# Линейная регрессия. L1 и L2 регуляризации.

# Вспомним термины

- R - что обозначает?
- Мат ожидание
- Дисперсия
- Скалярное произведение
- Линейная комбинация
- Линейная функция
- 1 и 2 норма вектора
- Градиент
- Линейная зависимость

# Задача регрессии

Если просто: $\mathbb{X} \to \mathbb{R}.$

Из мн-ва объектов (X), мы хотим найти верное отображение во мн-во вещественных чисел (R)

Если сложно:

Для качественного решения задачи регрессии мы хотим подобрать такую функцию, которая минимизирует функцию потерь (Loss) , но при этом остаётся достаточно простой, чтобы следовать “общей тенденции” данных

# Как мы это делаем?

Формула линейной регрессии: $$f = \sum_i w_i f_i.$$

w - коэффициент, который мы хотим подобрать

f - базисная функция, которая может быть нелинейной

ВАЖНО: линейная регрессия - это линейная комбинация базисных функций
=> линейная регрессия - это единая функция (модель)

# Метод наименьших квадратов

Принцип обучения линейной регрессии:

Подбираем такие коэффициенты перед  X и свободный член, которые максимально минимизируют MSE


Ошибка

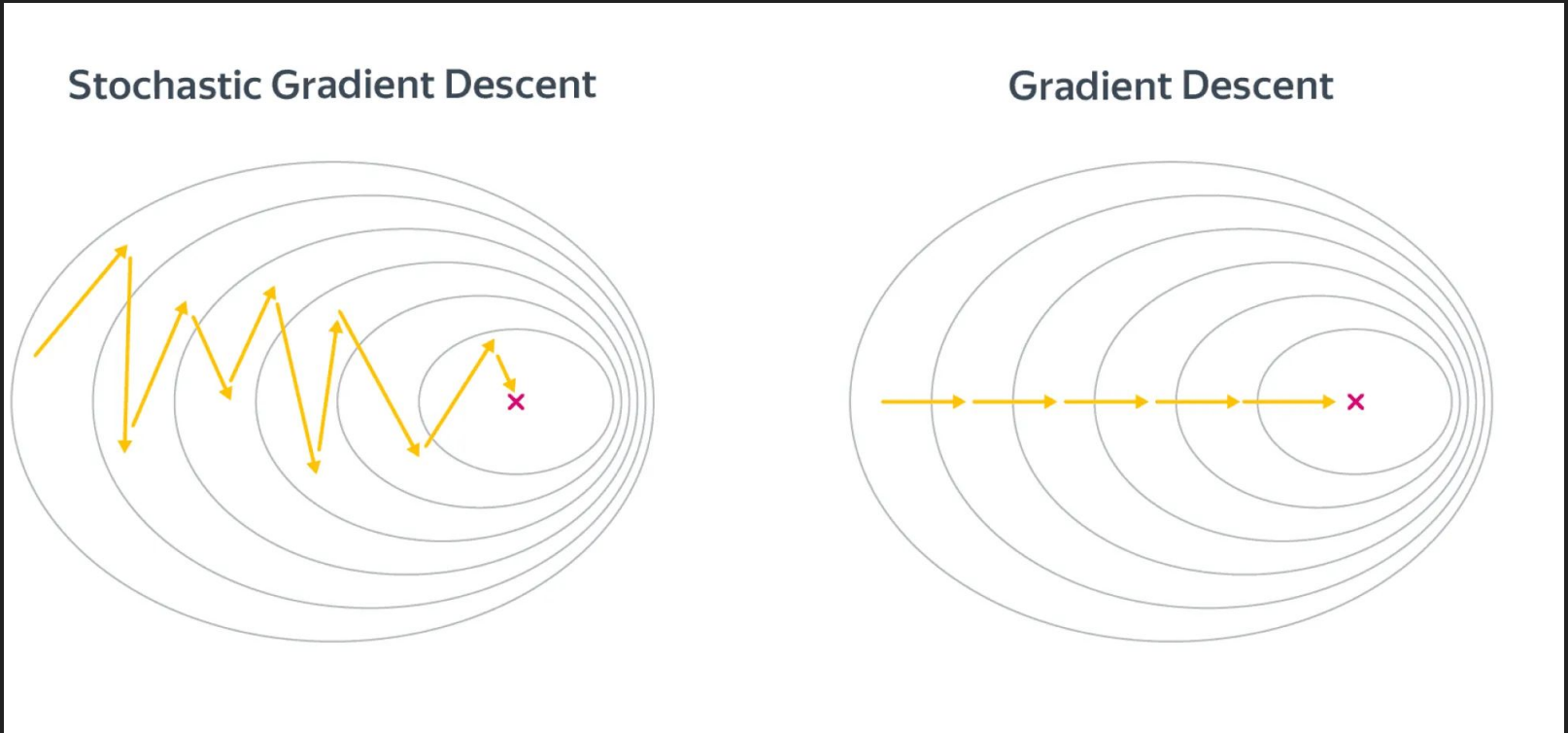
$$\varepsilon_i = y_i - \hat{y}_i$$

MSE

$$S(\beta_0, \beta_1, \ldots, \beta_n) = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{m} \varepsilon_i^2$$

Как алгоритм понимает насколько ему менять свои предсказания?

# Градиентный спуск



# Почему это работает?

Теорема Гаусса-Маркова описывает почему МНК оптимален при следующих условиях:

**Gauss–Markov assumptions:**

- $\mathbb{E}(\varepsilon_i) = 0 \quad \forall i$
- $\mathrm{Var}(\varepsilon_i) = \sigma^2 < \inf \quad \forall i$
- $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$

# Основные положения теоремы Гаусса-Маркова

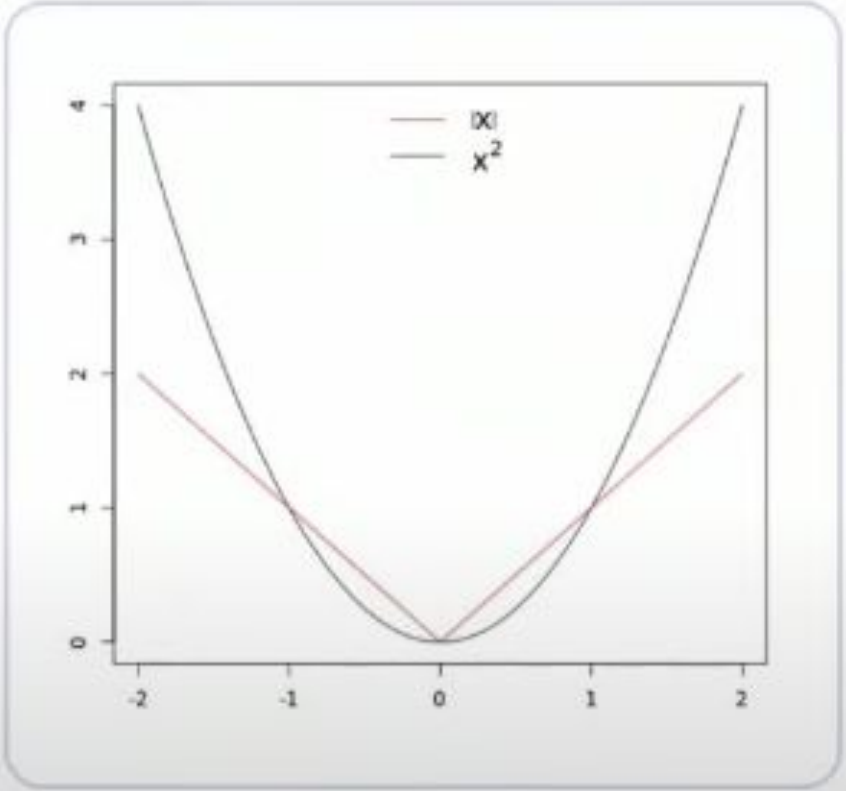
При выполнении вышеперечисленных условий:

1. Оценки коэффициентов $\beta 0, \beta 1, ..., \beta n$ полученные с помощью метода наименьших квадратов, являются **несмещёнными**. Это означает, что в среднем оценки будут равны истинным значениям коэффициентов.

2. Оценки имеют **наименьшую дисперсию** среди всех возможных линейных и несмещённых оценок. То есть оценки коэффициентов, полученные методом наименьших квадратов, являются самыми точными линейными оценками в смысле минимальной дисперсии.

Таким образом, при выполнении условий теоремы Гаусса-Маркова, OLS даёт наиболее точные и надёжные оценки среди всех линейных оценок, что делает этот метод основным инструментом для регрессионного анализа.

- **MSE ($L_2$)**
  - delivers BLUE according to Gauss-Markov theorem
  - differentiable
  - sensitive to noise
- **MAE ($L_1$)**
  - non-differentiable (not a problem)
  - much more prone to noise



# Итак, принцип работы линейной регрессии

1. Инициализируем веса
2. Считаем Loss
3. Находим градиент
4. Поправляем веса
5. Повторяем пункты 2 и 4 пока не будет достигнуто одно из условий условий остановки

Какие есть условия?

- Достигнут порог изменения функции ошибки
- достигнут минимум функции
- максимальное кол -во итераций

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

# Loss функции

$$MAPE(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{\hat{y}_i - y_i}{y_i} \right|$$

$$MSE = \frac{1}{N} \sum_{i=0}^{N} (\hat{y}_i - y_i)^2$$

$$MAE = \frac{\sum_{i=1}^{n} |y_i - y_i^p|}{n}$$

# Как мы можем улучшить показатели нашей модели?

1. Изменить масштаб данных. Для этого используются scaler-ы:

   - Min-Max scaler

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

   - Standard scaler

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

2. Применить регуляризацию

# L1 vs L2 регуляризации

L1

$$\text{Loss} = \text{MSE} + \lambda \sum |w_i|$$

Может занулять  веса =>
мы выкидываем некоторые признаки, которые мешают учить модель

L2

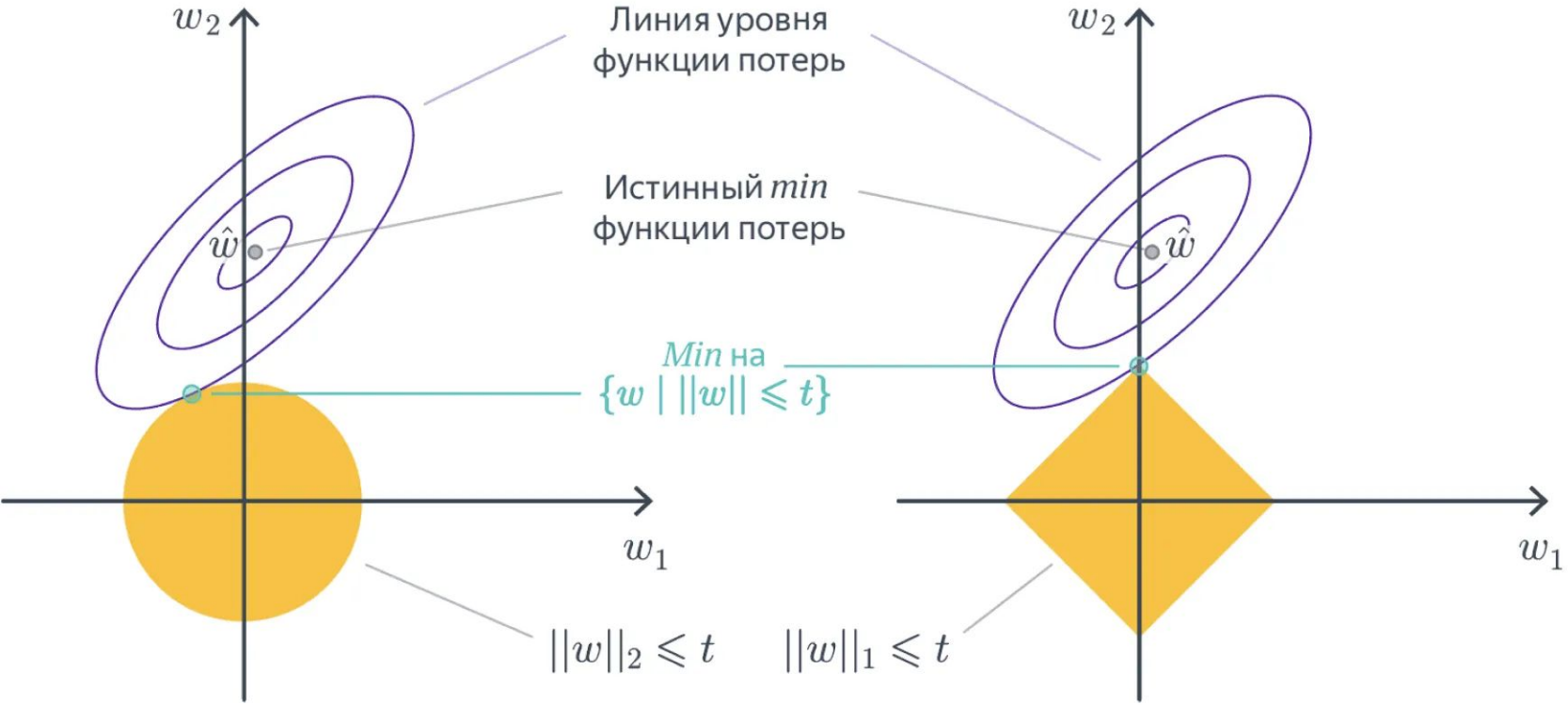
$$\text{Loss} = \text{MSE} + \lambda \sum w_i^2$$

Оставляет веса близкими к нулю,

что способствует более стабильному обучению

$L_2$–регуляризация

$L_1$–регуляризация



# ElasticNet

$$\text{Loss} = \frac{1}{2n}\sum_{i=1}^{n}(y_i - X_i w)^2 + \lambda_1 \sum_{j=1}^{p}|w_j| + \lambda_2 \sum_{j=1}^{p} w_j^2$$

**То есть это комбинация из l1 и l2 регуляризаций**

**L1-регуляризация** в Elastic Net способствует разреженности модели (обнуляет коэффициенты малозначимых признаков), как и Lasso (L1)

**L2-регуляризация** снижает влияние коррелированных признаков и стабилизирует решение, как Ridge (L2)

Elastic Net эффективен в случае, когда признаки сильно коррелированы между собой, так как Lasso может "выбросить" важные коррелированные признаки, а L2 помогает этому избежать.

# Плюсы линейной регрессии

1. Интерпретируемость
2. Быстро работает
3. Легко обучать
4. Есть много вариантов регуляризаций, метри и Loss функций
5. Хорошо работает на простых задачах и небольших датасетах
6. Может в экстраполяцию

# Минусы линейной регрессии

1. Теряет свою эффективность на больших выборках
2. Теряется при большом кол-ве коррелирующих признаков
3. Может быть неверное предположение о линейной зависимости между таргетом и фичами
4. Чувствительность к выбросам (при МНК)

# Метрики регрессии

$$MSE(y^{true}, y^{pred}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - f(x_i))^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}.$$

$$MAE(y^{true}, y^{pred}) = \frac{1}{N}\sum_{i=1}^{N}|y_i - f(x_i)|$$

$$MAPE(y^{true}, y^{pred}) = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - f(x_i)|}{|y_i|}$$

$$SMAPE(y^{true}, y^{pred}) = \frac{1}{N}\sum_{i=1}^{N}\frac{2\,|y_i - f(x_i)|}{y_i + f(x_i)}$$