Градиентные методы в задачах стохастической оптимизации

Юрий Владимирович Иванский

СПбГУ

Семинар по оптимизации, машинному обучению
и искусственному интеллекту

25 мая 2023

План доклада

1. Достижение консенсуса
2. Алгоритм стохастической аппроксимации с пробным одновременным возмущением (SPSA)
3. Ускоренные градиентные методы (метод Поляка, метод Нестерова)

Консенсус, протокол локального голосования

Определение

n агентов в сетевой системе достигают *консенсуса* в момент времени t, если $x_t^i = x_t^j \quad \forall i, j \in N, i \neq j$.

Определение

n агентов в сетевой системе достигают *асимптотического среднеквадратического ε-консенсуса* для $\varepsilon > 0$, если

$$\overline{\lim}_{t\to\infty} \mathrm{E}\|x_t^i - x_t^j\|^2 \leq \varepsilon.$$

Динамика состояния агента (узла):

$$x_{t+1}^i = x_t^i + u_t^i$$

Протокол локального голосования:

$$u_t^i = \gamma \sum_{j \in N^i} b^{i,j}(x_t^j - x_t^i)$$

N^i — множество соседей, $b^{i,j}$ — вес ребра между узлами i и j

Консенсус, протокол локального голосования

Динамика состояния агента (узла):

$$x_{t+1}^i = x_t^i + u_t^i + z_t^i$$

Протокол локального голосования:

$$u_t^i = \gamma \sum_{j \in N_t^i} b_t^{i,j} \left(x_{t-d_t^{i,j}}^j + w_t^{i,j} - x_t^i - w_t^{i,i} \right)$$

N_t^i — множество соседей узла i в момент времени t,
z_t^i — внешнее неконтролируемое воздействие в момент времени t,
$w_t^{i,j}$ — шум, возникающий при передаче данных по каналу связи от узла j на узел i в момент времени t,
$b_t^{i,j}$ — весовой коэффициент матрицы смежности графа связей сетевой системы
$d_t^{i,j}$ — целочисленная задержка, возникающая при передаче данных по каналу связи от узла j на узел i в момент времени t

Векторно-матричный вид

Динамика состояния агента (узла):

$$x_{t+1}^i = x_t^i + \gamma \sum_{j \in N^i} b^{i,j}(x_t^j - x_t^i)$$

Динамика состояния сетевой системы:

$$\mathrm{X}_{t+1} = \mathrm{X}_t + \gamma B \mathrm{X}_t - \gamma D(B)\mathrm{X}_t$$

$$X_{t+1} = (I - \gamma L)X_t$$

Лапласиан:

$$L = D - B,\ D = \mathrm{diag}\{B * 1_n\}$$

Обобщенная постановка задачи

Цель распределенной оптимизации (как правило) — вычислить минимум некоторой целевой функции за счет взаимодействия между агентами

$$\bar{F}(\mathrm{x}) = \sum_{i=1}^{n} F^i(\mathrm{x})$$

где $\mathrm{x} \in \mathbb{R}^d$, а $F^i(\mathrm{x}) : \mathbb{R}^d \to \mathbb{R}$ — целевая функция агента i, известная только ему

Оптимизация функционала среднего риска

Пусть $x_1, x_2, \ldots \in \mathbb{R}^d$ — точки измерений

$$y_t = f(x_t, w_t) + v_t,$$

где $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$,
w_t — случайный вектор, v_t — произвольная внешняя помеха.
Задача стохастической оптимизации:

$$F(x) = \int f(x, w) P(dw) \to \min_x .$$

Конечно-разностный подход

Процедура Кифера-Вольфовица
Число наблюдений на итерации $M = 2d$

$$\widehat{\theta}_0 \in \mathbb{R}^d$$

$$\widehat{\theta}_n = \widehat{\theta}_{n-1} - \frac{\alpha_n}{2\beta_n}(\mathrm{Y}_n^+ - \mathrm{Y}_n^-),$$

$$x_n^{(i,\pm)} = \widehat{\theta}_{n-1} \pm \beta_n e_i$$

$$\mathrm{Y}_n^{\pm} = \begin{pmatrix} f\left(x_n^{(1,\pm)}, w_n^{(1,\pm)}\right) + v_n^{(1,\pm)} \\ f\left(x_n^{(2,\pm)}, w_n^{(2,\pm)}\right) + v_n^{(2,\pm)} \\ \vdots \\ f\left(x_n^{(d,\pm)}, w_n^{(d,\pm)}\right) + v_n^{(d,\pm)} \end{pmatrix}$$

Рандомизированные алгоритмы стохастической аппроксимации (Simultaneous Perturbation Stochastic Approximation)

Сокращение числа наблюдений до 1 или 2 вместо $2d$ (!)

- Алгоритм с одним измерением

$$\mathrm{x}_n = \widehat{\theta}_{n-1} + \beta_n \Delta_n, \;\; \Delta_n = \begin{pmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{pmatrix}$$

$$y_n = f(\mathrm{x}_n, \mathrm{w}_n) + v_n$$

$$\widehat{\theta}_n = \widehat{\theta}_{n-1} - \frac{\alpha_n}{\beta_n} \mathscr{K}_n(\Delta_n) y_n$$

- Алгоритм с двумя измерениями

$$\mathrm{x}_n^{\pm} = \widehat{\theta}_{n-1} \pm \beta_n^{\pm} \Delta_n, \; y_n^{\pm} = f(\mathrm{x}_n^{\pm}, \mathrm{w}_n^{\pm}) + v_n^{\pm}$$

$$\widehat{\theta}_n = \widehat{\theta}_{n-1} - \frac{\alpha_n}{\beta_n^{+} + \beta_n^{-}} \mathscr{K}_n(\Delta_n)(y_n^{+} - y_n^{-})$$

Обоснование «псевдоградиентности» при произвольных внешних помехах

$$E\{\widehat{\theta}_{n-1} - \frac{\alpha_n}{\beta_n}\Delta_n y_n | \mathscr{F}_{n-1}\} =$$

$$= \widehat{\theta}_{n-1} - \frac{\alpha_n}{\beta_n}(E\{\Delta_n f(x_n) | \mathscr{F}_{n-1}\} + E\{\Delta_n\} E\{v_n | \mathscr{F}_{n-1}\}) =$$

$$= \widehat{\theta}_{n-1} - \frac{\alpha_n}{\beta_n}(E\{\Delta_n f(\widehat{\theta}_{n-1} + \beta_n \Delta_n) | \mathscr{F}_{n-1}\} \approx$$

$$\approx \widehat{\theta}_{n-1} - \frac{\alpha_n}{\beta_n}(E\{\Delta_n f(\widehat{\theta}_{n-1}) + \frac{\beta_n \Delta_n \Delta_n^{\mathrm{T}} \nabla f(\widehat{\theta}_{n-1})}{2} | \mathscr{F}_{n-1}\} =$$

$$= \widehat{\theta}_{n-1} - \frac{\alpha_n}{2} \nabla f(\widehat{\theta}_{n-1})$$

Преимущества алгоритма

- Состоятельность при почти произвольных внешних помехах [Граничин, 1989]
- Асимптотически-оптимальная скорость сходимости [Поляк, Цыбаков, 1990]
- Минимальное число наблюдений на итерации [Spall, 1992, 1997]
- Применимость в задаче об отслеживании изменений параметров [Granichin et al., 2009]

Отслеживание изменений параметров (трекинг)

Пусть Ξ — множество, $\{f_\xi(\mathrm{x},\mathrm{w})\}_{\xi\in\Xi}$ — семейство дифференцируемых функций

$$y_t = f_{\xi_t}(\mathrm{x}_t,\mathrm{w}_t) + v_t$$

$$F_t(\mathrm{x}) = \int f_{\xi_t}(\mathrm{x},\mathrm{w}_t)\mathrm{P}(d\mathrm{w}_t)\mathrm{P}(d\xi_t) \to \min_{\mathrm{x}}$$

Комбинированный алгоритм: SPSA + консенсус

[Granichin O. et al. 2021 Trans. Autom. Cont.]
Введем Δ_k^i, $k = 1, 2, \ldots$, $i \in N$ — последовательность независимых случайных векторов из $\mathbb{R}^d$, называемых *одновременным пробным возмущением*, с симметричными функциями распределения $\mathrm{P}_k^i(\cdot)$; и множество вектор-функций (ядер) $\mathscr{K}_k^i(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$, $k = 1, 2, \ldots$

Выберем начальное приближение $\widehat{\theta}_0^i \in \mathbb{R}^d$, шаг $\alpha > 0$, коэффициент усиления γ, и последовательности неотрицательных чисел $\{\beta_k^+\}$ и $\{\beta_k^-\}$ такие, что $\beta_k = \beta_k^+ + \beta_k^- > 0$. Рассмотрим алгоритм с двумя измерениями распределенных функций $f_{\xi_t}^i(\theta)$ для каждого агента $i \in N$, строящий последовательность точек измерения $\{\mathrm{x}_t^i\}$ и оценки $\{\widehat{\theta}_t^i\}$:

$$\begin{cases} \mathrm{x}_{2k}^i = \widehat{\theta}_{2k-2}^i + \beta_k^+ \Delta_k^i, \mathrm{x}_{2k-1}^i = \widehat{\theta}_{2k-2}^i - \beta_k^- \Delta_k^i, \\ \widehat{\theta}_{2k-1}^i = \widehat{\theta}_{2k-2}^i, \\ \widehat{\theta}_{2k}^i = \widehat{\theta}_{2k-1}^i - \alpha \left(\frac{y_{2k}^i - y_{2k-1}^i}{\beta_k} \mathscr{K}_k^i(\Delta_k^i) + \right. \\ \qquad \left. \gamma \sum_{j \in N_{2k-1}^i} b_{2k-1}^{i,j} (\tilde{\theta}_{2k-1}^{i,j} - \widehat{\theta}_{2k-1}^i) \right) \end{cases}$$

Основной результат

Предположения:

- Градиент $\nabla f_\xi^i(x)$ удовлетворяет условию Липшица $\forall x_1, x_2 \in \mathbb{R}^d$

$$\|\nabla f_\xi^i(x_1) - \nabla f_\xi^i(x_2)\| \leq M\|x_1 - x_2\|$$

- Последовательные разности $\tilde{v}_k^i = v_{2k}^i - v_{2k-1}^i$ помех наблюдения ограничены $|\tilde{v}_k^i| \leq c_v < \infty$, or $\mathbb{E}(\tilde{v}_k^i)^2 \leq c_v^2$ если посл-ть $\{\tilde{v}_k^i\}$ случайная
- Помехи в канале связи $w_t^{i,j}$ случайные, независимые, одинаково распределенные $\mathbb{E}w_t^{i,j} = 0$, $\mathbb{E}\|w_t^{i,j}\|^2 \leq \sigma_w^2$. Все случайные векторы и переменные $w_t^{i,j}$, $b_t^{i,j}$, ξ_t, and ξ_{t+1} взаимно независимы (если они случайные)

Теорема

Последовательность оценок $\{\bar{\theta}_{2k}^i\}$ имеет *асимптотически эффективную верхнюю границу* $\bar{L} > 0$ *невязки оценок:* $\forall \varepsilon > 0\ \exists \bar{k}$: $\forall k > \bar{k}$: $\sqrt{E\|\bar{\theta}_{2k}^i - 1_n \otimes \theta_{2k}\|^2} \leq \bar{L} + \varepsilon$.

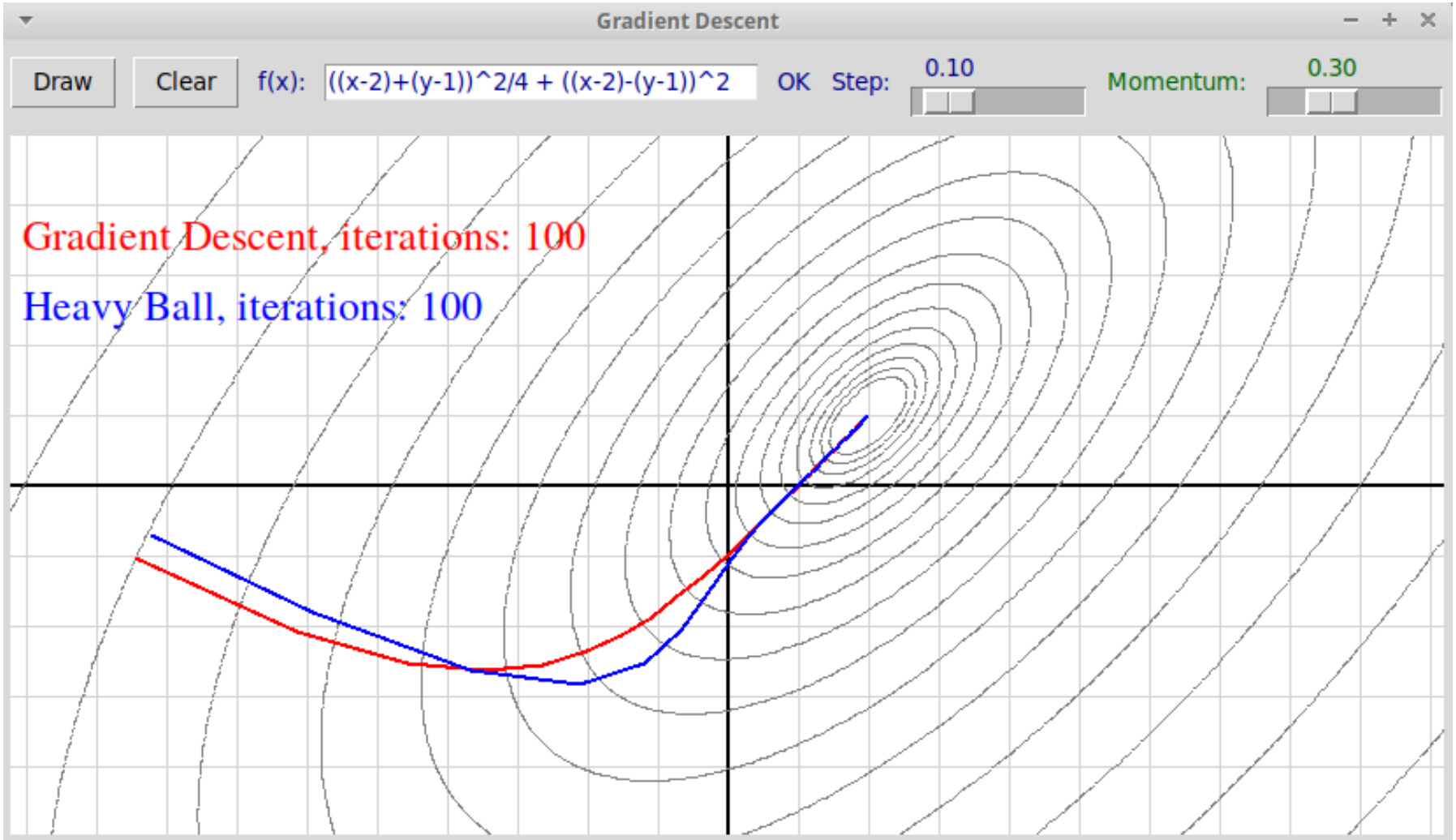
Метод Поляка

[Поляк, 1964, Ж. вычисл. матем. и матем. физ.]
Метод тяжелого шарика, метод инерции (momentum method, heavy ball method, momentum-gd)

Пусть $f(x)$ — непрерывно дифференцируемая функция, удовлетворяет условию Липшица для градиента с константой L:
$\|\nabla f(x_1) - \nabla f(x_2)\| \leq L\|x_1 - x_2\|$

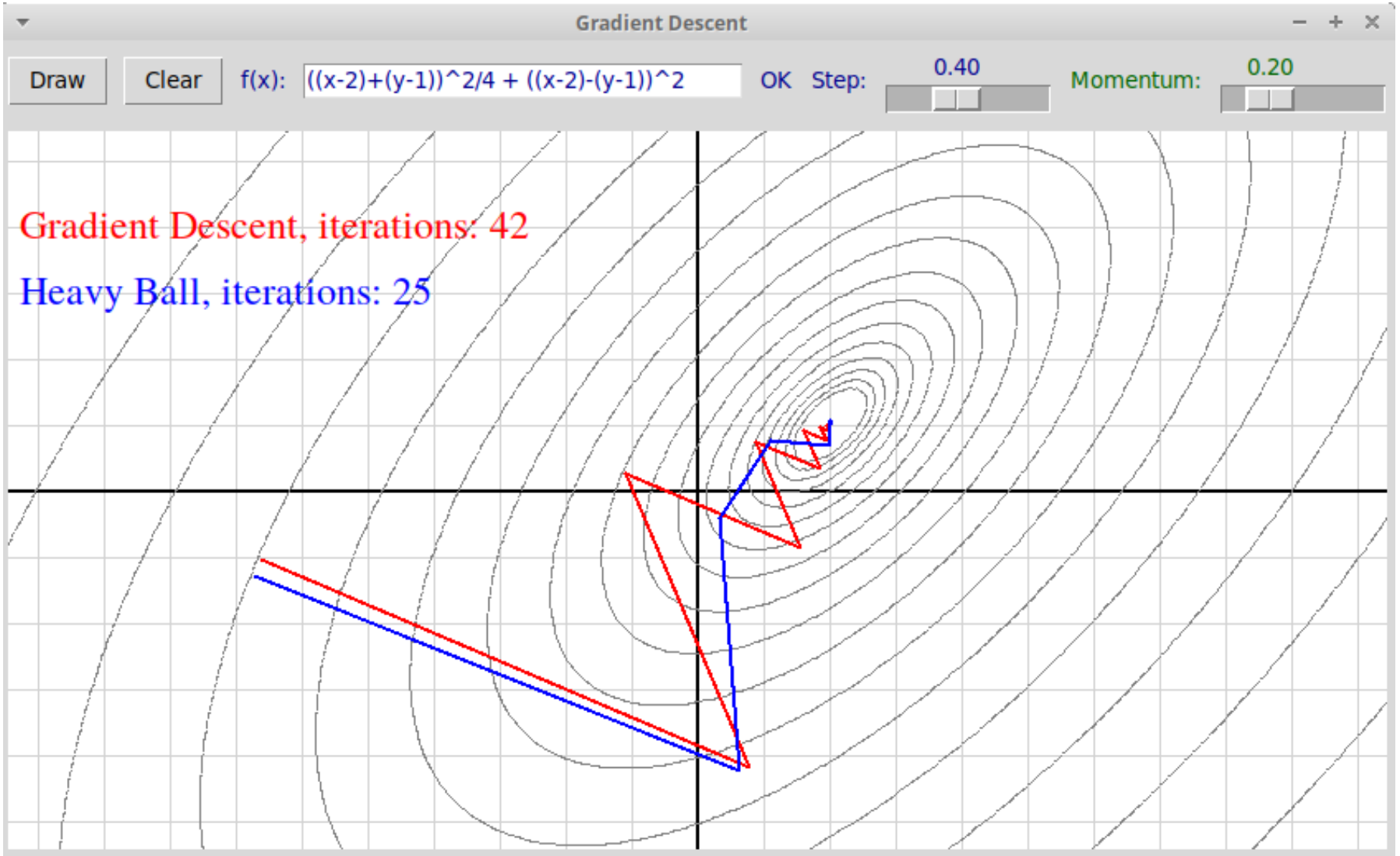
$$x_{k+1} = x_k - \alpha\nabla f(x_k) + \beta(x_k - x_{k-1}), \quad 0 \leq \beta < 1,\ 0 \leq \alpha < 2(1-\beta)/L$$

Метод Поляка



http://mech.math.msu.su/~vvb/MasterAI/GradientDescent.html

Метод Поляка

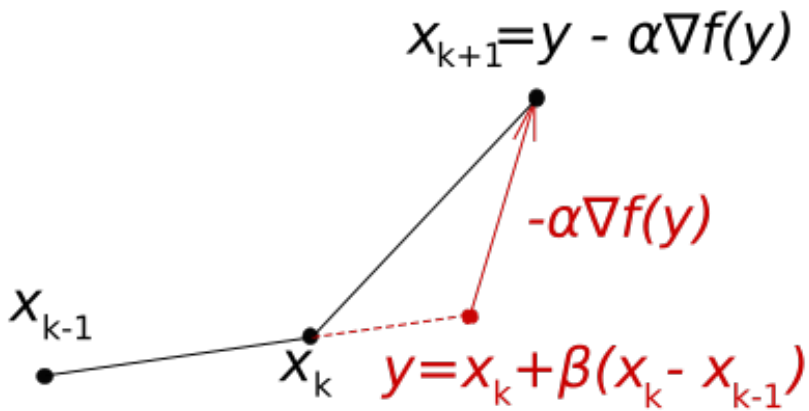


http://mech.math.msu.su/~vvb/MasterAI/GradientDescent.html

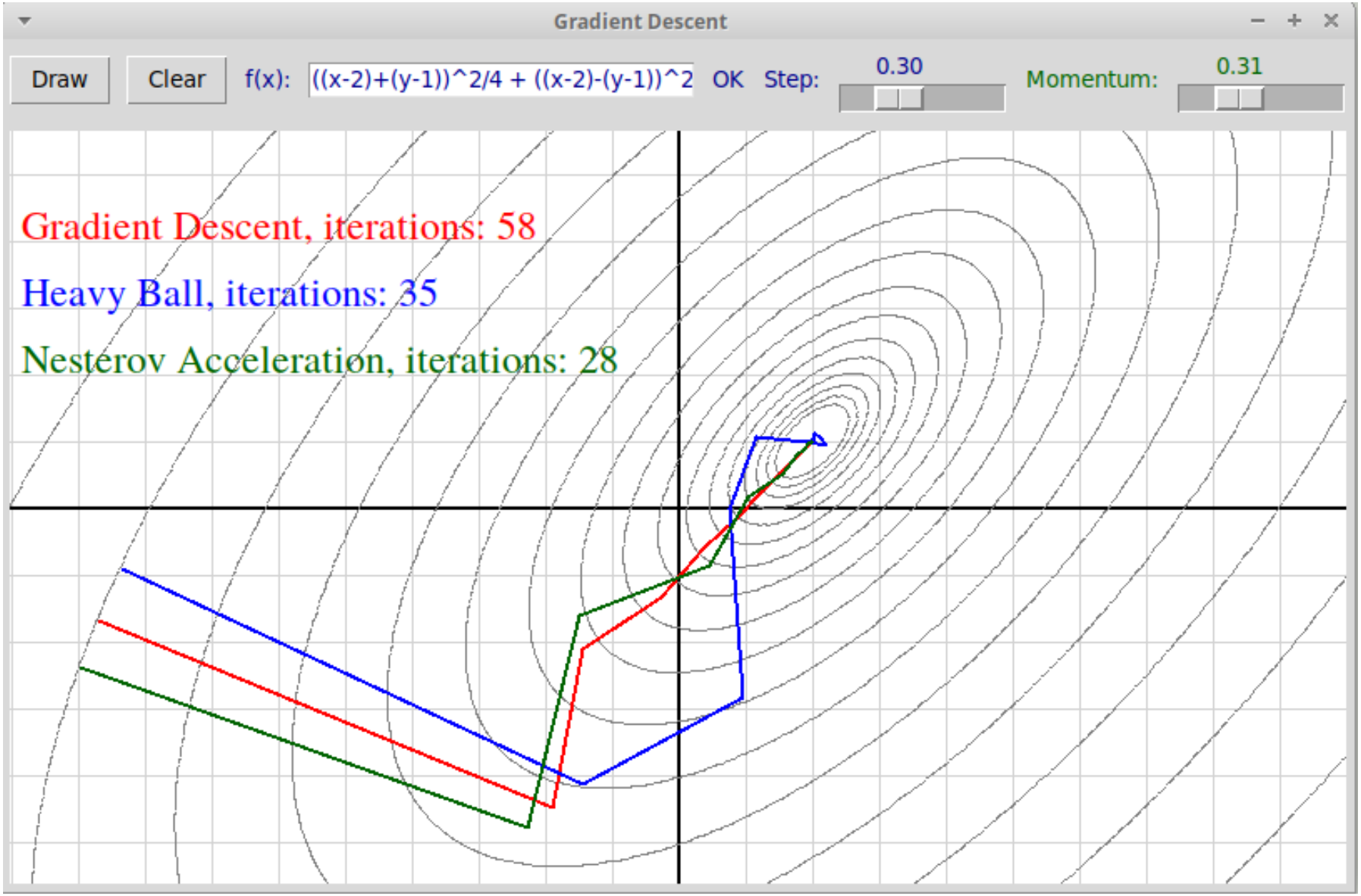
Метод Нестерова

[Нестеров, 1983, Докл. АН СССР]
Пусть $f(x)$ — непрерывно дифференцируемая функция, удовлетворяет условию Липшица для градиента с константой L:
$\|\nabla f(x_1) - \nabla f(x_2)\| \leq L\|x_1 - x_2\|$

$$y_0 = x_0$$
$$x_{k+1} = y_k - \alpha \nabla f(y_k)$$
$$y_{k+1} = x_{k+1} + \beta(x_{k+1} - x_k)$$
$$0 \leq \beta < 1,\ 0 \leq \alpha < 2(1-\beta)/L$$



Метод Нестерова



http://mech.math.msu.su/~vvb/MasterAI/GradientDescent.html

Распределенный алгоритм с ускорением

[Erofeeva et al. 2022. American Control Conference]

$$
\begin{cases}
\tilde{x}^i_{2k-2} = \frac{1}{\gamma^i_{k-1}+\alpha^i_k(\mu-\eta)}\left(\alpha^i_k\gamma^i_{k-1}z^i_{2k-2} + \gamma^i_k\widehat{\theta}^i_{2k-2}\right), \\
x^i_{2k} = \tilde{x}^i_{2k-2} + \beta\Delta^i_k,\ x^i_{2k-1} = \tilde{x}^i_{2k-2} - \beta\Delta^i_k, \\
\tilde{x}^i_{2k-1} = \tilde{x}^i_{2k-2},\ \widehat{\theta}^i_{2k-1} = \widehat{\theta}^i_{2k-2}, \\
g^i_{2k} = \Delta^i_k\frac{y^i_{2k}-y^i_{2k-1}}{2\beta} + \omega\sum_{j\in N^i} b^{i,j}(\tilde{x}^i_{2k-1} - \tilde{x}^{i,j}_{2k-1}), \\
\widehat{\theta}^i_{2k} = \tilde{x}^i_{2k-1} - hg^i_{2k}, \\
z^i_{2k} = \frac{1}{\gamma^i_k}\Big[(1-\alpha^i_k)\gamma^i_{k-1}z^i_{2k-2} + \\
\alpha^i_k(\mu-\eta)\tilde{x}^i_{2k-1} - \alpha^i_k g^i_{2k})\Big],
\end{cases}
$$

где $\alpha^i_k, \beta, \omega, \gamma^i_k, h, \mu, \eta$ – параметры и константы, x^i_t – точки наблюдения, Δ^i_k – пробное возмущение.

Заключение

- Выбор параметров итеративной процедуры
- Оценка скорости сходимости
- Оптимизация размера шага, распространение оценки в децентрализованной сети

Список источников

- Б.Т. Поляк, (1964) О некоторых способах ускорения сходимости итерационных методов // Журнал вычислительной математики и математической физики, т. 4, № 5, с. 791-803.
- Ю.Е. Нестеров, (1983) Метод минимизации выпуклых функций со скоростью сходимости O(1/k2) // Доклады Академии наук СССР, т. 269, № 3, с. 543-547.
- О.Н. Граничин, (1989) Об одной стохастической рекуррентной процедуре при зависимых помехах в наблюдении, использующей на входе пробные возмущения // Вестник Ленинградского университета. Серия 1: Математика, механика, астрономия, №4, с. 19-21.
- Б.Т. Поляк, А.Б. Цыбаков (1990) Оптимальные порядки точности поисковых алгоритмов стохастической оптимизации // Проблемы передачи информации, 26:2, с. 45–53.
- Spall, J.C. (1992) Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation // IEEE Transactions on Automatic Control, vol. 37, pp. 332-341.

Список источников

- Spall, J.C. (1997) A One-Measurement Form of Simultaneous Perturbation Stochastic Approximation // Automatica, vol. 33, pp. 109-112.
- Granichin, O., Gurevich, L., and Vakhitov, A. (2009). Discrete-time minimum tracking based on stochastic approximation algorithm with randomized differences // In Proceedings of the 48h IEEE Conference on Decision and Control (CDC), pp. 5763-5767. IEEE.
- Granichin, O., Erofeeva, V., Ivanskiy, Y., and Jiang, Y. (2020). Simultaneous perturbation stochastic approximation-based consensus for tracking under unknown-but-bounded disturbances. // IEEE Transactions on Automatic Control, 66(8), pp. 3710-3717.
- Erofeeva, V., Granichin, O., Tursunova, M., Sergeenko, A., and Jiang, Y. (2022). Accelerated simultaneous perturbation stochastic approximation for tracking under unknown-but-bounded disturbances. // In Proceedings of 2022 American Control Conference (ACC), pp. 1582-1587. IEEE.
- `https://distill.pub/2017/momentum/`