Седловые задачи и вариационные неравенства

Александр Безносиков

Управление, информация и оптимизация

12 июня 2021

План лекции

- Постановка задачи
- Основные методы
- Стохастические методы
- Практические применения

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: 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

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

- 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

- 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

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

- 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

- 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

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

- 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

- 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

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

- 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: билинейные седловые задачи

- Как квадратичная задача $\min_x x^T Ax$ является краеугольным камнем минимизации, так билинейная задача – база для седел:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} x^T Ay + b^T x + c^T y.$$

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

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

Если $\mu = 0$, то говорят, что оператор просто монотонен.

Аналогия: сильная выпуклость и выпуклость.
Для седел можно 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

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

Если $\mu = 0$, то говорят, что оператор просто монотонен.

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

А есть ли смысл рассматривать минимизацию и седла отдельно?

- ВН – обобщение, которое включает в себя и оптимизацию, и седловые задачи.
- Вопрос: а есть ли смысл отдельно рассматривать минимизацию и седловые задачи? Их же можно рассмотреть в общности ВН.
- Главная проблема: в случае ВН мы не знаем о существовании функции f и g, мы оперируем только с F. Это сильно ограничивает нас в теоретическом анализе.
- Из предположений с предыдущего слайда дополнительно ничего не следует (вспомните, как много свойств есть у Липшецевости градиента – см.книгу Ю.Нестерова стр.85)
- Теория для ВН и седел развита куда слабее, чем для оптимизации. В частности, ВН и седловые задачи в общем случае не ускоряются (вспомните, ускорение Нестерова и тяжелый шарик). Пока неизвестен, качественный анализ для покоординатных методов, методов с неограниченным шумом.

А есть ли смысл рассматривать минимизацию и седла отдельно?

- ВН – обобщение, которое включает в себя и оптимизацию, и седловые задачи.
- Вопрос: а есть ли смысл отдельно рассматривать минимизацию и седловые задачи? Их же можно рассмотреть в общности ВН.
- Главная проблема: в случае ВН мы не знаем о существовании функции f и g, мы оперируем только с F. Это сильно ограничивает нас в теоретическом анализе.
- Из предположений с предыдущего слайда дополнительно ничего не следует (вспомните, как много свойств есть у Липшецевости градиента – см.книгу Ю.Нестерова стр.85)
- Теория для ВН и седел развита куда слабее, чем для оптимизации. В частности, ВН и седловые задачи в общем случае не ускоряются (вспомните, ускорение Нестерова и тяжелый шарик). Пока неизвестен, качественный анализ для покоординатных методов, методов с неограниченным шумом.

А есть ли смысл рассматривать минимизацию и седла отдельно?

- ВН – обобщение, которое включает в себя и оптимизацию, и седловые задачи.
- Вопрос: а есть ли смысл отдельно рассматривать минимизацию и седловые задачи? Их же можно рассмотреть в общности ВН.
- Главная проблема: в случае ВН мы не знаем о существовании функции f и g, мы оперируем только с F. Это сильно ограничивает нас в теоретическом анализе.
- Из предположений с предыдущего слайда дополнительно ничего не следует (вспомните, как много свойств есть у Липшецевости градиента – см.книгу Ю.Нестерова стр.85)
- Теория для ВН и седел развита куда слабее, чем для оптимизации. В частности, ВН и седловые задачи в общем случае не ускоряются (вспомните, ускорение Нестерова и тяжелый шарик). Пока неизвестен, качественный анализ для покоординатных методов, методов с неограниченным шумом.

А есть ли смысл рассматривать минимизацию и седла отдельно?

- ВН – обобщение, которое включает в себя и оптимизацию, и седловые задачи.
- Вопрос: а есть ли смысл отдельно рассматривать минимизацию и седловые задачи? Их же можно рассмотреть в общности ВН.
- Главная проблема: в случае ВН мы не знаем о существовании функции f и g, мы оперируем только с F. Это сильно ограничивает нас в теоретическом анализе.
- Из предположений с предыдущего слайда дополнительно ничего не следует (вспомните, как много свойств есть у Липшецевости градиента – см.книгу Ю.Нестерова стр.85)
- Теория для ВН и седел развита куда слабее, чем для оптимизации. В частности, ВН и седловые задачи в общем случае не ускоряются (вспомните, ускорение Нестерова и тяжелый шарик). Пока неизвестен, качественный анализ для покоординатных методов, методов с неограниченным шумом.

А есть ли смысл рассматривать минимизацию и седла отдельно?

- ВН – обобщение, которое включает в себя и оптимизацию, и седловые задачи.
- Вопрос: а есть ли смысл отдельно рассматривать минимизацию и седловые задачи? Их же можно рассмотреть в общности ВН.
- Главная проблема: в случае ВН мы не знаем о существовании функции f и g, мы оперируем только с F. Это сильно ограничивает нас в теоретическом анализе.
- Из предположений с предыдущего слайда дополнительно ничего не следует (вспомните, как много свойств есть у Липшецевости градиента – см.книгу Ю.Нестерова стр.85)
- Теория для ВН и седел развита куда слабее, чем для оптимизации. В частности, ВН и седловые задачи в общем случае не ускоряются (вспомните, ускорение Нестерова и тяжелый шарик). Пока неизвестен, качественный анализ для покоординатных методов, методов с неограниченным шумом.

Интуиция

- Хотим построить итеративный метод по аналогии с методами спуска.
- В случае оптимизации $F(z) = \nabla f(z)$.
- Почему бы тогда не взять все оптимизационные методы и вместо $\nabla f(z)$ поставить $F(z)$?
- Для седел тоже кажется, что все натурально и естественно: по минимизируемой переменной – спускаемся, по максимизируемой – поднимаемся.
- С точки зрения практики 0 проблем. Перебираем все методы оптимизации и смотрим, какие хорошо работают для седел и ВН.
- Дьявол – в теории.

Интуиция

- Хотим построить итеративный метод по аналогии с методами спуска.
- В случае оптимизации $F(z) = \nabla f(z)$.
- Почему бы тогда не взять все оптимизационные методы и вместо $\nabla f(z)$ поставить $F(z)$?
- Для седел тоже кажется, что все натурально и естественно: по минимизируемой переменной – спускаемся, по максимизируемой – поднимаемся.
- С точки зрения практики 0 проблем. Перебираем все методы оптимизации и смотрим, какие хорошо работают для седел и ВН.
- Дьявол – в теории.

Интуиция

- Хотим построить итеративный метод по аналогии с методами спуска.
- В случае оптимизации $F(z) = \nabla f(z)$.
- Почему бы тогда не взять все оптимизационные методы и вместо $\nabla f(z)$ поставить $F(z)$?
- Для седел тоже кажется, что все натурально и естественно: по минимизируемой переменной – спускаемся, по максимизируемой – поднимаемся.
- С точки зрения практики 0 проблем. Перебираем все методы оптимизации и смотрим, какие хорошо работают для седел и ВН.
- Дьявол – в теории.

Интуиция

- Хотим построить итеративный метод по аналогии с методами спуска.
- В случае оптимизации $F(z) = \nabla f(z)$.
- Почему бы тогда не взять все оптимизационные методы и вместо $\nabla f(z)$ поставить $F(z)$?
- Для седел тоже кажется, что все натурально и естественно: по минимизируемой переменной – спускаемся, по максимизируемой – поднимаемся.
- С точки зрения практики 0 проблем. Перебираем все методы оптимизации и смотрим, какие хорошо работают для седел и ВН.
- Дьявол – в теории.

Интуиция

- Хотим построить итеративный метод по аналогии с методами спуска.
- В случае оптимизации $F(z) = \nabla f(z)$.
- Почему бы тогда не взять все оптимизационные методы и вместо $\nabla f(z)$ поставить $F(z)$?
- Для седел тоже кажется, что все натурально и естественно: по минимизируемой переменной – спускаемся, по максимизируемой – поднимаемся.
- С точки зрения практики 0 проблем. Перебираем все методы оптимизации и смотрим, какие хорошо работают для седел и ВН.
- Дьявол – в теории.

Интуиция

- Хотим построить итеративный метод по аналогии с методами спуска.
- В случае оптимизации $F(z) = \nabla f(z)$.
- Почему бы тогда не взять все оптимизационные методы и вместо $\nabla f(z)$ поставить $F(z)$?
- Для седел тоже кажется, что все натурально и естественно: по минимизируемой переменной – спускаемся, по максимизируемой – поднимаемся.
- С точки зрения практики 0 проблем. Перебираем все методы оптимизации и смотрим, какие хорошо работают для седел и ВН.
- Дьявол – в теории.

Как измерять сходимость и качество решения?

- По аргументу $\|z^k - z^*\|^2$ – все аналогично выпуклой оптимизации, самый надежный критерий.
- По функции. Напомню, что в оптимизации измеряли по $f(x^k) - f(x^*)$. Попробуем сконструировать что-то аналогичное для ВН и седловых задач.
- Попробуем: $g(x^k, y^k) - g(x^*, y^*)$. Хорош ли такой критерий?
  Рассмотрим самую простую седловую задачу $\min_x \max_y (x-1)\cdot(y+1)$. Решение этой задачи $x = 1, y = -1, g(1,-1) = 0$.
  Пусть начальная точка $x^0 = 0, y^0 = 0, g(0,0) = -1$. Тогда $g(x^0, y^0) - g(x^*, y^*)$ отрицательно. Такой критерий не подходит.
- Другой вариант: $g(x^k, y^*) - g(x^*, y^k)$.
  Рассмотрим самую простую седловую задачу $\min_x \max_y (x-1)\cdot(y+1)$. Решение этой задачи $x = 1, y = -1, g(1,-1) = 0$. Тогда $g(x, y^*) - g(x^*, y) = 0$. Такой критерий не подходит для выпукло-вогнутых седел, но подходит для сильно-выпукло–сильно-вогнутых.

Как измерять сходимость и качество решения?

- По аргументу $\|z^k - z^*\|^2$ – все аналогично выпуклой оптимизации, самый надежный критерий.
- По функции. Напомню, что в оптимизации измеряли по $f(x^k) - f(x^*)$. Попробуем сконструировать что-то аналогичное для ВН и седловых задач.
- Попробуем: $g(x^k, y^k) - g(x^*, y^*)$. Хорош ли такой критерий?
  Рассмотрим самую простую седловую задачу $\min_x \max_y (x-1)\cdot(y+1)$. Решение этой задачи $x = 1, y = -1, g(1,-1) = 0$.
  Пусть начальная точка $x^0 = 0, y^0 = 0, g(0,0) = -1$. Тогда $g(x^0, y^0) - g(x^*, y^*)$ отрицательно. Такой критерий не подходит.
- Другой вариант: $g(x^k, y^*) - g(x^*, y^k)$.
  Рассмотрим самую простую седловую задачу $\min_x \max_y (x-1)\cdot(y+1)$. Решение этой задачи $x = 1, y = -1, g(1,-1) = 0$. Тогда $g(x, y^*) - g(x^*, y) = 0$. Такой критерий не подходит для выпукло-вогнутых седел, но подходит для сильно-выпукло–сильно-вогнутых.

Как измерять сходимость и качество решения?

- По аргументу $\|z^k - z^*\|^2$ – все аналогично выпуклой оптимизации, самый надежный критерий.
- По функции. Напомню, что в оптимизации измеряли по $f(x^k) - f(x^*)$. Попробуем сконструировать что-то аналогичное для ВН и седловых задач.
- Попробуем: $g(x^k, y^k) - g(x^*, y^*)$. Хорош ли такой критерий?
  Рассмотрим самую простую седловую задачу $\min_x \max_y (x-1)\cdot(y+1)$. Решение этой задачи $x = 1, y = -1, g(1,-1) = 0$.
  Пусть начальная точка $x^0 = 0, y^0 = 0, g(0,0) = -1$. Тогда $g(x^0, y^0) - g(x^*, y^*)$ отрицательно. Такой критерий не подходит.
- Другой вариант: $g(x^k, y^*) - g(x^*, y^k)$.
  Рассмотрим самую простую седловую задачу $\min_x \max_y (x-1)\cdot(y+1)$. Решение этой задачи $x = 1, y = -1, g(1,-1) = 0$. Тогда $g(x, y^*) - g(x^*, y) = 0$. Такой критерий не подходит для выпукло-вогнутых седел, но подходит для сильно-выпукло–сильно-вогнутых.

Как измерять сходимость и качество решения?

- По аргументу $\|z^k - z^*\|^2$ – все аналогично выпуклой оптимизации, самый надежный критерий.
- По функции. Напомню, что в оптимизации измеряли по $f(x^k) - f(x^*)$. Попробуем сконструировать что-то аналогичное для ВН и седловых задач.
- Попробуем: $g(x^k, y^k) - g(x^*, y^*)$. Хорош ли такой критерий?
  Рассмотрим самую простую седловую задачу $\min_x \max_y (x-1)\cdot(y+1)$. Решение этой задачи $x = 1, y = -1, g(1,-1) = 0$.
  Пусть начальная точка $x^0 = 0, y^0 = 0, g(0,0) = -1$. Тогда $g(x^0, y^0) - g(x^*, y^*)$ отрицательно. Такой критерий не подходит.
- Другой вариант: $g(x^k, y^*) - g(x^*, y^k)$.
  Рассмотрим самую простую седловую задачу $\min_x \max_y (x-1)\cdot(y+1)$. Решение этой задачи $x = 1, y = -1, g(1,-1) = 0$. Тогда $g(x, y^*) - g(x^*, y) = 0$. Такой критерий не подходит для выпукло-вогнутых седел, но подходит для сильно-выпукло–сильно-вогнутых.

Как измерять сходимость и качество решения?

- По аргументу $\|z^k - z^*\|^2$ – все аналогично выпуклой оптимизации, самый надежный критерий.
- По функции. Напомню, что в оптимизации измеряли по $f(x^k) - f(x^*)$. Попробуем сконструировать что-то аналогичное для ВН и седловых задач.
- Попробуем: $g(x^k, y^k) - g(x^*, y^*)$. Хорош ли такой критерий?
  Рассмотрим самую простую седловую задачу $\min_x \max_y (x-1)\cdot(y+1)$. Решение этой задачи $x = 1, y = -1, g(1,-1) = 0$.
  Пусть начальная точка $x^0 = 0, y^0 = 0, g(0,0) = -1$. Тогда $g(x^0, y^0) - g(x^*, y^*)$ отрицательно. Такой критерий не подходит.
- Другой вариант: $g(x^k, y^*) - g(x^*, y^k)$.
  Рассмотрим самую простую седловую задачу $\min_x \max_y (x-1)\cdot(y+1)$. Решение этой задачи $x = 1, y = -1, g(1,-1) = 0$. Тогда $g(x, y^*) - g(x^*, y) = 0$. Такой критерий не подходит для выпукло-вогнутых седел, но подходит для сильно-выпукло–сильно-вогнутых.

Как измерять сходимость и качество решения?

- По аргументу $\|z^k - z^*\|^2$ – все аналогично выпуклой оптимизации, самый надежный критерий.
- По функции. Напомню, что в оптимизации измеряли по $f(x^k) - f(x^*)$. Попробуем сконструировать что-то аналогичное для ВН и седловых задач.
- Попробуем: $g(x^k, y^k) - g(x^*, y^*)$. Хорош ли такой критерий?
  Рассмотрим самую простую седловую задачу $\min_x \max_y (x-1)\cdot(y+1)$. Решение этой задачи $x = 1, y = -1, g(1,-1) = 0$.
  Пусть начальная точка $x^0 = 0, y^0 = 0, g(0,0) = -1$. Тогда $g(x^0, y^0) - g(x^*, y^*)$ отрицательно. Такой критерий не подходит.
- Другой вариант: $g(x^k, y^*) - g(x^*, y^k)$.
  Рассмотрим самую простую седловую задачу $\min_x \max_y (x-1)\cdot(y+1)$. Решение этой задачи $x = 1, y = -1, g(1,-1) = 0$. Тогда $g(x, y^*) - g(x^*, y) = 0$. Такой критерий не подходит для выпукло-вогнутых седел, но подходит для сильно-выпукло–сильно-вогнутых.

Как измерять сходимость и качество решения?

- Лучший вариант: $\max_y g(x^k, y) - \min_x g(x, y^k)$.

  $\max_y g(x^k, y) \geq g(x^k, y^*) \geq g(x^*, y^*)$
  $\min_x g(x, y^k) \leq g(x^*, y^k) \leq g(x^*, y^*)$
  Тогда $\max_y g(x^k, y) - \min_x g(x, y^k) \geq g(x^k, y^*) - g(x^*, y^k) \geq 0$.
  Если $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} (x-1) \cdot (y+1)$ $\mathcal{X} = \mathbb{R}, \mathcal{Y} = \mathbb{R}$, то
  $\max_y g(x^k, y) - \min_x g(x, y^k) = +\infty$, поэтому вводят еще одно предположение:

Ограниченность множества

$\mathcal{Z}$ – ограниченное, т.е. для любых $z, z' \in \mathcal{Z}$

$$\|z - z'\| \leq D_z.$$

Такое предположение нам понадобится для монотонных неравенств и выпукло-вогнутых седел. В сильно-монотонном случае от него можно отказаться.

Как измерять сходимость и качество решения?

- Лучший вариант: $\max_y g(x^k, y) - \min_x g(x, y^k)$.
  $\max_y g(x^k, y) \geq g(x^k, y^*) \geq g(x^*, y^*)$
  $\min_x g(x, y^k) \leq g(x^*, y^k) \leq g(x^*, y^*)$
  Тогда $\max_y g(x^k, y) - \min_x g(x, y^k) \geq g(x^k, y^*) - g(x^*, y^k) \geq 0$.
  Если $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} (x-1) \cdot (y+1)$ $\mathcal{X} = \mathbb{R}, \mathcal{Y} = \mathbb{R}$, то
  $\max_y g(x^k, y) - \min_x g(x, y^k) = +\infty$, поэтому вводят еще одно предположение:

Ограниченность множества

$\mathcal{Z}$ – ограниченное, т.е. для любых $z, z' \in \mathcal{Z}$

$$\|z - z'\| \leq D_z.$$

Такое предположение нам понадобится для монотонных неравенств и выпукло-вогнутых седел. В сильно-монотонном случае от него можно отказаться.

Как измерять сходимость и качество решения?

- Лучший вариант: $\max_y g(x^k, y) - \min_x g(x, y^k)$.
  $\max_y g(x^k, y) \geq g(x^k, y^*) \geq g(x^*, y^*)$
  $\min_x g(x, y^k) \leq g(x^*, y^k) \leq g(x^*, y^*)$
  Тогда $\max_y g(x^k, y) - \min_x g(x, y^k) \geq g(x^k, y^*) - g(x^*, y^k) \geq 0$.
  Если $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} (x-1) \cdot (y+1)$ $\mathcal{X} = \mathbb{R}, \mathcal{Y} = \mathbb{R}$, то
  $\max_y g(x^k, y) - \min_x g(x, y^k) = +\infty$, поэтому вводят еще одно предположение:

Ограниченность множества

$\mathcal{Z}$ – ограниченное, т.е. для любых $z, z' \in \mathcal{Z}$

$$\|z - z'\| \leq D_z.$$

Такое предположение нам понадобится для монотонных неравенств и выпукло-вогнутых седел. В сильно-монотонном случае от него можно отказаться.

Как измерять сходимость и качество решения?

- Лучший вариант: $\max_y g(x^k, y) - \min_x g(x, y^k)$.
  $\max_y g(x^k, y) \geq g(x^k, y^*) \geq g(x^*, y^*)$
  $\min_x g(x, y^k) \leq g(x^*, y^k) \leq g(x^*, y^*)$
  Тогда $\max_y g(x^k, y) - \min_x g(x, y^k) \geq g(x^k, y^*) - g(x^*, y^k) \geq 0$.
  Если $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} (x-1) \cdot (y+1)$ $\mathcal{X} = \mathbb{R}, \mathcal{Y} = \mathbb{R}$, то
  $\max_y g(x^k, y) - \min_x g(x, y^k) = +\infty$, поэтому вводят еще одно предположение:

Ограниченность множества

$\mathcal{Z}$ – ограниченное, т.е. для любых $z, z' \in \mathcal{Z}$

$$\|z - z'\| \leq D_z.$$

Такое предположение нам понадобится для монотонных неравенств и выпукло-вогнутых седел. В сильно-монотонном случае от него можно отказаться.

Как измерять сходимость и качество решения?

- Лучший вариант: $\max_y g(x^k, y) - \min_x g(x, y^k)$.
  $\max_y g(x^k, y) \geq g(x^k, y^*) \geq g(x^*, y^*)$
  $\min_x g(x, y^k) \leq g(x^*, y^k) \leq g(x^*, y^*)$
  Тогда $\max_y g(x^k, y) - \min_x g(x, y^k) \geq g(x^k, y^*) - g(x^*, y^k) \geq 0$.
  Если $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} (x-1) \cdot (y+1)$ $\mathcal{X} = \mathbb{R}, \mathcal{Y} = \mathbb{R}$, то
  $\max_y g(x^k, y) - \min_x g(x, y^k) = +\infty$, поэтому вводят еще одно предположение:

Ограниченность множества

$\mathcal{Z}$ – ограниченное, т.е. для любых $z, z' \in \mathcal{Z}$

$$\|z - z'\| \leq D_z.$$

Такое предположение нам понадобится для монотонных неравенств и выпукло-вогнутых седел. В сильно-монотонном случае от него можно отказаться.

Как измерять сходимость и качество решения?

- Для седел критерий нашли. Как записать его же только для ВН?
  Рассмотрим следующую цепочку рассуждений

$$
\begin{aligned}
\max_{y\in\mathcal{Y}} g\left(x^k, y\right) - \min_{x\in\mathcal{X}} g\left(x, y^k\right) & \\
&= \max_{(x,y)\in\mathcal{Z}} \left(g(x^k, y) - g(x, y^k)\right) \\
&= \max_{(x,y)\in\mathcal{Z}} \left(g(x^k, y) - g(x^k, y^k) + g(x^k, y^k) - g(x, y^k)\right) \\
&\leq \max_{(x,y)\in\mathcal{Z}} \left(\langle \nabla_y g(x^k, y^k), y - y^k\rangle + \langle \nabla_x g(x^k, y^k), x^k - x\rangle\right) \\
&= \max_{z\in\mathcal{Z}} \langle F(z^k), z^k - z\rangle.
\end{aligned}
$$

  Тогда в качестве критерия можно воспользоваться $\max\limits_{z\in\mathcal{Z}} \langle F(z^k), z^k - z\rangle$ (смысл более менее понятен: в случае если z^k – решение, имеем $\max\limits_{z\in\mathcal{Z}} \langle F(z^*), z^* - z\rangle \leq 0$).
- В теоретическом анализе пользуются критерием: $\max\limits_{z\in\mathcal{Z}} \langle F(z), z^k - z\rangle$.

Как измерять сходимость и качество решения?

- Для седел критерий нашли. Как записать его же только для ВН?
  Рассмотрим следующую цепочку рассуждений

$$
\begin{aligned}
\max_{y \in \mathcal{Y}} g\left(x^k, y\right) - \min_{x \in \mathcal{X}} g\left(x, y^k\right) & \\
&= \max_{(x,y) \in \mathcal{Z}} \left(g(x^k, y) - g(x, y^k)\right) \\
&= \max_{(x,y) \in \mathcal{Z}} \left(g(x^k, y) - g(x^k, y^k) + g(x^k, y^k) - g(x, y^k)\right) \\
&\leq \max_{(x,y) \in \mathcal{Z}} \left(\langle \nabla_y g(x^k, y^k), y - y^k \rangle + \langle \nabla_x g(x^k, y^k), x^k - x \rangle\right) \\
&= \max_{z \in \mathcal{Z}} \langle F(z^k), z^k - z \rangle.
\end{aligned}
$$

Тогда в качестве критерия можно воспользоваться $\max\limits_{z \in \mathcal{Z}} \langle F(z^k), z^k - z \rangle$ (смысл более менее понятен: в случае если z^k – решение, имеем $\max\limits_{z \in \mathcal{Z}} \langle F(z^*), z^* - z \rangle \leq 0$).

- В теоретическом анализе пользуются критерием: $\max\limits_{z \in \mathcal{Z}} \langle F(z), z^k - z \rangle$.

Как измерять сходимость и качество решения?

- Для седел критерий нашли. Как записать его же только для ВН?
  Рассмотрим следующую цепочку рассуждений

$$
\begin{aligned}
\max_{y\in\mathcal{Y}} g\left(x^k, y\right) - \min_{x\in\mathcal{X}} g\left(x, y^k\right) & \\
&= \max_{(x,y)\in\mathcal{Z}} \left(g(x^k, y) - g(x, y^k)\right) \\
&= \max_{(x,y)\in\mathcal{Z}} \left(g(x^k, y) - g(x^k, y^k) + g(x^k, y^k) - g(x, y^k)\right) \\
&\leq \max_{(x,y)\in\mathcal{Z}} \left(\langle \nabla_y g(x^k, y^k), y - y^k\rangle + \langle \nabla_x g(x^k, y^k), x^k - x\rangle\right) \\
&= \max_{z\in\mathcal{Z}} \langle F(z^k), z^k - z\rangle.
\end{aligned}
$$

  Тогда в качестве критерия можно воспользоваться $\max\limits_{z\in\mathcal{Z}} \langle F(z^k), z^k - z\rangle$ (смысл более менее понятен: в случае если z^k – решение, имеем $\max\limits_{z\in\mathcal{Z}} \langle F(z^*), z^* - z\rangle \leq 0$).

- В теоретическом анализе пользуются критерием: $\max\limits_{z\in\mathcal{Z}} \langle F(z), z^k - z\rangle$.

Как измерять сходимость и качество решения?

- Для седел критерий нашли. Как записать его же только для ВН?
  Рассмотрим следующую цепочку рассуждений

$$
\begin{aligned}
\max_{y\in\mathcal{Y}} g\left(x^k, y\right) - \min_{x\in\mathcal{X}} g\left(x, y^k\right) & \\
&= \max_{(x,y)\in\mathcal{Z}} \left(g(x^k, y) - g(x, y^k)\right) \\
&= \max_{(x,y)\in\mathcal{Z}} \left(g(x^k, y) - g(x^k, y^k) + g(x^k, y^k) - g(x, y^k)\right) \\
&\leq \max_{(x,y)\in\mathcal{Z}} \left(\langle \nabla_y g(x^k, y^k), y - y^k\rangle + \langle \nabla_x g(x^k, y^k), x^k - x\rangle\right) \\
&= \max_{z\in\mathcal{Z}} \langle F(z^k), z^k - z\rangle.
\end{aligned}
$$

  Тогда в качестве критерия можно воспользоваться $\max\limits_{z\in\mathcal{Z}} \langle F(z^k), z^k - z\rangle$ (смысл более менее понятен: в случае если z^k – решение, имеем $\max\limits_{z\in\mathcal{Z}} \langle F(z^*), z^* - z\rangle \leq 0$).

- В теоретическом анализе пользуются критерием: $\max\limits_{z\in\mathcal{Z}} \langle F(z), z^k - z\rangle$.

Базовые методы: а ля градиентный спуск

Рассмотрим следующие подходы из выпуклой оптимизации:

- Классический спуск(-подъем):

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

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

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

Базовые методы: а ля градиентный спуск

Рассмотрим следующие подходы из выпуклой оптимизации:

- Классический спуск(-подъем):

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

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

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

Базовые методы: посмотрим, как работают

Билинейная задача:

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

Подбирается шаг, обеспечивающий наилучшую сходимость.

Не совсем честный момент: задача выпукло-вогнутая, а сходимость измеряется по аргументу – теория не гарантирует, что она вообще есть (для обоих методов). Но теория не гарантирует и сходимость по функции для метода без дополнительного шага.

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".

Беремся 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

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

Без проекций это один и тот же метод.

Беремся 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

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

Без проекций это один и тот же метод.

Беремся 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

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

Без проекций это один и тот же метод.

Сравним исходный метод и модификации

Билинейная задача:

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

По количеству вызовов "градиента" обгоняем исходный метод примерно в 2 раза (чуть меньше).

Сравним исходный метод и модификации

Видно, что наилучший шаг для модификации меньше, чем для обычного Extra Step. При $\gamma = 1/L$ модификация вообще расходится.

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):

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

Подставим в лемму $u = z^*$:

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

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

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

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):

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

Подставим в лемму $u = z^*$:

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

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

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

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):

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

Подставим в лемму $u = z^*$:

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

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

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

Сходимость: сильно-монотонный случай

Далее применяем сильную монотонность и Липшецевость:

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

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

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

Сгруппируем:

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

С $\gamma \leq \frac{1}{4L}$ имеем

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

Откуда и следует требуемое утверждение.

Сходимость: сильно-монотонный случай

Далее применяем сильную монотонность и Липшецевость:

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

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

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

Сгруппируем:

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

С $\gamma \leq \frac{1}{4L}$ имеем

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

Откуда и следует требуемое утверждение.

Сходимость: сильно-монотонный случай

Далее применяем сильную монотонность и Липшецевость:

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

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

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

Сгруппируем:

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

С $\gamma \leq \frac{1}{4L}$ имеем

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

Откуда и следует требуемое утверждение.

Сходимость: сильно-монотонный случай

Далее применяем сильную монотонность и Липшецевость:

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

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

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

Сгруппируем:

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

С $\gamma \leq \frac{1}{4L}$ имеем

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

Откуда и следует требуемое утверждение.

Сходимость: монотонный случай

Теорема для Extra Step

В монотонном случае для Extra Step метода с $\gamma \leq \frac{1}{L}$ справедлива следующая оценка сходимости:

$$\max_{u \in \mathcal{Z}} \left[\frac{1}{K} \sum_{k=0}^{K-1} \langle F(z^{k+1/2}), z^{k+1/2} - u \rangle \right] \leq \max_{u \in \mathcal{Z}} \frac{\|z^0 - u\|^2}{2\gamma K} \leq \frac{D_Z^2}{2\gamma K}.$$

Интерпретируем в терминах критериев сходимости:

- Для ВН.
  Воспользуемся монотонностью:
  $\frac{1}{K} \sum\limits_{k=0}^{K-1} \langle F(z^{k+1/2}), z^{k+1/2} - u \rangle \geq \frac{1}{K} \sum\limits_{k=0}^{K-1} \langle F(u), z^{k+1/2} - u \rangle = \langle F(u), \bar{z}^K - u \rangle$. Здесь введено $\bar{z}^K = \frac{1}{K} \sum\limits_{k=0}^{K-1} z^{k+1/2}$.
  Итого: $\max_{u \in \mathcal{Z}} \langle F(u), \bar{z}^K - u \rangle \leq \frac{D_Z^2}{2\gamma K}$.

Сходимость: монотонный случай

Теорема для Extra Step

В монотонном случае для Extra Step метода с $\gamma \leq \frac{1}{L}$ справедлива следующая оценка сходимости:

$$\max_{u \in \mathcal{Z}} \left[\frac{1}{K} \sum_{k=0}^{K-1} \langle F(z^{k+1/2}), z^{k+1/2} - u \rangle \right] \leq \max_{u \in \mathcal{Z}} \frac{\|z^0 - u\|^2}{2\gamma K} \leq \frac{D_Z^2}{2\gamma K}.$$

Интерпретируем в терминах критериев сходимости:

- Для ВН.
  Воспользуемся монотонностью:
  $\frac{1}{K} \sum\limits_{k=0}^{K-1} \langle F(z^{k+1/2}), z^{k+1/2} - u \rangle \geq \frac{1}{K} \sum\limits_{k=0}^{K-1} \langle F(u), z^{k+1/2} - u \rangle = \langle F(u), \bar{z}^K - u \rangle$. Здесь введено $\bar{z}^K = \frac{1}{K} \sum\limits_{k=0}^{K-1} z^{k+1/2}$.
  Итого: $\max_{u \in \mathcal{Z}} \langle F(u), \bar{z}^K - u \rangle \leq \frac{D_Z^2}{2\gamma K}$.

Сходимость: монотонный случай

- Для седел.
  Используя неравенство Йенсена:

$$\max_{y\in\mathcal{Y}} g\left(\frac{1}{K}\left(\sum_{k=0}^{K-1} x^{k+1/2}\right), y\right) - \min_{x\in\mathcal{X}} g\left(x, \frac{1}{K}\left(\sum_{k=0}^{K-1} y^{k+1/2}\right)\right)$$
$$\leq \max_{y\in\mathcal{Y}} \frac{1}{K}\sum_{k=0}^{K-1} g(x^{k+1/2}, y) - \min_{x\in\mathcal{X}} \frac{1}{K}\sum_{k=0}^{K-1} g(x, y^{k+1/2}).$$

Аналогично слайду 11:

$$\begin{aligned} & g\left(x^{k+1/2}, y\right) - g\left(x, y^{k+1/2}\right) \\ &= g(x^{k+1/2}, y) - g(x^{k+1/2}, y^{k+1/2}) + g(x^{k+1/2}, y^{k+1/2}) - g(x, y^{k+1/2}) \\ &\leq \langle \nabla_y g(x^{k+1/2}, y^{k+1/2}), y - y^{k+1/2}\rangle + \langle \nabla_x g(x^{k+1/2}, y^{k+1/2}), x^{k+1/2} - x\rangle \\ &= \langle F(z^{k+1/2}), z^{k+1/2} - u\rangle. \end{aligned}$$

Итого:
$\max_{y\in\mathcal{Y}} g\left(\frac{1}{K}\left(\sum_{k=0}^{K-1} x^{k+1/2}\right), y\right) - \min_{x\in\mathcal{X}} g\left(x, \frac{1}{K}\left(\sum_{k=0}^{K-1} y^{k+1/2}\right)\right) \leq \frac{D_Z^2}{2\gamma K}.$

Сходимость: монотонный случай

- Для седел.
  Используя неравенство Йенсена:

$$\max_{y\in\mathcal{Y}} g\left(\frac{1}{K}\left(\sum_{k=0}^{K-1} x^{k+1/2}\right), y\right) - \min_{x\in\mathcal{X}} g\left(x, \frac{1}{K}\left(\sum_{k=0}^{K-1} y^{k+1/2}\right)\right)$$
$$\leq \max_{y\in\mathcal{Y}} \frac{1}{K}\sum_{k=0}^{K-1} g(x^{k+1/2}, y) - \min_{x\in\mathcal{X}} \frac{1}{K}\sum_{k=0}^{K-1} g(x, y^{k+1/2}).$$

Аналогично слайду 11:

$$\begin{aligned} & g\left(x^{k+1/2}, y\right) - g\left(x, y^{k+1/2}\right) \\ &= g(x^{k+1/2}, y) - g(x^{k+1/2}, y^{k+1/2}) + g(x^{k+1/2}, y^{k+1/2}) - g(x, y^{k+1/2}) \\ &\leq \langle \nabla_y g(x^{k+1/2}, y^{k+1/2}), y - y^{k+1/2}\rangle + \langle \nabla_x g(x^{k+1/2}, y^{k+1/2}), x^{k+1/2} - x\rangle \\ &= \langle F(z^{k+1/2}), z^{k+1/2} - u\rangle. \end{aligned}$$

Итого:
$\max_{y\in\mathcal{Y}} g\left(\frac{1}{K}\left(\sum_{k=0}^{K-1} x^{k+1/2}\right), y\right) - \min_{x\in\mathcal{X}} g\left(x, \frac{1}{K}\left(\sum_{k=0}^{K-1} y^{k+1/2}\right)\right) \leq \frac{D_Z^2}{2\gamma K}.$

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 по аргументу и по функции. В недавних работах появились оценки с ускорением по $\|F(z)\|^2$.
- Напомню, для выпуклой оптимизации оптимальным методом является ускоренный метод Нестерова. В том числе из-за этого ее рассматривают отдельно от ВН.

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.

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.

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.

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: m батчей, 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: m батчей, 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: m батчей, выбирается случайный на каждой итерации.

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

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

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

Метод: сходимость

Билинейная задача:

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

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.

Метод: сходимость

- Оценки сходимости:

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

- В предположении: $\mathbb{E}_\xi F(z,\xi) = F(z)$, $\mathbb{E}_\xi \|F(z,\xi) - F(z)\|^2 \leq D\|F(z)\|^2 + \sigma^2$, нет качественного анализа и "правильных" оценок.
- Для покоординатных методов есть теория только для билинейных задач.
- Только сейчас появился хороший вариант метода с Variance Reduction. Методов типа SAGA нет.

Метод: сходимость

- Оценки сходимости:

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

- В предположении: $\mathbb{E}_\xi F(z,\xi) = F(z)$, $\mathbb{E}_\xi \|F(z,\xi) - F(z)\|^2 \leq D\|F(z)\|^2 + \sigma^2$, нет качественного анализа и "правильных" оценок.
- Для покоординатных методов есть теория только для билинейных задач.
- Только сейчас появился хороший вариант метода с Variance Reduction. Методов типа SAGA нет.

Метод: сходимость

- Оценки сходимости:

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

- В предположении: $\mathbb{E}_\xi F(z,\xi) = F(z)$, $\mathbb{E}_\xi \|F(z,\xi) - F(z)\|^2 \leq D\|F(z)\|^2 + \sigma^2$, нет качественного анализа и "правильных" оценок.
- Для покоординатных методов есть теория только для билинейных задач.
- Только сейчас появился хороший вариант метода с Variance Reduction. Методов типа SAGA нет.

Метод: сходимость

- Оценки сходимости:

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

- В предположении: $\mathbb{E}_\xi F(z,\xi) = F(z)$, $\mathbb{E}_\xi \|F(z,\xi) - F(z)\|^2 \leq D\|F(z)\|^2 + \sigma^2$, нет качественного анализа и "правильных" оценок.
- Для покоординатных методов есть теория только для билинейных задач.
- Только сейчас появился хороший вариант метода с Variance Reduction. Методов типа SAGA нет.

Модификации: Revisiting Extra Step

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

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

- В Монте-Карло постановке:

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

Модификации: Revisiting Extra Step

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

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

- В Монте-Карло постановке:

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

Revisiting Extra Step: сходимость

Билинейная задача:

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

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

Variance Reduction

- Здесь рассматривается Монте-Карло постановка.
- Вспомним, как выглядят методы редукции дисперсии в обычной версии (два цикла) и loopless:

Algorithm `SVRG`

Input: learning rate $\gamma > 0$, epoch length m, starting point $x^0 \in \mathbb{R}^d$
  $\phi = x^0$
  for $s = 0, 1, 2, \ldots$ **do**
    for $k = 0, 1, 2, \ldots, m-1$ **do**
      Sample $i \in \{1, \ldots, n\}$ uniformly at random
      $g^k = \nabla f_i(x^k) - \nabla f_i(\phi) + \nabla f(\phi)$
      $x^{k+1} = x^k - \gamma g^k$
    end for
    $\phi = x^0 = \frac{1}{m} \sum_{k=1}^{m} x^k$
  end for

Algorithm `L-SVRG`

Input: learning rate $\gamma > 0$, probability $p \in (0, 1]$, starting point $x^0 \in \mathbb{R}^d$
  $w^0 = x^0$
  for $k = 0, 1, 2, \ldots$ **do**
    Sample $i \in \{1, \ldots, n\}$ uniformly at random
    $g^k = \nabla f_i(x^k) - \nabla f_i(w^k) + \nabla f(w^k)$
    $x^{k+1} = x^k - \gamma g^k$
    $w^{k+1} = \begin{cases} x^k & \text{with probability } p \\ w^k & \text{with probability } 1-p \end{cases}$
  end for

Variance Reduction: 4 подхода

- Стандартный:

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

- Extra Step:

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

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

- Extra Step Revisiting:

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

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

Variance Reduction: 4 подхода

- Стандартный:

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

- Extra Step:

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

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

- Extra Step Revisiting:

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

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

Variance Reduction: 4 подхода

- Стандартный:

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

- Extra Step:

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

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

- Extra Step Revisiting:

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

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

Variance Reduction: 4 подхода

- Но правильные и наилучшие оценки дает только метод:

$$\bar{z}^k = \alpha z^k + (1-\alpha) w^k$$

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

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

- Подсчет $\bar{z}^k$ является ключевым в теоретическом анализе.
- Этому результату всего несколько месяцев.

Variance Reduction: 4 подхода

- Но правильные и наилучшие оценки дает только метод:

$$\bar{z}^k = \alpha z^k + (1-\alpha) w^k$$

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

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

- Подсчет $\bar{z}^k$ является ключевым в теоретическом анализе.
- Этому результату всего несколько месяцев.

Variance Reduction: 4 подхода

- Но правильные и наилучшие оценки дает только метод:

$$\bar{z}^k = \alpha z^k + (1-\alpha) w^k$$

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

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

- Подсчет $\bar{z}^k$ является ключевым в теоретическом анализе.
- Этому результату всего несколько месяцев.

Variance Reduction: сравнение

Билинейная задача в Монте-Карло постановке:

$$\min_{x,y\in[-1;1]^n}\max \frac{1}{M}\sum_{m=1}^{M}\left(x^T A_m y + b_m^T x + c_m^T y\right),$$

Figure: Сравнение loopless версий представленных алгоритмов с одинаковыми вероятностями $p = \frac{2}{M}$ и одинаковым шагом.

Как выглядит седло

- Гиперболический параболоид: $g(x, y) = x^2 - y^2$

Figure: Гиперболический параболоид. Изображение с сайта wikimedia.org

- Обезьянье седло: $g(x, y) = x^3 - 3xy^2$

Классические задачи

- Билинейная задача на вероятностном симплексе:

$$\min_{x \in \Delta^{n_x}} \max_{y \in \Delta^{n_y}} x^T A y,$$

  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.

Классические задачи

- Билинейная задача на вероятностном симплексе:

$$\min_{x \in \Delta^{n_x}} \max_{y \in \Delta^{n_y}} x^T A y,$$

  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.

Билинейная задача: интерпретация

- Пусть играют 2 игрока X и Y. Игрок Y на каждом ходу выбирает 1 оружие и атакует игрока X и пытается нанести максимальный урон. Игрок X наоборот выбирает 1 защиту и этот урон сокращает.
- Пусть у игрока Y набор из n_y оружий, а у X – n_x защитных мер.
- Игроки знают какой урон будет нанесен, если игрок Y выберет оружие i, а игрок X оружие j. Все эти данные заносятся в матрицу A.
- Между игроками начинается война – считаем, что ударов идет очень большое количество.
- Игроки не знают какое действие выбирает соперник на очередном ходу.
- Найдите оптимальную стратегию игры X и Y.

Билинейная задача: интерпретация

- Самое легкое решение: пусть игрок Y возьмет самое мощное оружие – оружие, которое дает наибольший урон при какой-то защите (максимум в матрице A). Всегда ли это эффективно?
- Нет. Пусть такое оружие действительно против какой-то защиты даст урон 1000, но другие защиты полностью его блокируют. Тогда X просто выставит эти защиты и будет невредим всегда.
- Но пусть у игрока Y есть оружие, которое всегда дает урон 10. Да, это меньше, но он постоянный.
- Суть – найти равновесие между X и Y!

Билинейная задача: интерпретация

- Самое легкое решение: пусть игрок Y возьмет самое мощное оружие – оружие, которое дает наибольший урон при какой-то защите (максимум в матрице A). Всегда ли это эффективно?
- Нет. Пусть такое оружие действительно против какой-то защиты даст урон 1000, но другие защиты полностью его блокируют. Тогда X просто выставит эти защиты и будет невредим всегда.
- Но пусть у игрока Y есть оружие, которое всегда дает урон 10. Да, это меньше, но он постоянный.
- Суть – найти равновесие между X и Y!

Билинейная задача: интерпретация

- Самое легкое решение: пусть игрок Y возьмет самое мощное оружие – оружие, которое дает наибольший урон при какой-то защите (максимум в матрице A). Всегда ли это эффективно?
- Нет. Пусть такое оружие действительно против какой-то защиты даст урон 1000, но другие защиты полностью его блокируют. Тогда X просто выставит эти защиты и будет невредим всегда.
- Но пусть у игрока Y есть оружие, которое всегда дает урон 10. Да, это меньше, но он постоянный.
- Суть – найти равновесие между X и Y!

Билинейная задача: интерпретация

- Самое легкое решение: пусть игрок Y возьмет самое мощное оружие – оружие, которое дает наибольший урон при какой-то защите (максимум в матрице A). Всегда ли это эффективно?
- Нет. Пусть такое оружие действительно против какой-то защиты даст урон 1000, но другие защиты полностью его блокируют. Тогда X просто выставит эти защиты и будет невредим всегда.
- Но пусть у игрока Y есть оружие, которое всегда дает урон 10. Да, это меньше, но он постоянный.
- Суть – найти равновесие между X и Y!

Билинейная задача: интерпретация

- Перепишем в виде билинейной задачи на вероятностных симплексах:

$$\min_{x \in \Delta^{n_x}} \max_{y \in \Delta^{n_y}} x^T A y.$$

Матрица A – матрица уронов. Что такое означают векторы x и y?

- Векторы x и y – вероятности (частоты) выбора той или иной защиты/атаки.
- Такая постановка реализует самый простой случай, когда игроки случайно (согласно равновесному распределению) выбирают действия.
- В самом простом случае вектор x и y имеют одну 1 и все остальные 0. Случай, когда у матрицы одно равновесной состояние.
- Если один из игроков чуть умнее, то он может извлечь выгоду из того, что соперник просто играет по равновесию, но это уже совсем другая история...

Билинейная задача: интерпретация

- Перепишем в виде билинейной задачи на вероятностных симплексах:

$$\min_{x \in \Delta^{n_x}} \max_{y \in \Delta^{n_y}} x^T A y.$$

  Матрица A – матрица уронов. Что такое означают векторы x и y?
- Векторы x и y – вероятности (частоты) выбора той или иной защиты/атаки.
- Такая постановка реализует самый простой случай, когда игроки случайно (согласно равновесному распределению) выбирают действия.
- В самом простом случае вектор x и y имеют одну 1 и все остальные 0. Случай, когда у матрицы одно равновесной состояние.
- Если один из игроков чуть умнее, то он может извлечь выгоду из того, что соперник просто играет по равновесию, но это уже совсем другая история...

Билинейная задача: интерпретация

- Перепишем в виде билинейной задачи на вероятностных симплексах:

$$\min_{x \in \Delta^{n_x}} \max_{y \in \Delta^{n_y}} x^T A y.$$

  Матрица A – матрица уронов. Что такое означают векторы x и y?
- Векторы x и y – вероятности (частоты) выбора той или иной защиты/атаки.
- Такая постановка реализует самый простой случай, когда игроки случайно (согласно равновесному распределению) выбирают действия.
- В самом простом случае вектор x и y имеют одну 1 и все остальные 0. Случай, когда у матрицы одно равновесной состояние.
- Если один из игроков чуть умнее, то он может извлечь выгоду из того, что соперник просто играет по равновесию, но это уже совсем другая история...

Билинейная задача: интерпретация

- Перепишем в виде билинейной задачи на вероятностных симплексах:

$$\min_{x \in \Delta^{n_x}} \max_{y \in \Delta^{n_y}} x^T A y.$$

  Матрица A – матрица уронов. Что такое означают векторы x и y?
- Векторы x и y – вероятности (частоты) выбора той или иной защиты/атаки.
- Такая постановка реализует самый простой случай, когда игроки случайно (согласно равновесному распределению) выбирают действия.
- В самом простом случае вектор x и y имеют одну 1 и все остальные 0. Случай, когда у матрицы одно равновесной состояние.
- Если один из игроков чуть умнее, то он может извлечь выгоду из того, что соперник просто играет по равновесию, но это уже совсем другая история...

Билинейная задача: интерпретация

- Перепишем в виде билинейной задачи на вероятностных симплексах:

$$\min_{x \in \Delta^{n_x}} \max_{y \in \Delta^{n_y}} x^T A y.$$

  Матрица A – матрица уронов. Что такое означают векторы x и y?
- Векторы x и y – вероятности (частоты) выбора той или иной защиты/атаки.
- Такая постановка реализует самый простой случай, когда игроки случайно (согласно равновесному распределению) выбирают действия.
- В самом простом случае вектор x и y имеют одну 1 и все остальные 0. Случай, когда у матрицы одно равновесной состояние.
- Если один из игроков чуть умнее, то он может извлечь выгоду из того, что соперник просто играет по равновесию, но это уже совсем другая история...

Билинейная задача: интерпретация

- Перепишем в виде билинейной задачи на вероятностных симплексах:

$$\min_{x \in \Delta^{n_x}} \max_{y \in \Delta^{n_y}} x^T A y.$$

Матрица A – матрица уронов. Что такое означают векторы x и y?

- Векторы x и y – вероятности (частоты) выбора той или иной защиты/атаки.
- Такая постановка реализует самый простой случай, когда игроки случайно (согласно равновесному распределению) выбирают действия.
- В самом простом случае вектор x и y имеют одну 1 и все остальные 0. Случай, когда у матрицы одно равновесной состояние.
- Если один из игроков чуть умнее, то он может извлечь выгоду из того, что соперник просто играет по равновесию, но это уже совсем другая история...

Билинейная задача: интерпретация

- Перепишем в виде билинейной задачи на вероятностных симплексах:

$$\min_{x \in \Delta^{n_x}} \max_{y \in \Delta^{n_y}} x^T A y.$$

  Матрица A – матрица уронов. Что такое означают векторы x и y?
- Векторы x и y – вероятности (частоты) выбора той или иной защиты/атаки.
- Такая постановка реализует самый простой случай, когда игроки случайно (согласно равновесному распределению) выбирают действия.
- В самом простом случае вектор x и y имеют одну 1 и все остальные 0. Случай, когда у матрицы одно равновесной состояние.
- Если один из игроков чуть умнее, то он может извлечь выгоду из того, что соперник просто играет по равновесию, но это уже совсем другая история...

Билинейная задача: интерпретация

- Перепишем в виде билинейной задачи на вероятностных симплексах:

$$\min_{x \in \Delta^{n_x}} \max_{y \in \Delta^{n_y}} x^T A y.$$

  Матрица A – матрица уронов. Что такое означают векторы x и y?
- Векторы x и y – вероятности (частоты) выбора той или иной защиты/атаки.
- Такая постановка реализует самый простой случай, когда игроки случайно (согласно равновесному распределению) выбирают действия.
- В самом простом случае вектор x и y имеют одну 1 и все остальные 0. Случай, когда у матрицы одно равновесной состояние.
- Если один из игроков чуть умнее, то он может извлечь выгоду из того, что соперник просто играет по равновесию, но это уже совсем другая история...

Билинейная задача: интерпретация

- Перепишем в виде билинейной задачи на вероятностных симплексах:

$$\min_{x \in \Delta^{n_x}} \max_{y \in \Delta^{n_y}} x^T A y.$$

  Матрица A – матрица уронов. Что такое означают векторы x и y?
- Векторы x и y – вероятности (частоты) выбора той или иной защиты/атаки.
- Такая постановка реализует самый простой случай, когда игроки случайно (согласно равновесному распределению) выбирают действия.
- В самом простом случае вектор x и y имеют одну 1 и все остальные 0. Случай, когда у матрицы одно равновесной состояние.
- Если один из игроков чуть умнее, то он может извлечь выгоду из того, что соперник просто играет по равновесию, но это уже совсем другая история...

Билинейная задача: интерпретация

- Перепишем в виде билинейной задачи на вероятностных симплексах:

$$\min_{x \in \Delta^{n_x}} \max_{y \in \Delta^{n_y}} x^T A y.$$

  Матрица A – матрица уронов. Что такое означают векторы x и y?
- Векторы x и y – вероятности (частоты) выбора той или иной защиты/атаки.
- Такая постановка реализует самый простой случай, когда игроки случайно (согласно равновесному распределению) выбирают действия.
- В самом простом случае вектор x и y имеют одну 1 и все остальные 0. Случай, когда у матрицы одно равновесной состояние.
- Если один из игроков чуть умнее, то он может извлечь выгоду из того, что соперник просто играет по равновесию, но это уже совсем другая история...

Билинейная задача: вор и полицейский

- Пусть город представляет собой квадрат из $n \times n$ маленьких квадратиков. В каждом квадратике стоит дом и полицейская будка рядом с ним. Пусть так же известны ценности домов w_i.
- Каждую ночь вор выбирает, какой дом ограбить, а полицейский выбирает будку, в которой будет дежурить.
- Вероятность поимки вора, если вор грабит дом в квадрате i, а полисмен дежурит в квадрате j равна:

$$\exp(-\alpha \cdot \text{dist}(i, j)).$$

  Т.е. уменьшается с увеличением расстояния между квадратами.
- Вор хочет максимизировать свою ожидаемую прибыль:

$$w_i \left(1 - \exp(-\alpha \cdot \text{dist}(i, j))\right).$$

  Полицейский наоборот – минимизировать.

Билинейная задача: вор и полицейский

- Пусть город представляет собой квадрат из $n \times n$ маленьких квадратиков. В каждом квадратике стоит дом и полицейская будка рядом с ним. Пусть так же известны ценности домов w_i.
- Каждую ночь вор выбирает, какой дом ограбить, а полицейский выбирает будку, в которой будет дежурить.
- Вероятность поимки вора, если вор грабит дом в квадрате i, а полисмен дежурит в квадрате j равна:

$$\exp(-\alpha \cdot \text{dist}(i, j)).$$

  Т.е. уменьшается с увеличением расстояния между квадратами.
- Вор хочет максимизировать свою ожидаемую прибыль:

$$w_i \left(1 - \exp(-\alpha \cdot \text{dist}(i, j))\right).$$

  Полицейский наоборот – минимизировать.

Билинейная задача: вор и полицейский

- Пусть город представляет собой квадрат из $n \times n$ маленьких квадратиков. В каждом квадратике стоит дом и полицейская будка рядом с ним. Пусть так же известны ценности домов w_i.
- Каждую ночь вор выбирает, какой дом ограбить, а полицейский выбирает будку, в которой будет дежурить.
- Вероятность поимки вора, если вор грабит дом в квадрате i, а полисмен дежурит в квадрате j равна:

$$\exp(-\alpha \cdot \text{dist}(i, j)).$$

  Т.е. уменьшается с увеличением расстояния между квадратами.
- Вор хочет максимизировать свою ожидаемую прибыль:

$$w_i \left(1 - \exp(-\alpha \cdot \text{dist}(i, j))\right).$$

  Полицейский наоборот – минимизировать.

Билинейная задача: вор и полицейский

- Пусть город представляет собой квадрат из $n \times n$ маленьких квадратиков. В каждом квадратике стоит дом и полицейская будка рядом с ним. Пусть так же известны ценности домов w_i.
- Каждую ночь вор выбирает, какой дом ограбить, а полицейский выбирает будку, в которой будет дежурить.
- Вероятность поимки вора, если вор грабит дом в квадрате i, а полисмен дежурит в квадрате j равна:

$$\exp(-\alpha \cdot \text{dist}(i, j)).$$

  Т.е. уменьшается с увеличением расстояния между квадратами.
- Вор хочет максимизировать свою ожидаемую прибыль:

$$w_i \left(1 - \exp(-\alpha \cdot \text{dist}(i, j))\right).$$

  Полицейский наоборот – минимизировать.

Билинейная задача: вор и полицейский

- Эту задачу можно переписать в виде билинейной:

$$\min_{x \in \Delta^{n_x}} \max_{y \in \Delta^{n_y}} x^T A y.$$

  В A_{ij} хранится $w_i \left(1 - \exp(-\alpha \cdot \text{dist}(i, j))\right)$.

- Какие есть предположения насчет ответов задачи?

Билинейная задача: вор и полицейский

- Эту задачу можно переписать в виде билинейной:

$$\min_{x \in \Delta^{n_x}} \max_{y \in \Delta^{n_y}} x^T A y.$$

  В A_{ij} хранится $w_i \left(1 - \exp(-\alpha \cdot \text{dist}(i, j))\right)$.
- Какие есть предположения насчет ответов задачи?

Билинейная задача: вор и полицейский

Figure: Вор и полицейский с одинаковой ценностью домов на квадрате 200×200. Изображение отсюда.

Состязательный подход

- Имеется модель, которую мы хотим обучить.
- Решаем задачу минимизации функции потерь и получаем обученную модель.
- Но что если "мешать" исходной модели и поддерживать процесс обучения "в тонусе".
- Пример, почему это важно:

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

Состязательный подход

- Имеется модель, которую мы хотим обучить.
- Решаем задачу минимизации функции потерь и получаем обученную модель.
- Но что если "мешать" исходной модели и поддерживать процесс обучения "в тонусе".
- Пример, почему это важно:

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

Состязательный подход

- Имеется модель, которую мы хотим обучить.
- Решаем задачу минимизации функции потерь и получаем обученную модель.
- Но что если "мешать" исходной модели и поддерживать процесс обучения "в тонусе".
- Пример, почему это важно:

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

Состязательный подход

- Имеется модель, которую мы хотим обучить.
- Решаем задачу минимизации функции потерь и получаем обученную модель.
- Но что если "мешать" исходной модели и поддерживать процесс обучения "в тонусе".
- Пример, почему это важно:

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

Состязательный подход

- Рассмотрим следующий модифицированную функцию потерь:

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

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

Состязательный подход

- Рассмотрим следующий модифицированную функцию потерь:

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

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

GANs

- Такой подход в обучении называется состязательным – Adversarial.
- В последнее время стали популярны Adversarial Attacks. Их осуществляют за счет второй сетки, которая генерирует "плохие" примеры для основной модели.
- Далее разговор пойдет о GANs (Generative Adversarial Nets) – главном представителе Adversarial подхода. Изложение будет по первой статье про GANs.

GANs

- Такой подход в обучении называется состязательным – Adversarial.
- В последнее время стали популярны Adversarial Attacks. Их осуществляют за счет второй сетки, которая генерирует "плохие" примеры для основной модели.
- Далее разговор пойдет о GANs (Generative Adversarial Nets) – главном представителе Adversarial подхода. Изложение будет по первой статье про GANs.

GANs

- Такой подход в обучении называется состязательным – Adversarial.
- В последнее время стали популярны Adversarial Attacks. Их осуществляют за счет второй сетки, которая генерирует "плохие" примеры для основной модели.
- Далее разговор пойдет о GANs (Generative Adversarial Nets) – главном представителе Adversarial подхода. Изложение будет по первой статье про GANs.

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:

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

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:

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

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:

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

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:

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

GANs: процесс обучения

Figure: Процесс обучения: слева - направо. Черный пунктир – реальное распределение данных. Зеленая линия – распределение данных от генератора. Синяя линия – распределение дискриминатора. Нижняя диграмма со стрелками показывает, как генератор отображает равномерное распределение z в некоторое распределение на x (откуда и генерируются картинки).

GANs: выводы

- В хорошо обученном GAN дискриминатор гадает, а генератор воспроизводит картинки абсолютно идентично реальным.
- Главная цель такого GAN – увеличиение объема выборки. В отличие от Adversarial Attacks тут важно в первую очередь обучение сети, которая "атакует".
- Сейчас нашли массу других задач, которые решаются с помощью GAN. Архитектур GAN сейчас становится все больше и больше.

GANs: выводы

- В хорошо обученном GAN дискриминатор гадает, а генератор воспроизводит картинки абсолютно идентично реальным.
- Главная цель такого GAN – увеличиение объема выборки. В отличие от Adversarial Attacks тут важно в первую очередь обучение сети, которая "атакует".
- Сейчас нашли массу других задач, которые решаются с помощью GAN. Архитектур GAN сейчас становится все больше и больше.

GANs: выводы

- В хорошо обученном GAN дискриминатор гадает, а генератор воспроизводит картинки абсолютно идентично реальным.
- Главная цель такого GAN – увеличиение объема выборки. В отличие от Adversarial Attacks тут важно в первую очередь обучение сети, которая "атакует".
- Сейчас нашли массу других задач, которые решаются с помощью GAN. Архитектур GAN сейчас становится все больше и больше.

GANs: выводы

- GAN на практике неустойчивая и "капризная" модель.
- Часто не получается повторить эксперимент, описанные в статьях, и про это пишутся отдельные статьи :-)

Figure: Пример обучения различных архитектур GAN на датасете интерьеров.

- Популярные оптимизаторы: Adam, RMSProp. Но как их настраивать – отдельный разговор.
- Глобальный совет по обучению GANs: искать статьи и туториалы конкретно по вашей модели с похожим датасетом и смотреть, что предлагается сделать.

GANs: выводы

- GAN на практике неустойчивая и "капризная" модель.
- Часто не получается повторить эксперимент, описанные в статьях, и про это пишутся отдельные статьи :-)

Figure: Пример обучения различных архитектур GAN на датасете интерьеров.

- Популярные оптимизаторы: Adam, RMSProp. Но как их настраивать – отдельный разговор.
- Глобальный совет по обучению GANs: искать статьи и туториалы конкретно по вашей модели с похожим датасетом и смотреть, что предлагается сделать.

GANs: выводы

- GAN на практике неустойчивая и "капризная" модель.
- Часто не получается повторить эксперимент, описанные в статьях, и про это пишутся отдельные статьи :-)

Figure: Пример обучения различных архитектур GAN на датасете интерьеров.

- Популярные оптимизаторы: Adam, RMSProp. Но как их настраивать – отдельный разговор.
- Глобальный совет по обучению GANs: искать статьи и туториалы конкретно по вашей модели с похожим датасетом и смотреть, что предлагается сделать.

GANs: выводы

- GAN на практике неустойчивая и "капризная" модель.
- Часто не получается повторить эксперимент, описанные в статьях, и про это пишутся отдельные статьи :-)

Figure: Пример обучения различных архитектур GAN на датасете интерьеров.

- Популярные оптимизаторы: Adam, RMSProp. Но как их настраивать – отдельный разговор.
- Глобальный совет по обучению GANs: искать статьи и туториалы конкретно по вашей модели с похожим датасетом и смотреть, что предлагается сделать.

Денойзинг изображений

- Пусть имеется картинка g с шумом.
- Хотим получить картинку u уже без шума.
- Сформулируем задачу так (смотри статью):

$$\min_{u \in X} \int_{\Omega} |\nabla u| + \frac{\lambda}{2} \|u - g\|^2.$$

Здесь первый член отвечает за "плавность" картинки, второй – чтобы полученная картинка была похожа на исходную. λ варьируется.

- Ее можно переписать в виде седловой:

$$\min_{u \in X} \max_{p \in Y} -\langle u; \mathrm{div} p \rangle_X + \frac{\lambda}{2} \|u - g\|^2 + \delta_P(p).$$

Здесь $\delta_P(p) = 0$, если $p \in P$, иначе бесконечности.

Денойзинг изображений

- Пусть имеется картинка g с шумом.
- Хотим получить картинку u уже без шума.
- Сформулируем задачу так (смотри статью):

$$\min_{u \in X} \int_{\Omega} |\nabla u| + \frac{\lambda}{2} \|u - g\|^2.$$

Здесь первый член отвечает за "плавность" картинки, второй – чтобы полученная картинка была похожа на исходную. λ варьируется.

- Ее можно переписать в виде седловой:

$$\min_{u \in X} \max_{p \in Y} -\langle u; \mathrm{div} p \rangle_X + \frac{\lambda}{2} \|u - g\|^2 + \delta_P(p).$$

Здесь $\delta_P(p) = 0$, если $p \in P$, иначе бесконечности.

Денойзинг изображений

- Пусть имеется картинка g с шумом.
- Хотим получить картинку u уже без шума.
- Сформулируем задачу так (смотри статью):

$$\min_{u \in X} \int_{\Omega} |\nabla u| + \frac{\lambda}{2} \|u - g\|^2.$$

Здесь первый член отвечает за "плавность" картинки, второй – чтобы полученная картинка была похожа на исходную. λ варьируется.

- Ее можно переписать в виде седловой:

$$\min_{u \in X} \max_{p \in Y} -\langle u; \mathrm{div} p \rangle_X + \frac{\lambda}{2} \|u - g\|^2 + \delta_P(p).$$

Здесь $\delta_P(p) = 0$, если $p \in P$, иначе бесконечности.

Денойзинг изображений

- Пусть имеется картинка g с шумом.
- Хотим получить картинку u уже без шума.
- Сформулируем задачу так (смотри статью):

$$\min_{u \in X} \int_{\Omega} |\nabla u| + \frac{\lambda}{2} \|u - g\|^2.$$

Здесь первый член отвечает за "плавность" картинки, второй – чтобы полученная картинка была похожа на исходную. λ варьируется.

- Ее можно переписать в виде седловой:

$$\min_{u \in X} \max_{p \in Y} -\langle u; \mathrm{div} p \rangle_X + \frac{\lambda}{2} \|u - g\|^2 + \delta_P(p).$$

Здесь $\delta_P(p) = 0$, если $p \in P$, иначе бесконечности.

Денойзинг изображений

(a) Noisy image ($\sigma = 0.05$)

(b) Noisy image ($\sigma = 0.1$)

(c) Denoised image ($\lambda = 16$)

(d) Denoised image ($\lambda = 8$)

Figure: Пример денойзинга картинок с разным уровнем шума и параметром λ.

Спасибо за внимание!