№1 [2 балла] Докажите, что функция

$$f(x) = \log\left(\sum_{i=1}^{n} e^{x_i}\right)$$

выпуклая, используя любой дифференциальный критерий выпуклости.

Гессиан функции

$$\nabla^2 f(x) = \frac{1}{1^T z}\text{diag}(z) - \frac{1}{(1^T z)^2} z z^T \quad (z_k = e^{x_k})$$

чтобы показать $\nabla^2 f(x) \succeq 0$, мы должны проверить, что $v^T \nabla^2 f(x) v \geq 0$ для всех v:

$$v^T \nabla^2 f(x) v = \frac{(\sum_k z_k v_k^2)(\sum_k z_k) - (\sum_k v_k z_k)^2}{(\sum_k z_k)^2} \geq 0$$

так как $(\sum_k v_k z_k)^2 \leq (\sum_k z_k v_k^2)(\sum_k z_k)$ (из неравенства Коши-Буняковского-Шварца)

№2 [3 балла] Пусть $X \in \mathbb{R}^{m \times n}$, где $\text{rk} X = n$, $\Omega \in \mathbb{S}^m_{++}$, и $W \in \mathbb{R}^{k \times n}$. Найдите матрицу $G \in \mathbb{R}^{k \times m}$, являющуюся решением следующей задачи оптимизации:

$$f(G) = \text{tr}\left(G \Omega G^\top\right) \to \min_{GX = W}$$

Заметим, что задача - выпуклая, т.к. минимизируемая функция (квадратичная с положительно определенной матрицей Ω) и бюджетное множество выпуклы, а ограничение-равенство - афинно. Значит, необходимые условия оптимальности будут достаточными. Запишем лагранжиан (введя матрицу множителей Лагранжа Λ):

$$L(G, \Lambda) = \langle G\Omega, G \rangle + \langle \Lambda, GX - W \rangle$$

Дифференциал:

$$dL = \langle 2G\Omega + \Lambda X^T, dG \rangle$$

Необходимое условие оптимальности первого порядка (т.к. $\Omega \in \mathbb{S}^m_{++}$ - обратима):

$$2G\Omega + \Lambda X^T = 0 \to G = -\frac{1}{2}\Lambda X^T \Omega^{-1}$$

Подставляя это в условие нахождения в бюджетном множестве, получаем:

$$\begin{aligned} GX &= W \\ -\frac{1}{2}\Lambda X^T \Omega^{-1} X &= W \qquad \text{так как } rkX = n \\ \Lambda &= -2W(X^T \Omega^{-1} X)^{-1} \end{aligned}$$

Подставляя в G, получаем ответ (следует проверить размерности - они корректны)

$$\underset{k \times m}{G} = \underset{k \times n}{W} \Big(\underset{n \times m}{X^T} \underset{m \times m}{\Omega^{-1}} \underset{m \times n}{X} \Big)^{-1} \underset{n \times m}{X^T} \underset{m \times m}{\Omega^{-1}}$$

№3 [1 балл] Приведите пример μ-сильно выпуклой L-гладкой функции, для которой градиентный спуск сойдётся из любой стартовой точки ровно за одну итерацию. Укажите необходимый для этого шаг метода (он не должен зависеть от точки старта). Ответ обоснуйте.

$$f(x) = \frac{1}{2}x^T x$$

У этой функции константа сильной выпуклости $\mu = 1$, константа гладкости $L = 1$. Легко убедиться, что градиентный спуск с шагом $\alpha = 1$ будет сходится за один шаг из любой точки:

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$
$$x_1 = x_0 - x_0 = 0$$

Здесь часто забывают про $\frac{1}{2}$. Или указывают шаг, который зависит от точки старта.

№4 [2 балла] Упростите выражение:

$$\sum_{i=1}^{n} \langle X^{-1} w_i, w_i \rangle, \quad \text{где} \quad X = \sum_{i=1}^{n} w_i w_i^T, w_i \in \mathbb{R}^n, \det X \neq 0.$$

$$\sum_{i=1}^{n} \langle X^{-1} w_i, w_i \rangle = \sum_{i=1}^{n} \langle X^{-1}, w_i w_i^T \rangle = \left\langle X^{-1}, \sum_{i=1}^{n} w_i w_i^T \right\rangle = \langle X^{-1}, X \rangle = \operatorname{tr}\left(X^{-1} X\right) = n.$$

Отметим, что по определению матрицы X верно, что $X^T = X$.

№5 [2 балла] Предложите метод решения линейной системы уравнений большой размерности:

$$Ax = b,$$

где матрица A симметрична и положительно определена, с помощью одного из методов оптимизации. Пусть известны собственные значения матрицы $\lambda_{\min}(A), \lambda_{\max}(A)$. Приведите оценку скорости сходимости метода (здесь предполагается, что вы предложите метод из курса, тогда доказывать скорость сходимости не нужно).

Можно привести много вариантов, главное, чтобы ответ был правильный и полный. Например, можно поставить следующую задачу оптимизации:

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2.$$

Это квадратичная сильно выпуклая задача (из-за положительно определенной матрицы A). Для её решения можно использовать метод сопряженных градиентов, сходимость которого будет ускоренной линейной, т.е. скорость линейной сходимости будет определяться корнем из числа обусловленности матрицы $\kappa = \kappa(A^T A) = \frac{\lambda_{\max}^2(A)}{\lambda_{\min}^2(A)}$:

$$\|x_k - x^*\|_{A^T A} \leqslant 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x_0 - x^*\|_{A^T A}$$

Можно использовать метод градиентного спуска, ускоренный метод Нестерова, метод тяжелого шарика, квазиньютоновские методы. Здесь часто забывают, что у новой задачи матрица $A^T A$. В этой задаче ставить 0 баллов за ответ $x^* = A^{-1}b$ без формулировки задачи оптимизации и метода. За полную явную формулировку метода Ньютона ставить 1 балл, потому что одна итерация метода стоит столько же, сколько и решение линейной системы.

> **i** №6 [3 балла] Рассмотрите задачу оптимизации:
>
> $$\begin{split} f(x) \to \min_{x \in \mathbb{R}^n} \\ \text{s.t. } a \preceq x \preceq b \end{split}, \qquad a \in \mathbb{R}^n, b \in \mathbb{R}^n,\ a \prec b.$$
>
> Выпишите явно нетривиальную итерацию проксимального градиентного метода для неё.

Следует переписать задачу в эквивалентном виде:

$$f(x) + \mathbb{I}(x) \to \min_{x \in \mathbb{R}^n}, \qquad \mathbb{I}(x) = 0 \text{ если } a \preceq x \preceq b \text{ and } \mathbb{I}(x) = \infty \text{ иначе.}$$

Тогда для её решения можно применить проксимальный метод, который будет в точности совпадать с методом проекции градиента.

$$x_{k+1} = \text{prox}_{\alpha \mathbb{I}}\left(x_k - \alpha \nabla f(x_k)\right)$$

Проксимальный оператор:

$$\text{prox}_{\alpha \mathbb{I}}(x_k) = \arg\min_{x \in \mathbb{R}^n} \left[\frac{1}{2}\|x - x_k\|_2^2 + \alpha \mathbb{I}(x)\right] = \text{proj}_{x \in S}(x_k), \quad S = \{x \in \mathbb{R}^n \mid a \preceq x \preceq b\}$$

В данной задаче надо явно выписать оператор проекции. Это будет клиппинг по каждой из координат:

$$\text{proj}_{x \in S}(x_k^i) = \text{clip}_{[a^i; b^i]}(x_k^i) = \min\left(\max\left(a^i, x_k^i\right), b^i\right)$$

Таким образом, итоговая итерация метода выглядит так:

$$x_{k+1} = \min\left(\max\left(a, x_k - \alpha \nabla f(x_k)\right), b\right)$$

где все сравнения делаются покомпонентно.

> **i** №7 [2 балла] Сходимость алгоритма Франк-Вульфа в курсе показана только для гладких функций. В этой задаче мы рассмотрим, является ли гладкость необходимой. Рассмотрим следующую негладкую задачу с $f : \mathbb{R}^2 \to \mathbb{R}$, $f = \max\{x_1, x_2\}$:
>
> $$\min_{x_1^2 + x_2^2 \le 2} f(x_1, x_2).$$
>
> Предположим, что мы стартуем из точки $(0, 0)$ и запускаем алгоритм Франк-Вульфа (с любым правилом шага). Поскольку функция не является гладкой, мы будем использовать произвольный субградиент вместо градиента. Сходится ли этот алгоритм к оптимуму? Ответ обоснуйте.

Пусть на итерации k текущая итерация обозначается как x_k, а s_k является решением одномерной минимизации на бюджетном множестве (первый шаг алгоритма Франк-Вульфа). Теперь для любого $k \ge 0$ x_k является выпуклой комбинацией x_0 и $\{s_0, \dots, s_{k-1}\}$. Следующая итерация x_{k+1} вычисляется как

$$x_{k+1} = (1 - \gamma_k)x_k + \gamma_k s_k,$$

для некоторого шага $\gamma_k \in [0, 1]$. Таким образом, x_{k+1} является выпуклой комбинацией x_0 и $\{s_0, \dots, s_k\}$.

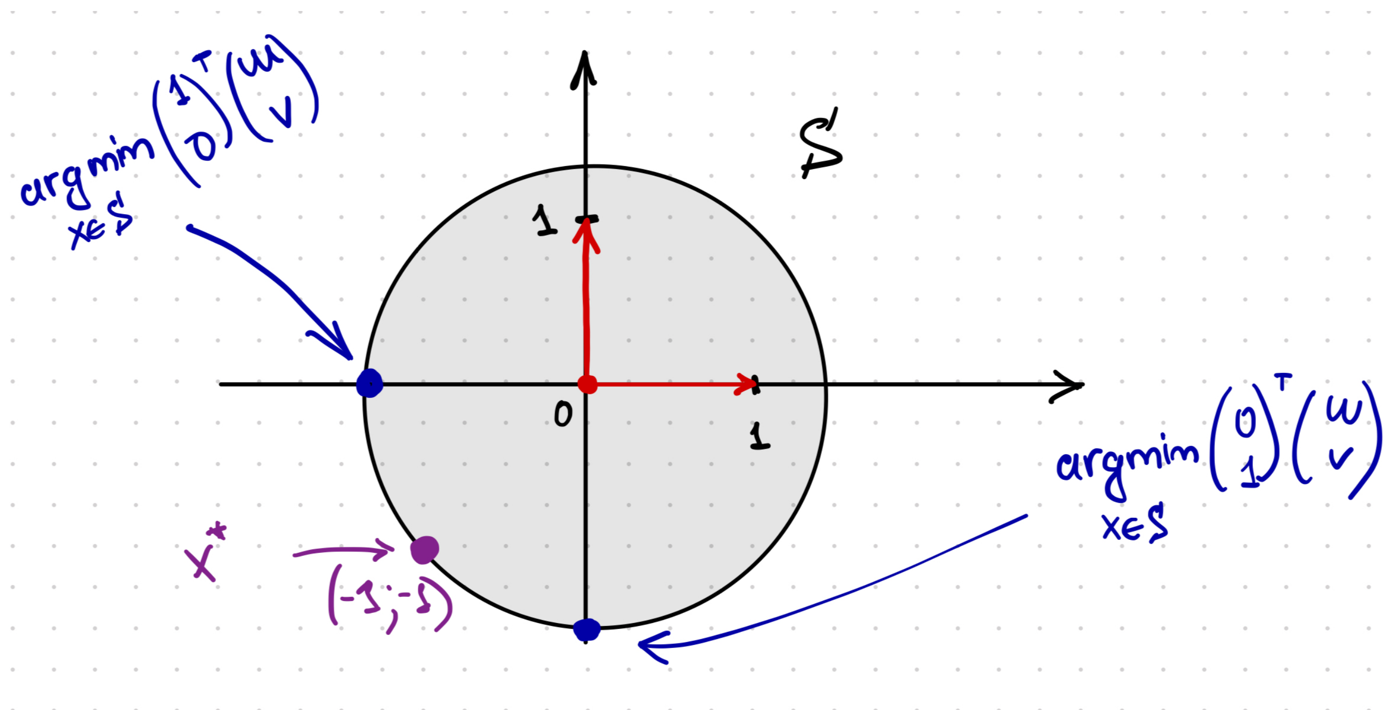
Вычислим s_k для нашей задачи. Субдифференциал в точке $(0, 0)$ представляет из себя отрезок между векторами $(0, 1)$ и $(1, 0)$. Однако, во всех остальных точках, не лежащих на прямой $x_1 = x_2$ он будет либо $(0, 1)$ либо $(1, 0)$.

$$\arg\min\{(1, 0) \cdot (w, v)^\top\} = (-\sqrt{2}, 0), \quad w^2 + v^2 \leq 2$$

и

$$\arg\min\{(0, 1) \cdot (w, v)^\top\} = (0, -\sqrt{2}), \quad w^2 + v^2 \leq 2.$$

Это означает, что на каждом шаге итерации Франк-Вульфа всегда будут лежать в выпуклой оболочке $(0, 0), (-\sqrt{2}, 0)$ и $(0, -\sqrt{2})$. Оптимум находится в точке $(-1, -1)$ и лежит вне этой выпуклой оболочки. Таким образом, алгоритм никогда не сойдется к оптимуму, даже если его запустить на очень долгое время.



Стоит, однако, обратить внимание, что здесь можно случайно угадать с субградиентом $(0.5, 0.5)$ и тогда метод сойдётся. Но, к сожалению, метод предполагает выбор произвольного субградиента. Поэтому следует отметить любознательность студента 1 баллом из двух за такое замечание. Однако, это решение не верно поскольку, когда мы говорим о сходимости алгоритма с произвольным выбором субградиента из субдифференциала - мы можем утверждать наличие сходимости только в случае сходимости при любом выборе субградиента. В противном случае поиск *нужного* субградиента может стоить столько же, сколько и решение самой задачи.

№8 [5 баллов] Рассмотрим выпуклую гладкую задачу минимизации конечной суммы:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) \to \min_{x \in \mathbb{R}^p}$$

Алгоритм SGD выбирает $i \in [n]$ равномерно и устанавливает $\nabla f_i(x_k)$ как стохастический градиент. Иногда можно ускорить SGD, выполняя сэмплирование не равномерно, а по значимости.

а. [1 балл] Рассмотрим произвольное распределение вероятностей $p = (p_1, \dots, p_n)$ с $p_i > 0$ и $\sum_{i=1}^{n} p_i = 1$. Мы выбираем i согласно распределению p и определяем g_k как:

$$g_k := \frac{1}{p_i n}\nabla f_i(x_k) \tag{IS}$$

Тогда покажите, что g_k является несмещенной оценкой градиента, то есть $\mathbb{E}[g_k|x_k] = \nabla f(x_k)$.

$$\mathbb{E}[g_k|x_k] = \sum_{i=1}^{n} p_i \frac{1}{p_i n}\nabla f_i(x_k) = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x) = \nabla f(x_k)$$

б. [2 балла] Напомним, что стандартный единичный симплекс определяется как

$$\Delta_n := \left\{ y \in \mathbb{R}^n : \sum_{i=1}^{n} y_i = 1, y_i \geq 0 \forall i \right\}.$$

Для некоторых фиксированных констант $c_i \in \mathbb{R}$ для $i \in [n]$. Пусть $y^\star$ будет оптимумом следующей задачи оптимизации:

$$\min_{y \in \Delta_n} \sum_{i=1}^{n} \frac{c_i^2}{y_i} \tag{P}$$

Используя общие условия локального экстремума первого порядка, докажите, что

$$y_i^\star = \frac{|c_i|}{\sum_{j=1}^{n} |c_j|}, \forall i \in [n]$$

является решением поставленной задачи (P).

Общие условия локального экстремума первого порядка для минимума функции $g(y) = \sum_{i=1}^{n} \frac{c_i^2}{y_i}$:

$$\nabla g(y^*)^T (y - y^*) \geq 0 \quad \forall y \in \Delta_n.$$

Отметим, что:

$$\frac{\partial g}{\partial y_k}(y^*) = -\frac{c_k^2}{y_k^{*2}} = -\frac{c_k^2}{c_k^2}\left(\sum_{j=1}^{n}|c_j|\right)^2 = -\left(\sum_{j=1}^{n}|c_j|\right)^2.$$

Подставляя полученное в условия оптимальности:

$$\sum_{i=1}^{n} -\left(\sum_{j=1}^{n}|c_j|\right)^2 \left(y_i - \frac{|c_i|}{\sum_{j=1}^{n}|c_j|}\right) = -\left(\sum_{j=1}^{n}|c_j|\right)^2 \sum_{i=1}^{n}\left(y_i - \frac{|c_i|}{\sum_{j=1}^{n}|c_j|}\right) = -\left(\sum_{j=1}^{n}|c_j|\right)^2 (1-1) = 0$$

в. [2 балла] Используя результат из предыдущего пункта, вычислите оптимальную вероятность сэмплирования $p^\star$ для того, чтобы минимизировать дисперсию $\mathbb{E}[\|g_k - \nabla f(x_k)\|^2]$ стохастического градиента g_k, определенного в (IS).

Запишем дисперсию стохастического градиента:

$$
\begin{aligned}
\mathbb{E}[\|g_k - \nabla f(x_k)\|^2] &= \mathbb{E}[\|g_k\|^2] - 2\mathbb{E}[\langle g_k, \nabla f(x_k)\rangle] + \mathbb{E}[\|\nabla f(x_k)\|^2] = \\
&= \mathbb{E}[\|g_k\|^2] - 2\langle \mathbb{E}[g_k], \nabla f(x_k)\rangle + \|\nabla f(x_k)\|^2 = \\
&= \mathbb{E}[\|g_k\|^2] - 2\langle \nabla f(x_k), \nabla f(x_k)\rangle + \|\nabla f(x_k)\|^2 = \mathbb{E}[\|g_k\|^2] - \|\nabla f(x_k)\|^2 \\
&= \sum_{i=1}^n p_i \left\| \frac{1}{p_i n} \nabla f_i(x_k) \right\|^2 - \|\nabla f(x_k)\|^2 = \sum_{i=1}^n \frac{\|\nabla f_i(x_k)\|^2}{p_i n^2} - \|\nabla f(x_k)\|^2.
\end{aligned}
$$

Теперь поставим задачу оптимизации дисперсии с помощью выбора распределения сэмплирования (игнорируя все части, не зависящие от p):

$$
p^\star = \arg\min_{p \in \Delta_n} \sum_{i=1}^n \frac{\|\nabla f_i(x_k)\|^2}{p_i}.
$$

Что в точности совпадает с задачей из предыдущего пункта ($c_i = \|\nabla f_i(x_k)\|$), а значит:

$$
p_k^\star = \frac{\|\nabla f_i(x_k)\|}{\sum_{j=1}^n \|\nabla f_j(x_k)\|}, \forall i \in [n].
$$