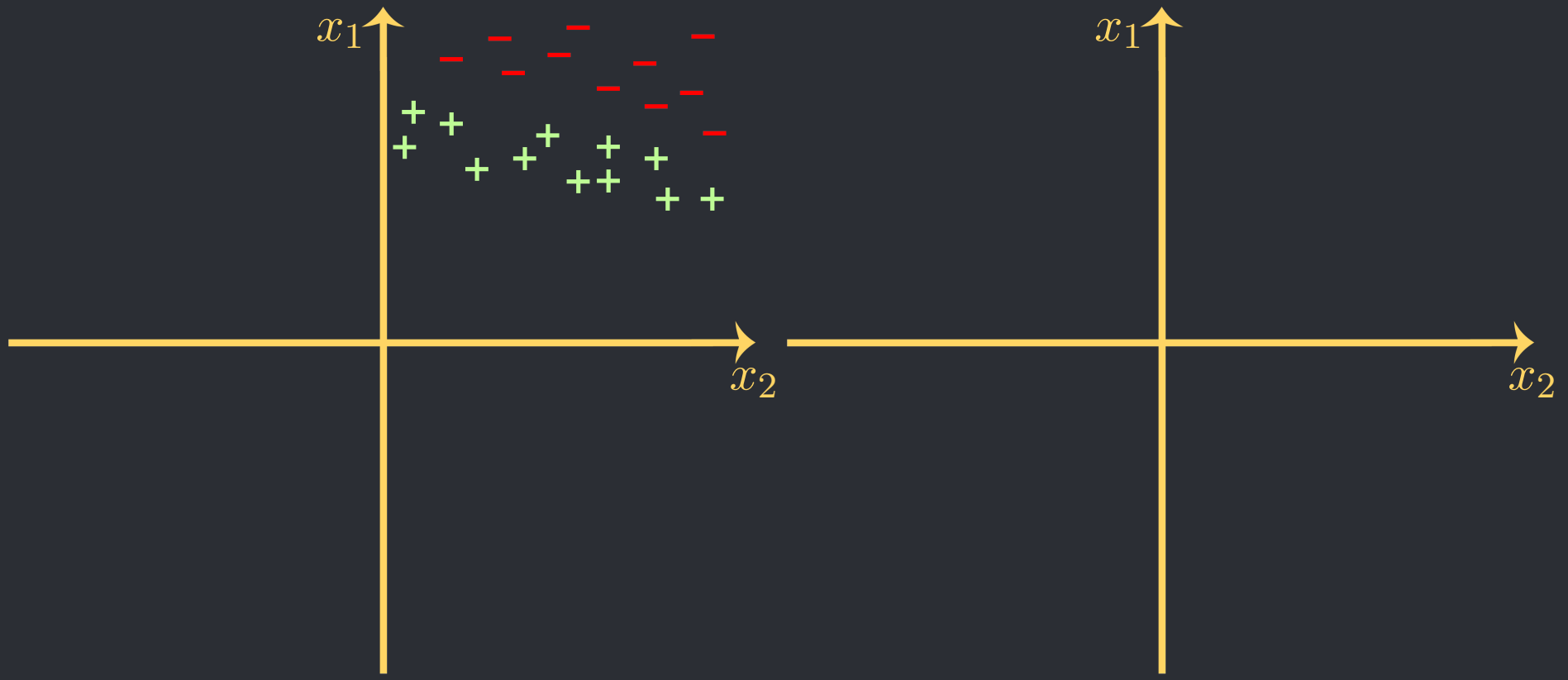
Часть 9: Нормализация

Романов Михаил, Игорь Слинько

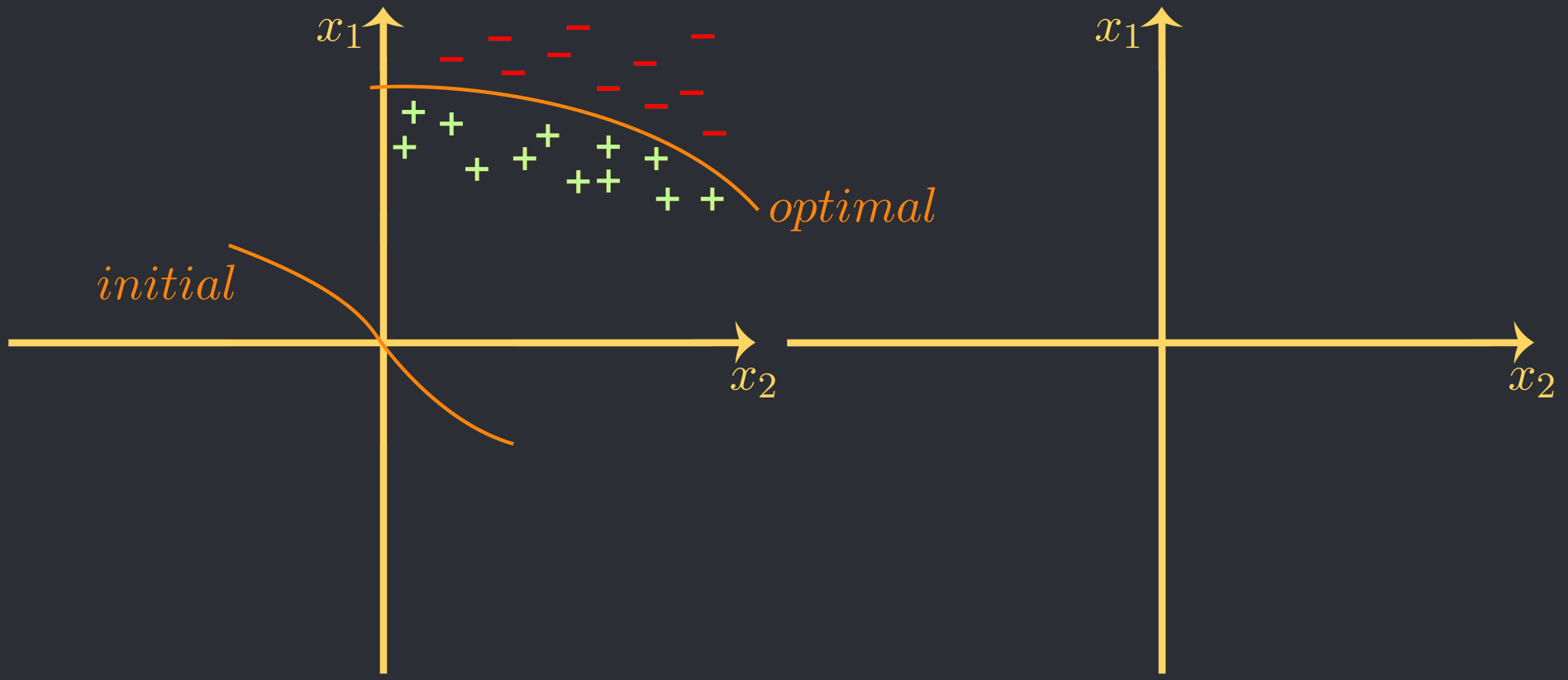
Центрирование данных

$$p = \sigma(w_1 x_1 + w_2 x_2 + b)$$



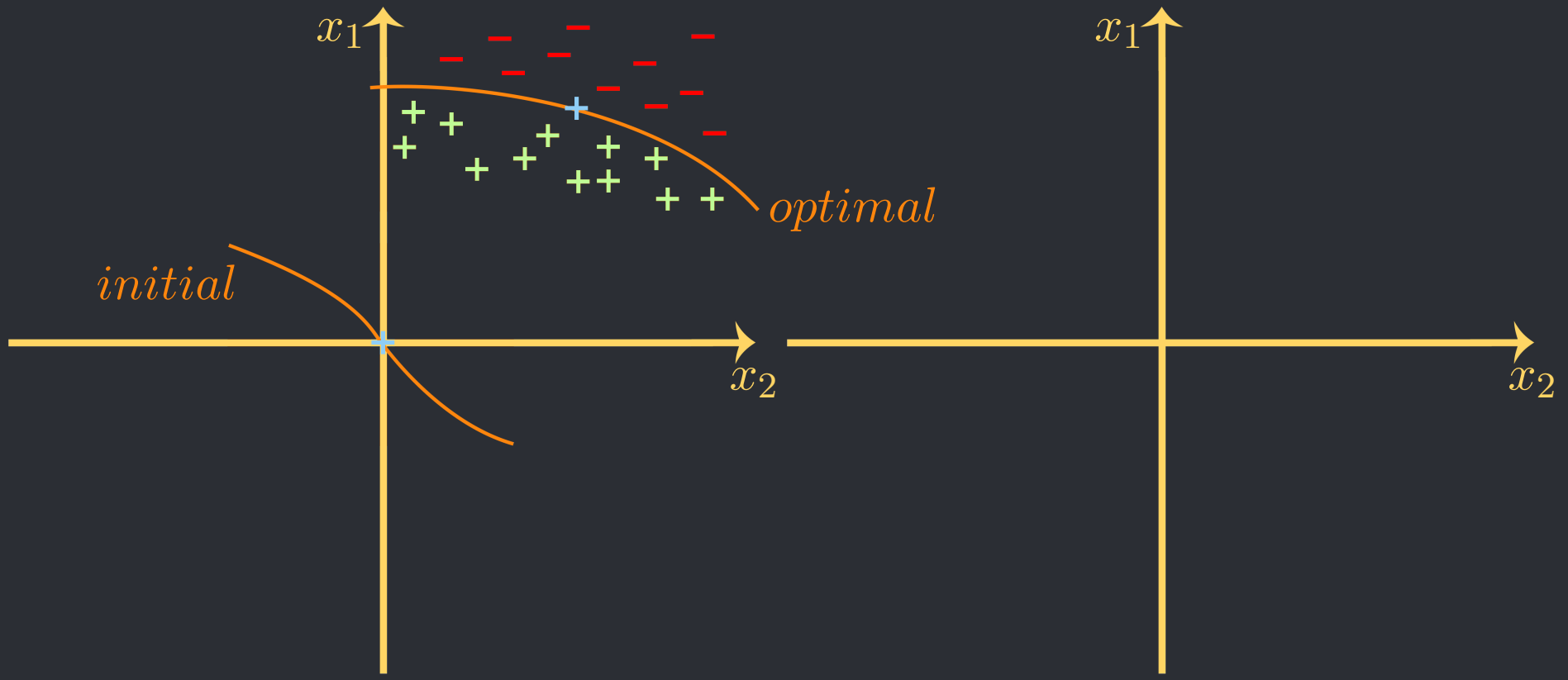
Центрирование данных

$$p = \sigma(w_1 x_1 + w_2 x_2 + b)$$



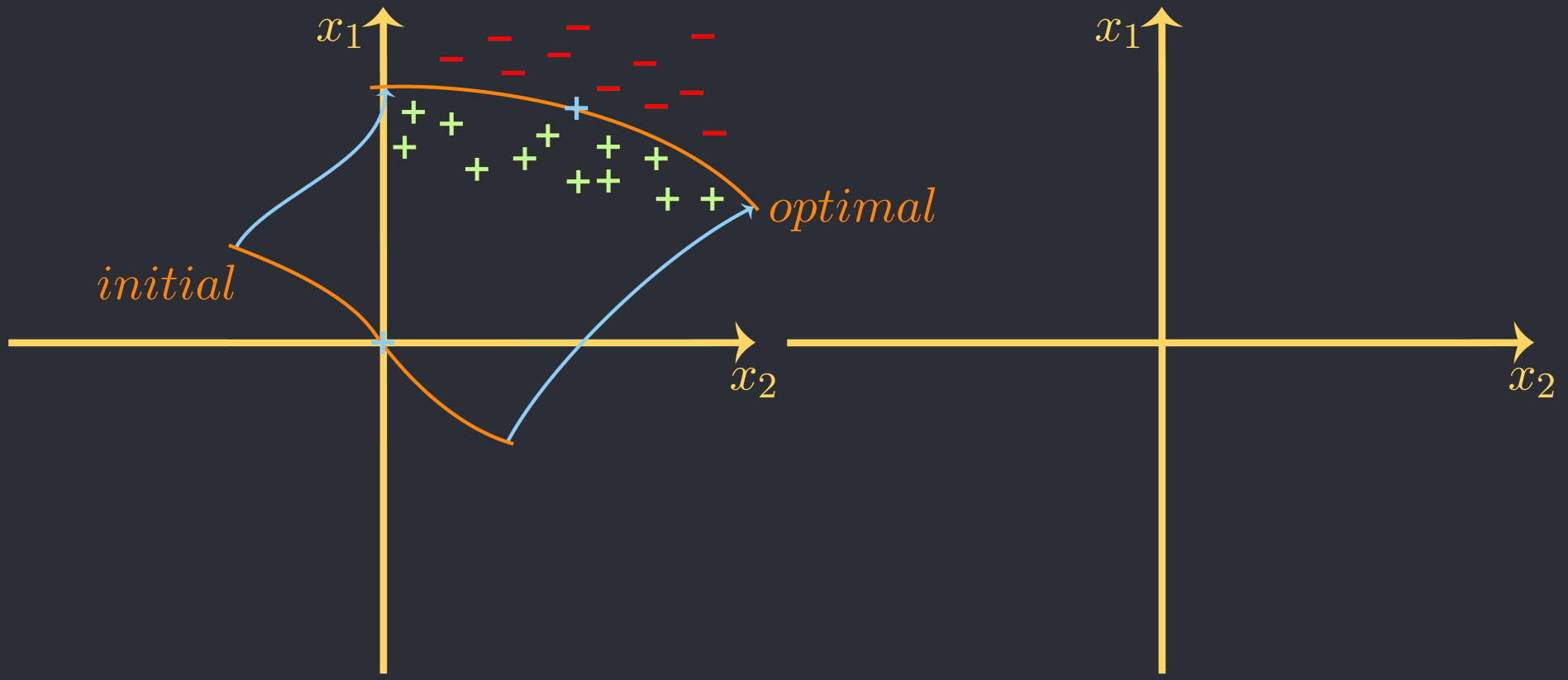
Центрирование данных

$$p = \sigma(w_1 x_1 + w_2 x_2 + b)$$



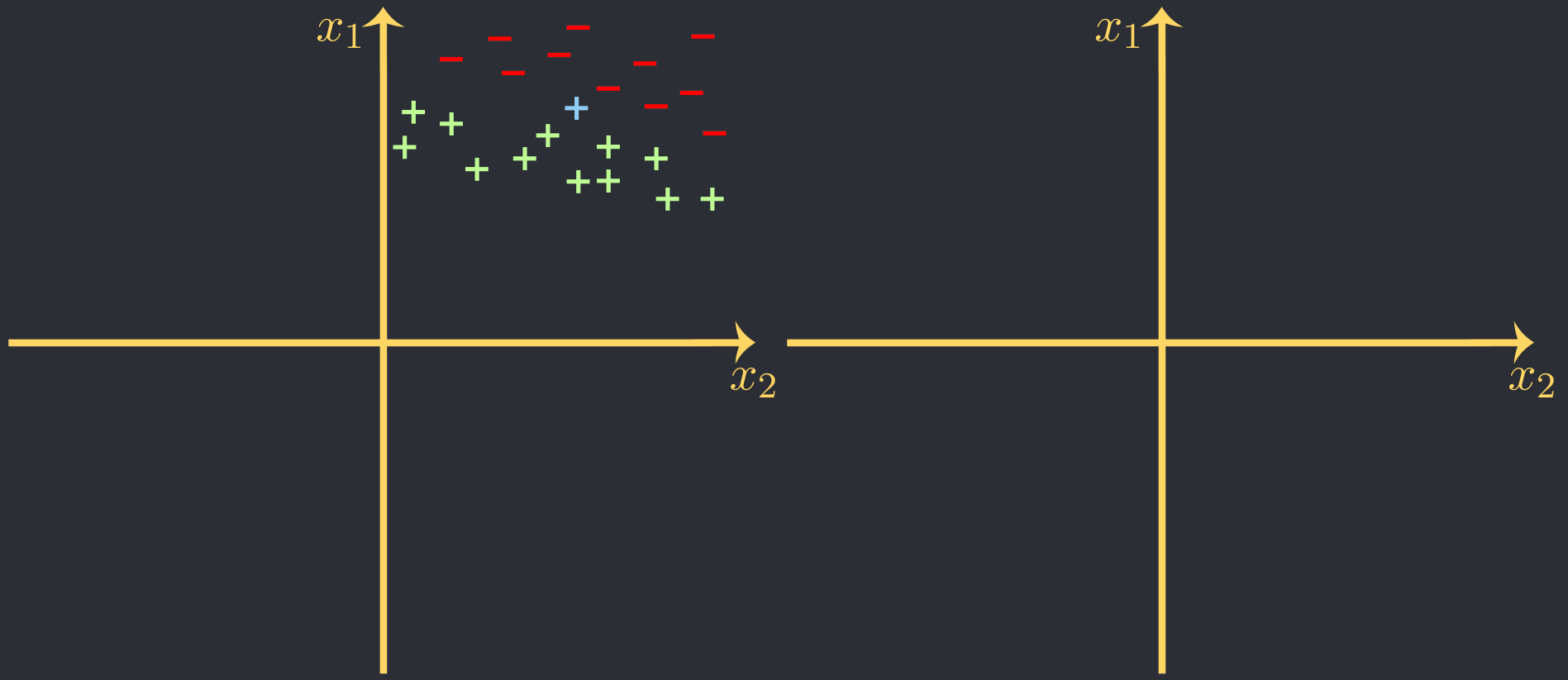
Центрирование данных

$$p = \sigma(w_1 x_1 + w_2 x_2 + b)$$



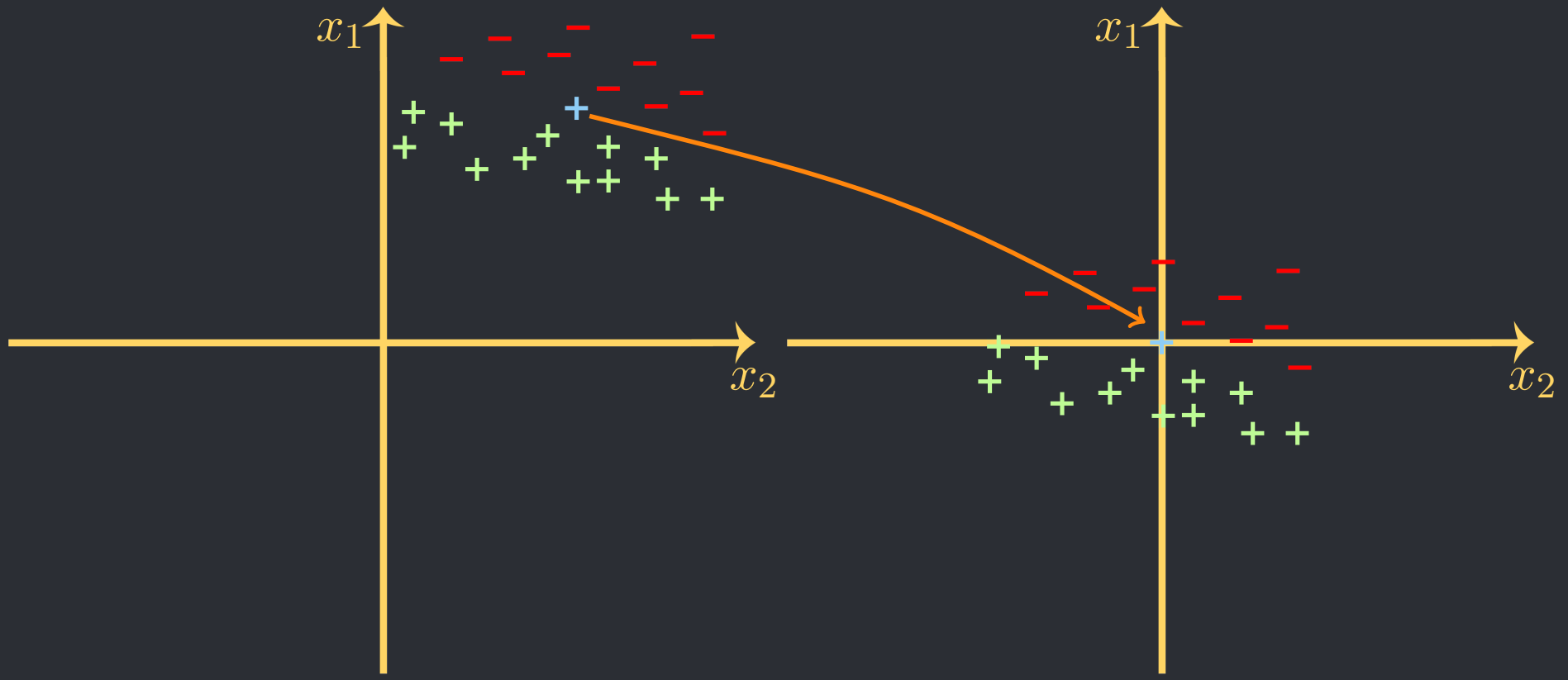
Центрирование данных

$$p = \sigma(w_1 x_1 + w_2 x_2 + b)$$



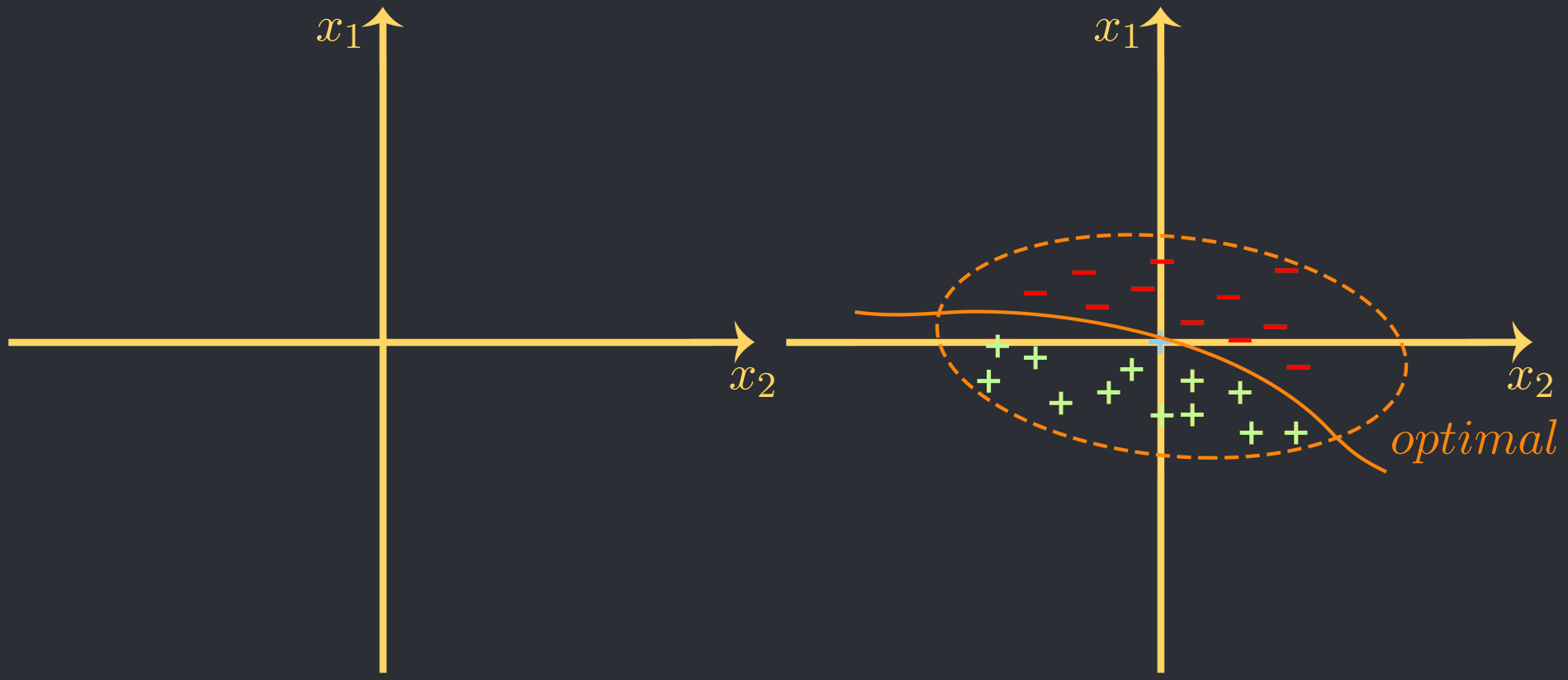
Центрирование данных

$$p = \sigma(w_1 x_1 + w_2 x_2 + b)$$



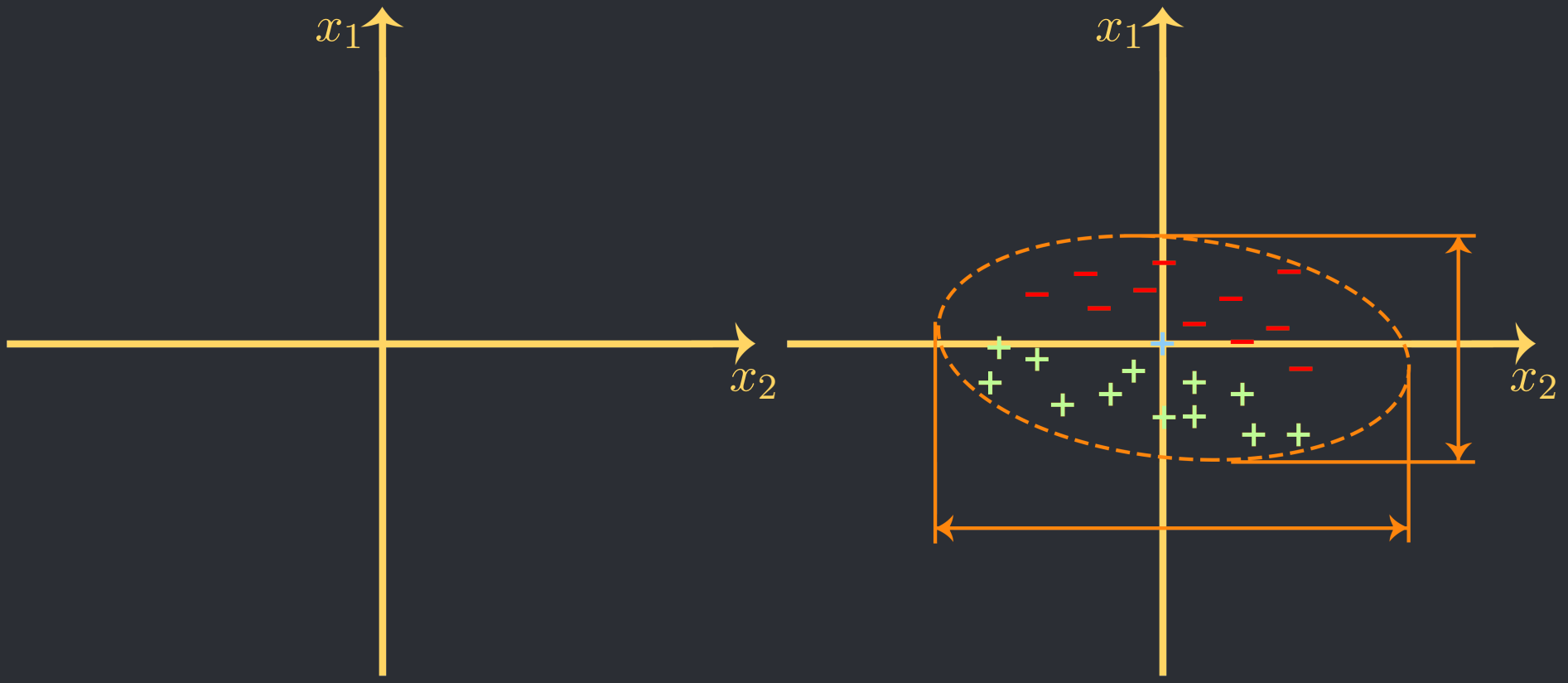
Нормализация данных

$$p = \sigma(w_1 x_1 + w_2 x_2 + b)$$



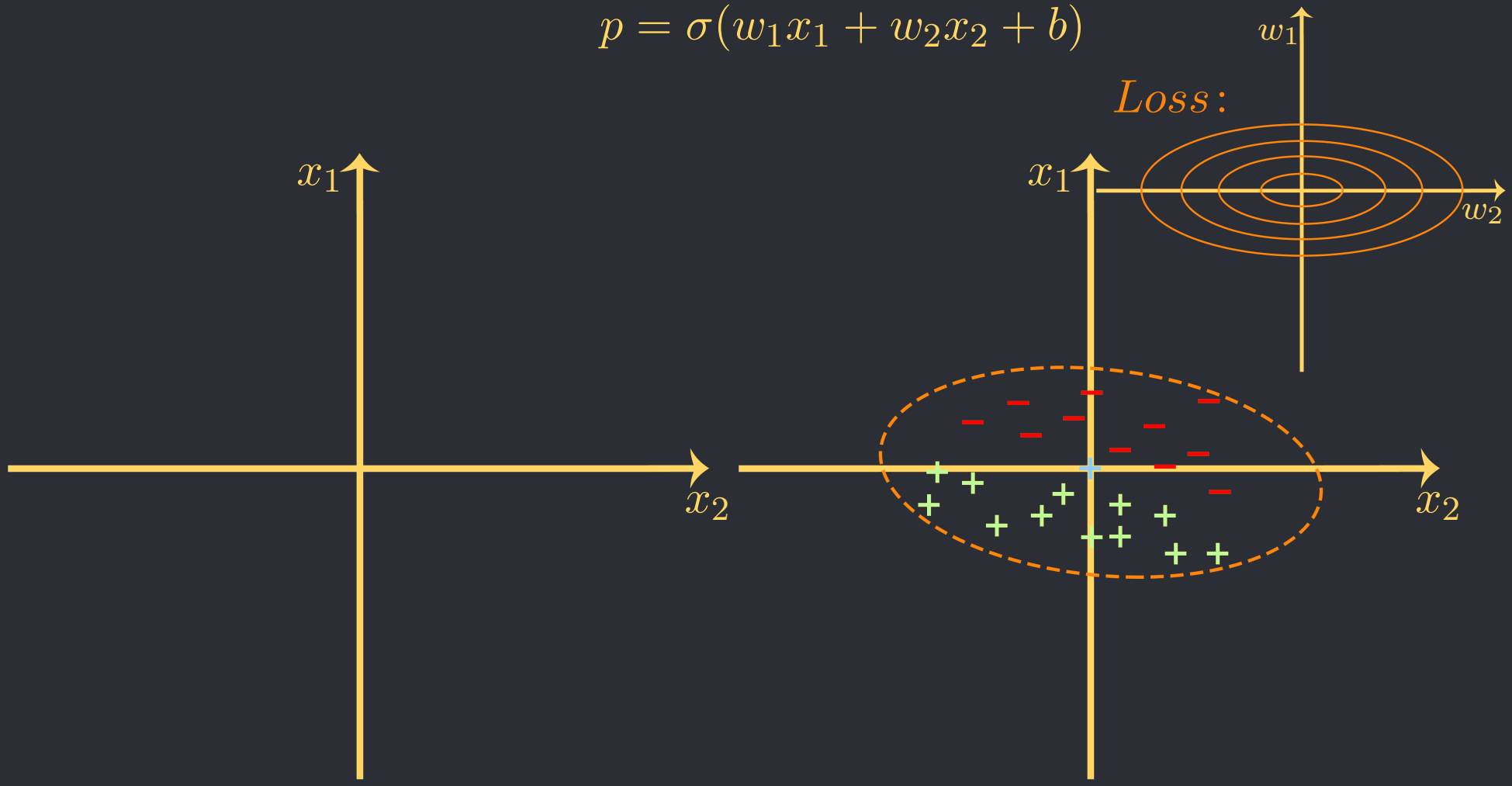
Нормализация данных

$$p = \sigma(w_1 x_1 + w_2 x_2 + b)$$

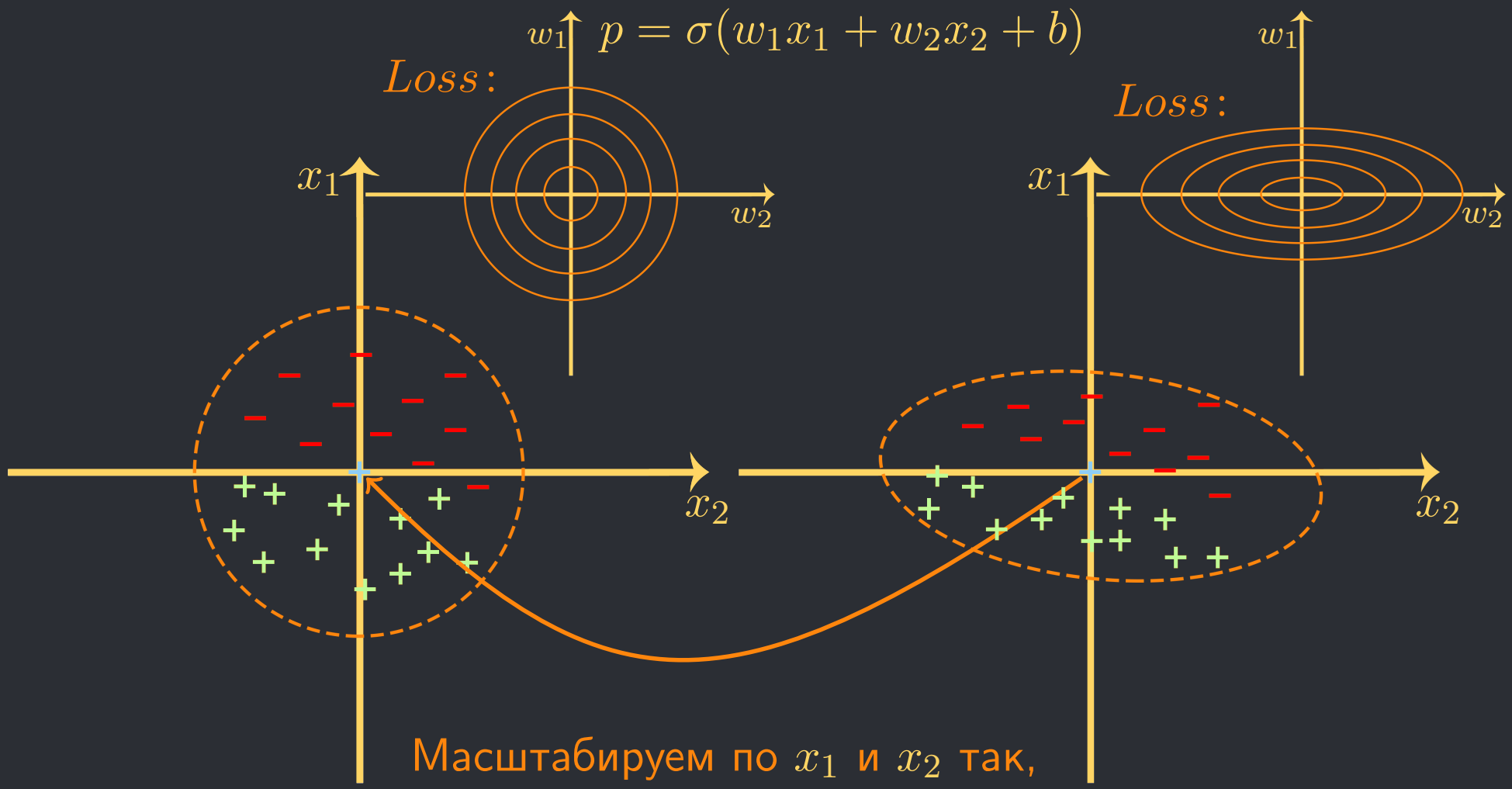


Нормализация данных

$$p = \sigma(w_1 x_1 + w_2 x_2 + b)$$



Нормализация данных

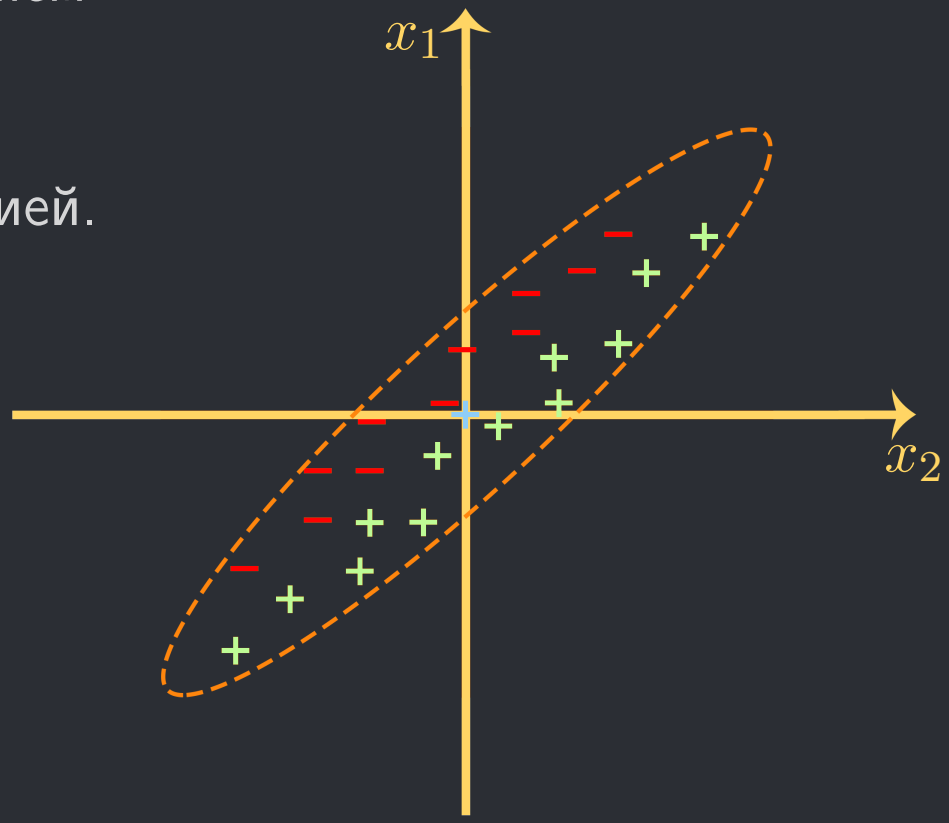


Масштабируем по x_1 и x_2 так,
чтобы в среднем x отклонялись от 0 на 1

Нормализация данных

$$p = \sigma(w_1 x_1 + w_2 x_2 + b)$$

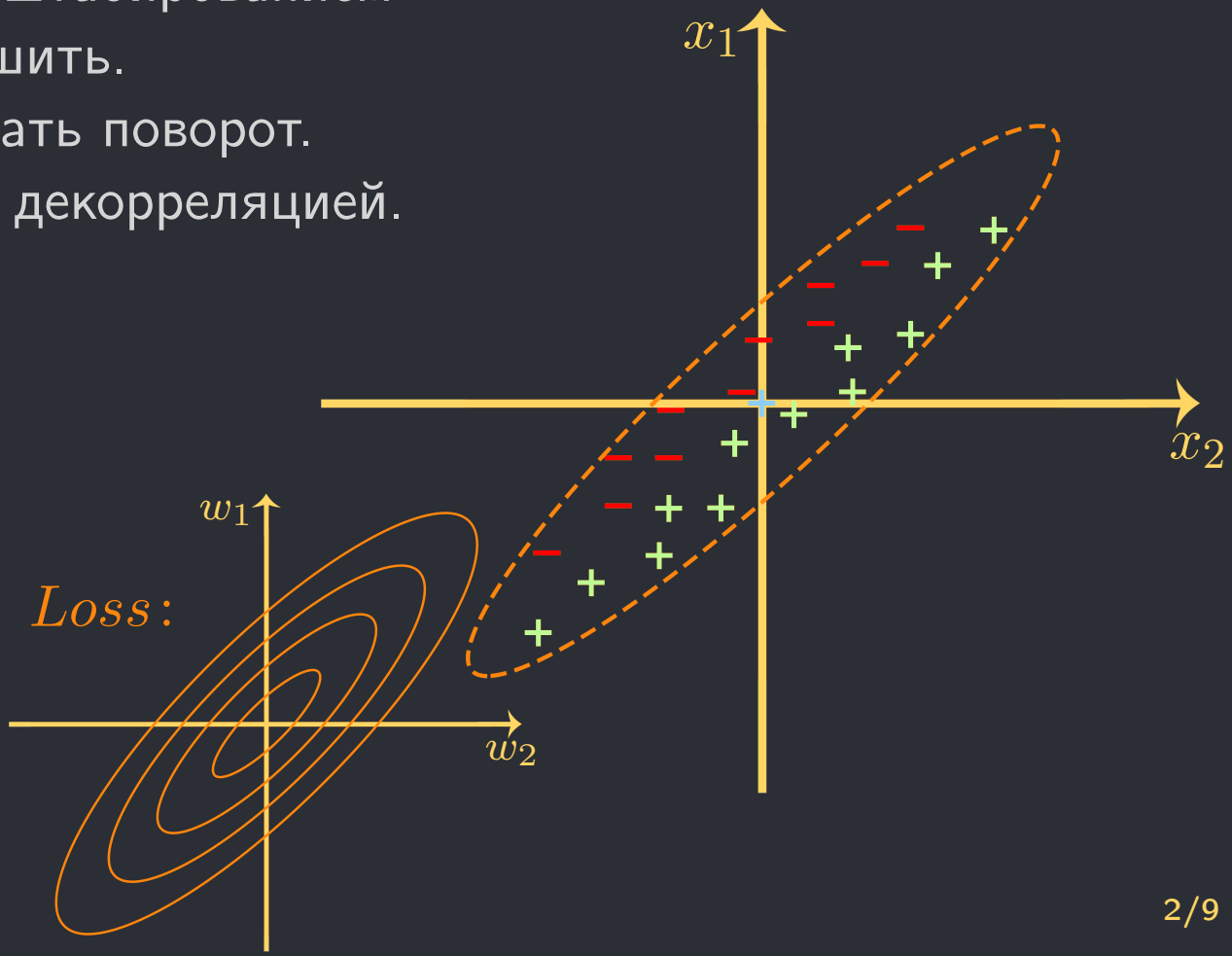
Такую ситуацию масштабированием по x_1 и x_2 не разрешить.
Нужно сначала сделать поворот.
Поворот называется декорреляцией.



Нормализация данных

$$p = \sigma(w_1 x_1 + w_2 x_2 + b)$$

Такую ситуацию масштабированием по x_1 и x_2 не разрешить.
Нужно сначала сделать поворот.
Поворот называется декорреляцией.



Нормализация данных

$$\widetilde{\mathbf{x}} = \frac{\mathbf{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}$$

Нормализация данных

$$\widetilde{\mathbf{x}} = \frac{\mathbf{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}$$

где

$$\boldsymbol{\mu} = \frac{\sum\limits_{i} \mathbf{x}_i}{N}$$

Нормализация данных

$$\widetilde{\mathbf{x}} = \frac{\mathbf{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}$$

где

$$\boldsymbol{\mu} = \frac{\sum\limits_i \mathbf{x}_i}{N}$$

$$\boldsymbol{\sigma} = \sqrt{\frac{\sum\limits_i \left(\mathbf{x}_i - \boldsymbol{\mu}_i\right)^2}{N-1}}$$

Нормализация данных

$$\widetilde{\mathbf{x}} = \frac{\mathbf{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}$$

где

$$\boldsymbol{\mu} = \frac{\sum\limits_i \mathbf{x}_i}{N}$$

$$\boldsymbol{\sigma} = \sqrt{\frac{\sum\limits_i (\mathbf{x}_i - \boldsymbol{\mu}_i)^2}{N-1}}$$

Так происходит нормализация данных

BatchNorm

В нейронных сетях мы можем нормализовать данные после каждого слоя для каждого батча

$$\widetilde{\mathbf{x}} = \frac{\mathbf{x} - \boldsymbol{\mu}_b}{\boldsymbol{\sigma}_b} \cdot \gamma + \beta$$

BatchNorm

В нейронных сетях мы можем нормализовать данные после каждого слоя для каждого батча

$$\widetilde{\mathbf{x}} = \frac{\mathbf{x} - \boldsymbol{\mu}_b}{\boldsymbol{\sigma}_b} \cdot \gamma + \beta$$

где

$$\boldsymbol{\mu}_b = \frac{\sum_{i}^{N_b} \mathbf{x}_i}{N_b}$$

BatchNorm

В нейронных сетях мы можем нормализовать данные после каждого слоя для каждого батча

$$\widetilde{\mathbf{x}} = \frac{\mathbf{x} - \boldsymbol{\mu}_b}{\boldsymbol{\sigma}_b} \cdot \gamma + \beta$$

где

$$\boldsymbol{\mu}_b = \frac{\sum_i^{N_b} \mathbf{x}_i}{N_b}$$

$$\boldsymbol{\sigma}_b = \sqrt{\frac{\sum_i^{N_b} (\mathbf{x}_i - \boldsymbol{\mu}_b)^2}{N_b - 1}}$$

BatchNorm

В нейронных сетях мы можем нормализовать данные после каждого слоя для каждого батча

$$\widetilde{\mathbf{x}} = \frac{\mathbf{x} - \boldsymbol{\mu}_b}{\boldsymbol{\sigma}_b} \cdot \gamma + \beta$$

где

$$\boldsymbol{\mu}_b = \frac{\sum_i^{N_b} \mathbf{x}_i}{N_b}$$

$$\boldsymbol{\sigma}_b = \sqrt{\frac{\sum_i^{N_b} (\mathbf{x}_i - \boldsymbol{\mu}_b)^2}{N_b - 1}}$$

Inference: $\hat{\boldsymbol{\mu}} = EMA\boldsymbol{\mu}_b$; $\hat{\boldsymbol{\sigma}} = EMA\boldsymbol{\sigma}_b$

BatchNorm

В нейронных сетях мы можем нормализовать данные после каждого слоя для каждого батча

$$\widetilde{\mathbf{x}} = \frac{\mathbf{x} - \boldsymbol{\mu}_b}{\boldsymbol{\sigma}_b} \cdot \gamma + \beta$$

где

$$\boldsymbol{\mu}_b = \frac{\sum_i^{N_b} \mathbf{x}_i}{N_b}$$

$$\boldsymbol{\sigma}_b = \sqrt{\frac{\sum_i^{N_b} (\mathbf{x}_i - \boldsymbol{\mu}_b)^2}{N_b - 1}}$$

На стадии тренировки оцениваем по батчу

На стадии валидации оцениваем – по истории

Inference: $\hat{\boldsymbol{\mu}} = EMA\boldsymbol{\mu}_b$; $\hat{\boldsymbol{\sigma}} = EMA\boldsymbol{\sigma}_b$

BatchRenorm

- BatchNorm вычисляет статистики по батчу

BatchRenorm

- BatchNorm вычисляет статистики по батчу
- Чтобы статистики были репрезентативными, нужно, чтобы батч был достаточно большим (16 изображений)

BatchRenorm

- BatchNorm вычисляет статистики по батчу
- Чтобы статистики были репрезентативными, нужно, чтобы батч был достаточно большим (16 изображений)
- Если батч маленький, то статистики будут нерепрезентативными

BatchRenorm

- BatchNorm вычисляет статистики по батчу
- Чтобы статистики были репрезентативными, нужно, чтобы батч был достаточно большим (16 изображений)
- Если батч маленький, то статистики будут нерепрезентативными
- Можно считать статистики по нескольким батчам

BatchRenorm

$$\mu_b \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$$

- BatchNorm вычисляет статистики по батчу
- Чтобы статистики были репрезентативными, нужно, чтобы батч был достаточно большим (16 изображений)
- Если батч маленький, то статистики будут нерепрезентативными
- Можно считать статистики по нескольким батчам

BatchRenorm

- BatchNorm вычисляет статистики по батчу
- Чтобы статистики были репрезентативными, нужно, чтобы батч был достаточно большим (16 изображений)
- Если батч маленький, то статистики будут нерепрезентативными
- Можно считать статистики по нескольким батчам

$$\mu_b \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\sigma_b \leftarrow \sqrt{\epsilon + \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_b)^2}$$

BatchRenorm

- BatchNorm вычисляет статистики по батчу
- Чтобы статистики были репрезентативными, нужно, чтобы батч был достаточно большим (16 изображений)
- Если батч маленький, то статистики будут нерепрезентативными
- Можно считать статистики по нескольким батчам

$$\mu_b \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\sigma_b \leftarrow \sqrt{\epsilon + \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_b)^2}$$

$$r \leftarrow stop\_gradient\left(clip_{[1/r_{max}, r_{max}]}\left(\frac{\sigma_b}{\sigma}\right)\right)$$

BatchRenorm

- BatchNorm вычисляет статистики по батчу
- Чтобы статистики были репрезентативными, нужно, чтобы батч был достаточно большим (16 изображений)
- Если батч маленький, то статистики будут нерепрезентативными
- Можно считать статистики по нескольким батчам

$$\mu_b \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\sigma_b \leftarrow \sqrt{\epsilon + \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_b)^2}$$

$$r \leftarrow stop\_gradient\left(clip_{[1/r_{max}, r_{max}]}\left(\frac{\sigma_b}{\sigma}\right)\right)$$

$$d \leftarrow stop\_gradient\left(clip_{[-d_{max}, d_{max}]}\left(\frac{\mu_b - \mu}{\sigma}\right)\right)$$

BatchRenorm

- BatchNorm вычисляет статистики по батчу
- Чтобы статистики были репрезентативными, нужно, чтобы батч был достаточно большим (16 изображений)
- Если батч маленький, то статистики будут нерепрезентативными
- Можно считать статистики по нескольким батчам

$$\mu_b \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\sigma_b \leftarrow \sqrt{\epsilon + \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_b)^2}$$

$$r \leftarrow stop\_gradient\left(clip_{[1/r_{max}, r_{max}]}\left(\frac{\sigma_b}{\sigma}\right)\right)$$

$$d \leftarrow stop\_gradient\left(clip_{[-d_{max}, d_{max}]}\left(\frac{\mu_b - \mu}{\sigma}\right)\right)$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_b}{\sigma_b} \cdot r + d$$

$$y_i \leftarrow \gamma \hat{x}_i + \beta$$

BatchRenorm

- BatchNorm вычисляет статистики по батчу
- Чтобы статистики были репрезентативными, нужно, чтобы батч был достаточно большим (16 изображений)
- Если батч маленький, то статистики будут нерепрезентативными
- Можно считать статистики по нескольким батчам

$$\mu_b \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\sigma_b \leftarrow \sqrt{\epsilon + \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_b)^2}$$

$$r \leftarrow stop\_gradient\left(clip_{[1/r_{max}, r_{max}]}\left(\frac{\sigma_b}{\sigma}\right)\right)$$

$$d \leftarrow stop\_gradient\left(clip_{[-d_{max}, d_{max}]}\left(\frac{\mu_b - \mu}{\sigma}\right)\right)$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_b}{\sigma_b} \cdot r + d$$

$$y_i \leftarrow \gamma \hat{x}_i + \beta$$

$$\mu := \mu + \alpha(\mu_b - \mu)$$

$$\sigma := \sigma + \alpha(\sigma_b - \sigma)$$

BatchRenorm

- BatchNorm вычисляет статистики по батчу
- Чтобы статистики были репрезентативными, нужно, чтобы батч был достаточно большим (16 изображений)
- Если батч маленький, то статистики будут нерепрезентативными
- Можно считать статистики по нескольким батчам

$$\mu_b \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\sigma_b \leftarrow \sqrt{\epsilon + \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_b)^2}$$

$$r \leftarrow stop\_gradient\left(clip_{[1/r_{max}, r_{max}]}\left(\frac{\sigma_b}{\sigma}\right)\right)$$

$$d \leftarrow stop\_gradient\left(clip_{[-d_{max}, d_{max}]}\left(\frac{\mu_b - \mu}{\sigma}\right)\right)$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_b}{\sigma_b} \cdot r + d$$

$$y_i \leftarrow \gamma \hat{x}_i + \beta$$

$$\mu := \mu + \alpha(\mu_b - \mu)$$

$$\sigma := \sigma + \alpha(\sigma_b - \sigma)$$

Inference: $y \leftarrow \gamma \cdot \frac{x - \mu}{\sigma} + \beta$

InstanceNorm

- Мы используем статистики к каждой картинке по некоторому батчу

InstanceNorm

- Мы используем статистики к каждой картинке по некоторому батчу
- Кажется, имеет смысл применять статистики к картинке, рассчитанные только на одной картинке

InstanceNorm

- Мы используем статистики к каждой картинке по некоторому батчу
- Кажется, имеет смысл применять статистики к картинке, рассчитанные только на одной картинке
- В режиме inference работает так же, как в режиме training

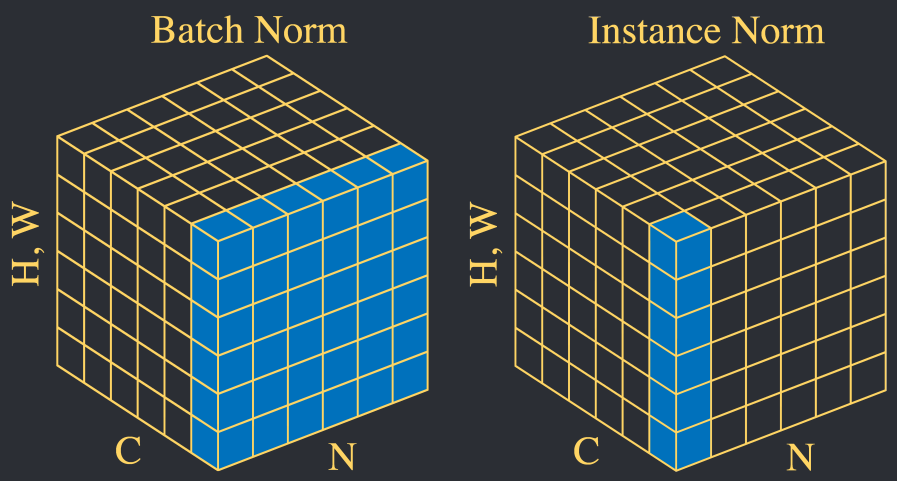
InstanceNorm

- Мы используем статистики к каждой картинке по некоторому батчу
- Кажется, имеет смысл применять статистики к картинке, рассчитанные только на одной картинке
- В режиме inference работает так же, как в режиме training
- Работает не очень, но спасает, если хватает ресурсов только на обучение с маленьким батчем

InstanceNorm

- Мы используем статистики к каждой картинке по некоторому батчу
- Кажется, имеет смысл применять статистики к картинке, рассчитанные только на одной картинке
- В режиме inference работает так же, как в режиме training

- Работает не очень, но спасает, если хватает ресурсов только на обучение с маленьким батчем



GroupNorm

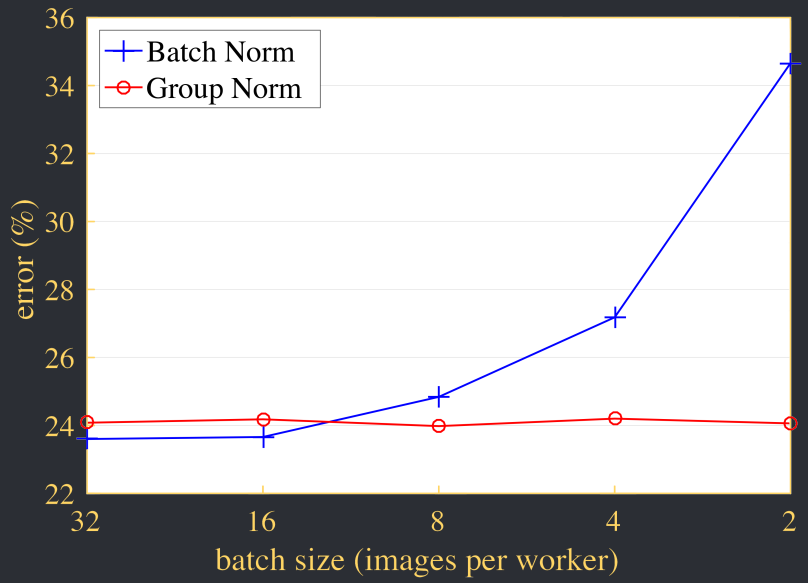
- Вместо того, чтобы считать статистики только по одному каналу (как в InstanceNorm), будем считать статистики по группам каналов

GroupNorm

- Вместо того, чтобы считать статистики только по одному каналу (как в InstanceNorm), будем считать статистики по группам каналов
- Работает значительно лучше, чем InstanceNorm

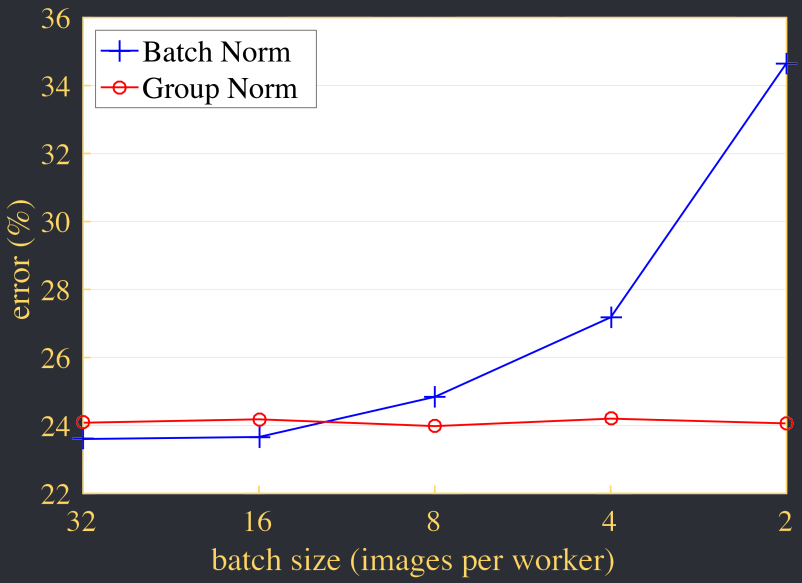
GroupNorm

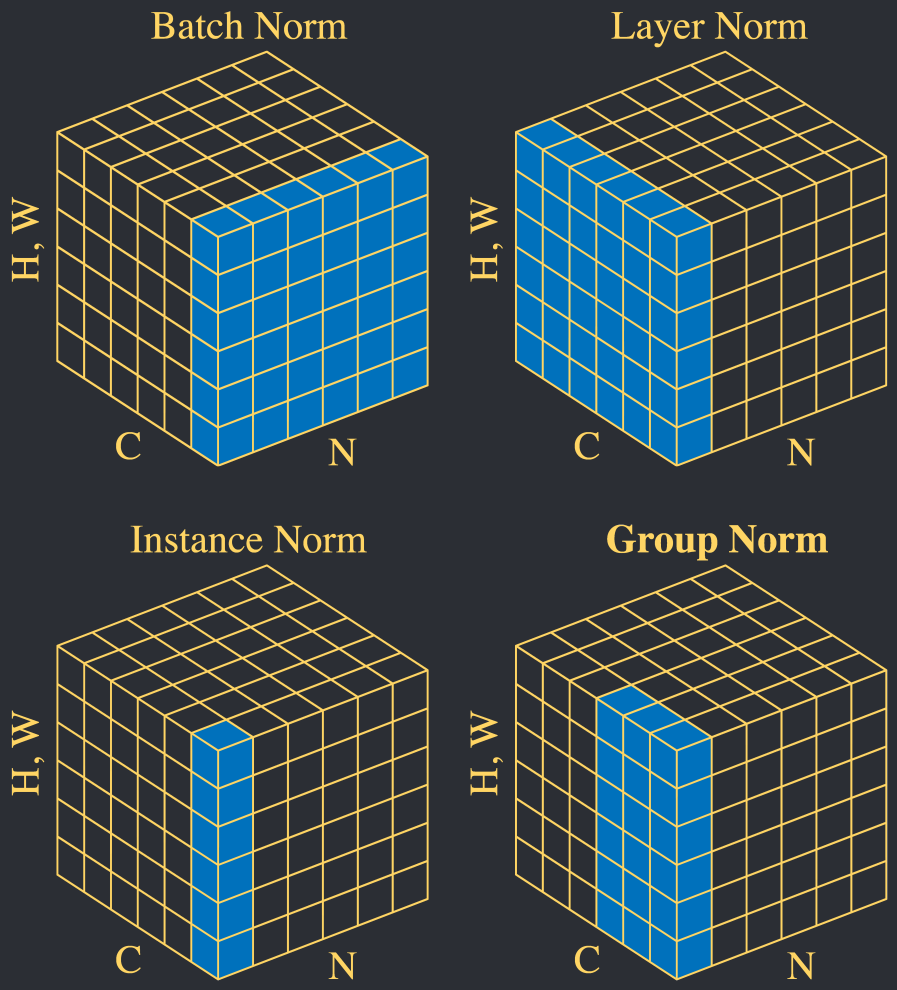
- Вместо того, чтобы считать статистики только по одному каналу (как в InstanceNorm), будем считать статистики по группам каналов
- Работает значительно лучше, чем InstanceNorm



GroupNorm

- Вместо того, чтобы считать статистики только по одному каналу (как в InstanceNorm), будем считать статистики по группам каналов
- Работает значительно лучше, чем InstanceNorm





Filter Responce Normalization

- Хоть GroupNorm и дает некоторый выигрыш, принцип его работы достаточно странный

Filter Responce Normalization

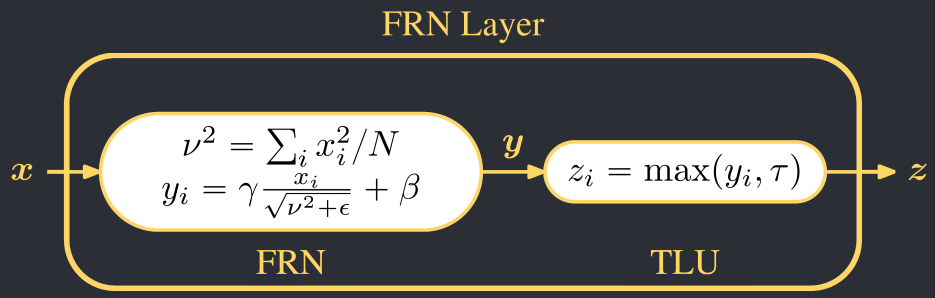
- Хоть GroupNorm и дает некоторый выигрыш, принцип его работы достаточно странный
- Вернемся к идее о том, что нормализовать нужно каждый канал отдельно

Filter Responce Normalization

- Хоть GroupNorm и дает некоторый выигрыш, принцип его работы достаточно странный
- Вернемся к идее о том, что нормализовать нужно каждый канал отдельно
- Попробуем просто не делать центровку

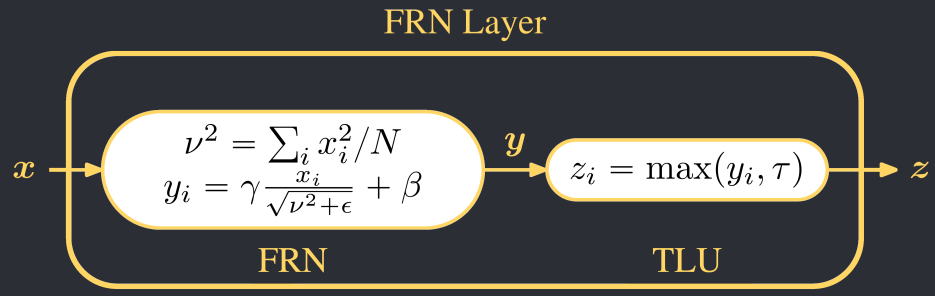
Filter Responce Normalization

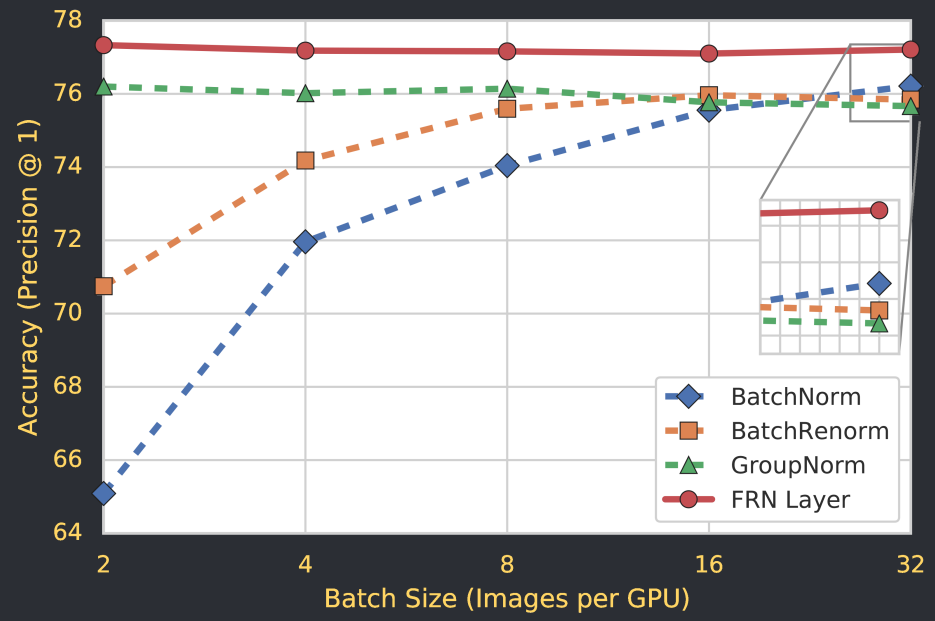
- Хоть GroupNorm и дает некоторый выигрыш, принцип его работы достаточно странный
- Вернемся к идее о том, что нормализовать нужно каждый канал отдельно
- Попробуем просто не делать центровку



Filter Responce Normalization

- Хоть GroupNorm и дает некоторый выигрыш, принцип его работы достаточно странный
- Вернемся к идее о том, что нормализовать нужно каждый канал отдельно
- Попробуем просто не делать центровку





Итог

- Центрирование данных

Итог

- Центрирование данных
- Нормализация данных

Итог

- Центрирование данных
- Нормализация данных
- BatchNorm

Итог

- Центрирование данных
- Нормализация данных
- BatchNorm
- Batch Renormalization

Итог

- Центрирование данных
- Нормализация данных
- BatchNorm
- Batch Renormalization
- Instance Normalization

Итог

- Центрирование данных
- Нормализация данных
- BatchNorm
- Batch Renormalization
- Instance Normalization
- Group Normalization

Итог

- Центрирование данных
- Нормализация данных
- BatchNorm
- Batch Renormalization
- Instance Normalization
- Group Normalization
- Filter Responce Normalization