■

Замечания

В общем случае скорость сходимости порядка $\mathcal{O}(1/\sqrt{k})$ оптимальна: пусть $f(x) = E(x-\omega)^2$, где ω – некоторая случайная величина. В силу

$$\nabla f(x) = \nabla \int_\Omega (x-\omega)^2 dp(\omega) = \int_\Omega \nabla[(x-\omega)^2] dp(\omega)$$
$$= \int_\Omega 2(x-\omega) dp(\omega) = 2(x - E\omega)$$

получаем, что $E\omega$ есть точка минимума f.

Метод покоординатного спуска

Теорема

Пусть f дважды дифференцируема, δ_{ij} – символ Кронекера,

$$\Delta_i = \left(\delta_{i1} \frac{\partial f}{\partial x_1}, \ldots, \delta_{ij} \frac{\partial f}{\partial x_j}, \ldots, \delta_{in} \frac{\partial f}{\partial x_n} \right)^T,$$

g_k принимает значение Δ_i с вероятностью p_i, $D = \mathrm{diag}\{p_1, \ldots, p_n\}$, для некоторых констант $0 < m < M$ выполняется

$$mI \preceq D\nabla^2 f(\cdot) \preceq MI$$

тогда при $\alpha_k \equiv \alpha$, $0 < \alpha \leq \frac{2}{M+m}$ последовательность (1) сходится в среднеквадратичном смысле, и при этом выполняется оценка

$$E||x_k - x^*||^2 \leq (1 - \alpha m)^k \, ||x_0 - x^*||^2$$

Метод покоординатного спуска

Док-во. Пусть $D_i = \text{diag}\{\delta_{1i}, \ldots, \delta_{ni}\}$, тогда $g_k = D_i \nabla f(x_k)$ с вероятностью p_i, $Eg_k = D\nabla f(x_k)$. Обозначив $A_k = \int_0^1 \nabla^2 f(x^* + t(x_k - x^*))dt$ получаем

$$x_{k+1} - x^* = x_k - x^* - \alpha g_k = x_k - x^* - \alpha D_i \nabla f(x_k) = (I - \alpha D_i A_k)(x_k - x^*)$$

Отметим, что $0 \preceq I - \alpha D_i A_k \preceq I$, а при $0 \preceq A \preceq I$ выполняется $A^T A \preceq A$ (unchecked!!!), таким образом

$$\begin{aligned} E||x_{k+1} - x^*||^2 &= E(x_k - x^*)^T (I - \alpha D_i A_k)^T (I - \alpha D_i A_k)(x_k - x^*) \\ &= (x_k - x^*)^T E\{(I - \alpha D_i A_k)^T (I - \alpha D_i A_k)\}(x_k - x^*) \\ &\leq (x_k - x^*)^T E\{I - \alpha D_i A_k\}(x_k - x^*) \\ &= (x_k - x^*)^T (I - \alpha D A_k)(x_k - x^*) \\ &\leq (1 - \alpha m)||x_k - x^*||^2 \end{aligned}$$

Последнее неравенство справедливо в силу выбора α. ■

Метод покоординатного спуска

Замечание 1. Так как вероятности можно выбирать произвольно, целесообразно подбирать их так, чтобы минимизировать число обусловленности M/m.

Замечание 2. Скорость сходимости покоординатного градиентного спуска соразмерна скорости сходимости обычного градиентного спуска, однако во многих задачах вычисление частных требует гораздо меньше вычислений, например для функций вида

$$f(x) = \sum_{i=1}^{n} g_i(x_i)$$

вычисление граиента всегда имеет стоимость последовательного вычисления всех частных производных.

Рандомизированный метод случайного поиска

Пусть $\Delta_k \in \mathbb{R}^n$ – ограниченная случайная величина, такая что

$$E\Delta_k = 0_n$$
$$E\Delta_k\Delta_k^T = I$$

Из формулы Тейлора получаем для некоторой точки ξ_k на отрезке $[x, x + \beta\Delta_k]$

$$E\{\Delta_k f(x + \beta\Delta_k)\} = E\Delta_k f(x) + \beta E\{\Delta_k\Delta_k^T\}\nabla f(x) + \frac{\beta^2}{2}E\{\Delta_k\Delta_k^T\nabla^2 f(\xi_k)\Delta_k\}$$

При ограниченности δ_k и $\nabla^2 f$ получаем

$$E\{\Delta_k f(x + \beta\Delta_k)\} = \beta\nabla f(x) + \mathcal{O}(\beta^2)$$

Рандомизированный метод случайного поиска

Рандомизированным алгоритмом случайного поиска называется процедура, генерирующая последовательность оценок по правилу

$$x_{k+1} = x_k - \alpha_k f(x_k + \beta \Delta_k), \tag{2}$$

где $\alpha_k, \beta_k > 0$ – некоторые числовые последовательности, Δ_k – последовательность независимых случайных величин, описанных ранее.

Анализ сходимости

Теорема

Пусть f – дважды дифференцируемая функция, сильно выпуклая с константой $0 < m$, $\nabla f(x)$ и $\nabla^2 f$ ограничены, $\alpha_k = \alpha > 0$, $\beta > \frac{1}{2m\alpha}$, $\beta_k = \frac{\beta}{k+1}$, то последовательность, заданная (2), сходится в среднеквадратичном смысле и при этом

$$E||x_k - x^*||^2 = \mathcal{O}(1/k)$$

Док-во. Обозначим $y_k = \Delta_k f(x_k + \beta_k \Delta_k)$. Из показанного ранее

$$\begin{aligned} E||x_{k+1} - x^*||^2 &= ||x_k - x^*||^2 - 2\alpha_k E y_k^T (x_k - x^*) + \alpha_k^2 E y_k^T y_k \\ &\leq ||x_k - x^*||^2 - 2\alpha_k \beta_k \nabla f(x_k)^T (x_k - x^*) \\ &\quad - 2\alpha_k (x_k - x^*)\mathcal{O}(\beta_k^2) + \alpha_k^2 \mathcal{O}(\beta_k^2) \\ &\leq (1 - 2\alpha_k \beta_k)||x_k - x^*||^2 + C\beta_k^2 \alpha_k \end{aligned}$$

где C – некоторая положительная константа. Оставшаяся часть док-ва – индукция (см. сходимость градиентного спуска). ■

Замечания

Замечание 1. Условия сильной выпуклости и ограниченности градиента одновременно выполнимы только на некотором ограниченном множестве. Данное доказательство использует то, что $||x_k - x^*||$ ограничено, что вообще говоря не так.

Замечание 2. Шаг проекции может быть добавлен в этот алгоритм (по аналогии с проективным градиентным спуском) с сохранением результата. Это также избавляет от проблемы с $||x_k - x^*||$.

Замечание 3. Варьирование параметра α_k также может быть использован для решения проблемы с $||x_k - x^*||$.

Рандомизированная стохастическая аппроксимация

Рассмотрим задачу минимизации функционала среднего риска:

$$f(x) = E_\omega F(x, \omega) = \int_\Omega F(x, \omega)\ dp(\omega) \to min$$

На практике сложность таких задач заключается в возможности измерения $F(x, \omega)$ при известном x и неизвестном случайном ω.

По аналогии со случайным поиском, если Δ независимо с ω, то

$$E\Delta F(x + \beta\Delta, \omega) = E_\Delta \Delta f(x + \beta\Delta) = \beta\nabla f(x) + \mathcal{O}(\beta^2)$$

В дополнении, если измерению доступна величина $F(x + \beta\Delta, \omega) + \xi$, где ξ – погрешность измерения, то при условии незивисимости ξ от Δ и ограниченности ξ в силу $E\Delta = 0$ также получаем

$$E\Delta(F(x + \beta\Delta, \omega) + \xi) = E_\Delta \Delta f(x + \beta\Delta) + E_\Delta \Delta\xi = \beta\nabla f(x) + \mathcal{O}(\beta^2)$$

Рандомизированная стохастическая аппроксимация

Рандомизированной стохастической аппроксимацией называется алгоритм построения последовательности оценок по следующей схеме

$$\begin{aligned} y_k &= F(x_k + \beta\Delta_k, \omega) + \xi_k \\ x_{k+1} &= x_k - \alpha_k y_k \end{aligned} \tag{3}$$

Как и в случае случайного поиска, при выборе $\beta_k = \mathcal{O}(1/k)$ мы получаем сходимость порядка $\mathcal{O}(1/k)$ для среднеквадратичного отклонения от точки минимума.