Ускорение сведением к седловым задачам с приложением к поиску барицентров Вассерштейна

Общероссийский семинар по оптимизации

Даниил Тяпкин

ФКН ВШЭ, HDI lab

2020

Структура рассказа

- Седловые задачи: общий обзор
- Стохастическая постановка
- Барицентры Вассертшейна: стохастический подход
- Разреженные матричные игры
- Area convexity
- Барицентры Вассерштейна: area convexity подход

Общая постановка седловых задачи

Дана функция $f\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, такая что $f(\cdot, y)$ - выпуклая $\forall y \in \mathcal{Y}$, $f(x, \cdot)$ – вогнутая $\forall x \in \mathcal{X}$. Тогда рассмотрим задачу

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y). \tag{1}$$

Общая постановка седловых задачи

Дана функция $f\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, такая что $f(\cdot, y)$ - выпуклая $\forall y \in \mathcal{Y}$, $f(x, \cdot)$ – вогнутая $\forall x \in \mathcal{X}$. Тогда рассмотрим задачу

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y). \tag{1}$$

Как правило, предполагают, что $\mathcal{X}$ и $\mathcal{Y}$ – компакты.

Общая постановка седловых задачи

Дана функция $f\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, такая что $f(\cdot, y)$ - выпуклая $\forall y \in \mathcal{Y}$, $f(x, \cdot)$ – вогнутая $\forall x \in \mathcal{X}$. Тогда рассмотрим задачу

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} f(x, y). \tag{1}$$

Как правило, предполагают, что $\mathcal{X}$ и $\mathcal{Y}$ – компакты.
Качество решения $(\tilde{x}, \tilde{y})$ можно измерять, пользуясь зазором двойственности:

$$\min_{x \in \mathcal{X}} f(x, \tilde{y}) - \max_{y \in \mathcal{Y}} f(\tilde{x}, y) \leq \varepsilon \tag{2}$$

Гладкие седловые задачи

Определение

$f(x, y)$ называется $(L_{xx}, L_{xy}, L_{yx}, L_{yy})$-гладкой, если для всяких $x, x' \in \mathcal{X}$ и $y, y' \in \mathcal{Y}$:

$$
\begin{aligned}
\|\nabla_x f(x, y) - \nabla_x f(x', y)\|_{\mathcal{X}^*} &\le L_{xx}\|x - x'\|_{\mathcal{X}}, \\
\|\nabla_x f(x, y) - \nabla_x f(x, y')\|_{\mathcal{X}^*} &\le L_{xy}\|y - y'\|_{\mathcal{Y}}, \\
\text{‘}\|\nabla_y f(x, y) - \nabla_y f(x, y')\|_{\mathcal{Y}^*} &\le L_{yy}\|y - y'\|_{\mathcal{Y}}, \\
\|\nabla_y f(x, y) - \nabla_y f(x', y)\|_{\mathcal{Y}^*} &\le L_{yx}\|x - x'\|_{\mathcal{X}}.
\end{aligned}
$$

Гладкие седловые задачи

Определение

$f(x, y)$ называется $(L_{xx}, L_{xy}, L_{yx}, L_{yy})$-гладкой, если для всяких $x, x' \in \mathcal{X}$ и $y, y' \in \mathcal{Y}$:

$$
\begin{aligned}
\|\nabla_x f(x, y) - \nabla_x f(x', y)\|_{\mathcal{X}^*} &\leq L_{xx}\|x - x'\|_{\mathcal{X}}, \\
\|\nabla_x f(x, y) - \nabla_x f(x, y')\|_{\mathcal{X}^*} &\leq L_{xy}\|y - y'\|_{\mathcal{Y}}, \\
`\|\nabla_y f(x, y) - \nabla_y f(x, y')\|_{\mathcal{Y}^*} &\leq L_{yy}\|y - y'\|_{\mathcal{Y}}, \\
\|\nabla_y f(x, y) - \nabla_y f(x', y)\|_{\mathcal{Y}^*} &\leq L_{yx}\|x - x'\|_{\mathcal{X}}.
\end{aligned}
$$

Также, как и в случае обычных задач оптимизации, гладкость позволяет добиться значительного ускорения.

Простейший пример: матричные игры

Пусть зафиксирована матрица $A \in \mathrm{Mat}_{n\times m}(\mathbb{R})$. Тогда рассмотрим следующую игру: игрок x выбирает строку матрицы (назовем его i), а игрок y – столбец ((j). Тогда первый игрок получает $-A_{ij}$ выигрыша, а второй – A_{ij}.

Простейший пример: матричные игры

Пусть зафиксирована матрица $A \in \mathrm{Mat}_{n\times m}(\mathbb{R})$. Тогда рассмотрим следующую игру: игрок x выбирает строку матрицы (назовем его i), а игрок y – столбец ((j). Тогда первый игрок получает $-A_{ij}$ выигрыша, а второй – A_{ij}.
Тогда задачу поиска равновесия Нэша в смешанных стратегиях можно записать как

$$\min_{x\in\Delta^n}\max_{y\in\Delta^m} x^\top Ay$$

Простейший пример: матричные игры

Пусть зафиксирована матрица $A \in \mathrm{Mat}_{n\times m}(\mathbb{R})$. Тогда рассмотрим следующую игру: игрок x выбирает строку матрицы (назовем его i), а игрок y – столбец ((j). Тогда первый игрок получает $-A_{ij}$ выигрыша, а второй – A_{ij}.
Тогда задачу поиска равновесия Нэша в смешанных стратегиях можно записать как

$$\min_{x\in\Delta^n} \max_{y\in\Delta^m} x^\top A y$$

Определение

Седловая задача (1) называется билинейной, если $f(x, y)$ – билинейная.

Прокс-функция и проксимальный оператор

Определение

Прокс-функцией называется сильно выпуклая (относительно нормы на $\mathcal{X}$) непрерывно-дифференцируемая функция $d_{\mathcal{X}}$.

Прокс-функция и проксимальный оператор

Определение

Прокс-функцией называется сильно выпуклая (относительно нормы на $\mathcal{X}$) непрерывно-дифференцируемая функция $d_{\mathcal{X}}$.

Определение

Дивергенцией Брегмана $B_d(x, x)$, соответствующей прокс-функцией d, называется следующая функция

$$B_d(x, x') = d(x) - d(x') - \nabla d(x')^\top (x - x')$$

Прокс-функция и проксимальный оператор

Определение

Прокс-функцией называется сильно выпуклая (относительно нормы на $\mathcal{X}$) непрерывно-дифференцируемая функция $d_{\mathcal{X}}$.

Определение

Дивергенцией Брегмана $B_d(x, x)$, соответствующей прокс-функцией d, называется следующая функция

$$B_d(x, x') = d(x) - d(x') - \nabla d(x')^\top (x - x')$$

Определение

Проксимальный операторомна нормированном пространстве $\mathcal{X}$ называется следующая функция

$$\mathrm{prox}_{\bar{x}}^{d}(v) = \arg\min_{x \in \mathcal{X}} \langle v, x \rangle + B_d(x, \bar{x})$$

Сетап для Mirror Descent / Mirror Prox

Зафиксируем прокс функции для $\mathcal{X}, \mathcal{Y} : d_{\mathcal{X}}, d_{\mathcal{Y}}$, которые согласованы с нормами на соотв. пространствах, а также значения $R^2 = \sup_x d(x) - \inf_x d(x)$ для $\mathcal{X}$ и $\mathcal{Y}$ ($R^2_{\mathcal{X}}$ и $R^2_{\mathcal{Y}}$ соответственно).

Сетап для Mirror Descent / Mirror Prox

Зафиксируем прокс функции для $\mathcal{X}, \mathcal{Y} : d_{\mathcal{X}}, d_{\mathcal{Y}}$, которые согласованы с нормами на соотв. пространствах, а также значения $R^2 = \sup_x d(x) - \inf_x d(x)$ для $\mathcal{X}$ и $\mathcal{Y}$ ($R^2_{\mathcal{X}}$ и $R^2_{\mathcal{Y}}$ соответственно).
Введем прокс функцию на $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ как взвешенную сумму прокс-функций на компонентах: $d_{\mathcal{Z}}(z) = a_1 d_{\mathcal{X}}(x) + a_2 d_{\mathcal{Y}}(y)$, где $z = (x, y)$.

Сетап для Mirror Descent / Mirror Prox

Зафиксируем прокс функции для $\mathcal{X}, \mathcal{Y} : d_{\mathcal{X}}, d_{\mathcal{Y}}$, которые согласованы с нормами на соотв. пространствах, а также значения $R^2 = \sup_x d(x) - \inf_x d(x)$ для $\mathcal{X}$ и $\mathcal{Y}$ ($R^2_{\mathcal{X}}$ и $R^2_{\mathcal{Y}}$ соответственно).
Введем прокс функцию на $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ как взвешенную сумму прокс-функций на компонентах: $d_{\mathcal{Z}}(z) = a_1 d_{\mathcal{X}}(x) + a_2 d_{\mathcal{Y}}(y)$, где $z = (x, y)$.
Введем градиентный оператор для седловой задачи (1):

$$G(z) = \begin{pmatrix} \nabla_x f(x, y) \\ -\nabla_y f(x, y) \end{pmatrix}$$

Сетап для Mirror Descent / Mirror Prox

Зафиксируем прокс функции для $\mathcal{X}, \mathcal{Y} : d_{\mathcal{X}}, d_{\mathcal{Y}}$, которые согласованы с нормами на соотв. пространствах, а также значения $R^2 = \sup_x d(x) - \inf_x d(x)$ для $\mathcal{X}$ и $\mathcal{Y}$ ($R^2_{\mathcal{X}}$ и $R^2_{\mathcal{Y}}$ соответственно).
Введем прокс функцию на $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ как взвешенную сумму прокс-функций на компонентах: $d_{\mathcal{Z}}(z) = a_1 d_{\mathcal{X}}(x) + a_2 d_{\mathcal{Y}}(y)$, где $z = (x, y)$.
Введем градиентный оператор для седловой задачи (1):

$$G(z) = \begin{pmatrix} \nabla_x f(x, y) \\ -\nabla_y f(x, y) \end{pmatrix}$$

Зафиксируем точку $\bar{z} = \arg\min d_{\mathcal{Z}}(z)$.

Сетап для Mirror Descent / Mirror Prox

Зафиксируем прокс функции для $\mathcal{X}, \mathcal{Y} : d_{\mathcal{X}}, d_{\mathcal{Y}}$, которые согласованы с нормами на соотв. пространствах, а также значения $R^2 = \sup_x d(x) - \inf_x d(x)$ для $\mathcal{X}$ и $\mathcal{Y}$ ($R^2_{\mathcal{X}}$ и $R^2_{\mathcal{Y}}$ соответственно).
Введем прокс функцию на $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ как взвешенную сумму прокс-функций на компонентах: $d_{\mathcal{Z}}(z) = a_1 d_{\mathcal{X}}(x) + a_2 d_{\mathcal{Y}}(y)$, где $z = (x, y)$.
Введем градиентный оператор для седловой задачи (1):

$$G(z) = \begin{pmatrix} \nabla_x f(x, y) \\ -\nabla_y f(x, y) \end{pmatrix}$$

Зафиксируем точку $\bar{z} = \arg\min d_{\mathcal{Z}}(z)$. Для негладких fзафиксируем константу Липшица по x как $B_{\mathcal{X}}$, по y как $B_{\mathcal{Y}}$

Mirror Descent / Mirror Prox

Тогда мы можем легко записать шаги Mirror Descent:

1. $z^{k+1} = \text{prox}_{z^k}^{d_\mathcal{Z}}(\eta G(z^k))$

Тогда для ε точности относительно зазора двойственности требуется в негладком случае:

$$N = O\left(\frac{(R_\mathcal{X} B_\mathcal{X} + R_\mathcal{Y} B_\mathcal{Y})^2}{\varepsilon^2}\right)$$

итераций.

Mirror Descent / Mirror Prox

Тогда мы можем легко записать шаги Mirror Descent:

1. $z^{k+1} = \text{prox}_{z^k}^{d_{\mathcal{Z}}}(\eta G(z^k))$

Тогда для ε точности относительно зазора двойственности требуется в негладком случае:

$$N = O\left(\frac{(R_{\mathcal{X}}B_{\mathcal{X}} + R_{\mathcal{Y}}B_{\mathcal{Y}})^2}{\varepsilon^2}\right)$$

итераций.

А также шаги Mirror Prox:

1. $u^{k+1} = \text{prox}_{z^k}^{d_{\mathcal{Z}}}(\eta G(z^k))$;
2. $z^{k+1} = \text{prox}_{z^k}^{d_{\mathcal{Z}}}(\eta G(u^{k+1}))$

С количеством итераций в гладком случае:

$$N = O\left(\frac{\max\{L_{xx}R_{\mathcal{X}}^2, L_{xy}R_{\mathcal{X}}R_{\mathcal{Y}}, L_{yx}R_{\mathcal{Y}}R_{\mathcal{X}}, L_{yy}R_{\mathcal{Y}}^2\}}{\varepsilon}\right)$$

Инструкция по применению

- Выбрать *хорошую* норму на $\mathcal{X}, \mathcal{Y}$, чтобы константы гладкости были *хорошо* ограничены;

Инструкция по применению

- Выбрать *хорошую* норму на $\mathcal{X}, \mathcal{Y}$, чтобы константы гладкости были *хорошо* ограничены;
- Выбрать *хорошую* прокс функцию, чтобы константы R^2 были *хорошо* ограничены и она была согласована с нормой;

Инструкция по применению

- Выбрать *хорошую* норму на $\mathcal{X}, \mathcal{Y}$, чтобы константы гладкости были *хорошо* ограничены;
- Выбрать *хорошую* прокс функцию, чтобы константы R^2 были *хорошо* ограничены и она была согласована с нормой;
- Научиться *быстро* считать прокс оператор.

Инструкция по применению

- Выбрать *хорошую* норму на $\mathcal{X}, \mathcal{Y}$, чтобы константы гладкости были *хорошо* ограничены;
- Выбрать *хорошую* прокс функцию, чтобы константы R^2 были *хорошо* ограничены и она была согласована с нормой;
- Научиться *быстро* считать прокс оператор.

НО: Большая проблема возникает с ℓ_∞ нормой – нет *хорошей* прокс структуры.

Классический пример: максимум гладких функций

Рассмотрим задачу оптимизации

$$\min_{x \in \mathcal{X}} F(x) = \max_{1 \le i \le m} f_i(x),$$

где всякая f_i – гладкая выпуклая функция. F – гладкой не является, упирается в нижнюю границу $\sim \varepsilon^{-2}$ итераций.

Классический пример: максимум гладких функций

Рассмотрим задачу оптимизации

$$\min_{x \in \mathcal{X}} F(x) = \max_{1 \le i \le m} f_i(x),$$

где всякая f_i – гладкая выпуклая функция. F – гладкой не является, упирается в нижнюю границу $\sim \varepsilon^{-2}$ итераций.
Заметим, что

$$\max_{1 \le i \le m} f_i(x) = \max_{y \in \Delta^m} \sum_{i=1}^{m} y_i f_i(x)$$

Классический пример: максимум гладких функций

Рассмотрим задачу оптимизации

$$\min_{x \in \mathcal{X}} F(x) = \max_{1 \le i \le m} f_i(x),$$

где всякая f_i – гладкая выпуклая функция. F – гладкой не является, упирается в нижнюю границу $\sim \varepsilon^{-2}$ итераций.
Заметим, что

$$\max_{1 \le i \le m} f_i(x) = \max_{y \in \Delta^m} \sum_{i=1}^{m} y_i f_i(x)$$

Тогда получаем гладкую седловую задачу, которая решается при помощи Mirror Prox за $\sim \varepsilon^{-1}$:

$$\min_{x \in \mathcal{X}} \max_{y \in \Delta^m} \sum_{i=1}^{m} y_i f_i(x)$$

Стохастическая постановка

Идея: вместо градиента использовать несмещенную оценку градиента:

$$\mathbb{E}[\tilde{G}(x, \xi)] = G(x)$$

Стохастическая постановка

Идея: вместо градиента использовать несмещенную оценку градиента:

$$\mathbb{E}[\tilde{G}(x, \xi)] = G(x)$$

Вместо зазора двойственности нашим функционалом качества будет его матожидание или ширина доверительного интервала.

Стохастическая постановка

Идея: вместо градиента использовать несмещенную оценку градиента:

$$\mathbb{E}[\tilde{G}(x, \xi)] = G(x)$$

Вместо зазора двойственности нашим функционалом качества будет его матожидание или ширина доверительного интервала.

Весь анализ Mirror Descent продолжает работать и давать ту же асимптотику для сходимости матожидания, но даже при более слабом условии – ограничении второго момента.

Простейший пример: матричные игры

Вернемся к примеру

$$\min_{x \in \Delta^n} \max_{y \in \Delta^m} x^\top A y$$

Простейший пример: матричные игры

Вернемся к примеру

$$\min_{x \in \Delta^n} \max_{y \in \Delta^m} x^\top A y$$

Если решать эту задачу при помощи Mirror Prox, то получаем число итераций $\tilde{O}(1/\varepsilon)$ при сложности итерации $O(mn)$.

Простейший пример: матричные игры

Вернемся к примеру

$$\min_{x \in \Delta^n} \max_{y \in \Delta^m} x^\top A y$$

Если решать эту задачу при помощи Mirror Prox, то получаем число итераций $\tilde{O}(1/\varepsilon)$ при сложности итерации $O(mn)$.

Если же использовать стохастический градиент и Mirror Descent, семплируя столбец/строку используя распределения x, y, то мы можем получить число итераций $\tilde{O}(1/\varepsilon^2)$ при сложности каждой итерации $O(m + n)$.

Расстояние Монжа-Канторовича

Перейдем к статье [6]: **Daniil Tiapkin, Alexander Gasnikov, and Pavel Dvurechensky**. *Stochastic saddle-point optimization for wasserstein barycenters*, 2020.
Расстояние, индуцированное задачей оптимального транспорта, называют расстоянием Монжа-Канторовича.

$$L_C(r, c) = \min_{X \in \mathcal{U}(r,c)} \langle C, X \rangle,$$

где X называется *транспортным планом*, а $\mathcal{U}(r, c)$ – *транспортным политопом*, определенным как

$$\mathcal{U}(r, c) = \{X \in \mathrm{Mat}_{n \times n}(\mathbb{R}_+) \mid X\mathbf{1} = r,\ X^T\mathbf{1} = c\}.$$

Расстояние Монжа-Канторовича

Перейдем к статье [6]: **Daniil Tiapkin, Alexander Gasnikov, and Pavel Dvurechensky**. *Stochastic saddle-point optimization for wasserstein barycenters*, 2020.

Расстояние, индуцированное задачей оптимального транспорта, называют расстоянием Монжа-Канторовича.

$$L_C(r, c) = \min_{X \in \mathcal{U}(r,c)} \langle C, X \rangle,$$

где X называется *транспортным планом*, а $\mathcal{U}(r, c)$ – *транспортным политопом*, определенным как

$$\mathcal{U}(r, c) = \{X \in \mathrm{Mat}_{n \times n}(\mathbb{R}_+) \mid X\mathbf{1} = r,\ X^T\mathbf{1} = c\}.$$

Для специального выбора матрицы $C : C_{i,j} = d^2(x_i, x_j)$ зафиксируем n-элементное метрическое пространство $(\{x_1, \ldots, x_n\}, d)$. Тогда мы можем определить 2-расстояние Вассертшейна:

$$\mathcal{W}_2(r, c) = \sqrt{\min_{X \in \mathcal{U}(r,c)} \langle C, X \rangle}.$$

Определение барицентров Вассерштейна

Зафиксируем распределение P_ξ над вероятностными мерами и случайную величину ξ. Тогда мы можем определить барицентр Вассерштейна::

$$r_* = \arg\min_{r \in \Delta^n} \mathbb{E}\mathcal{W}_2^2(r, \xi). \quad (3)$$

Определение барицентров Вассерштейна

Зафиксируем распределение P_ξ над вероятностными мерами и случайную величину ξ. Тогда мы можем определить барицентр Вассерштейна::

$$r_* = \arg\min_{r \in \Delta^n} \mathbb{E}\mathcal{W}_2^2(r, \xi). \tag{3}$$

Аналогия – для случайной величины X со значениями в $\mathbb{R}$:

$$\mathbb{E}X = \arg\min_{a \in \mathbb{R}} \mathbb{E}(X - a)^2 = \arg\min_{a \in \mathbb{R}} \mathbb{E}d^2(X, a)$$

Определение барицентров Вассерштейна

Зафиксируем распределение P_ξ над вероятностными мерами и случайную величину ξ. Тогда мы можем определить барицентр Вассерштейна::

$$r_* = \arg\min_{r\in\Delta^n} \mathbb{E}\mathcal{W}_2^2(r,\xi). \tag{3}$$

Аналогия – для случайной величины X со значениями в $\mathbb{R}$:

$$\mathbb{E}X = \arg\min_{a\in\mathbb{R}} \mathbb{E}(X-a)^2 = \arg\min_{a\in\mathbb{R}} \mathbb{E}d^2(X,a)$$

Для решения задач такого вида традициоенно существует два подхода: Stochastic Approximation (SA) и Stochastic Average Approximation (SAA). SA работает с задачей 3 и может рассматриваться как *онлайн* подход, в то время как SAA работает со следующим приближением исходной задачи:

$$r_* = \arg\min_{r\in\Delta^n} \frac{1}{m}\sum_{i=1}^{m} \mathcal{W}_2^2(r,c_i). \tag{4}$$

SAA может считаться *оффлайн* подходом.

Двойственная задача

Изначальная задача оптимального транспорта может быть записана, используя двойственность Канторовича:

$$L_C(r, c) = \max_{\substack{\lambda, \mu \in \mathbb{R}^n \\ -C_{i,j} - \lambda_i - \mu_j \leq 0}} -\langle \lambda, r \rangle - \langle \mu, c \rangle, \tag{5}$$

что экивалетно

$$L_C(r, c) = \max_{\mu \in \mathbb{R}^n} -\langle \lambda^*(\mu, C), r \rangle - \langle \mu, c \rangle, \tag{6}$$

где $\lambda^* \colon \mathbb{R}^n \times \mathrm{Mat}_{n \times n}(\mathbb{R}) \to \mathbb{R}^n$ определена поэлементно: $\lambda_i^*(\mu, C) = \max_{j \in [n]}(-C_{i,j} - \mu_j)$.

Двойственная задача

Изначальная задача оптимального транспорта может быть записана, используя двойственность Канторовича:

$$L_C(r,c) = \max_{\substack{\lambda,\mu\in\mathbb{R}^n \\ -C_{i,j}-\lambda_i-\mu_j\le 0}} -\langle\lambda, r\rangle - \langle\mu, c\rangle, \tag{5}$$

что экивалетно

$$L_C(r,c) = \max_{\mu\in\mathbb{R}^n} -\langle\lambda^*(\mu, C), r\rangle - \langle\mu, c\rangle, \tag{6}$$

где $\lambda^* \colon \mathbb{R}^n \times \mathrm{Mat}_{n\times n}(\mathbb{R}) \to \mathbb{R}^n$ определена поэлементно: $\lambda_i^*(\mu, C) = \max_{j\in[n]}(-C_{i,j} - \mu_j)$.
Теперь, подставив это все в онлайн подход и введя некоторые дополнительные ограничения, получаем:

$$\min_{r\in\Delta^n} \mathbb{E}\mathcal{W}_2^2(r,\xi) = \min_{r\in\Delta^n} \sup_{f_\mu\in\mathcal{F}^b} \mathbb{E}\left[-\langle\lambda^*(f_\mu(\xi), C), r\rangle - \langle f_\mu(\xi), \xi\rangle\right], \tag{7}$$

где $\mathcal{F}^b = \{f \colon \Delta^n \to \mathbb{R}^n \mid f$ – P_ξ-измеримая и ограниченная $\|f(\xi)\|_\infty \le \|C\|_\infty\}$

Первый алгоритм

Первое предположение – пусть распределение ξ на симплексе может принимать только конечный набор значений. Тогда задача превращается в конечномерную:

$$\min_{r \in \Delta^n} \mathbb{E}\mathcal{W}_p^p(r, \xi) = \min_{r \in \Delta_\delta^n} \max_{\substack{f_\mu \in \mathrm{Mat}_{m \times n}(\mathbb{R}) \\ \|M_\mu\|_\infty \le \|C\|_\infty}} \mathbb{E}\left[-\langle \lambda^*(f_\mu(\xi), D_p), r\rangle - \langle f_\mu(\xi), \xi\rangle\right]. \tag{8}$$

Первый алгоритм

Первое предположение – пусть распределение ξ на симплексе может принимать только конечный набор значений. Тогда задача превращается в конечномерную:

$$\min_{r\in\Delta^n} \mathbb{E}\mathcal{W}_p^p(r,\xi) = \min_{r\in\Delta^n_\delta} \max_{\substack{f_\mu\in \mathrm{Mat}_{m\times n}(\mathbb{R}) \\ \|M_\mu\|_\infty \le \|C\|_\infty}} \mathbb{E}\left[-\langle \lambda^*(f_\mu(\xi), D_p), r\rangle - \langle f_\mu(\xi), \xi\rangle\right]. \tag{8}$$

Используя стохастический градиент вида

$$G_{\mathcal{X}}(r, f_\mu, c_t, s) = -n \cdot \max_{j\in[n]}(-(D_p)_{s,j} - (M_\mu)_{t,j}); \tag{9}$$

$$G_{\mathcal{Y}}(r, f_\mu, c_t, q) = -e_{J_q(f_\mu, c_t)} + c_t, \tag{10}$$

где c_t семплируется из P_ξ, q из распределения, ассоциированного с r, а s равномерно среди $\{1,\ldots,n\}$, получаем общее число итераций для ε-сходимости матожидания:

$$O\left(\frac{mn^2\|D_p\|_\infty^2}{\varepsilon^2}\right)$$

Reproducing Kernel Hilbert Space

Второе предположение – пусть оптимальное решение лежит в пространстве специального вида, называемого Reproducing Kernel Hilbert Space.

Определение

A Hilbert space $\mathcal{H}$ of functions $f\colon \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel Hilbert space if there is a symmetric positive-defined function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ called a kernel, such that

1. $\forall x \in \mathcal{X} : k(\cdot, x) \in \mathcal{H}$;
2. $\forall f \in \mathcal{H} : \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$;
3. $\forall x, y \in \mathcal{X} : \langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}} = k(x, y)$.

Reproducing Kernel Hilbert Space

Второе предположение – пусть оптимальное решение лежит в пространстве специального вида, называемого Reproducing Kernel Hilbert Space.

Определение

A Hilbert space $\mathcal{H}$ of functions $f\colon \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel Hilbert space if there is a symmetric positive-defined function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ called a kernel, such that

1. $\forall x \in \mathcal{X} : k(\cdot, x) \in \mathcal{H}$;
2. $\forall f \in \mathcal{H} : \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$;
3. $\forall x, y \in \mathcal{X} : \langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}} = k(x, y)$.

Зачем? Чтобы считать градиент $f(x)$ по f:

$$\nabla_f f(x) = \nabla_f \langle f, k_x\rangle = k_x,$$

где $k_x(y) = k(x, y)$.

SA-подход

Вспомним, как выглядит наша задача:

$$\min_{r \in \Delta^n} \mathbb{E}\mathcal{W}_2^2(r, \xi) = \min_{r \in \Delta^n} \sup_{f_\mu \in \mathcal{F}^b} \mathbb{E}\left[-\langle \lambda^*(f_\mu(\xi), C), r\rangle - \langle f_\mu(\xi), \xi\rangle\right], \tag{11}$$

SA-подход

Вспомним, как выглядит наша задача:

$$\min_{r \in \Delta^n} \mathbb{E}\mathcal{W}_2^2(r, \xi) = \min_{r \in \Delta^n} \sup_{f_\mu \in \mathcal{F}^b} \mathbb{E}\left[-\langle \lambda^*(f_\mu(\xi), C), r\rangle - \langle f_\mu(\xi), \xi\rangle\right], \qquad (11)$$

Самое интересное – это градиент по f_μ:

$$G_{\mathcal{Y}}(r, f_\mu, c) = \left(\partial_{f_\mu}(-\psi(r, f_\mu(c), c))\right)_t = k(\cdot, c)\left(c_t - \sum_{i=1}^{n} r_i I\{t = J_i(c)\}\right),$$

где $J_i(c)$– один из индексов, на котором λ_i^* достигает максимума, а c – новый семпл.

Kernel Mirror Descent

При выборе гауссовского ядра $k(x, y) = \exp(-1/2\sigma^2 \|x - y\|_2^2)$ получилась следующая оценка:

Теорема

Kernel Mirror Descent вычисляет 2-Вассерштейн барицентр относительно произвольного распределения, используя

$$N = O\left(\frac{n^2 R^2}{\varepsilon^2} \log^2\left(\frac{1}{\sigma}\right)\right)$$

семплов, и имеет итоговую сложность

$$O\left(\frac{n^5 R^4}{\varepsilon^4} \log^4\left(\frac{1}{\sigma}\right)\right).$$

Разреженные матричные игры

Другой эффект от стохастики был получен в недавней статье [1].
Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian.*Coordinate methods for matrix games*, 2020.

Разреженные матричные игры

Другой эффект от стохастики был получен в недавней статье [1].
Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian.*Coordinate methods for matrix games*, 2020.
Рассматриваем билинейную седловую задачу

$$\max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} x^\top A y + x^\top b + y^\top c$$

Основная идея: при $\mathcal{X}$ и $\mathcal{Y}$ ограниченных по ℓ_1 и ℓ_2 норме можно выбрать стохастические градиенты, которые будут максимально возможно учитывать разреженность матрицы A.

Разреженные матричные игры

Другой эффект от стохастики был получен в недавней статье [1]. **Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian**.*Coordinate methods for matrix games*, 2020.

Рассматриваем билинейную седловую задачу

$$\max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} x^\top A y + x^\top b + y^\top c$$

Основная идея: при $\mathcal{X}$ и $\mathcal{Y}$ ограниченных по ℓ_1 и ℓ_2 норме можно выбрать стохастические градиенты, которые будут максимально возможно учитывать разреженность матрицы A.

Помимо этого, в работе также рассматриваются техники понижения дисперсии, которые дают также ускорение и не будут рассмотрены в этом докладе.

Area convexity

Уходя от стохастики, рассмотрим подход, предложенный в [5] и улучшенный в [4]. При этом остается в сеттинге билинейный седловых задач.

Area convexity

Уходя от стохастики, рассмотрим подход, предложенный в [5] и улучшенный в [4]. При этом остается в сеттинге билинейный седловых задач.
Основная идея: семейство сильно-выпуклых прокс функций слишком бедное (проблема ℓ_∞-регуляризации), семейство просто выпуклых прокс функций не обладает нужными свойствами.

Area convexity

Уходя от стохастики, рассмотрим подход, предложенный в [5] и улучшенный в [4]. При этом остается в сеттинге билинейный седловых задач.

Основная идея: семейство сильно-выпуклых прокс функций слишком бедное (проблема ℓ_∞-регуляризации), семейство просто выпуклых прокс функций не обладает нужными свойствами.

Будем рассматривать новое семейство:

Определение

Регуляризатор r называется κ-area convex относительно градиентного оператора Gесли для любых трех точек $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathcal{Z}$ верно:

$$\kappa\left(r(\boldsymbol{a}) + r(\boldsymbol{b}) + r(\boldsymbol{c}) - 3r\left(\frac{\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{c}}{3}\right)\right) \geq \langle G(\boldsymbol{a}) - G(\boldsymbol{b}), \boldsymbol{b} - \boldsymbol{c}\rangle$$

Area convexity

Уходя от стохастики, рассмотрим подход, предложенный в [5] и улучшенный в [4]. При этом остается в сеттинге билинейный седловых задач.

Основная идея: семейство сильно-выпуклых прокс функций слишком бедное (проблема ℓ_∞-регуляризации), семейство просто выпуклых прокс функций не обладает нужными свойствами.

Будем рассматривать новое семейство:

Определение

Регуляризатор r называется κ-area convex относительно градиентного оператора Gесли для любых трех точек $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c} \in \mathcal{Z}$ верно:

$$\kappa \left(r(\boldsymbol{a}) + r(\boldsymbol{b}) + r(\boldsymbol{c}) - 3r\left(\frac{\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{c}}{3}\right)\right) \geq \langle G(\boldsymbol{a}) - G(\boldsymbol{b}), \boldsymbol{b} - \boldsymbol{c}\rangle$$

Из area-convexity следует обычная выпуклость.

Комментарии

Оказалось,для такого семейства можно доказать, что Mirror Prox-подобная процедура для решения седловых задач, называемая Dual Extrapolation, работает, используя $O(\kappa\Theta/\varepsilon)$ вызовов прокс-оператора с κ-area convex регуляризатром вместо прокс функции.

Комментарии

Оказалось,для такого семейства можно доказать, что Mirror Prox-подобная процедура для решения седловых задач, называемая Dual Extrapolation, работает, используя $O(\kappa\Theta/\varepsilon)$ вызовов прокс-оператора с κ-area convex регуляризатром вместо прокс функции.

Идейно, использование обычной прокс функции никак не учитывает «связь» пространств $\mathcal{X}$ и $\mathcal{Y}$.

Комментарии

Оказалось,для такого семейства можно доказать, что Mirror Prox-подобная процедура для решения седловых задач, называемая Dual Extrapolation, работает, используя $O(\kappa\Theta/\varepsilon)$ вызовов прокс-оператора с κ-area convex регуляризатром вместо прокс функции.

Идейно, использование обычной прокс функции никак не учитывает «связь» пространств $\mathcal{X}$ и $\mathcal{Y}$.

Проблема: как быстро считать вызовы прокс оператора, когда нет сильной выпуклости?

Комментарии

Оказалось,для такого семейства можно доказать, что Mirror Prox-подобная процедура для решения седловых задач, называемая Dual Extrapolation, работает, используя $O(\kappa\Theta/\varepsilon)$ вызовов прокс-оператора с κ-area convex регуляризатром вместо прокс функции.

Идейно, использование обычной прокс функции никак не учитывает «связь» пространств $\mathcal{X}$ и $\mathcal{Y}$.

Проблема: как быстро считать вызовы прокс оператора, когда нет сильной выпуклости?

Проблема 2: а как подобрать хороший регуляризатор?

Перейдем к работа [3]: **Darina Dvinskikh and Daniil Tiapkin**. *Improved complexity bounds in wasserstein barycenter problem*,2020. Рассматривался исключительно SAA-подход.

Барицентры Вассертшейна: area convexity подход

Перейдем к работа [3]: **Darina Dvinskikh and Daniil Tiapkin**. *Improved complexity bounds in wasserstein barycenter problem*,2020. Рассматривался исключительно SAA-подход.

Пусть d – растянутая в вектор матрица C, $b = \begin{pmatrix} p \\ q \end{pmatrix}$, а матрица $A = \{0,1\}^{2n \times n^2}$ – матрица инцендентности полного двудольного графа. Так как $\sum_{i,j=1}^{n} X_{ij} = 1$, то, пользуясь [4], перепишем задачу поиска квадрата расстояния Вассертшейна:

$$\min_{x \in \Delta_{n^2}} \max_{y \in [-1,1]^{2n}} \{d^\top x + 2\|d\|_\infty (y^\top Ax - b^\top y)\}.$$

Барицентры Вассертшейна: area convexity подход

Тогда можно переписать задачу поиска барицентров в виде билинейной седловой задачи.
Для этого определим $\mathcal{X} \triangleq \prod^m \Delta_{n^2} \times \Delta_n$ и $\mathcal{Y} \triangleq [-1,1]^{2mn}$. Тогда для $\mathbf{x} = (x_1^\top, \ldots, x_m^\top, p^\top)^\top \in \mathcal{X}$ и $\mathbf{y} = (y_1^\top, \ldots, y_m^\top)^\top \in \mathcal{Y}$ получаем:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{x},\mathbf{y}) \qquad \triangleq \frac{1}{m}\left\{\boldsymbol{d}^\top \mathbf{x} + 2\|d\|_\infty \left(\mathbf{y}^\top \boldsymbol{A}\mathbf{x} - \top \mathbf{y}\right)\right\}, \qquad (12)$$

где $\boldsymbol{d} = (d^\top, \ldots, d^\top, \mathbf{0}_n^\top)^\top, = (\mathbf{0}_n^\top, q_1^\top, \ldots, \mathbf{0}_n^\top, q_m^\top)^\top$ and $\boldsymbol{A} \in \{-1,0,1\}^{2mn\times(mn^2+n)}$ – почти блочно-диагональная матрица:

$$\boldsymbol{A} = \begin{pmatrix} A & 0_{2n\times n^2} & \cdots & 0_{2n\times n^2} & \begin{pmatrix} -I_n \\ 0_{n\times n} \end{pmatrix} \\ 0_{2n\times n^2} & A & \cdots & 0_{2n\times n^2} & \begin{pmatrix} -I_n \\ 0_{n\times n} \end{pmatrix} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_{2n\times n^2} & 0_{2n\times n^2} & \cdots & A & \begin{pmatrix} -I_n \\ 0_{n\times n} \end{pmatrix} \end{pmatrix}$$

Mirror Prox

Как первый алгоритм, был использован обычный Mirror Prox.

Mirror Prox

Как первый алгоритм, был использован обычный Mirror Prox. Для пространства $\mathcal{Y} \triangleq [-1, 1]^{2nm}$использовался стандартный Евклидов сетап, а вот для $\mathcal{X} \triangleq \prod^m \Delta_{n^2} \times \Delta_n$ была использована специфичная норма $\|\mathbf{x}\|_{\mathcal{X}} = \sqrt{\sum_{i=1}^m \|x_i\|_1^2 + m\|p\|_1^2}$ for $\mathbf{x} = (x_1, \dots, x_m, p)^T$, где $\| \cdot \|_1$ – Манхэттенская норма ($a \in \mathbb{R}^n, \|a\|_1 = \sum_{i=1}^n |a_i|$).

Mirror Prox

Как первый алгоритм, был использован обычный Mirror Prox. Для пространства $\mathcal{Y} \triangleq [-1,1]^{2nm}$использовался стандартный Евклидов сетап, а вот для $\mathcal{X} \triangleq \prod^m \Delta_{n^2} \times \Delta_n$ была использована специфичная норма $\|\mathbf{x}\|_{\mathcal{X}} = \sqrt{\sum_{i=1}^m \|x_i\|_1^2 + m\|p\|_1^2}$ for $\mathbf{x} = (x_1, \ldots, x_m, p)^T$, где $\|\cdot\|_1$ – Манхэттенская норма ($a \in \mathbb{R}^n, \|a\|_1 = \sum_{i=1}^n |a_i|$).

Для $\mathcal{X}$ будем использовать прокс функцию $d_{\mathcal{X}}(\mathbf{x}) = \sum_{i=1}^m \langle x_i, \ln x_i \rangle + m\langle p, \ln p \rangle$ и следующую дивергенцию Брегмана:

$$B_{\mathcal{X}}(\mathbf{x}, \breve{\mathbf{x}}) = \sum_{i=1}^m \langle x_i, \ln(x_i/\breve{x}_i) \rangle - \sum_{i=1}^m \mathbf{1}^\top (x_i - \breve{x}_i)$$

$$+m\langle p, \ln(p/\breve{p}) \rangle - m\mathbf{1}^\top (p - \breve{p}).$$

Dual Extrapolation

Для второго алгоритма использовался уже area-convex регуляризатор (доказательство area-convexity которого нетривиально, но легко следует из рассуждений [4]):

$$
\begin{aligned}
r(\mathbf{x}, \mathbf{y}) = \frac{2\|d\|_\infty}{m}\Bigg(&\sum_{i=1}^{m}\Big[10\langle x_i, \log x_i\rangle + \langle Ax_i, (y_i)^2\rangle\Big] \\
&+ \sum_{i=1}^{m}\Big[5\langle p, \log p\rangle + \langle B_{\mathcal{E}}p, (y_i)^2\rangle\Big]\Bigg).
\end{aligned}
$$

Dual Extrapolation

Для второго алгоритма использовался уже area-convex регуляризатор (доказательство area-convexity которого нетривиально, но легко следует из рассуждений [4]):

$$r(\mathbf{x}, \mathbf{y}) = \frac{2\|d\|_\infty}{m}\left(\sum_{i=1}^{m}\Big[10\langle x_i, \log x_i\rangle + \langle Ax_i, (y_i)^2\rangle\Big] + \sum_{i=1}^{m}\Big[5\langle p, \log p\rangle + \langle B_{\mathcal{E}}p, (y_i)^2\rangle\Big]\right).$$

Также пришлось дополнительно доказывать, что процедура блочной оптимизации будет сходиться достаточно быстро для быстрого подсчета прокс-оператора.

Сравнение сложности

Table: Algorithms for WB problem and their rates of convergence

Approach	**Complexity**
IBP	$\widetilde{O}\left(\frac{mn^2\|C\|_\infty^2}{\varepsilon^2}\right)$
Accelerated IBP	$\widetilde{O}\left(\frac{mn^2\sqrt{n}\|C\|_\infty}{\varepsilon}\right)$
FastIBP	$\widetilde{O}\left(\frac{mn^2\sqrt[3]{n}\|C\|_\infty^{4/3}}{\varepsilon\sqrt[3]{\varepsilon}}\right)$
Mirror descent	$\widetilde{O}\left(\frac{mn^2\|C\|_\infty^2}{\varepsilon^2}\right)$
Mirror prox with specific norm	$\widetilde{O}\left(\frac{mn^2\sqrt{n}\|C\|_\infty}{\varepsilon}\right)$
Dual extrapolation with area-convexity	$\widetilde{O}\left(\frac{mn^2\|C\|_\infty}{\varepsilon}\right)$

Список источников

Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian.
Coordinate methods for matrix games, 2020.

Darina Dvinskikh.
Stochastic approximation versus sample average approximation for population wasserstein barycenter calculation, 2020.

Darina Dvinskikh and Daniil Tiapkin.
Improved complexity bounds in wasserstein barycenter problem, 2020.

Arun Jambulapati, Aaron Sidford, and Kevin Tian.
A direct $\tilde{O}(1/\varepsilon)$ iteration parallel algorithm for optimal transport, 2019.

Jonah Sherman.
Area-convexity, l∞ regularization, and undirected multicommodity flow.
In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, page 452–460, New York, NY, USA, 2017. Association for Computing Machinery.

Daniil Tiapkin, Alexander Gasnikov, and Pavel Dvurechensky.
Stochastic saddle-point optimization for wasserstein barycenters, 2020.

Спасибо за внимание!

Контакты:

- E-mail: unkoll@yandex.ru
- Telegram: @unkoll