Стохастические вариационные неравенства и седловые задачи

Александр Безносиков

Обучение, понимание и оптимизация в моделях искусственного интеллекта, г. Пушкин

25 июня 2022

Вариационное Неравенство (ВН)

Пусть дано выпуклое множество $\mathcal{Z} \subseteq \mathbb{R}^n$ и оператор $F : \mathbb{R}^n \to \mathbb{R}^n$. Тогда сформулируем задачу поиска решения ВН:

Задача ВН

Найти $z^* \in \mathcal{Z}$ такую, что:

$$\langle F(z^*), z - z^* \rangle \geq 0, \ \forall z \in \mathcal{Z}.$$

Частные случаи: минимизация

- Пусть $F(z) = \nabla f(z)$ – градиент функции f, $Z = \mathbb{R}^n$.
- z^* – решение соответствующего ВН $\Leftrightarrow \nabla f(z^*) = 0$.
  $\Leftarrow$ Очевидно.
  $\Rightarrow$ Пусть z^* – решение ВН и $\nabla f(z^*) \neq 0$, тогда должно быть выполнено
  $$\langle \nabla f(z^*), z - z^* \rangle \geq 0, \ \forall z \in \mathcal{Z}.$$
  Рассмотрим $z = z^* - \nabla f(z^*)$, тогда
  $$\langle \nabla f(z^*), z - z^* \rangle = -\|\nabla f(z^*)\|^2 < 0,$$
  что противоречит предположению о том, что z^* – решение ВН.
- Суть: если функция f выпуклая, то z^* – минимум; если функция f невыпуклая, то z^* – стационарная точка.
- Аналогично, можно поступить и с задачами условной оптимизации ($Z \neq \mathbb{R}^n$).

Частные случаи: минимизация

- Пусть $F(z) = \nabla f(z)$ – градиент функции f, $Z = \mathbb{R}^n$.
- z^* – решение соответствующего ВН $\Leftrightarrow \nabla f(z^*) = 0$.

  $\Leftarrow$ Очевидно.

  $\Rightarrow$ Пусть z^* – решение ВН и $\nabla f(z^*) \neq 0$, тогда должно быть выполнено

  $$\langle \nabla f(z^*), z - z^* \rangle \geq 0, \ \forall z \in \mathcal{Z}.$$

  Рассмотрим $z = z^* - \nabla f(z^*)$, тогда

  $$\langle \nabla f(z^*), z - z^* \rangle = -\|\nabla f(z^*)\|^2 < 0,$$

  что противоречит предположению о том, что z^* – решение ВН.
- Суть: если функция f выпуклая, то z^* – минимум; если функция f невыпуклая, то z^* – стационарная точка.
- Аналогично, можно поступить и с задачами условной оптимизации ($Z \neq \mathbb{R}^n$).

Частные случаи: минимизация

- Пусть $F(z) = \nabla f(z)$ – градиент функции f, $Z = \mathbb{R}^n$.
- z^* – решение соответствующего ВН $\Leftrightarrow \nabla f(z^*) = 0$.
  $\Leftarrow$ Очевидно.
  $\Rightarrow$ Пусть z^* – решение ВН и $\nabla f(z^*) \neq 0$, тогда должно быть выполнено
  $$\langle \nabla f(z^*), z - z^* \rangle \geq 0, \ \forall z \in \mathcal{Z}.$$
  Рассмотрим $z = z^* - \nabla f(z^*)$, тогда
  $$\langle \nabla f(z^*), z - z^* \rangle = -\|\nabla f(z^*)\|^2 < 0,$$
  что противоречит предположению о том, что z^* – решение ВН.
- Суть: если функция f выпуклая, то z^* – минимум; если функция f невыпуклая, то z^* – стационарная точка.
- Аналогично, можно поступить и с задачами условной оптимизации ($Z \neq \mathbb{R}^n$).

Частные случаи: минимизация

- Пусть $F(z) = \nabla f(z)$ – градиент функции f, $Z = \mathbb{R}^n$.
- z^* – решение соответствующего ВН $\Leftrightarrow \nabla f(z^*) = 0$.
  $\Leftarrow$ Очевидно.
  $\Rightarrow$ Пусть z^* – решение ВН и $\nabla f(z^*) \neq 0$, тогда должно быть выполнено
  $$\langle \nabla f(z^*), z - z^* \rangle \geq 0, \ \forall z \in \mathcal{Z}.$$
  Рассмотрим $z = z^* - \nabla f(z^*)$, тогда
  $$\langle \nabla f(z^*), z - z^* \rangle = -\|\nabla f(z^*)\|^2 < 0,$$
  что противоречит предположению о том, что z^* – решение ВН.
- Суть: если функция f выпуклая, то z^* – минимум; если функция f невыпуклая, то z^* – стационарная точка.
- Аналогично, можно поступить и с задачами условной оптимизации ($Z \neq \mathbb{R}^n$).

Частные случаи: минимизация

- Решение ВН неравенства:

$$\langle F(z^*), z - z^* \rangle \geq 0.$$

- Подставим $F(z^*) = \nabla f(z^*)$:

$$\langle \nabla f(z^*), z - z^* \rangle = \langle F(z^*), z - z^* \rangle \geq 0.$$

- Воспользуемся определением выпуклости:

$$\begin{aligned} f(z) - f(z^*) &\geq \langle \nabla f(z^*), z - z^* \rangle \\ &= \langle F(z^*), z - z^* \rangle \\ &\geq 0. \end{aligned}$$

Частные случаи: минимизация

- Решение ВН неравенства:

$$\langle F(z^*), z - z^* \rangle \geq 0.$$

- Подставим $F(z^*) = \nabla f(z^*)$:

$$\langle \nabla f(z^*), z - z^* \rangle = \langle F(z^*), z - z^* \rangle \geq 0.$$

- Воспользуемся определением выпуклости:

$$\begin{aligned} f(z) - f(z^*) &\geq \langle \nabla f(z^*), z - z^* \rangle \\ &= \langle F(z^*), z - z^* \rangle \\ &\geq 0. \end{aligned}$$

Частные случаи: минимизация

- Решение ВН неравенства:

$$\langle F(z^*), z - z^* \rangle \geq 0.$$

- Подставим $F(z^*) = \nabla f(z^*)$:

$$\langle \nabla f(z^*), z - z^* \rangle = \langle F(z^*), z - z^* \rangle \geq 0.$$

- Воспользуемся определением выпуклости:

$$\begin{aligned} f(z) - f(z^*) &\geq \langle \nabla f(z^*), z - z^* \rangle \\ &= \langle F(z^*), z - z^* \rangle \\ &\geq 0. \end{aligned}$$

Частные случаи: седловые задачи

- Пусть $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, где $\mathcal{X} \subseteq \mathbb{R}^{n_x}$ и $\mathcal{Y} \subseteq \mathbb{R}^{n_y}$, также дана выпукло-вогнутая функция $g(x, y)$.
- Определим

$$F(z) = F(x, y) = \begin{pmatrix} \nabla_x g(x, y) \\ -\nabla_y g(x, y) \end{pmatrix}.$$

Частные случаи: седловые задачи

- Решение соответствующего ВН:

$$\left\langle \begin{pmatrix} \nabla_x g(x^*, y^*) \\ -\nabla_y g(x^*, y^*) \end{pmatrix}, \begin{pmatrix} x - x^* \\ y - y^* \end{pmatrix} \right\rangle = \langle F(z^*), z - z^* \rangle \geq 0.$$

- Перепишем удобнее:

$$\begin{aligned} \langle \nabla_x g(x^*, y^*), x - x^* \rangle &+ \langle -\nabla_y g(x^*, y^*), y - y^* \rangle \\ &= \left\langle \begin{pmatrix} \nabla_x g(x^*, y^*) \\ -\nabla_y g(x^*, y^*) \end{pmatrix}, \begin{pmatrix} x - x^* \\ y - y^* \end{pmatrix} \right\rangle \\ &= \langle F(z^*), z - z^* \rangle \geq 0. \end{aligned}$$

- Выпуклость по x, вогнутость по y дают

$$\begin{gathered} [g(x, y^*) - g(x^*, y^*)] + [g(x^*, y^*) - g(x^*, y)] \\ \geq \langle F(z^*), z - z^* \rangle \geq 0. \end{gathered}$$

Частные случаи: седловые задачи

- Решение соответствующего ВН:

$$\left\langle \begin{pmatrix} \nabla_x g(x^*, y^*) \\ -\nabla_y g(x^*, y^*) \end{pmatrix}, \begin{pmatrix} x - x^* \\ y - y^* \end{pmatrix} \right\rangle = \langle F(z^*), z - z^* \rangle \geq 0.$$

- Перепишем удобнее:

$$\begin{aligned} \langle \nabla_x g(x^*, y^*), x - x^* \rangle + \langle -\nabla_y g(x^*, y^*), y - y^* \rangle \\ = \left\langle \begin{pmatrix} \nabla_x g(x^*, y^*) \\ -\nabla_y g(x^*, y^*) \end{pmatrix}, \begin{pmatrix} x - x^* \\ y - y^* \end{pmatrix} \right\rangle \\ = \langle F(z^*), z - z^* \rangle \geq 0. \end{aligned}$$

- Выпуклость по x, вогнутость по y дают

$$\begin{aligned} [g(x, y^*) - g(x^*, y^*)] + [g(x^*, y^*) - g(x^*, y)] \\ \geq \langle F(z^*), z - z^* \rangle \geq 0. \end{aligned}$$

Частные случаи: седловые задачи

- Решение соответствующего ВН:

$$\left\langle \begin{pmatrix} \nabla_x g(x^*, y^*) \\ -\nabla_y g(x^*, y^*) \end{pmatrix}, \begin{pmatrix} x - x^* \\ y - y^* \end{pmatrix} \right\rangle = \langle F(z^*), z - z^* \rangle \geq 0.$$

- Перепишем удобнее:

$$\begin{aligned} \langle \nabla_x g(x^*, y^*), x - x^* \rangle + \langle -\nabla_y g(x^*, y^*), y - y^* \rangle \\ = \left\langle \begin{pmatrix} \nabla_x g(x^*, y^*) \\ -\nabla_y g(x^*, y^*) \end{pmatrix}, \begin{pmatrix} x - x^* \\ y - y^* \end{pmatrix} \right\rangle \\ = \langle F(z^*), z - z^* \rangle \geq 0. \end{aligned}$$

- Выпуклость по x, вогнутость по y дают

$$\begin{aligned} [g(x, y^*) - g(x^*, y^*)] + [g(x^*, y^*) - g(x^*, y)] \\ \geq \langle F(z^*), z - z^* \rangle \geq 0. \end{aligned}$$

Частные случаи: седловые задачи

- Тогда решение соответствующего вариационного неравенства $z^* = (x^*, y^*)$ есть **седловая точка**:

$$g(x^*, y) \leq g(x^*, y^*) \leq g(x, y^*), \ \forall x \in \mathcal{X}, y \in \mathcal{Y}.$$

- Другая формулировка седловой задачи:

Задача поиска седловой точки

Найти $x^* \in \mathcal{X}$, $y^* \in \mathcal{Y}$ решение задачи:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} g(x, y).$$

Частные случаи: седловые задачи

- Тогда решение соответствующего вариационного неравенства $z^* = (x^*, y^*)$ есть **седловая точка**:

$$g(x^*, y) \leq g(x^*, y^*) \leq g(x, y^*), \ \forall x \in \mathcal{X}, y \in \mathcal{Y}.$$

- Другая формулировка седловой задачи:

Задача поиска седловой точки

Найти $x^* \in \mathcal{X}$, $y^* \in \mathcal{Y}$ решение задачи:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} g(x, y).$$

Частные случаи: задача поиска стационарной точки оператора

- Найти $z^* \in \mathcal{Z}$ такую, что:
$$T(z^*) = z^*,$$
где оператор $T : \mathbb{R}^n \to \mathbb{R}^n$.
- Определим
$$F(z) = T(z) - z.$$
- Тогда
$$F(z^*) = 0$$
или
$$\langle F(z^*), z - z^* \rangle \geq 0.$$

Классические задачи

- Билинейная задача на вероятностном симплексе:

$$\min_{x \in \mathbb{R}} \max_{y \in \mathbb{R}} \; x^T A y,$$

  Δ^{n_x} – вероятностный симплекс размера n_x (все компоненты вектора неотрицательные и в сумме дают 1 – интуиция: распределение вероятности).
- Имеет массу применений, в первую очередь в экономике и теории игр. Еще эту задачу называют матричной игрой или задачей поиска равновесия – Нэшевского эквилибриума.

Примеры: минимизация с ограничениями и множители Лагранжа

- Задача минимизации с выпуклыми ограничениями:

$$\min_{\substack{x \in \mathbb{R}^n \\ \mathbf{g}(x) = \mathbf{0}}} f(x).$$

- Запись в виде задачи с множителями Лагранжа

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^{\mathbf{len}(\mathbf{g})}} f(x) + y^T \mathbf{g}(x).$$

Примеры: робастная оптимизация/состязательный подход

- Рассмотрим следующий модифицированную функцию потерь:

$$\max_{\|\delta_n\|_2 \le e} f(w, \delta) := \max_{\|\delta\|_2 \le e} \frac{1}{N} \sum_{n=1}^{N} l(g(w, x_n + \delta_n), y_n).$$

- Здесь δ_n – является обучаемым параметром, которые и отвечает за небольшой шум входных данных.

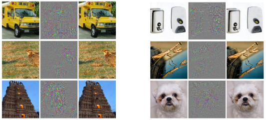

Figure: Слева картинка, которую AleXNet классифицирует правильно, по центру – небольшой шум, который мы вносим, справа – картинка, которую распознают неправильно. Картинка из статьи.

GANs

- GAN представляет собой две модели генератор G и дискриминатор D.
- D принимает на вход элемент x и определяет, является ли этот элемент реальным (из выборки данных) или искусственно созданным генератором.
- На вход генератор подается некоторый случайный вектор z, по которому генератор строит "фейковый" экземпляр, похожий на реальную выборку.
- Формально задачу GAN формулируют в виде седловой:

$$\min_G \max_D V(D, G) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

Основные предположения

Липшицевость

Оператор F называется Липшицев с константой L, если

$$\|F(z_1) - F(z_2)\| \leq L\|z_1 - z_2\|, \;\forall z_1, z_2 \in \mathcal{Z}.$$

Аналогия: Липшецевость градиента.

Сильная монотонность

Оператор F называется сильно монотонным с константой μ, если

$$\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq \mu\|z_1 - z_2\|^2, \;\forall z_1, z_2 \in \mathcal{Z}.$$

Аналогия: сильная выпуклость.
Для седел: писать сильную-выпуклость/выпуклость по x и сильную-вогнутость/вогнутость по y.

Основные предположения

Липшицевость

Оператор F называется Липшицев с константой L, если

$$\|F(z_1) - F(z_2)\| \leq L\|z_1 - z_2\|, \ \forall z_1, z_2 \in \mathcal{Z}.$$

Аналогия: Липшецевость градиента.

Сильная монотонность

Оператор F называется сильно монотонным с константой μ, если

$$\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq \mu\|z_1 - z_2\|^2, \ \forall z_1, z_2 \in \mathcal{Z}.$$

Аналогия: сильная выпуклость.
Для седел: писать сильную-выпуклость/выпуклость по x и сильную-вогнутость/вогнутость по y.

Базовые методы: а ля градиентный спуск

- Классический спуск(-подъем):

$$z^{k+1} = \mathrm{proj}_Z(z^k - \gamma F(z^k)).$$

Базовые методы: а ля градиентный спуск

- В сильно монотонном случае:

$$
\begin{aligned}
\|z^{k+1} - z^*\|^2 &= \|\mathrm{proj}_Z(z^k - \gamma F(z^k)) - \mathrm{proj}_Z(z^*)\|^2 \\
&\leq \|z^k - \gamma F(z^k) - z^*\|^2 \\
&= \|z^k - z^*\|^2 - 2\gamma\langle z^k - z^*, F(z^k)\rangle + \gamma^2\|F(z^k)\|^2 \\
&\leq \|z^k - z^*\|^2 - 2\gamma\langle z^k - z^*, F(z^k)\rangle + 2\gamma^2\|F(z^k) - F(z^*)\|^2 \\
&\quad +2\gamma^2\|F(z^*)\|^2 \\
&\leq \|z^k - z^*\|^2 - 2\gamma\langle z^k - z^*, F(z^k)\rangle + 2\gamma^2 L^2\|z^k - z^*\|^2 \\
&\quad +2\gamma^2\|F(z^*)\|^2 \\
&\leq \|z^k - z^*\|^2 - 2\gamma\langle z^k - z^*, F(z^k) - F(z^*)\rangle \\
&\quad +2\gamma^2 L^2\|z^k - z^*\|^2 + 2\gamma^2\|F(z^*)\|^2 \\
&\leq \|z^k - z^*\|^2 - 2\gamma\mu\|z^k - z^*\|^2 \\
&\quad +2\gamma^2 L^2\|z^k - z^*\|^2 + 2\gamma^2\|F(z^*)\|^2.
\end{aligned}
$$

Базовые методы: а ля градиентный спуск

- Выбор $\gamma \leq \frac{\mu}{2L^2}$:

$$\|z^{k+1} - z^*\|^2 \quad \leq \quad (1 - \mu\gamma)\|z^k - z^*\|^2 + \gamma^2\|F(z^*)\|^2.$$

- Линейная (но медленная) сходимость в случае $F(z^*) = 0$. Сложность:

$$O\left(\frac{L^2}{\mu^2}\log\frac{1}{\varepsilon}\right).$$

Базовые методы: а ля градиентный спуск

- Выбор $\gamma \leq \frac{\mu}{2L^2}$:

$$\|z^{k+1} - z^*\|^2 \quad \leq \quad (1 - \mu\gamma)\|z^k - z^*\|^2 + \gamma^2\|F(z^*)\|^2.$$

- Линейная (но медленная) сходимость в случае $F(z^*) = 0$. Сложность:

$$O\left(\frac{L^2}{\mu^2}\log\frac{1}{\varepsilon}\right).$$

Базовые методы: а ля градиентный спуск

- Классический спуск(-подъем):

$$z^{k+1} = \text{proj}_Z(z^k - \gamma F(z^k)).$$

Базовые методы: а ля градиентный спуск

- Билинейная задача:

$$\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} xy$$

  Где у нее решение?

- Решение: $x^* = 0$, $y^* = 0$ (из условия оптимальности)
- Тогда

$$\begin{aligned}
\|z^{k+1} - z^*\|^2 &= (x^{k+1} - x^*)^2 + (y^{k+1} - y^*)^2 \\
&= (x^k - \gamma y^k)^2 + (y^k + \gamma x^k)^2 \\
&= (1+\gamma^2)\left[(x^k - x^*)^2 + (y^k - y^*)^2\right] \\
&= (1+\gamma^2)\|z^{k+1} - z^*\|^2.
\end{aligned}$$

Базовые методы: а ля градиентный спуск

- Билинейная задача:

$$\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} xy$$

  Где у нее решение?

- Решение: $x^* = 0$, $y^* = 0$ (из условия оптимальности)
- Тогда

$$\begin{aligned}
\|z^{k+1} - z^*\|^2 &= (x^{k+1} - x^*)^2 + (y^{k+1} - y^*)^2 \\
&= (x^k - \gamma y^k)^2 + (y^k + \gamma x^k)^2 \\
&= (1+\gamma^2)\left[(x^k - x^*)^2 + (y^k - y^*)^2\right] \\
&= (1+\gamma^2)\|z^{k+1} - z^*\|^2.
\end{aligned}$$

Базовые методы: а ля градиентный спуск

- Билинейная задача:

$$\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} xy$$

  Где у нее решение?
- Решение: $x^* = 0$, $y^* = 0$ (из условия оптимальности)
- Тогда

$$\begin{aligned}
\|z^{k+1} - z^*\|^2 &= (x^{k+1} - x^*)^2 + (y^{k+1} - y^*)^2 \\
&= (x^k - \gamma y^k)^2 + (y^k + \gamma x^k)^2 \\
&= (1+\gamma^2)\left[(x^k - x^*)^2 + (y^k - y^*)^2\right] \\
&= (1+\gamma^2)\|z^{k+1} - z^*\|^2.
\end{aligned}$$

Базовые методы: а ля градиентный спуск

- Классический спуск(-подъем):

$$z^{k+1} = \mathrm{proj}_Z(z^k - \gamma_k F(z^k)).$$

- Спуск(-подъем) с дополнительным шагом (Extra Step):

$$\begin{array}{rcl} z^{k+1/2} & = & \mathrm{proj}_Z(z^k - \gamma_k F(z^k)), \\ z^{k+1} & = & \mathrm{proj}_Z(z^k - \gamma_k F(z^{k+1/2})). \end{array}$$

[Korpelevich, 1976, Nemirovski, 2004]

Сходимость: лемма спуска

Лемма спуска для Extra Step

Для одной итерации Extra Step метода справедливо следующее равенство (для любого $u \in \mathcal{Z}$):

$$\begin{aligned} \|z^{k+1} - u\|^2 &= \|z^k - u\|^2 - \|z^{k+1/2} - z^k\|^2 \\ &\quad -2\gamma\langle F(z^{k+1/2}), z^{k+1/2} - u\rangle + \gamma^2\|F(z^{k+1/2}) - F(z^k)\|^2. \end{aligned}$$

Докажем:

$$\begin{aligned} \|z^{k+1} - u\|^2 &= \|z^{k+1} - z^k + z^k - u\|^2 \\ &= \|z^k - u\|^2 + 2\langle z^{k+1} - z^k, z^k - u\rangle + \|z^{k+1} - z^k\|^2 \\ &= \|z^k - u\|^2 + 2\langle z^{k+1} - z^k, z^{k+1} - u\rangle \\ &\quad -2\langle z^{k+1} - z^k, z^{k+1} - z^k\rangle + \|z^{k+1} - z^k\|^2 \\ &= \|z^k - u\|^2 + 2\langle z^{k+1} - z^k, z^{k+1} - u\rangle - \|z^{k+1} - z^k\|^2 \\ &= \|z^k - u\|^2 - 2\gamma\langle F(z^{k+1/2}), z^{k+1} - u\rangle - \|z^{k+1} - z^k\|^2. \end{aligned}$$

Сходимость: лемма спуска

Лемма спуска для Extra Step

Для одной итерации Extra Step метода справедливо следующее равенство (для любого $u \in \mathcal{Z}$):

$$\begin{aligned}\|z^{k+1} - u\|^2 &= \|z^k - u\|^2 - \|z^{k+1/2} - z^k\|^2 \\ &\quad -2\gamma\langle F(z^{k+1/2}), z^{k+1/2} - u\rangle + \gamma^2\|F(z^{k+1/2}) - F(z^k)\|^2.\end{aligned}$$

Докажем:

$$\begin{aligned}\|z^{k+1} - u\|^2 &= \|z^{k+1} - z^k + z^k - u\|^2 \\ &= \|z^k - u\|^2 + 2\langle z^{k+1} - z^k, z^k - u\rangle + \|z^{k+1} - z^k\|^2 \\ &= \|z^k - u\|^2 + 2\langle z^{k+1} - z^k, z^{k+1} - u\rangle \\ &\quad -2\langle z^{k+1} - z^k, z^{k+1} - z^k\rangle + \|z^{k+1} - z^k\|^2 \\ &= \|z^k - u\|^2 + 2\langle z^{k+1} - z^k, z^{k+1} - u\rangle - \|z^{k+1} - z^k\|^2 \\ &= \|z^k - u\|^2 - 2\gamma\langle F(z^{k+1/2}), z^{k+1} - u\rangle - \|z^{k+1} - z^k\|^2.\end{aligned}$$

Сходимость: лемма спуска

Аналогично:

$$
\begin{aligned}
\|z^{k+1/2} - z^{k+1}\|^2 &= \|z^{k+1/2} - z^k + z^k - z^{k+1}\|^2 \\
&= \|z^k - z^{k+1}\|^2 + 2\langle z^{k+1/2} - z^k, z^k - z^{k+1}\rangle + \|z^{k+1/2} - z^k\|^2 \\
&= \|z^k - z^{k+1}\|^2 + 2\langle z^{k+1/2} - z^k, z^{k+1/2} - z^{k+1}\rangle \\
&\quad -2\langle z^{k+1/2} - z^k, z^{k+1/2} - z^k\rangle + \|z^{k+1/2} - z^k\|^2 \\
&= \|z^k - z^{k+1}\|^2 + 2\langle z^{k+1/2} - z^k, z^{k+1/2} - z^{k+1}\rangle - \|z^{k+1/2} - z^k\|^2 \\
&= \|z^k - z^{k+1}\|^2 - 2\gamma\langle F(z^k), z^{k+1/2} - z^{k+1}\rangle - \|z^{k+1/2} - z^k\|^2.
\end{aligned}
$$

Сложим два полученных равенства:

$$
\begin{aligned}
\|z^{k+1} - u\|^2 + \|z^{k+1/2} - z^{k+1}\|^2 &= \|z^k - u\|^2 - \|z^{k+1/2} - z^k\|^2 \\
&\quad -2\gamma\langle F(z^{k+1/2}), z^{k+1} - u\rangle - 2\gamma\langle F(z^k), z^{k+1/2} - z^{k+1}\rangle.
\end{aligned}
$$

Далее немного реорганизуем правую часть:

$$
\begin{aligned}
\|z^{k+1} - u\|^2 + \|z^{k+1/2} - z^{k+1}\|^2 &= \|z^k - u\|^2 - \|z^{k+1/2} - z^k\|^2 \\
&\quad -2\gamma\langle F(z^{k+1/2}), z^{k+1/2} - u\rangle \\
&\quad +2\gamma\langle F(z^{k+1/2}) - F(z^k), z^{k+1/2} - z^{k+1}\rangle.
\end{aligned}
$$

Сходимость: лемма спуска

Аналогично:

$$\begin{aligned}
\|z^{k+1/2} - z^{k+1}\|^2 &= \|z^{k+1/2} - z^k + z^k - z^{k+1}\|^2 \\
&= \|z^k - z^{k+1}\|^2 + 2\langle z^{k+1/2} - z^k, z^k - z^{k+1}\rangle + \|z^{k+1/2} - z^k\|^2 \\
&= \|z^k - z^{k+1}\|^2 + 2\langle z^{k+1/2} - z^k, z^{k+1/2} - z^{k+1}\rangle \\
&\quad -2\langle z^{k+1/2} - z^k, z^{k+1/2} - z^k\rangle + \|z^{k+1/2} - z^k\|^2 \\
&= \|z^k - z^{k+1}\|^2 + 2\langle z^{k+1/2} - z^k, z^{k+1/2} - z^{k+1}\rangle - \|z^{k+1/2} - z^k\|^2 \\
&= \|z^k - z^{k+1}\|^2 - 2\gamma\langle F(z^k), z^{k+1/2} - z^{k+1}\rangle - \|z^{k+1/2} - z^k\|^2.
\end{aligned}$$

Сложим два полученных равенства:

$$\begin{aligned}
\|z^{k+1} - u\|^2 + \|z^{k+1/2} - z^{k+1}\|^2 &= \|z^k - u\|^2 - \|z^{k+1/2} - z^k\|^2 \\
&\quad -2\gamma\langle F(z^{k+1/2}), z^{k+1} - u\rangle - 2\gamma\langle F(z^k), z^{k+1/2} - z^{k+1}\rangle.
\end{aligned}$$

Далее немного реорганизуем правую часть:

$$\begin{aligned}
\|z^{k+1} - u\|^2 + \|z^{k+1/2} - z^{k+1}\|^2 &= \|z^k - u\|^2 - \|z^{k+1/2} - z^k\|^2 \\
&\quad -2\gamma\langle F(z^{k+1/2}), z^{k+1/2} - u\rangle \\
&\quad +2\gamma\langle F(z^{k+1/2}) - F(z^k), z^{k+1/2} - z^{k+1}\rangle.
\end{aligned}$$

Сходимость: лемма спуска

Аналогично:

$$
\begin{aligned}
\|z^{k+1/2} - z^{k+1}\|^2 &= \|z^{k+1/2} - z^k + z^k - z^{k+1}\|^2 \\
&= \|z^k - z^{k+1}\|^2 + 2\langle z^{k+1/2} - z^k, z^k - z^{k+1}\rangle + \|z^{k+1/2} - z^k\|^2 \\
&= \|z^k - z^{k+1}\|^2 + 2\langle z^{k+1/2} - z^k, z^{k+1/2} - z^{k+1}\rangle \\
&\quad -2\langle z^{k+1/2} - z^k, z^{k+1/2} - z^k\rangle + \|z^{k+1/2} - z^k\|^2 \\
&= \|z^k - z^{k+1}\|^2 + 2\langle z^{k+1/2} - z^k, z^{k+1/2} - z^{k+1}\rangle - \|z^{k+1/2} - z^k\|^2 \\
&= \|z^k - z^{k+1}\|^2 - 2\gamma\langle F(z^k), z^{k+1/2} - z^{k+1}\rangle - \|z^{k+1/2} - z^k\|^2.
\end{aligned}
$$

Сложим два полученных равенства:

$$
\begin{aligned}
\|z^{k+1} - u\|^2 + \|z^{k+1/2} - z^{k+1}\|^2 &= \|z^k - u\|^2 - \|z^{k+1/2} - z^k\|^2 \\
&\quad -2\gamma\langle F(z^{k+1/2}), z^{k+1} - u\rangle - 2\gamma\langle F(z^k), z^{k+1/2} - z^{k+1}\rangle.
\end{aligned}
$$

Далее немного реорганизуем правую часть:

$$
\begin{aligned}
\|z^{k+1} - u\|^2 + \|z^{k+1/2} - z^{k+1}\|^2 &= \|z^k - u\|^2 - \|z^{k+1/2} - z^k\|^2 \\
&\quad -2\gamma\langle F(z^{k+1/2}), z^{k+1/2} - u\rangle \\
&\quad +2\gamma\langle F(z^{k+1/2}) - F(z^k), z^{k+1/2} - z^{k+1}\rangle.
\end{aligned}
$$

Сходимость: лемма спуска

Воспользуемся $2ab \leq a^2 + b^2$:

$$\begin{aligned} \|z^{k+1} - u\|^2 + \|z^{k+1/2} - z^{k+1}\|^2 \quad \leq \quad & \|z^k - u\|^2 - \|z^{k+1/2} - z^k\|^2 \\ & -2\gamma\langle F(z^{k+1/2}), z^{k+1/2} - u\rangle \\ & +\gamma^2\|F(z^{k+1/2}) - F(z^k)\|^2 + \|z^{k+1/2} - z^{k+1}\|^2, \end{aligned}$$

что и доказывает лемму.
Произвольное u в данной лемме нужно для монотонных ВН (выпукло-вогнутых седловых задач), в сильно-монотонном случае достаточно взять $u = z^*$.

Сходимость: сильно-монотонный случай

Теорема для Extra Step

В сильно-монотонном случае для Extra Step метода с $\gamma \leq \frac{1}{4L}$ справедлива следующая оценка сходимости (z^K – последняя точка алгоритма, z^0 – начальная):

$$\|z^K - z^*\|^2 \leq (1-\gamma\mu)^K \|z^0 - z^*\|^2.$$

Подставим в лемму $u = z^*$:

$$\begin{aligned}\|z^{k+1} - z^*\|^2 \leq\ & \|z^k - z^*\|^2 - \|z^{k+1/2} - z^k\|^2 \\ & -2\gamma\langle F(z^{k+1/2}), z^{k+1/2} - z^*\rangle + \gamma^2\|F(z^{k+1/2}) - F(z^k)\|^2.\end{aligned}$$

Воспользуемся свойством решения: $\langle F(z^*), z^{k+1/2} - z^*\rangle \geq 0$,

$$\begin{aligned}\|z^{k+1} - z^*\|^2 \leq\ & \|z^k - z^*\|^2 - \|z^{k+1/2} - z^k\|^2 \\ & -2\gamma\langle F(z^{k+1/2}) - F(z^*), z^{k+1/2} - z^*\rangle + \gamma^2\|F(z^{k+1/2}) - F(z^k)\|^2.\end{aligned}$$

Сходимость: сильно-монотонный случай

Теорема для Extra Step

В сильно-монотонном случае для Extra Step метода с $\gamma \leq \frac{1}{4L}$ справедлива следующая оценка сходимости (z^K – последняя точка алгоритма, z^0 – начальная):

$$\|z^K - z^*\|^2 \leq (1 - \gamma\mu)^K \|z^0 - z^*\|^2.$$

Подставим в лемму $u = z^*$:

$$\begin{aligned} \|z^{k+1} - z^*\|^2 \leq\ & \|z^k - z^*\|^2 - \|z^{k+1/2} - z^k\|^2 \\ & -2\gamma\langle F(z^{k+1/2}), z^{k+1/2} - z^*\rangle + \gamma^2\|F(z^{k+1/2}) - F(z^k)\|^2. \end{aligned}$$

Воспользуемся свойством решения: $\langle F(z^*), z^{k+1/2} - z^*\rangle \geq 0$,

$$\begin{aligned} \|z^{k+1} - z^*\|^2 \leq\ & \|z^k - z^*\|^2 - \|z^{k+1/2} - z^k\|^2 \\ & -2\gamma\langle F(z^{k+1/2}) - F(z^*), z^{k+1/2} - z^*\rangle + \gamma^2\|F(z^{k+1/2}) - F(z^k)\|^2. \end{aligned}$$

Сходимость: сильно-монотонный случай

Теорема для Extra Step

В сильно-монотонном случае для Extra Step метода с $\gamma \leq \frac{1}{4L}$ справедлива следующая оценка сходимости (z^K – последняя точка алгоритма, z^0 – начальная):

$$\|z^K - z^*\|^2 \leq (1 - \gamma\mu)^K \|z^0 - z^*\|^2.$$

Подставим в лемму $u = z^*$:

$$\begin{aligned}\|z^{k+1} - z^*\|^2 \leq \; & \|z^k - z^*\|^2 - \|z^{k+1/2} - z^k\|^2 \\ & -2\gamma\langle F(z^{k+1/2}), z^{k+1/2} - z^*\rangle + \gamma^2\|F(z^{k+1/2}) - F(z^k)\|^2.\end{aligned}$$

Воспользуемся свойством решения: $\langle F(z^*), z^{k+1/2} - z^*\rangle \geq 0$,

$$\begin{aligned}\|z^{k+1} - z^*\|^2 \leq \; & \|z^k - z^*\|^2 - \|z^{k+1/2} - z^k\|^2 \\ & -2\gamma\langle F(z^{k+1/2}) - F(z^*), z^{k+1/2} - z^*\rangle + \gamma^2\|F(z^{k+1/2}) - F(z^k)\|^2.\end{aligned}$$

Сходимость: сильно-монотонный случай

Далее применяем сильную монотонность и Липшецевость:

$$\|z^{k+1} - z^*\|^2 \leq \|z^k - z^*\|^2 - \|z^{k+1/2} - z^k\|^2 - 2\gamma\mu\|z^{k+1/2} - z^*\|^2 + \gamma^2 L^2 \|z^{k+1/2} - z^k\|^2.$$

Используем следующее неравенство: $-2\|z^{k+1/2} - z^*\|^2 \leq -\|z^k - z^*\|^2 + 2\|z^{k+1/2} - z^k\|^2$ (модификация $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$):

$$\|z^{k+1} - z^*\|^2 \leq \|z^k - z^*\|^2 - \|z^{k+1/2} - z^k\|^2 - \gamma\mu\|z^k - z^*\|^2 + 2\gamma\mu\|z^{k+1/2} - z^k\|^2 + \gamma^2 L^2 \|z^{k+1/2} - z^k\|^2.$$

Сгруппируем:

$$\|z^{k+1} - z^*\|^2 \leq (1 - \gamma\mu)\|z^k - z^*\|^2 - (1 - 2\gamma\mu + \gamma^2 L^2)\|z^{k+1/2} - z^k\|^2.$$

С $\gamma \leq \frac{1}{4L}$ имеем

$$\|z^{k+1} - z^*\|^2 \leq (1 - \gamma\mu)\|z^k - z^*\|^2.$$

Откуда и следует требуемое утверждение.

Сходимость: сильно-монотонный случай

Далее применяем сильную монотонность и Липшецевость:

$$\begin{aligned}\|z^{k+1} - z^*\|^2 &\leq \|z^k - z^*\|^2 - \|z^{k+1/2} - z^k\|^2 \\ &\quad -2\gamma\mu\|z^{k+1/2} - z^*\|^2 + \gamma^2 L^2 \|z^{k+1/2} - z^k\|^2.\end{aligned}$$

Используем следующее неравенство: $-2\|z^{k+1/2} - z^*\|^2 \leq -\|z^k - z^*\|^2 + 2\|z^{k+1/2} - z^k\|^2$ (модификация $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$):

$$\begin{aligned}\|z^{k+1} - z^*\|^2 &\leq \|z^k - z^*\|^2 - \|z^{k+1/2} - z^k\|^2 \\ &\quad -\gamma\mu\|z^k - z^*\|^2 + 2\gamma\mu\|z^{k+1/2} - z^k\|^2 + \gamma^2 L^2 \|z^{k+1/2} - z^k\|^2.\end{aligned}$$

Сгруппируем:

$$\|z^{k+1} - z^*\|^2 \leq (1-\gamma\mu)\|z^k - z^*\|^2 - (1 - 2\gamma\mu + \gamma^2 L^2)\|z^{k+1/2} - z^k\|^2.$$

С $\gamma \leq \frac{1}{4L}$ имеем

$$\|z^{k+1} - z^*\|^2 \leq (1-\gamma\mu)\|z^k - z^*\|^2.$$

Откуда и следует требуемое утверждение.

Сходимость: сильно-монотонный случай

Далее применяем сильную монотонность и Липшецевость:

$$\|z^{k+1} - z^*\|^2 \leq \|z^k - z^*\|^2 - \|z^{k+1/2} - z^k\|^2 - 2\gamma\mu\|z^{k+1/2} - z^*\|^2 + \gamma^2 L^2\|z^{k+1/2} - z^k\|^2.$$

Используем следующее неравенство: $-2\|z^{k+1/2} - z^*\|^2 \leq -\|z^k - z^*\|^2 + 2\|z^{k+1/2} - z^k\|^2$ (модификация $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$):

$$\|z^{k+1} - z^*\|^2 \leq \|z^k - z^*\|^2 - \|z^{k+1/2} - z^k\|^2 - \gamma\mu\|z^k - z^*\|^2 + 2\gamma\mu\|z^{k+1/2} - z^k\|^2 + \gamma^2 L^2\|z^{k+1/2} - z^k\|^2.$$

Сгруппируем:

$$\|z^{k+1} - z^*\|^2 \leq (1-\gamma\mu)\|z^k - z^*\|^2 - (1 - 2\gamma\mu + \gamma^2 L^2)\|z^{k+1/2} - z^k\|^2.$$

С $\gamma \leq \frac{1}{4L}$ имеем

$$\|z^{k+1} - z^*\|^2 \leq (1-\gamma\mu)\|z^k - z^*\|^2.$$

Откуда и следует требуемое утверждение.

Сходимость: сильно-монотонный случай

Далее применяем сильную монотонность и Липшецевость:

$$\begin{aligned}\|z^{k+1} - z^*\|^2 &\leq \|z^k - z^*\|^2 - \|z^{k+1/2} - z^k\|^2 \\ &\quad -2\gamma\mu\|z^{k+1/2} - z^*\|^2 + \gamma^2 L^2 \|z^{k+1/2} - z^k\|^2.\end{aligned}$$

Используем следующее неравенство: $-2\|z^{k+1/2} - z^*\|^2 \leq -\|z^k - z^*\|^2 + 2\|z^{k+1/2} - z^k\|^2$ (модификация $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$):

$$\begin{aligned}\|z^{k+1} - z^*\|^2 &\leq \|z^k - z^*\|^2 - \|z^{k+1/2} - z^k\|^2 \\ &\quad -\gamma\mu\|z^k - z^*\|^2 + 2\gamma\mu\|z^{k+1/2} - z^k\|^2 + \gamma^2 L^2 \|z^{k+1/2} - z^k\|^2.\end{aligned}$$

Сгруппируем:

$$\|z^{k+1} - z^*\|^2 \leq (1 - \gamma\mu)\|z^k - z^*\|^2 - (1 - 2\gamma\mu + \gamma^2 L^2)\|z^{k+1/2} - z^k\|^2.$$

С $\gamma \leq \frac{1}{4L}$ имеем

$$\|z^{k+1} - z^*\|^2 \leq (1 - \gamma\mu)\|z^k - z^*\|^2.$$

Откуда и следует требуемое утверждение.

Сходимости: сравнение с выпуклой минимизацией

Суммируем результаты в виде таблицы:

Случай	Верхняя оценка ВН - Extra Step	Нижняя оценка (лучше не получится)	Верхняя оценка Оптимизация - Нестеров
Сильно-вып.	$\mathcal{O}\left(\left(1-\frac{\mu}{L}\right)^K \|z^0 - z^*\|^2\right)$	$\Omega\left(\left(1-\frac{8\mu}{L}\right)^K \|z^0 - z^*\|^2\right)$	$\mathcal{O}\left(\left(1-\sqrt{\frac{\mu}{L}}\right)^K \|z^0 - z^*\|^2\right)$
Выпуклый	$\mathcal{O}\left(\frac{LD_Z^2}{K}\right)$	$\Omega\left(\frac{LD_Z^2}{K}\right)$	$\mathcal{O}\left(\frac{LD_Z^2}{K^2}\right)$

- Extra Step – оптимальный метод для седел и ВН.
- Седла нельзя ускорить (если не углубляться в спицифику константа L, μ).
- Напомню, для выпуклой оптимизации оптимальным методом является, например, ускоренный метод Нестерова. В том числе из-за этого ее рассматривают отдельно от ВН.

Метод с дополнительным шагом

Плюсы:

- Взгляд в "будущие": делает шаг по "градиенту" в "будущей" точке.
- Оптимальный теоретический анализ.
- Показывает лучшие результаты на практике (по количеству итераций).

Минусы:

- Два вызова "градиента" за итерацию.
- Часто уступает на практике обычному спуску по количеству вызовов "градиента".

Метод с дополнительным шагом

Плюсы:

- Взгляд в "будущие": делает шаг по "градиенту" в "будущей" точке.
- Оптимальный теоретический анализ.
- Показывает лучшие результаты на практике (по количеству итераций).

Минусы:

- Два вызова "градиента" за итерацию.
- Часто уступает на практике обычному спуску по количеству вызовов "градиента".

Боремся с минусом

Рассмотрим следующие модификации:

- Past Extra Step Method

$$\begin{aligned} z^{k+1/2} &= \text{proj}_Z(z^k - \gamma F(z^{k-1/2})), \\ z^{k+1} &= \text{proj}_Z(z^k - \gamma F(z^{k+1/2})). \end{aligned}$$

- Optimistic Extra Step Method

$$\begin{aligned} z^{k+1/2} &= \text{proj}_Z(z^k - \gamma F(z^{k-1/2})), \\ z^{k+1} &= z^{k+1/2} - \gamma F(z^{k+1/2}) + \gamma F(z^{k-1/2}). \end{aligned}$$

- Reflected Extra Step Method

$$\begin{aligned} z^{k+1/2} &= 2 \cdot z^k - z^{k-1}, \\ z^{k+1} &= \text{proj}_Z(z^k - \gamma F(z^{k+1/2})). \end{aligned}$$

[Popov, 1980, Hsieh et al., 2019]

Немного об ускорениях

- Еще раз: ВН и седла в теории не ускоряются!
- Рассмотрим все же моментные методы: $z^{k+1} = z^k - \gamma_k F(z^k) + \beta(z^k - z^{k-1})$
- Но есть интуиция, что для седел лучше использовать отрицательные моменты ($\beta < 0$):

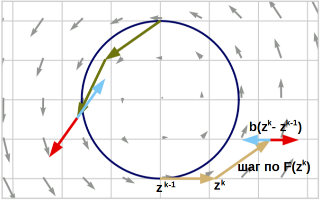

Figure: Рассматривается вихревое, спиралевидное поле, равное 0 в центре – решение. Коричневым показана траектория метода, красным и синим способы выбора моментума: красный – положительным, синий – отрицательным. Видно, что отрицательный лучше - он тянет в центр к решению. Зеленая траектория – альтернативный подход к выбору момента. Картинка отсюда.

Немного об ускорениях

- Еще раз: ВН и седла в теории не ускоряются!
- Рассмотрим все же моментные методы:
  $z^{k+1} = z^k - \gamma_k F(z^k) + \beta(z^k - z^{k-1})$
- Но есть интуиция, что для седел лучше использовать отрицательные моменты ($\beta < 0$):

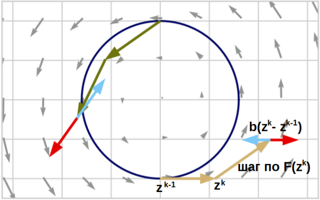

Figure: Рассматривается вихревое, спиралевидное поле, равное 0 в центре – решение. Коричневым показана траектория метода, красным и синим способы выбора моментума: красный – положительным, синий – отрицательным. Видно, что отрицательный лучше - он тянет в центр к решению. Зеленая траектория – альтернативный подход к выбору момента. Картинка отсюда.

Немного об ускорениях

- Еще раз: ВН и седла в теории не ускоряются!
- Рассмотрим все же моментные методы:
  $z^{k+1} = z^k - \gamma_k F(z^k) + \beta(z^k - z^{k-1})$
- Но есть интуиция, что для седел лучше использовать отрицательные моменты ($\beta < 0$):

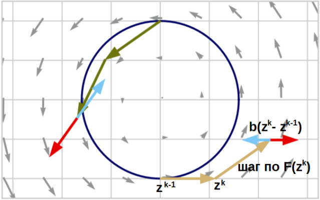

Figure: Рассматривается вихревое, спиралевидное поле, равное 0 в центре – решение. Коричневым показана траектория метода, красным и синим способы выбора моментума: красный – положительным, синий – отрицательным. Видно, что отрицательный лучше - он тянет в центр к решению. Зеленая траектория – альтернативный подход к выбору момента. Картинка отсюда.

Стохастические методы

Решаются те же самые проблемы ВН или седловой точки: 1) Найти $z^* \in \mathcal{Z}$ такую, что $\langle F(z^*), z - z^* \rangle \geq 0, \ \forall z \in \mathcal{Z}$ или 2) $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} g(x, y)$.
Но теперь мы не имеем доступа к точным значениям $F(z)$ или $g(x, y)$.
Выделим, две популярные стохастические постановки:

- Мы имеем доступ к некоторой стохастической реализации $F(z, \xi)$ или $g(x, y, \xi)$. ξ может отвечать за некоторый шум, например: $F(z, \xi) = F(z) + \xi$; или ξ – это номер батча, который выбирается случайно.
- Монте-Карло постановка:

$$F(z) = \frac{1}{M} \sum_{m=1}^{M} F_m(z) \text{ или } g(x, y) = \frac{1}{M} \sum_{m=1}^{M} g_m(x, y).$$

Интерпретация: батч, выбирается случайный на каждой итерации.

Стохастические методы

Решаются те же самые проблемы ВН или седловой точки: 1) Найти $z^* \in \mathcal{Z}$ такую, что $\langle F(z^*), z - z^* \rangle \geq 0, \ \forall z \in \mathcal{Z}$ или 2) $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} g(x, y)$.
Но теперь мы не имеем доступа к точным значениям $F(z)$ или $g(x, y)$.
Выделим, две популярные стохастические постановки:

- Мы имеем доступ к некоторой стохастической реализации $F(z, \xi)$ или $g(x, y, \xi)$. ξ может отвечать за некоторый шум, например: $F(z, \xi) = F(z) + \xi$; или ξ – это номер батча, который выбирается случайно.
- Монте-Карло постановка:

$$F(z) = \frac{1}{M} \sum_{m=1}^{M} F_m(z) \text{ или } g(x, y) = \frac{1}{M} \sum_{m=1}^{M} g_m(x, y).$$

Интерпретация: батч, выбирается случайный на каждой итерации.

Стохастические методы

Решаются те же самые проблемы ВН или седловой точки: 1) Найти $z^* \in \mathcal{Z}$ такую, что $\langle F(z^*), z - z^* \rangle \geq 0, \ \forall z \in \mathcal{Z}$ или 2) $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} g(x, y)$.
Но теперь мы не имеем доступа к точным значениям $F(z)$ или $g(x, y)$.
Выделим, две популярные стохастические постановки:

- Мы имеем доступ к некоторой стохастической реализации $F(z, \xi)$ или $g(x, y, \xi)$. ξ может отвечать за некоторый шум, например: $F(z, \xi) = F(z) + \xi$; или ξ – это номер батча, который выбирается случайно.
- Монте-Карло постановка:

$$F(z) = \frac{1}{M} \sum_{m=1}^{M} F_m(z) \text{ или } g(x, y) = \frac{1}{M} \sum_{m=1}^{M} g_m(x, y).$$

Интерпретация: батч, выбирается случайный на каждой итерации.

Базовый Метод: все тот же

- Стохастический спуск(-подъем) с дополнительным шагом (Extra Step):

$$
\begin{aligned}
z^{k+1/2} &= \text{proj}_Z(z^k - \gamma_k F(z^k, \xi^k)), \\
z^{k+1} &= \text{proj}_Z(z^k - \gamma_k F(z^{k+1/2}, \xi^{k+1/2})).
\end{aligned}
$$

- Предположения: $\mathbb{E}_\xi F(z,\xi) = F(z)$, $\mathbb{E}_\xi \|F(z,\xi) - F(z)\|^2 \le \sigma^2$.
- Сходимость: при постоянном шаге к окрестности решения.

[Juditsky et al., 2011]

Метод: сходимость

- Оценки сходимости:

Случай	Верхняя оценка ВН - Extra Step	Верхняя оценка Оптимизация - Нестеров
Сильно-вып.	$\mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^K \lVert z^0 - z^* \rVert^2 + \frac{\sigma^2}{\mu K}\right)$	$\mathcal{O}\left(\left(1 - \sqrt{\frac{\mu}{L}}\right)^K \lVert z^0 - z^* \rVert^2 + \frac{\sigma^2}{\mu K}\right)$
Выпуклый	$\mathcal{O}\left(\frac{LD_Z^2}{K} + \frac{\sigma D_Z}{\sqrt{K}}\right)$	$\mathcal{O}\left(\frac{LD_Z^2}{K^2} + \frac{\sigma D_Z}{\sqrt{K}}\right)$

Метод: сходимость

Билинейная задача:

$$\min_{x,y\in[-1;1]^n}\max \left(x^T A y + b^T x + c^T y\right),$$

Концепция: $F(z,\xi) = F(z) + \xi$, где $\mathbb{E}\xi = 0$, $\mathbb{E}\|\xi\|^2 = \sigma^2$. Подбирается шаг, обеспечивающий наилучшую сходимость.

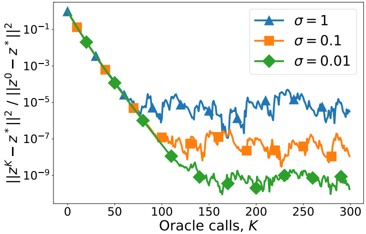

Figure: Сходимость стохастического Extra Step метода с разным уровнем шума σ.

Модификации: Revisiting Extra Step

- Идея использовать ту же случайность ξ^k на обоих шагах:

$$\begin{aligned} z^{k+1/2} &= \mathrm{proj}_Z(z^k - \gamma_k F(z^k, \xi^k)), \\ z^{k+1} &= \mathrm{proj}_Z(z^k - \gamma_k F(z^{k+1/2}, \xi^k)). \end{aligned}$$

- В Монте-Карло постановке:

$$\begin{aligned} z^{k+1/2} &= \mathrm{proj}_Z(z^k - \gamma_k F_{\pi_k}(z^k)), \\ z^{k+1} &= \mathrm{proj}_Z(z^k - \gamma_k F_{\pi_k}(z^{k+1/2})). \end{aligned}$$

[Mishchenko et al., 2020, Gorbunov et al., 2021]

Модификации: Revisiting Extra Step

- Идея использовать ту же случайность ξ^k на обоих шагах:

$$\begin{aligned} z^{k+1/2} &= \text{proj}_Z(z^k - \gamma_k F(z^k, \xi^{\text{k}})), \\ z^{k+1} &= \text{proj}_Z(z^k - \gamma_k F(z^{k+1/2}, \xi^k)). \end{aligned}$$

- В Монте-Карло постановке:

$$\begin{aligned} z^{k+1/2} &= \text{proj}_Z(z^k - \gamma_k F_{\pi_k}(z^k)), \\ z^{k+1} &= \text{proj}_Z(z^k - \gamma_k F_{\pi_k}(z^{k+1/2})). \end{aligned}$$

[Mishchenko et al., 2020, Gorbunov et al., 2021]

Revisiting Extra Step: сходимость

Билинейная задача:

$$\min_{x,y\in[-1;1]^n}\max \left(x^T A y\right),$$

Концепция №1: $F(z,\xi) = F(z) + \xi$, где $\mathbb{E}\xi = 0$, $\mathbb{E}\|\xi\|^2 = \sigma^2$.

Концепция №2 (Монте-Карло): $A = \frac{1}{M}\sum\limits_{m=1}^{M} A_m$. Выбирается случайная матрица A_m.

Подбирается шаг, обеспечивающий наилучшую сходимость. Более того, в этом эксперименте он оказался одинаковым.

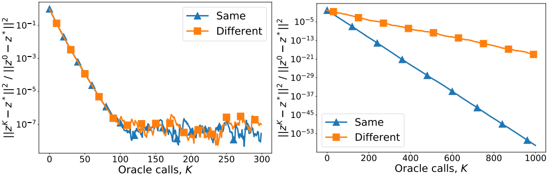

Figure: Сравнение обычного Extra Step с разными ξ с Revisiting Extra Step: левый график – для концепции $F(z,\xi) = F(z) + \xi$, правый график – для Монте-Карло концепции

Variance Reduction: 5 подходов

- Стандартный:

$$z^{k+1} = \text{proj}_Z(z^k - \gamma(F_{\pi_k}(z^k) - F_{\pi_k}(w^k) + F(w^k))),$$

- Extra Step:

$$z^{k+1/2} = \text{proj}_Z(z^k - \gamma(F_{\pi_k}(z^k) - F_{\pi_k}(w^k) + F(w^k))),$$

$$z^{k+1} = \text{proj}_Z(z^k - \gamma(F_{\pi_{k+1/2}}(z^k) - F_{\pi_{k+1/2}}(w^k) + F(w^k))),$$

- Extra Step Revisiting:

$$z^{k+1/2} = \text{proj}_Z(z^k - \gamma(F_{\pi_k}(z^k) - F_{\pi_k}(w^k) + F(w^k)),$$

$$z^{k+1} = z^k - \gamma(F_{\pi_k}(z^k) - F_{\pi_k}(w^k) + F(w^k))).$$

Variance Reduction: 5 подходов

- Стандартный:

$$z^{k+1} = \mathrm{proj}_Z(z^k - \gamma(F_{\pi_k}(z^k) - F_{\pi_k}(w^k) + F(w^k))),$$

- Extra Step:

$$z^{k+1/2} = \mathrm{proj}_Z(z^k - \gamma(F_{\pi_k}(z^k) - F_{\pi_k}(w^k) + F(w^k))),$$

$$z^{k+1} = \mathrm{proj}_Z(z^k - \gamma(F_{\pi_{k+1/2}}(z^k) - F_{\pi_{k+1/2}}(w^k) + F(w^k))),$$

- Extra Step Revisiting:

$$z^{k+1/2} = \mathrm{proj}_Z(z^k - \gamma(F_{\pi_k}(z^k) - F_{\pi_k}(w^k) + F(w^k)),$$

$$z^{k+1} = z^k - \gamma(F_{\pi_k}(z^k) - F_{\pi_k}(w^k) + F(w^k))).$$

Variance Reduction: 5 подходов

- Стандартный:

$$z^{k+1} = \text{proj}_Z(z^k - \gamma(F_{\pi_k}(z^k) - F_{\pi_k}(w^k) + F(w^k))),$$

- Extra Step:

$$z^{k+1/2} = \text{proj}_Z(z^k - \gamma(F_{\pi_k}(z^k) - F_{\pi_k}(w^k) + F(w^k))),$$

$$z^{k+1} = \text{proj}_Z(z^k - \gamma(F_{\pi_{k+1/2}}(z^k) - F_{\pi_{k+1/2}}(w^k) + F(w^k))),$$

- Extra Step Revisiting:

$$z^{k+1/2} = \text{proj}_Z(z^k - \gamma(F_{\pi_k}(z^k) - F_{\pi_k}(w^k) + F(w^k)),$$

$$z^{k+1} = z^k - \gamma(F_{\pi_k}(z^k) - F_{\pi_k}(w^k) + F(w^k))).$$

Variance Reduction: 5 подходов

- Но правильные и наилучшие оценки дает:

$$z^{k+1/2} = \mathrm{proj}_Z(z^k - \alpha(z^k - w^k) - \gamma F(w^k)),$$

$$z^{k+1} = \mathrm{proj}_Z(z^k - \alpha(z^k - w^k) - \gamma(F_{\pi_k}(z^k) - F_{\pi_k}(w^k) + F(w^k)))$$

$$w^{k+1} = \begin{cases} z^{k+1} & \text{with small probability } p, \\ w^k & \text{with probability } 1 - p. \end{cases}$$

- Негативный моментум является ключевым в теоретическом анализе.

[Alacaoglu and Malitsky, 2021]

Variance Reduction: 5 подходов

- А также:

$$\Delta^k = \frac{1}{b} \sum_{j \in S^k} \left(F_{\pi_j}(z^k) - F_{\pi_j}(w^{k-1}) + \eta(F_{\pi_j}(z^k) - F_{\pi_j}(z^{k-1})) \right) + F(w^{k-1}),$$

$$z^{k+1} = \mathrm{proj}_Z(z^k - \alpha(z^k - w^k) - \gamma\Delta^k))$$

$$w^{k+1} = \begin{cases} z^{k+1} & \text{with small probability } p, \\ w^k & \text{with probability } 1 - p. \end{cases}$$

[Kovalev et al., 2022]

Table: Summary complexities for finding an ε-solution for strongly monotone stochastic (finite-sum) variational inequality.

Reference	Complexity	Weaknesses
[Palaniappan and Bach, 2016]	$\mathcal{O}\left(n + \sqrt{bn}\frac{L}{\mu}\log^2\frac{1}{\varepsilon}\right)$	envelope acceleration batching
[Chavdarova et al., 2019]	$\mathcal{O}\left(n + b\frac{L^2}{\mu^2}\log\frac{1}{\varepsilon}\right)$	bad complexity batching
[Carmon et al., 2019]	$\mathcal{O}\left(n + \sqrt{bn}\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$	for games only batching
[Yang et al., 2020]	$\mathcal{O}\left(b^{\frac{1}{3}}n^{\frac{2}{3}}\frac{L^3}{\mu^3}\log\frac{1}{\varepsilon}\right)$	bad complexity batching
[Alacaoglu and Malitsky, 2021]	$\mathcal{O}\left(n + \sqrt{bn}\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$	batching
[Alacaoglu et al., 2021]	$\mathcal{O}\left(n + n\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$	bad rates
[Tominin et al., 2021]	$\mathcal{O}\left(n + \sqrt{bn}\frac{L}{\mu}\log^2\frac{1}{\varepsilon}\right)$	envelope acceleration batching
[Kovalev et al., 2022]	$\mathcal{O}\left(n + \sqrt{n}\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$	
Нижняя оценка	$\Omega\left(n + \sqrt{n}\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$	

Распределенные ВН с сжатие через VR

- Распределенная задача:

$$F(z) = \frac{1}{M}\sum_{m=1}^{M} F_m(z),$$

  где F_m распределены по вычислительным устройствам.
- Коммуникации стоят значительно дороже!
- Для удешевления коммуникаций можно сжимать информацию:

$$\mathbb{E}Q(z) = z, \quad \mathbb{E}\|Q(z)\|^2 \leq q\|z\|^2.$$

Распределенные ВН с сжатие через VR

- Распределенная задача:

$$F(z) = \frac{1}{M} \sum_{m=1}^{M} F_m(z),$$

где F_m распределены по вычислительным устройствам.

- Коммуникации стоят значительно дороже!
- Для удешевления коммуникаций можно сжимать информацию:

$$\mathbb{E} Q(z) = z, \quad \mathbb{E}\|Q(z)\|^2 \leq q\|z\|^2.$$

Распределенные ВН с сжатие через VR

- Распределенная задача:

$$F(z) = \frac{1}{M}\sum_{m=1}^{M} F_m(z),$$

  где F_m распределены по вычислительным устройствам.
- Коммуникации стоят значительно дороже!
- Для удешевления коммуникаций можно сжимать информацию:

$$\mathrm{E}Q(z) = z, \quad \mathrm{E}\|Q(z)\|^2 \le q\|z\|^2.$$

Распределенные ВН с сжатие через VR

- Локально считаем

$$z^{k+1/2} = z^k - \alpha(z^k - w^k) - \gamma F(w^k)$$

- Посылаем на сервер

$$g_m^k = Q_m^{\text{dev}}(F_m(z^{k+1/2}) - F_m(w^k))$$

- Получаем усредненный ответ

$$g^k = Q^{\text{serv}}\left[\frac{1}{M}\sum_{m=1}^{M} g_m^k\right]$$

- Локально обновляемся

$$z^{k+1} = z^{k+1/2} - \gamma g^k$$

$$w^{k+1} = \begin{cases} z^{k+1} & \text{with small probability } p, \\ w^k & \text{with probability } 1-p. \end{cases}$$

[Beznosikov et al., 2021]

Alacaoglu, A. and Malitsky, Y. (2021).
Stochastic variance reduction for variational inequality methods.
arXiv preprint arXiv:2102.08352.

Alacaoglu, A., Malitsky, Y., and Cevher, V. (2021).
Forward-reflected-backward method with variance reduction.
Computational Optimization and Applications, 80.

Beznosikov, A., Richtárik, P., Diskin, M., Ryabinin, M., and Gasnikov, A. (2021).
Distributed methods with compressed communication for solving variational inequalities, with theoretical guarantees.
arXiv preprint arXiv:2110.03313.

Carmon, Y., Jin, Y., Sidford, A., and Tian, K. (2019).
Variance reduction for matrix games.
arXiv preprint arXiv:1907.02056.

Chavdarova, T., Gidel, G., Fleuret, F., and Lacoste-Julien, S. (2019).
Reducing noise in gan training with variance reduced extragradient.
arXiv preprint arXiv:1904.08598.

Gorbunov, E., Berard, H., Gidel, G., and Loizou, N. (2021).
Stochastic extragradient: General analysis and improved rates.
arXiv preprint arXiv:2111.08611.

Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. (2019).
On the convergence of single-call stochastic extra-gradient methods.
Advances in Neural Information Processing Systems, 32.

Juditsky, A., Nemirovski, A., and Tauvel, C. (2011).
Solving variational inequalities with stochastic mirror-prox algorithm.
Stochastic Systems, 1(1):17–58.

Korpelevich, G. M. (1976).
The extragradient method for finding saddle points and other problems.
Matecon, 12:747–756.

Kovalev, D., Beznosikov, A., Sadiev, A., Persiianov, M., Richtárik, P., and Gasnikov, A. (2022).
Optimal algorithms for decentralized stochastic variational inequalities.
arXiv preprint arXiv:2202.02771.

Mishchenko, K., Kovalev, D., Shulgin, E., Richtárik, P., and Malitsky, Y. (2020).
Revisiting stochastic extragradient.
In International Conference on Artificial Intelligence and Statistics, pages 4573–4582. PMLR.

Nemirovski, A. (2004).
Prox-method with rate of convergence $o(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems.
SIAM Journal on Optimization, 15:229–251.

Palaniappan, B. and Bach, F. (2016).
Stochastic variance reduction methods for saddle-point problems.
In Advances in Neural Information Processing Systems, pages 1416–1424.

Popov, L. D. (1980).
A modification of the arrow-hurwicz method for search of saddle points.
Mathematical notes of the Academy of Sciences of the USSR, 28(5):845–848.

Tominin, V., Tominin, Y., Borodich, E., Kovalev, D., Gasnikov, A., and Dvurechensky, P. (2021).
On accelerated methods for saddle-point problems with composite structure.
arXiv preprint arXiv:2103.09344.

Yang, J., Kiyavash, N., and He, N. (2020).
Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems.
arXiv preprint arXiv:2002.09621.