

Методы оптимизации

Лекция 9: Введение в стохастическую оптимизацию

Александр Катруца

Факультет инноваций и высоких технологий
Физтех-школа прикладной математики и информатики



16 декабря 2018 г.

На прошлой лекции

- ▶ Метод сопряжённых градиентов
- ▶ Метод тяжёлого шарика
- ▶ Ускоренный градиентный метод Нестерова

Что нам известно?

- ▶ Детерминированные методы первого порядка
- ▶ Нижние оценки для методов первого порядка
- ▶ Методы, которые достигают этих оценок

Вопросы

- ▶ Как изменятся методы при введении стохастичности в задачу?
- ▶ Как оценивать сходимость в таком случае?

Зачем вводить стохастику?

- ▶ Для большого числа переменных может быть очень долго вычислять точный градиент
- ▶ Стохастического градиента может быть достаточно для решения задачи
- ▶ Иногда параметры задачи измеряются с естественной погрешностью

Как ввести случайность в задачу?

- ▶ Известные параметры задачи – случайные величины с известным распределением

$$\begin{aligned} \min \quad & x_1 + x_2 \\ \text{s.t.} \quad & w_1 x_1 + x_2 \geq 0 \\ & w_2 x_1 + x_2 \geq 0 \\ & x_{1,2} \geq 0, \end{aligned}$$

где $w_1 \sim \mathcal{U}[0, 4]$, $w_2 \sim \mathcal{U}[2, 3]$

- ▶ Целевая функция – матожидание некоторой другой функции

$$\min f(x) := \mathbb{E}_\omega[F(x, \omega)]$$

- ▶ Частный случай

$$\min \frac{1}{N} \sum_{i=1}^N f(x_i)$$

SAA vs SA

- ▶ Стохастическая аппроксимация (SA)
 - ▶ Сгенерировать ω^k iid
 - ▶ Вычислить стохастический градиент $G(x, \omega^k)$
 - ▶ Использовать его в стохастическом градиентном спуске
- ▶ Усреднённая по сэмплам аппроксимация (SAA)
 - ▶ Сгенерировать N сэмплов $\omega_1, \dots, \omega_N$
 - ▶ Вычислить эмпирическую оценку целевой функции
$$\hat{f}_N = \frac{1}{N} \sum_{i=1}^N F(x, \omega_i)$$
 - ▶ Минимизировать \hat{f}_N вместо исходной функции f

Постановка задачи

$$\min_{x \in \mathbb{R}^n} \frac{1}{N} \sum_{i=1}^N f_i(x)$$

- ▶ $f_i(x)$ могут быть невыпуклыми
- ▶ n может быть порядка 10^6 и больше
- ▶ N также может быть очень большим

Пример 1

- ▶ Стохастическая оценка следа матрицы

$$\text{trace}(A) = \text{trace}(AI) = \text{trace}(A\mathbb{E}_z z z^\top) = \mathbb{E}_z(z^\top A z),$$

где z - вектор из стандартного нормального распределения

- ▶ Матожидание заменим на несмещённую оценку \hat{f}_N как в SAA подходе
- ▶ Минимизируем \hat{f}_N для фиксированных z_i из предыдущего пункта

Пример 2

- ▶ Задача классификации
- ▶ Функция ошибки ℓ аддитивна по объектам обучающей выборки

$$\min_w \frac{1}{N} \sum_{i=1}^N \ell(w|x_i)$$

- ▶ Интерпретация через эмпирический риск и приближение истинного распределения данных

Стохастический градиентный спуск (SGD)

$$x_{k+1} = x_k - \alpha_k h_k,$$

где

- ▶ $h_k = f'_{i_k}(x_k)$, $i_k \in \{1, \dots, N\}$ выбирается случайно
- ▶ $h_k = \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} f'_i(x_k)$, $\mathcal{I}_k \subset \{1, \dots, N\}$ – некоторое подмножество индексов, обычно, фиксированной мощности $|\mathcal{I}_k| = m$

Свойства

1. Несмещённая оценка градиента

$$\mathbb{E}[h_k] = f'(x_k)$$

2. Большая дисперсия

Сходимость

Теорема

Пусть f выпуклая функция с Липшицевым градиентом с константой L . Тогда если SGD генерирует h_k такие что $\text{Var}(h_k) \leq \sigma^2$ и $\alpha_k \leq \frac{1}{L}$ тогда

$$\mathbb{E}[f(\bar{x}_k)] - f^* \leq \frac{\|x_0 - x^*\|_2^2}{\alpha_k k} + \frac{\alpha_k \sigma^2}{2}$$

В частности после $k = \frac{(\sigma^2 + L\|x^* - x_0\|_2^2)}{\varepsilon^2}$ итераций при условии $\alpha_k = \frac{1}{\sqrt{k}}$ получим решение с точностью 2ε

Общий подход к уменьшению дисперсии

- ▶ Пусть X_ω даёт несмещённую оценку для параметра x :
 $\mathbb{E}_\omega[X_\omega] = x$
- ▶ Пусть $Z_\omega = X_\omega - Y_\omega$ так что $\mathbb{E}_\omega[Y_\omega] \approx 0$
- ▶ Тогда $\mathbb{E}_\omega[X_\omega] = \mathbb{E}_\omega[Z_\omega] = x$
- ▶ $\text{Var}(Z_\omega) = \text{Var}(X_\omega) + \text{Var}(Y_\omega) - 2\text{Cov}(X_\omega, Y_\omega) \ll \text{Var}(X_\omega)$
если Y_ω сильно коррелирует с X_ω

Рецепт по уменьшению дисперсии

Найти такую оценку Y , что

1. Матожидание близко к 0
2. Сильно коррелирует с данной оценкой X

Stochastic average gradient (Schmidt, Le Roux, Bach 2013)

- ▶ Инициализация x_0 и $g_i^0 = x_0, i = \{1, \dots, N\}$
- ▶ На k -ой итерации выбираем i_k и обновляем $g_{i_k}^k = f'_{i_k}(x_k)$
- ▶ $x_{k+1} = x_k - \alpha_k \frac{1}{N} \sum_{i=1}^N g_i^k$
- ▶ Более наглядная запись

$$x_{k+1} = x_k - \alpha_k \left(\frac{1}{N} g_{i_k}^{(k+1)} - \frac{1}{N} g_{i_k}^k + \frac{1}{N} \sum_{i=1}^N g_i^k \right)$$

Уменьшение дисперсии

- ▶ $X = g_{i_k}^{(k+1)}$ и $\mathbb{E}_\omega[X] = f'(x_k)$
- ▶ $Y = g_{i_k}^k - \sum_{i=1}^N g_i^k$ и $\mathbb{E}_\omega[Y] \neq 0$
- ▶ $\|X - Y\|_2 = \|(g_{i_k}^{(k+1)} - g_{i_k}^k) + \sum_{i=1}^N g_i^k\|_2 \rightarrow 0, k \rightarrow \infty$
- ▶ Дисперсия итоговой оценки стремится к 0

Сходимость для L -выпуклой функции

Теорема

Пусть f_i дифференцируемы с Липшицевым градиентом,
 $\bar{x}^{(k)} = \frac{1}{k} \sum_{i=0}^{k-1} x_i$, $\alpha_k = \frac{1}{16L}$ и инициализация

$$g_i^0 = f'_i(x_0) - f'(x_0), \quad i = 1, \dots, N$$

даёт

$$\mathbb{E}[f(\bar{x}^{(k)})] - f(x^*) \leq \frac{48n}{k}(f(x_0) - f^*) + \frac{128L}{k}\|x_0 - x^*\|_2^2$$

Сравнение

- ▶ SAG

$$\frac{48n}{k}(f(x_0) - f^*) + \frac{128L}{k}\|x_0 - x^*\|_2^2$$

Зависимость от n в первом слагаемом!

- ▶ GD

$$\frac{L\|x_0 - x^*\|_2^2}{k}$$

- ▶ SGD

$$\frac{\|x_0 - x^*\|_2^2 + \sigma^2}{2\sqrt{k}}$$

Сходимость для L -выпуклой и μ -сильно выпуклой функции

Теорема

При тех же предположениях, что и в теореме для L -выпуклой функции выполнено следующее

$$\mathbb{E}[f(\bar{x}^{(k)})] - f(x^*) \leq \left(1 - \min\left\{\frac{m}{16L}, \frac{1}{8n}\right\}\right)^k \left(\frac{3}{2}(f(x_0) - f^*) + \frac{4L}{n}\|x^* - x_0\|_2^2\right)$$

- ▶ Адаптируется к сильной выпуклости
- ▶ Аналогичен GD
- ▶ SGD даёт только $\mathcal{O}(1/k)$

Комментарии

- ▶ Для SAG необходима аккуратная настройка
- ▶ Начальное приближение лучше получать после одной эпохи SGD + сохранение g_i^0
- ▶ Выбор α_k

SAGA (Defazio, Bach, Lacoste-Julien 2014)

Аналогично SAG, но

$$x_{k+1} = x_k - \alpha_k \left(g_{i_k}^{(k+1)} - g_{i_k}^k + \frac{1}{N} \sum_{i=1}^N g_i^k \right)$$

- ▶ Несмещённая оценка: $\mathbb{E}[Y] = 0$
- ▶ Дисперсия выше, чем у SAG
- ▶ Аналогичный анализ уменьшения дисперсии
- ▶ Обобщается на комбинированные задачи
- ▶ Оценки сходимости для двух классов выпуклых функций аналогичны SAG
- ▶ Детали реализации аналогичны SAG

SVRG (Johnson, Zhang 2013)

- ▶ Инициализация \bar{x}_0

- ▶ Цикл $k = 1, 2, \dots$

- ▶ $\bar{x} = \bar{x}_0$

- ▶ $\bar{\mu} = f'(\bar{x})$

- ▶ $x_0 = \bar{x}_0$

- ▶ Цикл $m = 1, \dots, l$

- ▶ Случайно выбрать $i_m \in \{1, \dots, N\}$

- ▶

$$x_{m+1} = x_m - \alpha(f'_{i_m}(x_m) - f'_{i_m}(\bar{x}) + \bar{\mu})$$

- ▶ $\bar{x}_0 = x_l$

Особенности

- ▶ Аналог SAGA
- ▶ Гораздо более простое доказательство
- ▶ Зависит от размера эпох

Недостатки методов уменьшения дисперсии

- ▶ Требуют вычисления точного градиента
- ▶ Зависят от дополнительных параметров
- ▶ Нет универсального способа запуска

- ▶ Различные способы появления случайности
- ▶ SAA vs SA
- ▶ Наличие шума приводит к замедлению SGD
- ▶ Различные способы усреднения градиентов приводят к ускорению
- ▶ Завышенные требования для реальных задач