Теория оптимизации: подробный конспект

1. Общая формулировка задачи оптимизации и аналитическая сложность минимизации невыпуклых липшицевых функций

Формулировка задачи оптимизации:

$$\min_{x \in \mathbb{R}^n} f(x),$$

где $f(x)$ — целевая функция, заданная на $\mathbb{R}^n$, и может быть выпуклой/невыпуклой, гладкой/негладкой.

Нижняя граница сложности: Для невыпуклых L-липшицевых функций (f с L-липшицевым градиентом) методам требуется $\Omega(\epsilon^{-2})$ итераций для нахождения ϵ-оптимального решения ($\|\nabla f(x)\| \leq \epsilon$).

2. Выпуклая оптимизация и примеры в машинном обучении

Определение: Функция $f(x)$ выпукла, если

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \quad \forall x, y \text{ и } \lambda \in [0,1].$$

Примеры в машинном обучении:

- **Регрессия (Lasso, Ridge):**
$$\min_w \|Xw - y\|_2^2 + \lambda \|w\|_1.$$
- **Логистическая регрессия:**
$$\min_w \frac{1}{N} \sum_{i=1}^{N} \log(1 + e^{-y_i w^\top x_i}) + \lambda \|w\|_2^2.$$
- **SVM с линейным ядром:**
$$\min_w \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{N} \max(0, 1 - y_i w^\top x_i).$$

3. Адаптивный градиентный спуск и наискорейший спуск

Адаптивный градиентный спуск:

- Приспосабливает шаг t к характеристикам задачи, учитывая историю градиентов.
- Пример: Adam, AdaGrad.

Наискорейший спуск:

- На каждой итерации минимизируется $f(x - t\nabla f(x))$ по t.
- Пример: Применяется в выпуклой оптимизации, где возможно аналитическое нахождение оптимального t.

4. Градиентный метод для задач с градиентным доминированием (Поляка-Лоясиевича)

Условие Поляка-Лоясиевича:

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*), \quad \mu > 0.$$

Пример: Обучение глубокой нейронной сети через нелинейные перепараметризованные слои; ускоренная сходимость благодаря свойству градиентного доминирования.

5. Стохастический градиентный метод (SGD)

Суть: Замена градиента $\nabla f(x)$ его приближением на случайной выборке:

$$\nabla f_i(x).$$

Применение: Эффективен для задач с большими размерами данных (например, обучение нейронных сетей).

6. Неточный оракул и минибатчинг

Неточный оракул: Предоставляет приближенные градиенты, обеспечивая:

$$\|\nabla f(x) - g(x)\| \leq \epsilon.$$

Минибатчинг: Используется для уменьшения шума стохастических градиентов; вычисляется на небольшой выборке данных.

7. Ускоренные градиентные методы и метод подобных треугольников

Метод Нестерова:

- **Ускорение:** $O(1/k^2)$ для выпуклых задач.
- **Схема:**

$$y_{k+1} = x_k + \frac{k-1}{k+2}(x_k - x_{k-1}).$$

Метод подобных треугольников: Использует геометрическое представление, применим к гладким выпуклым задачам.

8. Метод сопряжённых градиентов

Цель: Решение задач квадратичной оптимизации:

$$\min_x \frac{1}{2}x^\top Qx - b^\top x.$$

Сходимость: Зависит от числа обусловленности $\kappa(Q)$.

9. Метод Франк-Вульфа

Суть: Решение задач:

$$\min f(x) \text{ при ограничении } x \in \mathcal{D},$$

где $\mathcal{D}$ — выпуклое множество.

Сходимость: $O(1/k)$ для гладких функций.

10. Субградиентный метод

Оценка сходимости:

$$O(1/\sqrt{k})$$

для выпуклых задач.

11. Универсальные методы градиентного типа

Суть: Применимы к широкому классу задач (гладкие/негладкие). Пример: ускоренный универсальный метод.

12. Стохастический субградиентный метод

Особенность: Применим к негладким задачам. Использует выборку для оценки субградиента.

13. AdaGrad

Идея: Регулировка шагов t на основе накопленной информации о градиентах:

$$t \propto \frac{1}{\sqrt{\sum g_t^2}}.$$

Пример: Обучение моделей с разреженными градиентами.

14. Метод Ньютона

Схема: Использует гессиан H:

$$x_{k+1} = x_k - H^{-1}\nabla f(x_k).$$

Сходимость: Квадратичная для выпуклых функций.

15. Квазиньютоновские методы

Суть: Аппроксимация гессиана H без явного вычисления. **Пример:** Метод BFGS, L-BFGS.