Необходимые условия локального минимума. Теорема ККТ

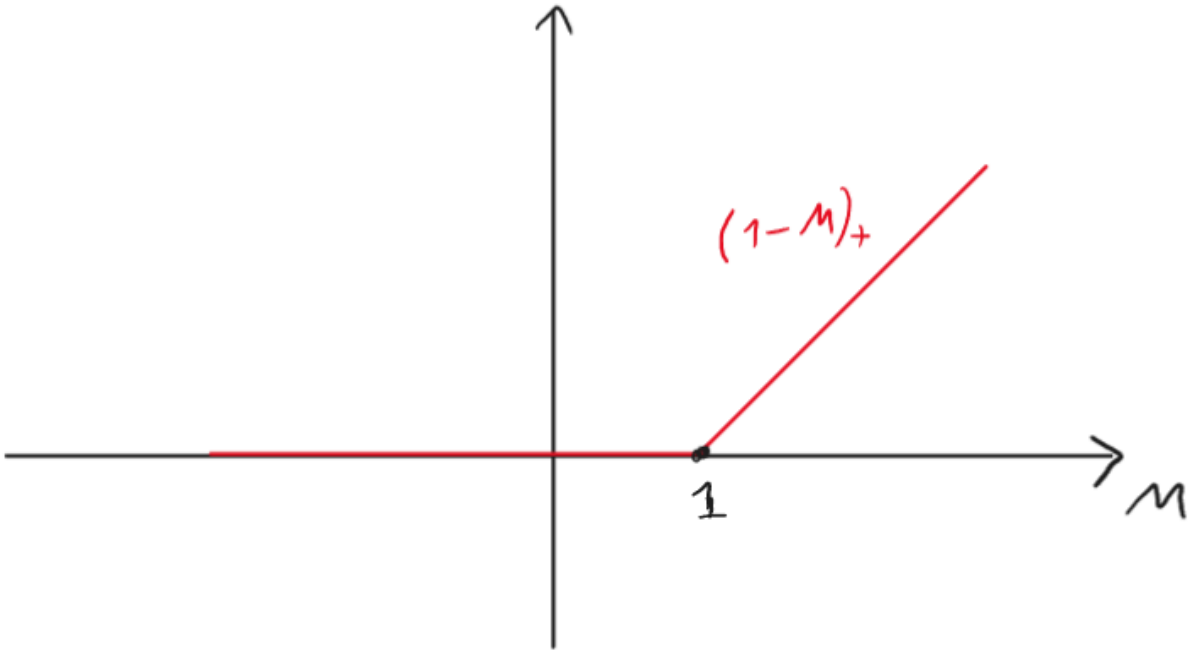
На прошлых семинарах мы уже сталкивались с постановками задач оптимизации: фиксировалась функция потерь $\mathcal{L}(y, \hat{y})$, фиксировалось параметрическое семейство моделей $a(x, y, w)$, где $w \in \mathbb{R}^d$ --- вектор весов. Причем любому выбору вектора весов $w \in \mathbb{R}^d$ соответствовала корректная модель. Для выбора оптимального вектора весов решалась задача оптимизации:

$$\mathcal{L}(y, a(x, y, \omega)) \to \min_{w \in \mathcal{R}^d}$$

Описанный выше пример относился к классу задач безусловной оптимизации. Однако далее нам потребуется научиться решать задачи условной оптимизации. На следующих семинарах будет показано, что возможно сводить некоторые недифференцируемые задачи безусловной оптимизации к гладким задачам условной оптимизации.

В ближайшее время на лекции будет использован аппарат условной оптимизации для обучения классификатора с недифференцируемой функцией потерь. Пусть $\{(x_i, y_i)\}_{i=1}^l$
.
Для SVM-классификатора оптимизационная задача ставится следующим образом:

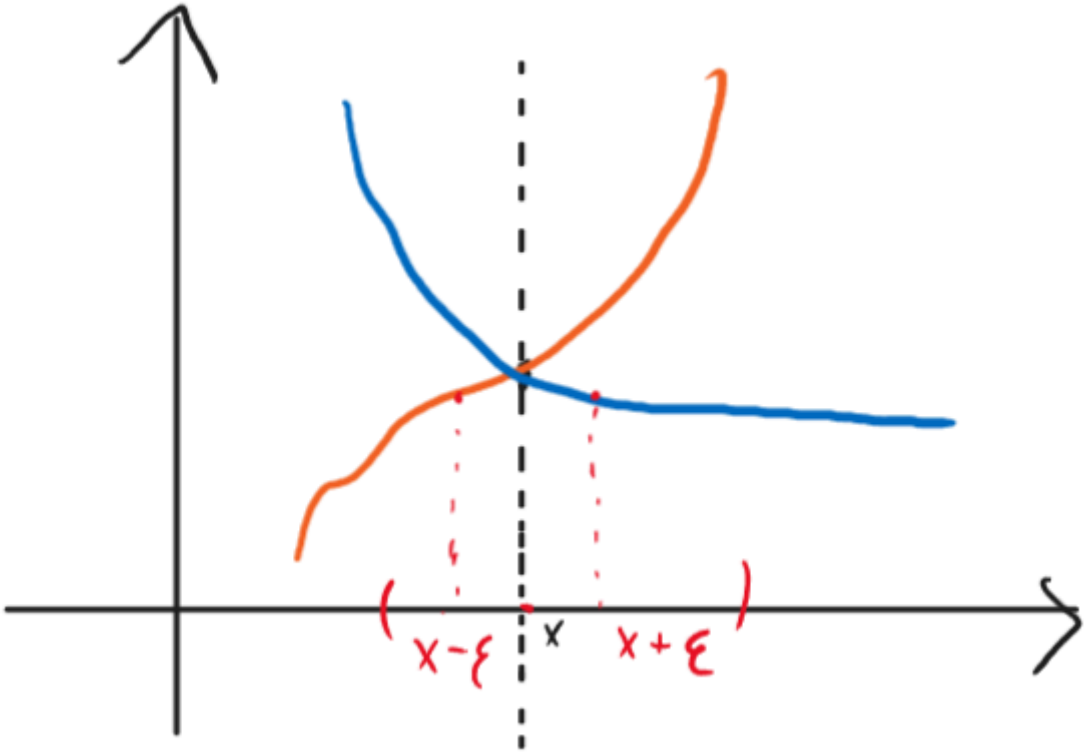
$$M_i(w, w_0) = (\langle x_i, w \rangle - w_0) y_i$$
$$\sum_{i=1}^{l} (1 - M_i)_+ + \frac{1}{2C} \|w\|^2 \to \min$$



В конце текущего семинара мы обсудим причину, по которой l_1 регуляризация осуществляет отбор признаков

От безусловной оптимизации к условной

Для начала вспомним необходимое условие локального минимума для функций одного переменного $f : \mathbb{R} \to \mathbb{R}$, $f \in C^1(\mathbb{R})$
Пусть x_0 --- локальный минимум $f(x_0)$, тогда по определению найдется ε-окрестность x_0, в которой $f(x) \geq f(x_0) \forall x \in B_\varepsilon(x_0)$. Если $\frac{d}{dx} f(x_0) > 0$, то функция возрастает, значит в точке $x - \varepsilon$ будет принимать меньшее значение, если $\frac{d}{dx} f(x_0) < 0$, то в точке $x + \varepsilon$ будет принимать меньшее значение. Отсюда следует, что $\frac{d}{dx} f(x_0) = 0$.



Данное рассуждение обобщается на случай функций многих переменных $f : \mathbb{R}^d \to \mathbb{R}$, $f \in C^1(\mathbb{R}^d)$.
Пусть x_0 --- локальный минимум $f(x_0)$, тогда по определению найдется ε-окрестность x_0, в которой $f(x) \geq f(x_0) \forall x \in B_\varepsilon(x_0)$. Если $\nabla f(x_0) \neq 0$, то в точке $x - \gamma \nabla f(x_0)$ функция будет принимать меньшее значение (γ достаточно мало и положительно), поэтому необходимым условием локального минимума является $\nabla f(x_0) = 0$.

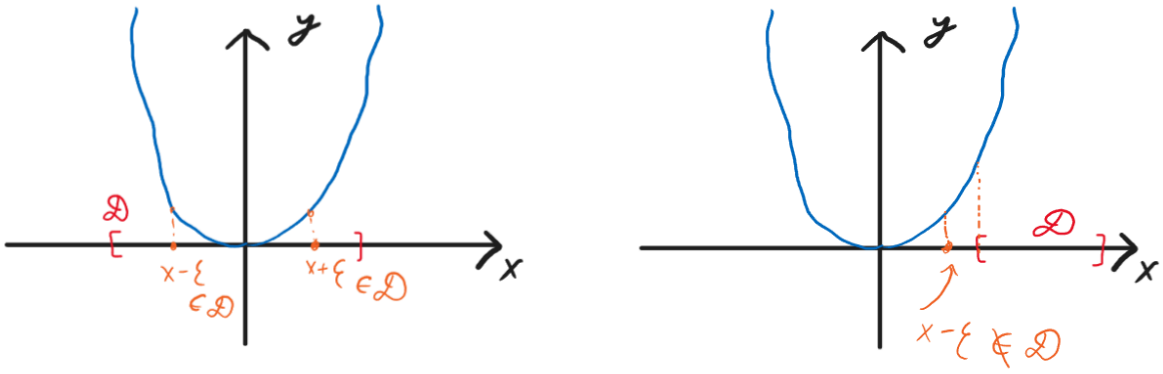
Мы хотим получить необходимое условие для локального минимума в задаче условной оптимизации. Для начала определим класс рассматриваемых задач:

$$
\begin{aligned}
&f : \mathbb{R}^d \to \mathbb{R} \\
&D = \{x \in \mathbb{R}^d : \phi_i(x) \leq 0;\ \psi_j(x) = 0;\ i = \overline{1, m};\ j = \overline{1, n}\} \\
&f,\ \phi_i,\ \psi_j \in C(\mathbb{R}^d)
\end{aligned}
$$

Попробуем понять, почему же условие $\nabla f(x_0) = 0$ не будет являться необходимым условием локального минимума.
Рассмотрим 2 задачи:

1. $$f(x) = x^2 \to \min$$
$$s.t. \;\; x \le 1$$

2. $$f(x) = x^2 \to \min$$
$$s.t. \;\; -x \le -1 \iff x \ge 1$$



В первом случае локальным минимумом будет точка $x_0 = 0$ и градиент $\nabla f = 2x$ будет равен 0 в данной точке.
Во втором случае локальным минимумом будет точка $x_0 = 1$, но градиент в данной точке не обращается в 0.

Если внимательно посмотреть на доказательство теоремы о необходимом условии локального минимума, то можно увидеть, что мы существенно пользовались тем, что нам "доступна" вся окрестность точки x_0. Если градиент не обращается в 0, то найдется направление в котором функция будет убывать. В случае безусловной оптимизации мы могли сделать достаточно малый шаг в данном направлении и получить уменьшение целевой функции. В случае условной оптимизации найденное направление может не принадлежать области допустимых значений D.

Формулировка теоремы Каруша-Куна-Таккера

$$f : \mathbb{R}^d \to \mathbb{R}$$
$$D = \{x \in \mathbb{R}^d : \phi_i(x) \le 0;\; \psi_j(x) = 0;\; i = \overline{1,m};\; j = \overline{1,n}\}$$
$$f,\, \phi_i,\, \psi_j \in C(\mathbb{R}^d)$$

Пусть в рассматриваемой задаче все функции непрерывно дифференцируемы. Пусть выполняется одно из достаточных условий регулярности. Рассмотрим функцию Лагранжа:

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i \varphi_i(x) + \sum_{j=1}^{n} \mu_j \psi_j(x)$$

Если x_0 --- локальный минимум в данной задаче, то найдутся двойственные переменные $\lambda \in \mathbb{R}^m$ и $\mu \in \mathbb{R}^n$. Для тройки (x_0, λ, μ) выполняются следующие условия:

1. $\nabla_x \mathcal{L}(x_0, \lambda, \mu) = 0$
2. $\lambda \geq 0$
3. $\lambda_i \varphi_i(x_0) = 0$
4. $\psi_j(x_0) = 0$

Условие 2 называют двойственной допустимостью, природу данного названия мы обсудим позже. Условие 3 называют дополняющей нежёсткостью.

Если все функции $f(x), \varphi_i(x), \psi_j(x)$ --- выпуклые, то данное условие является ещё и достаточным. Таким образом, для выпуклых задач теорема ККТ является критерием локального минимума.

Если присмотреться, то можно понять, что стационарность функции Лагранжа --- вполне естественное условие. Если $x \notin D$, то $\exists i : \varphi_i(x) > 0$ или $\exists j : |\psi_j(x)| > 0$. Тогда на аддитивную добавку к $f(x)$ в виде $\sum_{i=1}^{m} \lambda_i \varphi_i(x) + \sum_{j=1}^{n} \mu_j \psi_j(x)$ можно смотреть как на штраф за выход из множества. Ожидается, что при достаточно больших значениях $\lambda_i, |\mu_j|$ все стационарные точки будут лежать в множестве D. Если же $\varphi_i(x) < 0$, то данная функция не должна входить в суммарный штраф, поэтому естественно требовать $\lambda_i = 0$ для подобных ограничений.

Далее мы подробно рассмотрим принцип работы данной теоремы

Принцип работы условий ККТ

Ранее мы поняли, что нужно научиться проверять только допустимые направления. Для этого будем рассматривать гладкие кривые $\gamma : [0, 1] \to D$, $\gamma(0) = x_0$. Если x_0 --- локальный минимум, то для любой гладкой кривой $\gamma(t)$ будет выполнено

$$\frac{d}{dt}(f(\gamma(t)))\Big|_{t=0} \geq 0$$

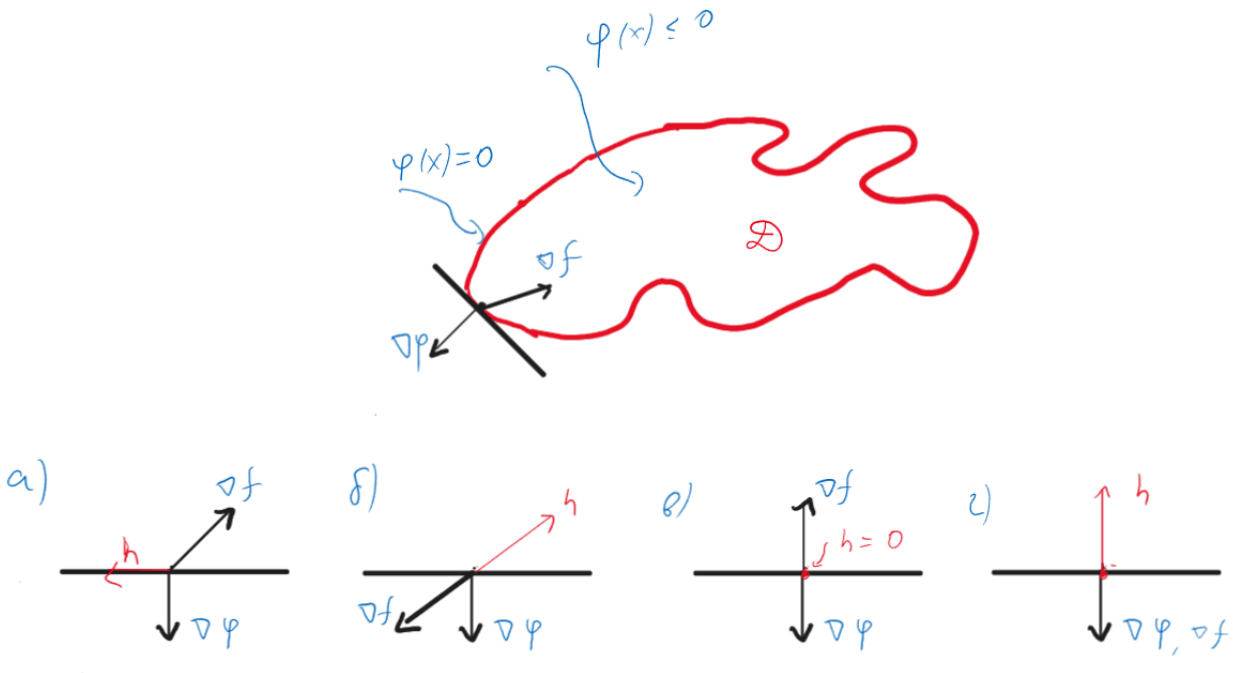
Это условие означает, что функция не убывает по всем допустимым направлениям.

Ранее был описан класс задач для которых мы хотим получить необходимое условие оптимальности. Рассмотрим случай $m = 1,\ n = 0$, с единственным ограничением вида $\varphi(x) \leq 0$.
Пусть x_0 --- точка локального минимума в задаче условной оптимизации. Возможно 2 ситуации:

1. $x_0 \in intD$, тогда необходимым условием будет $\nabla f(x_0) = 0$.
2. $x_0 \in \partial D$. Попробуем получить необходимое условие для данного случая

Если $x_0 \in \partial D$, то $\varphi(x_0) = 0$. Для $\forall x \notin D \Rightarrow \varphi(x) > 0$, для $\forall x \in D \Rightarrow \varphi(x) \leq 0$, значит для любой гладкой кривой $\gamma(t)$ будет выполнено соотношение $\frac{d}{dt}(\varphi(\gamma(t)))\Big|_{t=0} \leq 0$. Это равносильно $\langle \nabla \varphi(x_0), \dot{\gamma}(0) \rangle \leq 0$.

Точка x_0 --- локальный минимум $f(x)$ на множестве D, тогда для любой гладкой кривой будет выполнено $\frac{d}{dt}(f(\gamma(t)))\Big|_{t=0} \geq 0$. Это равносильно $\langle \nabla f(x_0), \dot{\gamma}(0) \rangle \geq 0$.



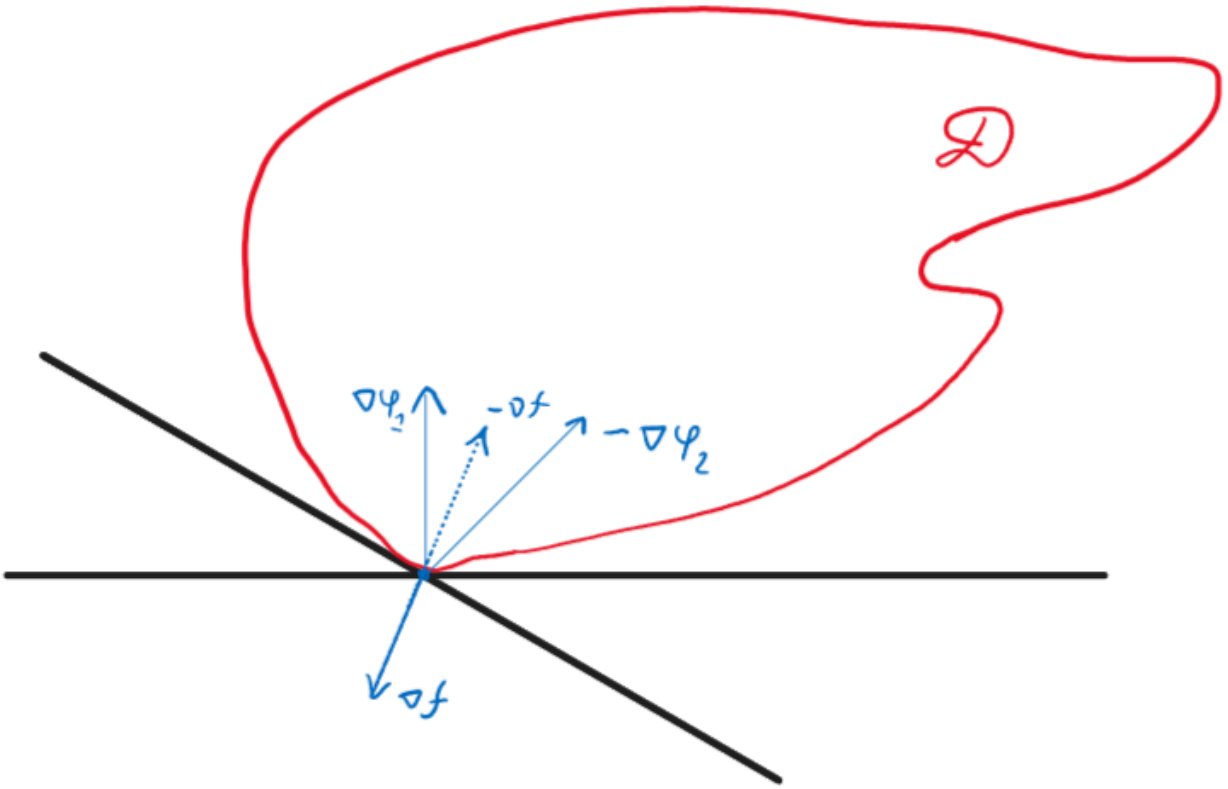
Такое возможно (ситуация 'в') $\iff$ $\nabla f(x_0)$ и $\nabla \varphi(x_0)$ противонаправлены $\iff$

$$\exists \lambda_0 \geq 0, \lambda \geq 0, (\lambda_0, \lambda) \neq 0 : \lambda_0 \nabla f(x_0) + \lambda \varphi(x_0) = 0$$

Далее рассмотрим ситуацию $m > 1$ и $\varphi_i(x_0) = 0 \quad \forall\, i = \overline{1, m}$.
Необходимым условием является противонаправленность $\nabla f(x_0)$ некоторой конической комбинации $\nabla \varphi_1(x_0), \ldots, \nabla \varphi_m(x_0)$:

$$\exists (\lambda_0, \lambda_1, \ldots, \lambda_m) \neq 0 : (\lambda_i \geq 0) \;\&\; (\lambda_0 \nabla f(x_0) + \sum_{i=1}^{m} \lambda_i \nabla \varphi_i(x_0) = 0)$$

Интуиция данного утверждения аналогична интуиции $m = 1$. Строгое доказательство приведено в аппендиксе и может быть рассказано желающим по завершении основной части.



Можно заметить, что полученное выражение в правой части эквивалентно следующим:

$$\lambda_0 \nabla f(x_0) + \sum_{i=1}^{m} \lambda_i \nabla \varphi_i(x_0) = 0 \iff \nabla(\lambda_0 f(x_0) + \sum_{i=1}^{m} \lambda_i \varphi_i(x_0) = 0) \iff \nabla_x \mathcal{L}(x, \lambda) = 0$$

Если точка $x \in intD$, то применимо необходимое условие локального безусловного максимума: $\nabla f(x_0) = 0$. В данной ситуации искомым вектором двойственных переменных является $\lambda = (1, 0, \dots, 0)$. Тогда $\nabla f(x_0) = 0 \iff \nabla_x \mathcal{L}(x, \lambda) = 0$.

Рассматривая ситуацию $m > 1$ мы предполагали обращение всех неравенств в равенство:

$$\varphi_i(x_0) = 0 \quad \forall\, i = \overline{1, m}$$

Если же некоторое неравенство выполняется строго $\varphi_i(x_0) < 0$, то в силу непрерывности любое малое приращение $x = x_0 + \varepsilon h$ не будет выводить нас из множества D. Таким образом, строгие неравенства не накладывают никаких дополнительных ограничений относительно ситуации безусловного экстремума. Если $\varphi_i(x_0) < 0$, то мы можем положить $\lambda_i = 0$. Если же $\varphi_i(x_0) = 0$, то значение $\lambda_i \geq 0$ определяется из условия стационарности функции Лагранжа. Следовательно, будет верно следующее соотношение:

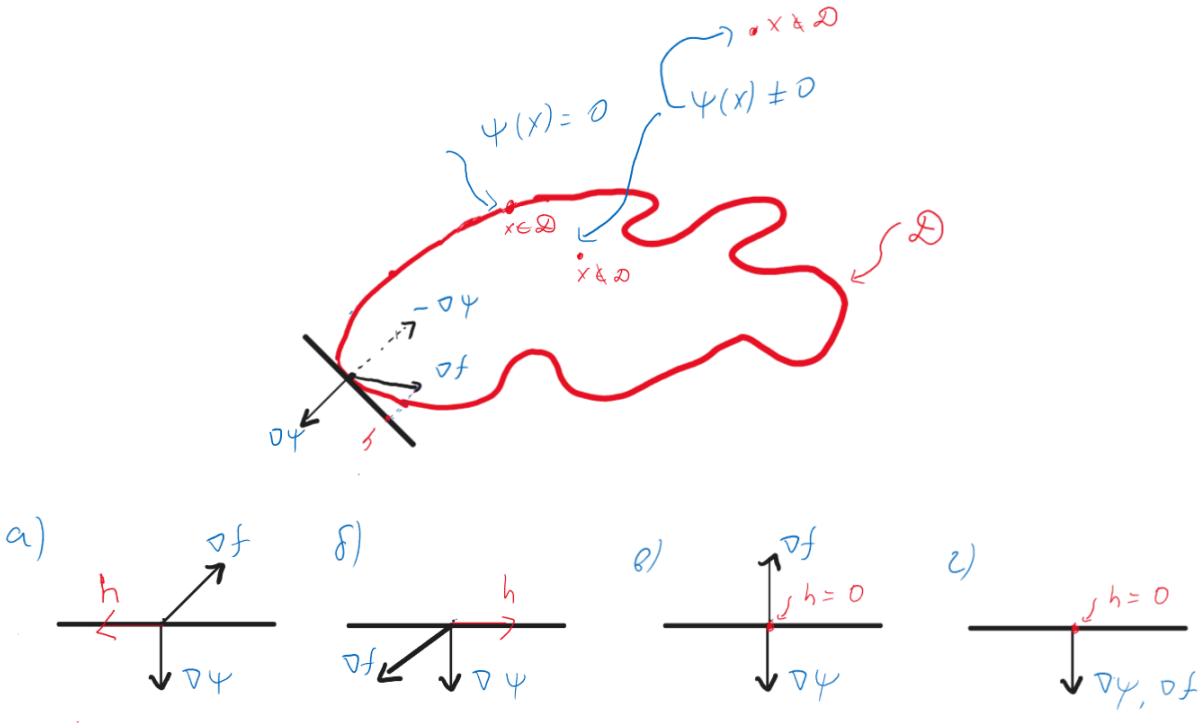
$$(\lambda_i = 0) \vee (\varphi_i(x_0) = 0) \iff \lambda_i \varphi_i(x_0) = 0$$

Важное замечание: возможна ситуация $\lambda_0 = 0$. Регулярность задачи --- суть запрет данной ситуации. Достаточные условия регулярности мы обсудим немного позже

Важное замечание: Все $\lambda_i \geq 0$, это связано с тем, что покинуть множество D можно только увеличив значение некоторой $\varphi_i(x)$

По аналогии рассмотрим ситуацию $m = 0, n = 1$. Нас будет интересовать лишь ситуация $x_0 \in \partial D$, так как в случае $x_0 \in intD$ необходимым условием будет являться $\nabla f(x_0) = 0$

Если $x_0 \in \partial D$, то $\psi(x_0) = 0$. Тогда для любой гладкой кривой $\gamma(t) : [0, 1] \to D$, $\gamma(0) = x_0$ будет выполнено $\frac{d}{dt}\psi(\gamma(t))\Big|_{t=0} = 0$. Это равносильно $\langle \nabla\psi(x_0), \dot\gamma(0)\rangle = 0$.
Так как x_0 --- точка локального минимума $f(x)$, то для любой гладкой кривой $\gamma(t)$ будет выполнено $\frac{d}{dt}(f(\gamma(t)))\Big|_{t=0} \geq 0$. Это равносильно $\langle \nabla f(x_0), \dot\gamma(0)\rangle = 0$.



Из условий выше (ситуации в, г на рисунке выше) следует, что $\nabla\psi(x_0)$ и $\nabla f(x_0)$ линейно зависимы.
Тогда существуют коэффициенты λ_0, μ не равные 0 одновременно, такие что

$$\lambda_0\nabla f(x_0) + \mu\nabla\psi(x_0) = 0$$

В данном случае знак λ_0, μ не определен. Для согласованности с предыдущими размышлениями можно считать, что $\lambda_0 \geq 0$. Это не влияет на общность рассуждений, так как в противном случае, в силу произвольности знака μ, можно умножить равенство на -1.

Далее рассмотрим ситуацию $n > 1$. Проводя аналогичные рассуждения, можно получить следующее условие: если $x_0 \in \partial D$ --- точка локального минимума, то $\nabla f(x_0), \nabla\psi_1(x_0), \ldots, \nabla\psi_n(x_0)$ --- линейно зависимы. Тогда существуют коэффициенты $\lambda_0 \geq 0, \mu_1, \ldots, \mu_n$ не равные 0 одновременно, такие что

$$\lambda_0\nabla f(x_0) + \sum_{j=1}^{n}\mu_j\nabla\psi_j(x_0) = 0$$

Можно заметить, что полученное выражение эквивалентно следующим:

$$\lambda_0 \nabla f(x_0) + \sum_{j=1}^{n} \mu_j \nabla \psi_j(x_0) = 0 \iff \nabla(\lambda_0 f(x_0) + \sum_{j=1}^{n} \mu_j \psi_j(x_0) = 0) \iff \nabla_x \mathcal{L}(x, \mu) = 0$$

Важное замечание: В данном случае всегда выполняется равенство $\psi_j(x_0) = 0, \ \forall j$. А неотрицательность μ_j не требуется. Таким образом, условия двойственной допустимости и дополняющей нежесткости.

Собрав полученные условия воедино мы получаем теорему ККТ.

Регулярные задачи

Будем называть ограничение $\varphi_i(x) \leq 0$ активным в точке x_0, если $\varphi_i(x_0) = 0$.
Далее будем считать, что $\{i_1, \dots, i_k\}$ --- множество индексов всех активных ограничений в точке x_0.

Необходимое условие локального условного минимума формулировалось в терминах линейной зависимости $\nabla f(x_0)$ и $\nabla \varphi_{i_1}(x_0), \dots, \nabla \varphi_{i_k}(x_0), \nabla \psi_1(x_0), \dots \nabla \psi_n(x_0)$. Однако может оказаться так, что коэффициент перед $\nabla f(x_0)$ в линейной комбинации равен 0. Мы хотим найти условия, которые исключают подобную ситуацию.

Равенство коэффициента $\lambda_0 = 0$ означает линейную зависимость $\nabla \varphi_{i_1}(x_0), \dots, \nabla \varphi_{i_k}(x_0), \nabla \psi_1(x_0), \dots \nabla \psi_n(x_0)$. Поэтому если $\nabla \varphi_{i_1}(x_0), \dots, \nabla \varphi_{i_k}(x_0), \nabla \psi_1(x_0), \dots \nabla \psi_n(x_0)$ линейно независимы, то гарантируется $\lambda_0 > 0$.

Если все ограничения $\varphi_1(x), \dots, \varphi_m(x), \psi_1(x), \dots, \psi_n(x)$ --- аффинные функции с линейно независимыми направляющими векторами, то гарантируется $\lambda_0 > 0$. Это напрямую следует из предыдущего пункта.

Условие Слейтера (док-во можно вынести в бонусную часть). Пусть $f(x), \varphi_1(x), \dots, \varphi_m(x)$ --- выпуклые функции, $\psi_1(x), \dots, \psi_n(x)$ --- аффинные функции. Пусть существует $\tilde{x} : \varphi_i(x) < 0, \psi_j(x) = 0$. Тогда гарантируется $\lambda_0 > 0$.

Достаточные условия регулярности

1. Градиенты активных ограничений линейно независимы
2. Все ограничения --- линейно независимые аффинные функции
3. Все ограничения типа равенств --- аффинные функции, все ограничения типа неравенств --- выпуклые функции, существует $\tilde{x} : \varphi_i(x) < 0, \psi_j(x) = 0$
4. Все ограничения типа равенств --- аффинные функции. Функции $\varphi_1(x), \dots, \varphi_k(x)$ --- аффинные, $\varphi_{k+1}(x), \dots, \varphi_m(x)$ --- выпуклые. $\tilde{x} : \varphi_i(\tilde{x}) \leq 0 \ \forall i = \overline{1, k}$, $\varphi_i(\tilde{x}) < 0 \ \forall i = \overline{k+1, m}$, $\psi_j(x) = 0$

Двойственные задачи

В предыдущей главе мы ввели функцию Лагранжа:

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i \varphi_i(x) + \sum_{j=1}^{n} \mu_j \psi(x)$$

Далее мы будем считать, что задача регулярна, поэтому мы $f(x)$ входит в функцию Лагранжа без множителя λ_0.

Если $x \in D$, то для любого набора двойственных переменных $\lambda_i \geq 0$ мы получим следующее соотношение:

$$\mathcal{L}(x, \lambda, \mu) \leq f(x)$$

Пусть (x_*, λ_*, μ_*) --- тогда

$$\mathcal{L}(x_*, \lambda, \mu) \leq \mathcal{L}(x_*, \lambda_*, \mu_*) \leq \mathcal{L}(x, \lambda_*, \mu_*)$$

Тогда получим следующие соотношения:

$$\mathcal{L}(x_*, \lambda_*, \mu_*) = \sup_{\lambda \geq 0, \mu} \mathcal{L}(x_*, \lambda, \mu)$$

$$\mathcal{L}(x_*, \lambda_*, \mu_*) = \inf_{x \in D} \mathcal{L}(x, \lambda_*, \mu_*)$$

Рассмотрим более внимательно второе соотношение: $x \in D$, поэтому справедливо неравенство $\mathcal{L}(x, \lambda, \mu) \leq f(x)$. Попробуем найти $\sup\limits_{\lambda \geq 0, \mu} \mathcal{L}(x, \lambda, \mu)$. Если $x \in D$, то в силу неравенства $\sup\limits_{\lambda \geq 0, \mu} \mathcal{L}(x, \lambda, \mu) \leq f(x)$, причём равенство достигается при $\lambda = 0, \mu = 0$. Если $x \notin D$, то либо $\exists i : \varphi_i(x) > 0$ и тогда $\lim_{\lambda_i \to +\infty} \mathcal{L}(x, \lambda, \mu) = +\infty$, либо $\exists j : \psi_j(x) \neq 0$, тогда $\limsup_{\mu_j \to \infty} \mathcal{L}(x, \lambda, \mu) = +\infty$.
Таким образом, мы получили следующее соотношение:

$$\sup_{\lambda \geq 0, \mu} \mathcal{L}(x, \lambda, \mu) = f(x), x \in D \sup_{\lambda \geq 0, \mu} \mathcal{L}(x, \lambda, \mu) = +\infty, x \notin D$$

Таким образом, исходная задача условной оптимизации переписывается в виде:

$$\mathcal{L}(x_*, \lambda_*, \mu_*) = \inf_{x \in R^n} \sup_{\lambda \geq 0, \mu} \mathcal{L}(x, \lambda, \mu)$$

Данное соотношение наталкивает на мысль, что можно попробовать переставить $\inf$ и $\sup$ местами, перейдя к задаче

$$\mathcal{L}(x^*, \lambda^*, \mu^*) = \sup_{\lambda \geq 0, \mu} \inf_{x \in R^n} \mathcal{L}(x, \lambda, \mu)$$

Далее мы посмотрим к каким результатам приведёт новая постановка задачи, а также обсудим условия, при которых данный переход будет эквивалентен (
$\mathcal{L}(x_*, \lambda_*, \mu_*) = \mathcal{L}(x^*, \lambda^*, \mu^*)$)

Введем обозначение $g(\lambda, \mu) = \inf\limits_{x \in R^n} \mathcal{L}(x, \lambda, \mu)$, в терминах которого формулируется задача двойственной оптимизации:

$$g(\lambda, \mu) \to \max$$
$$\text{s. t. } \lambda \geq 0$$

Функция $g(\lambda, \mu)$ является выпуклой в силу линейности $\mathcal{L}(x, \lambda, \mu)$ по λ и μ

$$\mathcal{L}(x, \theta\lambda_1 + (1-\theta)\lambda_2, \theta\mu_1 + (1-\theta)\lambda_2) = \theta\mathcal{L}(x, \lambda_1, \mu_1) + (1-\theta)\mathcal{L}(x, \lambda_2, \mu_2)$$

$$\sup_{\lambda \geq 0, \mu} [\theta\mathcal{L}(x, \lambda_1, \mu_1) + (1-\theta)\mathcal{L}(x, \lambda_2, \mu_2)] \leq \theta \sup_{\lambda \geq 0, \mu} \mathcal{L}(x, \lambda_1, \mu_1) + (1-\theta) \sup_{\lambda \geq 0, \mu} \mathcal{L}(x, \lambda_2, \mu_2)$$

Таким образом, новая задача является задачей выпуклой условной оптимизации с ограничениями вида $\lambda \geq 0$. В ходе решения задач мы убедимся, что полученные задачи решаются проще. В том числе, переход к двойственной задаче будет использован при постановке задачи обучения SVM.

Существенным недостатком данного перехода является жёсткость достаточных условий его эквивалентности. Если $f(x), \varphi_i(x), \psi_j(x)$ --- выпуклые функции, то
$\mathcal{L}(x_*, \lambda_*, \mu_*) \ = \ \mathcal{L}(x^*, \lambda^*, \mu^*)$
В противном случае гарантируется лишь неравенство:

$$\mathcal{L}(x_*, \lambda_*, \mu_*) \ \geq \ \mathcal{L}(x^*, \lambda^*, \mu^*)$$

А величина $\mathcal{L}(x_*, \lambda_*, \mu_*) - \mathcal{L}(x^*, \lambda^*, \mu^*)$ называется зазором двойственности. Некоторые задачи могут оказаться толерантными к малым величинам зазора двойственности, поэтому двойственную постановку используют и для некоторых невыпуклых задач условной оптимизации.