ВЛИЯНИЕ МОМЕНТА В АЛГОРИТМАХ ОПТИМИЗАЦИИ

РАЗБОР РАБОТЫ GOH G. WHY MOMENTUM REALLY WORKS //DISTILL. – 2017. – Т. 2. – №. 4. – С. E6.

Докладчик: Денишева Рушана

24.04.23, MIPT

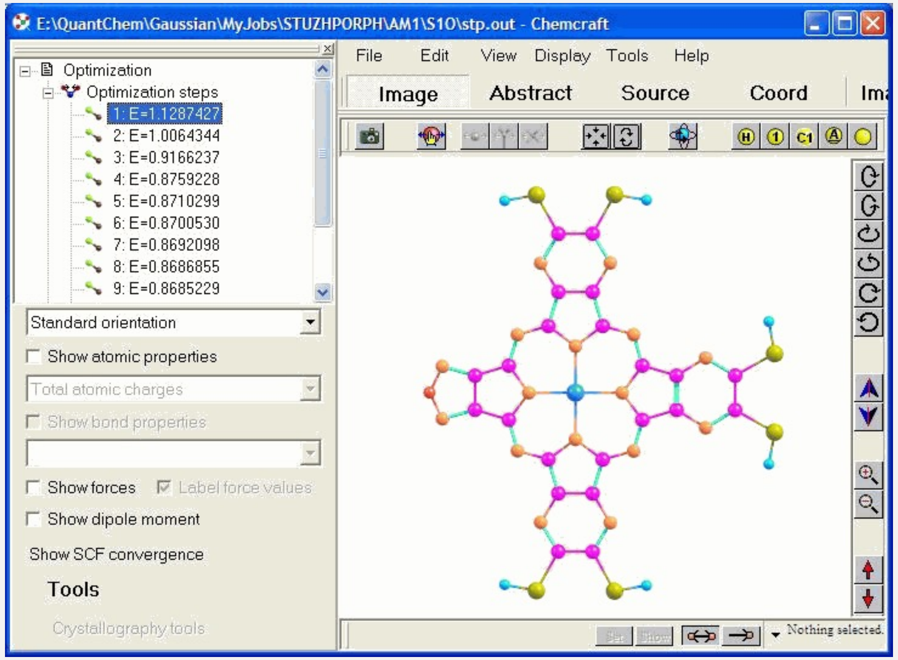
СОДЕРЖАНИЕ

- Описание задачи оптимизации
  - Описание метода градиентного спуска
- Описание решения задачи путем добавления момента
- Преимущества градиентного спуска с моментом
- Сильные и слабые стороны метода

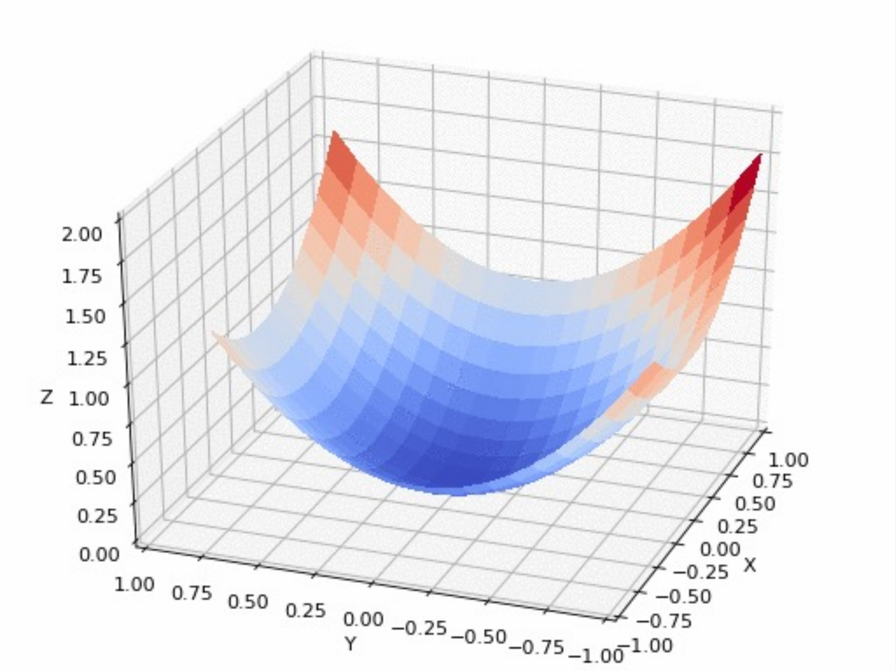
ЗАДАЧА ОПТИМИЗАЦИИ

Задача оптимизации – это задача нахождения экстремума целевой функции с учетом ограничений на управляемые переменные.

ЗАДАЧА ОПТИМИЗАЦИИ



Оптимизация в медицине

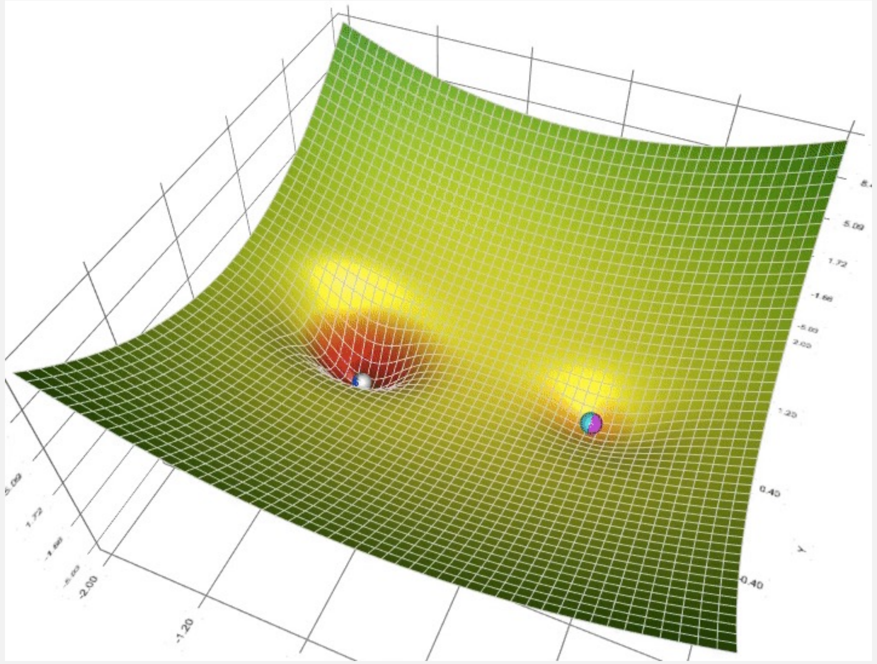


Оптимизация функции потерь

https://www.chemcraftprog.com/ https://blog.paperspace.com/intro-to-optimization-in-deep-learning-gradient-descent/

ГРАДИЕНТНЫЙ СПУСК

Градиентный спуск – численный метод нахождения *локального* минимума или максимума функции с помощью движения вдоль градиента, один из основных численных методов современной оптимизации.

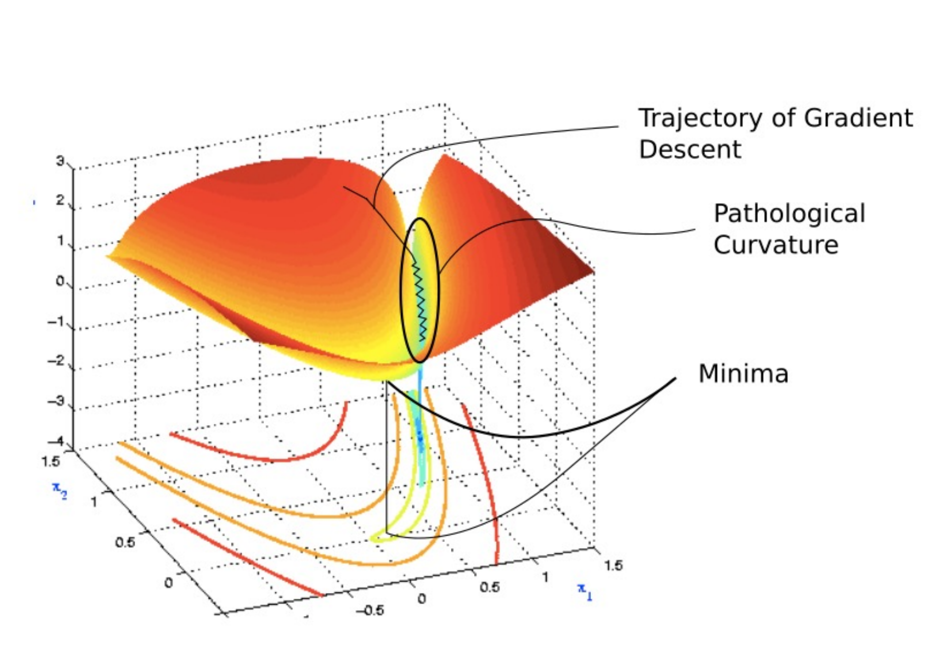


Анимация 5 методов градиентного спуска на поверхности: градиентный спуск (голубой), импульс (пурпурный), AdaGrad (белый), RMSProp (зеленый), Adam (синий). Левая лунка - глобальный минимум; правая лунка - локальный минимум.

https://towardsdatascience.com/

ПРОБЛЕМА

- “Застрять” в локальном минимуме, седле целевой функции или патологической кривизне



https://blog.paperspace.com/intro-to-optimization-momentum-rmsprop-adam/

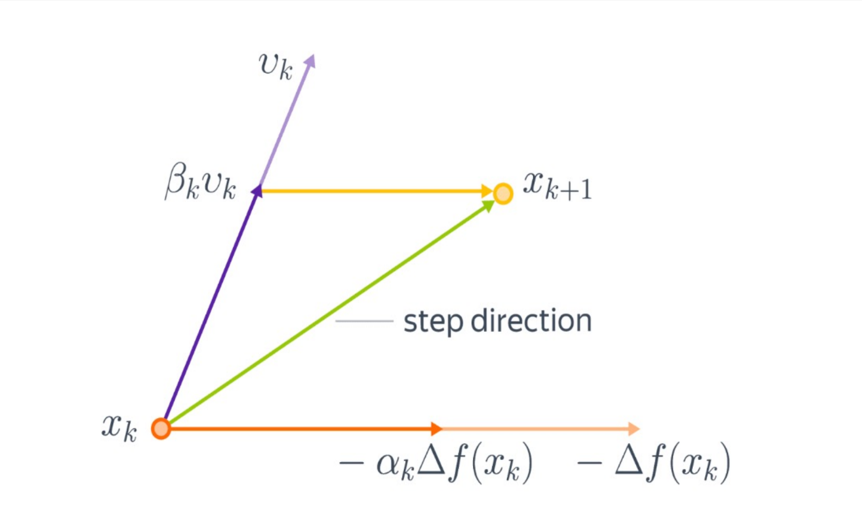
РЕШЕНИЕ – МЕТОД ИНЕРЦИИ

$$x_{k+1} = x_k - \alpha \nabla f(x_k),$$

Градиентный спуск

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

Градиентный спуск с добавлением “момента”



Momentum

ПРЕИМУЩЕСТВА МОМЕНТА

- Позволяет использовать больший диапазон размеров шага и создает собственные колебания
- Дает квадратичное ускорение на многих функциях

ОЦЕНКА СКОРОСТИ СХОДИМОСТИ К МИНИМУМУ И ШАГА

ВЫПУКЛАЯ КВАДРАТИЧНАЯ ФУНКЦИЯ

Без момента

$$f(w) = \tfrac{1}{2} w^T A w - b^T w, \qquad w \in \mathbf{R}^n.$$

$$A = Q \operatorname{diag}(\lambda_1, \ldots, \lambda_n)\, Q^T, \qquad Q = [q_1, \ldots, q_n],$$

$$f(w^k) - f(w^\star) = \sum (1 - \alpha\lambda_i)^{2k} \lambda_i [x_i^0]^2$$

С моментом

$$\begin{aligned} z^{k+1} &= \beta z^k + (A w^k - b) \\ w^{k+1} &= w^k - \alpha z^{k+1}. \end{aligned}$$

$$\begin{pmatrix} y_i^k \\ x_i^k \end{pmatrix} = R^k \begin{pmatrix} y_i^0 \\ x_i^0 \end{pmatrix} \qquad R = \begin{pmatrix} \beta & \lambda_i \\ -\alpha\beta & 1 - \alpha\lambda_i \end{pmatrix}.$$

где $x^k = Q(w^k - w^\star)$ **и** $y^k = Q z^k$

ОЦЕНКА СКОРОСТИ СХОДИМОСТИ К МИНИМУМУ И ШАГА

ВЫПУКЛАЯ КВАДРАТИЧНАЯ ФУНКЦИЯ

Без момента

$$0 < \alpha\lambda_i < 2.$$

С моментом

$$0 < \alpha\lambda_i < 2 + 2\beta \qquad \text{for} \qquad 0 \leq \beta < 1$$

Момент позволяет увеличить размер шага в 2 раза перед расхождением!

ОЦЕНКА СКОРОСТИ СХОДИМОСТИ К МИНИМУМУ И ШАГА

ВЫПУКЛАЯ КВАДРАТИЧНАЯ ФУНКЦИЯ

Без момента

$$\text{optimal } \alpha \;=\; \underset{\alpha}{\text{argmin}} \text{ rate}(\alpha) \;=\; \frac{2}{\lambda_1 + \lambda_n}$$

$$\text{optimal rate} \;=\; \min_{\alpha} \text{ rate}(\alpha) \;=\; \frac{\lambda_n/\lambda_1 - 1}{\lambda_n/\lambda_1 + 1}$$

$$\text{condition number} := \kappa := \frac{\lambda_n}{\lambda_1}$$

С моментом

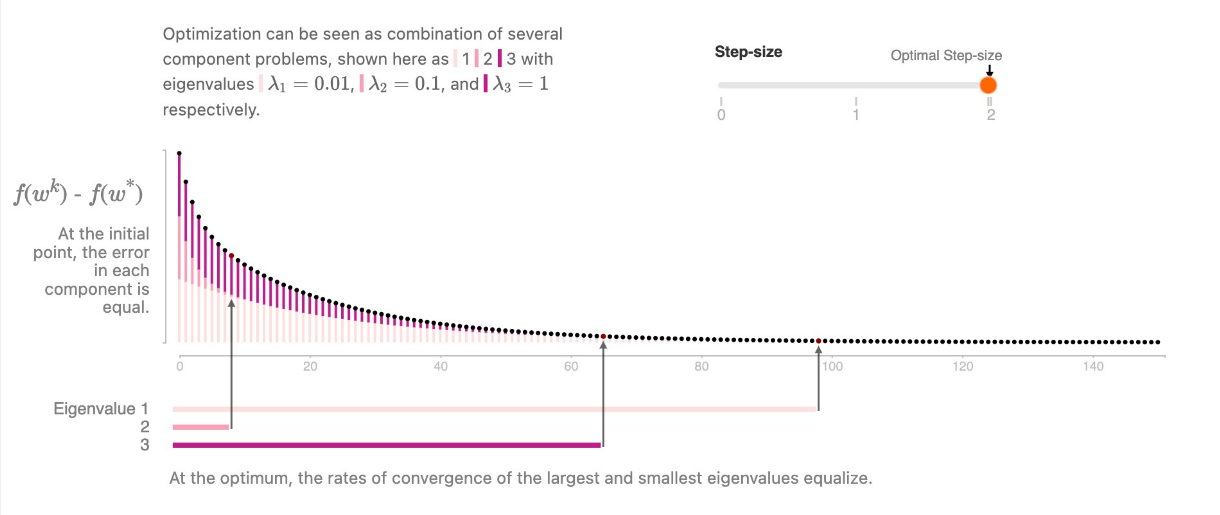
$$\alpha = \left(\frac{2}{\sqrt{\lambda_1} + \sqrt{\lambda_n}}\right)^2 \quad \beta = \left(\frac{\sqrt{\lambda_n} - \sqrt{\lambda_1}}{\sqrt{\lambda_n} + \sqrt{\lambda_1}}\right)^2$$

$$\text{optimal rate} \;=\; \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}$$

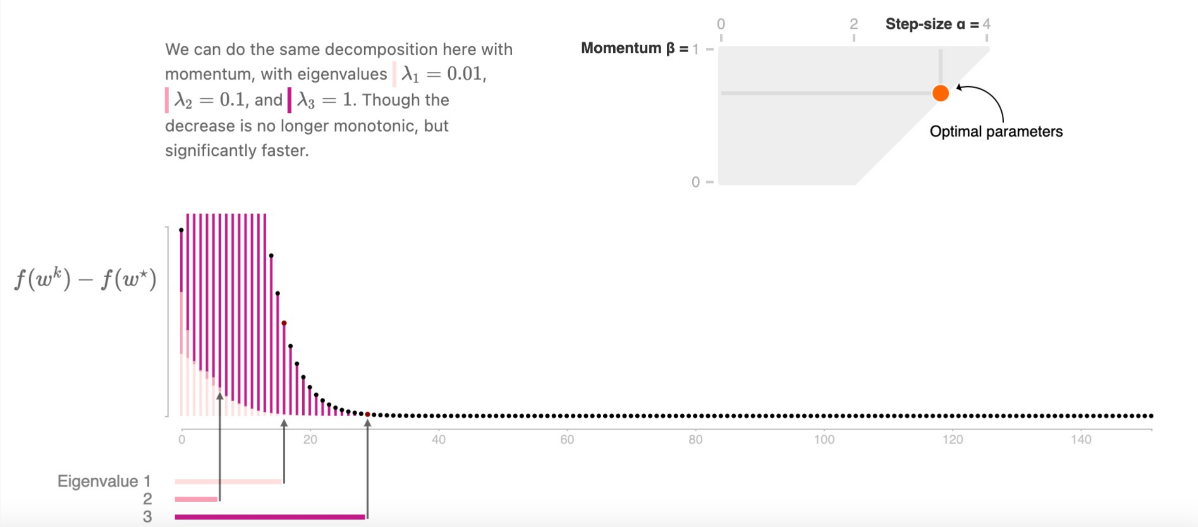
Момент дал квадратичное ускорение на нашей функции!

ОЦЕНКА СКОРОСТИ СХОДИМОСТИ К МИНИМУМУ И ШАГА

ВЫПУКЛАЯ КВАДРАТИЧНАЯ ФУНКЦИЯ



Без момента



С моментом

ПРИМЕР: ПРОБЛЕМА КОЛОРИЗАЦИИ

G - граф с вершинами в виде пикселей,

E - множество ребер, соединяющих каждый пиксель с четырьмя соседними пикселями,

D - небольшое множество из нескольких выделенных вершин

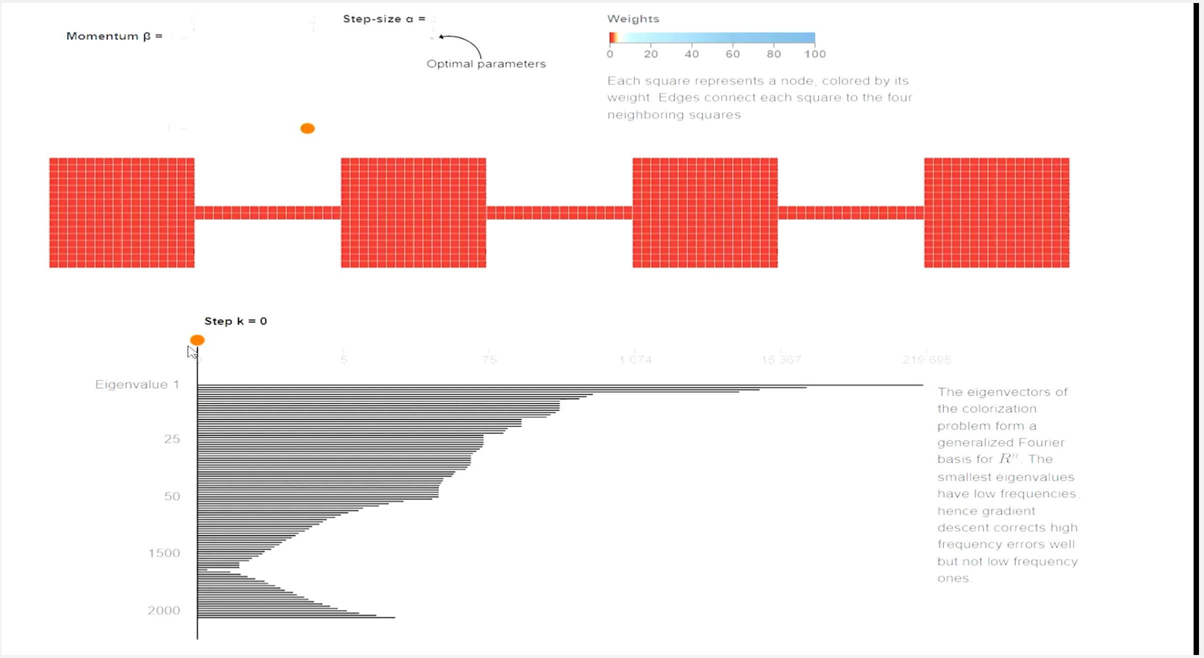
$$\text{minimize} \quad \frac{1}{2}\sum_{i \in D}(w_i - 1)^2 \; + \; \frac{1}{2}\sum_{i,j \in E}(w_i - w_j)^2.$$

The **colorizer** pulls distinguished pixels towards 1

The **smoother** spreads out the color

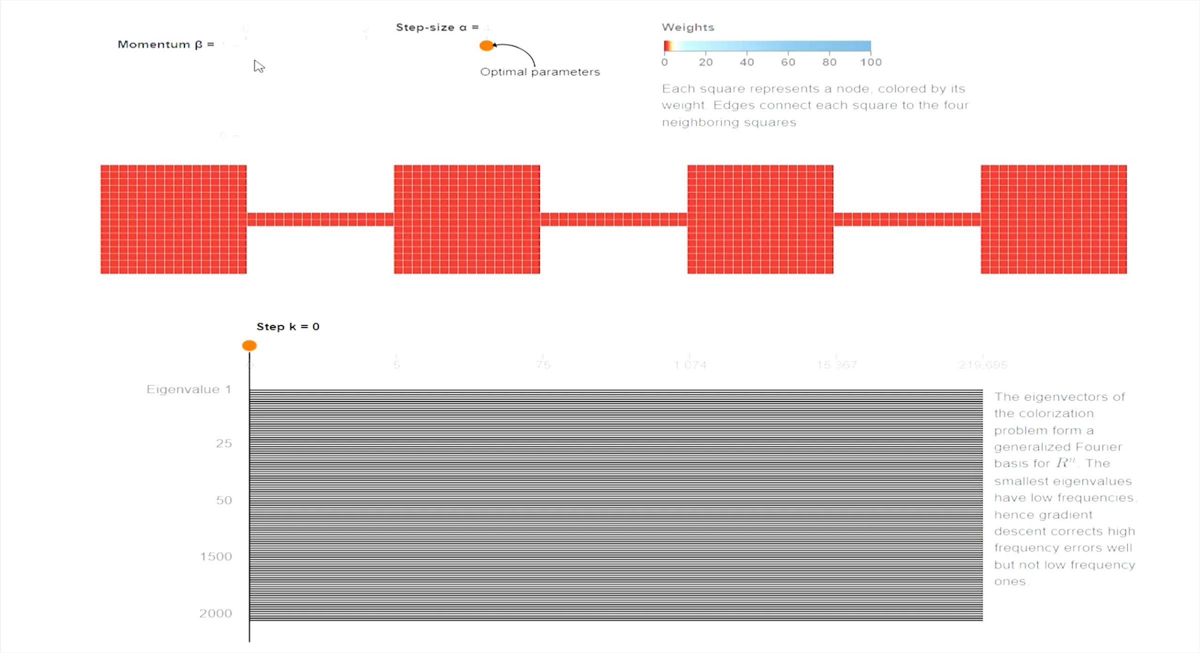
ПРИМЕР: ПРОБЛЕМА КОЛОРИЗАЦИИ

Без момента



ПРИМЕР: ПРОБЛЕМА КОЛОРИЗАЦИИ

С моментом



ГРАНИЦЫ УЛУЧШЕНИЯ СХОДИМОСТИ ГРАДИЕНТНОГО СПУСКА

$$w^{k+1} \;=\; w^0 \;+\; \sum_i^k \Gamma_i^k \nabla f(w^i) \quad \text{for some diagonal matrix } \Gamma_i^k.$$

Алгоритмическое пространство

Рассмотрим граф из одного пути

$$f^n(w) \;=\; \frac{1}{2}(w_1 - 1)^2 \;+\; \frac{1}{2}\sum_{i=1}^{n}(w_i - w_{i+1})^2 \;+\; \frac{2}{\kappa - 1}\|w\|^2.$$

with a colorizer of one node

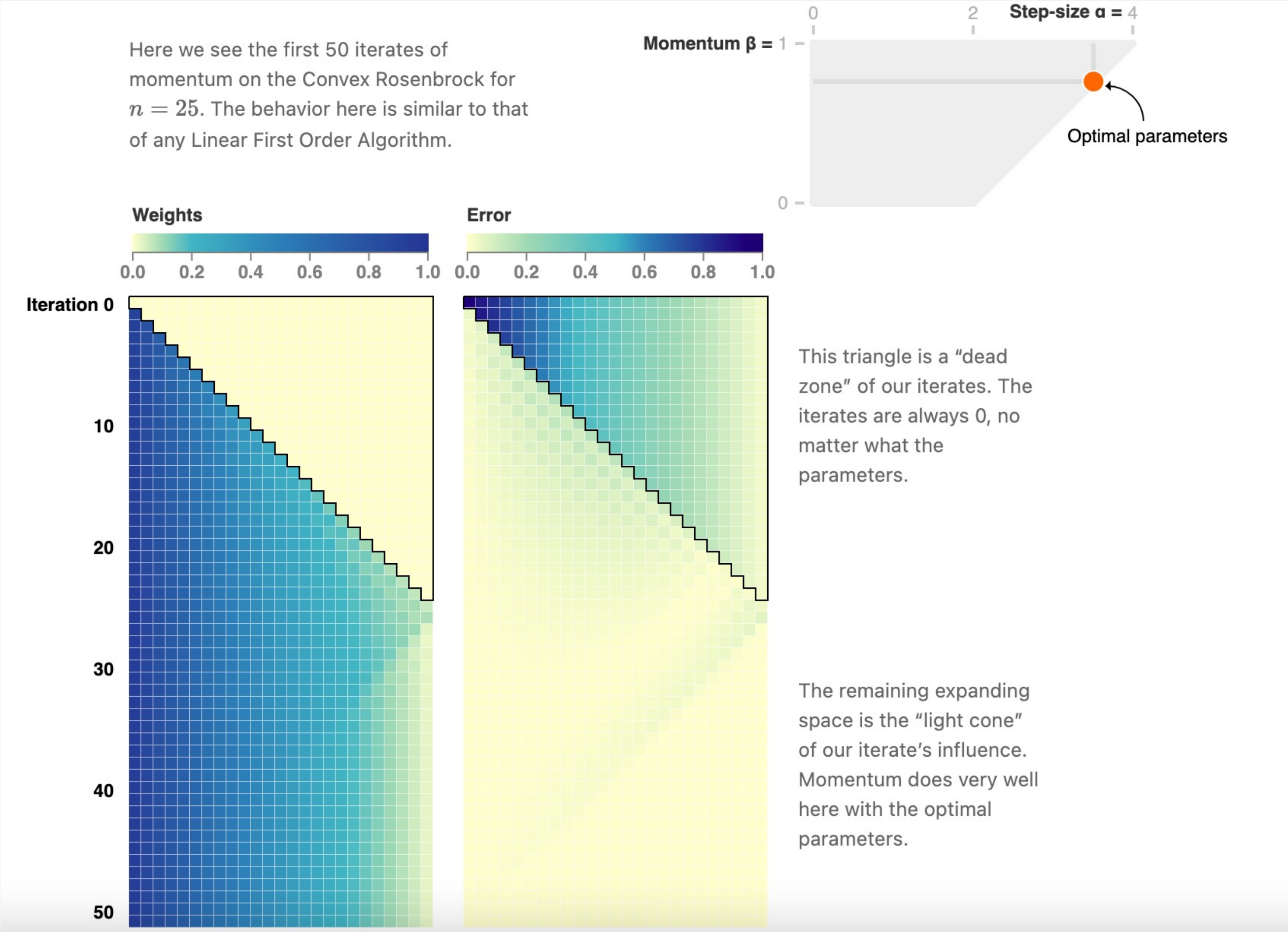
strong couplings of adjacent nodes in the path,

and a small regularization term.

$$w_i^\star = \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^i$$

Оптимальное решение

ГРАНИЦЫ УЛУЧШЕНИЯ СХОДИМОСТИ ГРАДИЕНТНОГО СПУСКА



Для выпуклой квадратичной функции:

$$w^k - w^\star = Qx^k = \sum_i^n x_i^0 (1 - \alpha\lambda_i)^k q_i$$

$$\begin{aligned}\|w^k - w^\star\|_\infty &\geq \max_{i \geq k+1} \{|w_i^\star|\} \\ &= \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^{k+1} \\ &= \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^{k} \|w^0 - w^\star\|_\infty.\end{aligned}$$

Максимальное улучшение сходимости градиентного спуска на квадратичный коэффициент!

ЗАКЛЮЧЕНИЕ

- **Сильные стороны метода инерции**:

1. выход из локального минимума целевой функции или патологической кривизны
2. позволяет использовать больший диапазон размеров шага и создает собственные колебания
3. дает квадратичное ускорение на многих функциях

- **Слабые стороны метода инерции**:

1. фиксированный размер шага

ЗАКЛЮЧЕНИЕ

- **Практическая ценность статьи** – на методе градиентного спуска с моментом основаны более мощные алгоритмы

- **Научная ценность статьи** - получена численная оценка скорости сходимости к минимуму функции

СПАСИБО ЗА ВНИМАНИЕ!

denisheva.rr@phystech.edu