Расширенный конспект по теории оптимизации

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:

Теорема 1. *Для функций $f(x)$, имеющих L-липшицев градиент ($\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$), методам первого порядка требуется $\Omega(\epsilon^{-2})$ итераций для нахождения ϵ-оптимального решения ($\|\nabla f(x)\| \leq \epsilon$).*

Доказательство: Строится контрпример на основе квадратичной функции $f(x) = \frac{L}{2}x^2$, где вычисление градиентов показывает невозможность достижения точности за меньшее число итераций. Полное доказательство опускается в виду сложности.

2. Выпуклая оптимизация и примеры в машинном обучении

Определение выпуклой функции: 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 выпуклых задач:

- **Регрессия (Lasso):**

  $$\min_w \|Xw - y\|_2^2 + \lambda \|w\|_1.$$

  Здесь $\|w\|_1$ является выпуклой, но негладкой нормой.

- **Логистическая регрессия:**

  $$\min_w \frac{1}{N} \sum_{i=1}^{N} \log(1 + e^{-y_i w^\top x_i}) + \lambda \|w\|_2^2.$$

- **Поддерживающие векторы (SVM):**

  $$\min_w \frac{1}{2}\|w\|_2^2 + C \sum_{i=1}^{N} \max(0, 1 - y_i w^\top x_i).$$

3. Адаптивный градиентный спуск и наискорейший спуск

Адаптивный градиентный спуск: Методы семейства AdaGrad, RMSProp и Adam используют информацию о предыдущих градиентах для масштабирования шагов:

$$t_t \propto \frac{1}{\sqrt{\sum_{i=1}^{t} g_i^2}}.$$

Наискорейший спуск:

Определение 1. *На каждом шаге выбирается t_k, минимизирующий $f(x_k - t\nabla f(x_k))$ по t.*

Теорема 2. *При L-липшицевом градиенте наискорейший спуск обладает линейной скоростью сходимости для строго выпуклых функций.*

4. Градиентный метод при условии градиентного доминирования (Поляка-Лоясиевича)

Определение условия Поляка-Лоясиевича:

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

Теорема 3. *Градиентный метод для функций, удовлетворяющих условию Поляка-Лоясиевича, сходится со скоростью:*

$$f(x_k) - f^* \leq \frac{C}{k^2}.$$

Пример: Обучение глубокой нейронной сети, где перепараметризация вводит скрытую регуляризацию, улучшая поведение градиентов.

5. Стохастический градиентный метод (SGD)

Идея: Вместо вычисления полного градиента используется приближенный градиент на случайной выборке:

$$\nabla f_i(x).$$

Теорема 4. *При уменьшении шага по правилу $t_k = O(1/\sqrt{k})$, SGD достигает сходимости:*

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

Применение: Эффективен для обучения больших моделей.

6. Неточный оракул и минибатчинг

Неточный оракул: Приближает градиенты с заданной точностью:

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

Минибатчинг: Выбирается подмножество данных, что снижает шум, но сохраняет стохастичность.

7. Ускоренные градиентные методы

Метод Нестерова:

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

Теорема 5. *Скорость сходимости для выпуклых функций:*

$$f(x_k) - f^* \leq O(1/k^2).$$

8. Метод сопряжённых градиентов

Для задач:

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

Используется структура матрицы Q для ускорения сходимости.

9. Метод Франк-Вульфа

Теорема 6. *Скорость сходимости метода:*

$$f(x_k) - f^* \leq O(1/k).$$

10. Субградиентный метод

Теорема 7. *Для выпуклых задач:*

$$f(x_k) - f^* \leq O(1/\sqrt{k}).$$

11. Универсальные методы

Методы, работающие как для гладких, так и для негладких задач. Пример: ускоренные универсальные схемы Нестерова.

12. Стохастический субградиентный метод

Работает аналогично SGD, но применим для негладких задач.

13. AdaGrad

$$t_t \propto \frac{1}{\sqrt{\sum_{i=1}^{t} g_i^2}}.$$

14. Метод Ньютона

Использует гессиан для квадратичной сходимости.

15. Квазиньютоновские методы

Метод BFGS: аппроксимирует гессиан с помощью информации о градиентах.