Оптимизация — это о чём? Ну, о том, что мы ищем минимум (или максимум) какой-то функции на каком-то множестве. Причём минимум ищется обычно локальный, потому что глобальный искать очень сложно.
Рассмотрим непрерывную оптимизацию: наша функция непрерывна. В таких методах применяется принцип чёрного ящика: мы не анализируем то, что делает функция, нас интересуют только её значения. Иногда нас также интересует градиент функции или гессиан. В зависимости от того, производная какого порядка нам нужно в нашем методе, говорят о том, что наш метод — это метод соответствующего порядка. Методы второго порядка — это, например, методы Ньютона. Также есть методы квази-Ньютона, которые пытаются как-то аппроксимировать гессиан, нь сейчас не о них.
Зачем нам вообще градиент и гессиан? Потому что они позволяют разложить функцию в ряд Тейлора, и выглядеть он будет так:

$$f(x+\Delta x) = f(x) + \Delta x^T \nabla f(x) + \frac{1}{2}\Delta x^T H_f(x)\Delta x + o(\|\Delta x\|^2)$$

Более высокие порядки не используются обычно, потому что там возникают такие операции как взятие обратной матрицы, которые вводят численную нестабильность.

Пример. Тернарный поиск. Все мы его знаем. Это метод нулевого порядка. Берём функцию, которая сначала убывает, потом возрастает и итеративно делим отрезок на три части, на каждом этапе отбрасывая одну треть.
Можно немного упростить жизнь себе, использовав метод золотого сечения. А именно отрезок $[a; b]$ мы делим на три части точками $x_1 = b - \frac{b-a}{\Phi}$, $x_2 = a + \frac{b-a}{\Phi}$, где $\Phi = \frac{1+\sqrt{5}}{2}$. Остальное как в обычном тернарном поиске. При таком разделении значение в одной из двух точек будет переиспользовано, ведь x_1 делит $[a; x_2]$ в соотношении золотого сечения и x_2 делит $[x_1; b]$ в соотношении золотого сечения.

Нулевого порядка в целом больше ничего не придумать.
Для методов первого порядка обычно итеративно выбирают последовательность точек, значения в которых должны уменьшаться. Как выбирать — рассмотрим позже. А сейчас подумаем, когда останавливаться.

Определение 1. Пусть $\{x_k\}$ — последовательность точек, которые выбирает метод, а x^* — локальный минимум, к которому метод стремится. Тогда если выполнено

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^\mu} < r$$

то r называют **скоростью сходимости**, а μ — **степенью сходимости**.

Определение 2. **Контурными линиями** называются множества

$$L_f(a) = \{x \mid f(x) = a\}$$

Утверждение. *Контурные линии ортогональны градиенту функции.*

Условия оптимальности. Хочется найти какие-нибудь критерии оптимальности точки в какой-то окрестности. У нас могут быть достаточные условия, могу быть необходимые, и все они следуют из квадратичной аппроксимации функции.
Ну, точка оптимальна, если при добавлении произвольного Δx функция возрастёт. Заметим, что $\nabla f(x)$ и $H_f(x)$ никак не зависят от x. Для методов первого порядка есть только необходимое условие:

Теорема 1. *Если точка оптимальна, то* $\nabla f(x) = 0$.

Доказательство. Ну, если он не ноль, то можно найти достаточно маленькое Δx, противонаправленное $\nabla f(x)$, и функция уменьшится. А значит точка не оптимальна была. □

Самое интересное, что для хороших функций можно решить уравнение $\nabla f(x) = 0$ и получить аналитическое решение. Например, так отлично решается линейная регрессия.
А вот если у нас есть гессиан, то мы и достаточное условие можем сформулировать. Это условие доказывается через спектральное разложение гессиана.

Теорема 2. *Если все собственные числа гессиана в точке положительные, то эта точка — минимум.*
Если точка — минимум, то все собственные числа гессиана неотрицательны.

Доказательство.

$$\begin{aligned} f(x+\Delta x) &\approx f(x) + \frac{1}{2}\Delta x^T H_f(x)\Delta x = \\ &= f(x) + \frac{1}{2}\Delta x^T Q \wedge Q^T \Delta x = \\ &= f(x) + \frac{1}{2}(Q^T\Delta x)^T \wedge Q^T \Delta x = \\ &= f(x) + \frac{1}{2}\sum_i \lambda_i \|(Q^T\Delta x)_i\| \end{aligned}$$

□

Линейные ограничения. Давайте добавим какие-нибудь ограничения. например, линейные ограничения: минимизировать функцию f при условии $Ax = b$. Тогда $A\Delta x = 0$. Соотвественно, проверяя условия необходимости и/или достаточности, нам надо проверять только те направления, для которых верно $A\Delta x = 0$. Как с этим работать? Ну, довольно легко: надо рассмотреть базис $\ker A$ (назовём N матрицу проекции на $\ker A$). И все Δx надо заменить на $N\Delta x$. Тогда наш Δx всегда будет правильным. Что будет с нашим рядом Тейлора?

$$f(x + N\Delta x) = f(x) + \Delta x^T N^T \nabla f(x) + \frac{1}{2}\Delta x^T N^T H_f(x) N \Delta x + o(\|\Delta x\|^2)$$

В итоге можно только домножать градиент и гессиан на N.