

Впрацюючи зі прикладами, де і буде позначати арифметичну середню,
(і не мішати)

ф-ція лінійного
визначення

def. $f: C \rightarrow \mathbb{R}^k$, де $C \subset \mathbb{R}^d$, зветься ρ -лінійною, де $\rho > 0$,

якщо $\forall w_1, w_2 \in C \quad |f(w_1) - f(w_2)| \leq \rho |w_1 - w_2|$.

Рл Якщо $f: C \rightarrow \mathbb{R}$ диференційовна, де $C \subset \mathbb{R}^d$ опукла і
фігурна, і $|\nabla f|$ обмежений на C : $\exists \rho > 0: |\nabla f| \leq \rho$,
то f - ρ -лінійна на C .

⇒ Зл (Болановичного версії) Тл. що середнє значення ма
опуклість C , $\forall w_1, w_2 \in C \quad \exists \alpha \in [0, 1]$:

$$f(w_1) - f(w_2) = \langle \nabla f(\alpha w_1 + (1-\alpha)w_2), w_1 - w_2 \rangle.$$

Погі за нек. К.-Б.-М.

$$|f(w_1) - f(w_2)| \leq |\nabla f(\alpha w_1 + (1-\alpha)w_2)| |w_1 - w_2| \leq \rho |w_1 - w_2|. \triangle$$

Ex. 1. $f(w) = |w|$ 1-линейна (maxi p-гii не было неограниченно) в силу пер-сти тр-ка: $\forall w_1, w_2 \in \mathbb{R}^d$

$$|w_1| - |w_2| = |w_1 - w_2 + w_2| - |w_2| \leq |w_1 - w_2| + |w_2| - |w_2| = |w_1 - w_2|,$$

(так само $|w_2| - |w_1| \leq |w_2 - w_1| = |w_1 - w_2| \Rightarrow ||w_1| - |w_2|| \leq |w_1 - w_2|$).

2. $f(x) = \ln(1 + e^x)$ 1-линейна на $\mathbb{R}$, до ii неограни

$$f'(x) = \frac{e^x}{1+e^x} = \frac{1}{e^{-x}+1} < 1 \quad (i > 0) \quad - \text{ограничена.}$$

3. $f(x) = x^2$ не $\in$ линейна на $\mathbb{R}$: $\forall \delta > 0, x_1 := 0$ и $x_2 := 1 + \delta$

$$|f(x_2) - f(x_1)| = (1 + \delta)^2 > \delta(1 + \delta) = \delta |x_2 - x_1|$$

Тем самым $\forall \delta > 0$ была δ -линейна на $C = [-\frac{\delta}{2}, \frac{\delta}{2}]$: $\forall x_1, x_2 \in C$

$$|f(x_2) - f(x_1)| = |x_1^2 - x_2^2| = |x_1 + x_2| |x_1 - x_2| \leq (|x_1| + |x_2|) |x_1 - x_2| \leq \delta |x_1 - x_2|.$$

4. Априори p-гii $f(w) = \langle v, w \rangle + b$ (где фикс. $v \in \mathbb{R}^d, b \in \mathbb{R}$)

$|v|$ -линейна на $\mathbb{R}^d$: за пер. К-Б-М., $\forall w_1, w_2 \in \mathbb{R}^d$

$$|f(w_1) - f(w_2)| = |\langle v, w_1 - w_2 \rangle| \leq |v| |w_1 - w_2|.$$

Pr. Если f_1 ρ_1 -линейна, а f_2 ρ_2 -линейна, то $f_1 \circ f_2 \in$

$\rho_1 \rho_2$ -линейна.

$\Rightarrow$ Таким же образом, что $f_1 \circ f_2$ булн, можно $f_2: \mathbb{R}^d \rightarrow \mathbb{R}^k$,
 $f_1: \mathbb{R}^k \rightarrow \mathbb{R}^l$. Тогдаме, относительно связувае з лев.: $\forall w_1, w_2$
 $|f_2(f_1(w_1)) - f_2(f_1(w_2))| \leq \rho_2 |f_1(w_1) - f_1(w_2)| \leq \rho_2 \rho_1 |w_1 - w_2|$. Δ
Ex. $\forall$ ρ -линем. $f: \mathbb{R} \rightarrow \mathbb{R}$ $w \mapsto f(\langle \sigma, w \rangle + b)$ $\rho|\sigma|$ -линем. за
Ex. 4. булн $\in \underline{Pr}$. ($\forall \sigma \in \mathbb{R}^d, b \in \mathbb{R}$).

лев. Функция $f: \mathbb{R}^d \rightarrow \mathbb{R}$ будем называть β -мажн., $\beta > 0$,
 якщо она дифференцируема, а її градиент $\in \beta$ -линейное
 бигн. $\mathbb{R}^d \rightarrow \mathbb{R}^d$: $\forall w_1, w_2 \in \mathbb{R}^d$ $|\nabla f(w_1) - \nabla f(w_2)| \leq \beta |w_1 - w_2|$.

Рем. З линейности булн, непрерывности, могу чей бариант
 мажн. - посленна "хачайн" C^1 -мажн. з анализа.

За Тн. про сер. знач., $\forall w_1, w_2 \exists \alpha \in [0, 1]$:

$$\begin{aligned}
 f(w_1) - f(w_2) &= \langle \nabla f(\underbrace{\alpha w_1 + (1-\alpha)w_2}_{r: w_\alpha}), w_1 - w_2 \rangle = \langle \nabla f(w_2), w_1 - w_2 \rangle + \\
 &+ \langle \nabla f(w_\alpha) - \nabla f(w_2), w_1 - w_2 \rangle \leq [M - \beta \cdot M] \leq \langle f(w_2), w_1 - w_2 \rangle + |\nabla f(w_\alpha) - \\
 &- \nabla f(w_2)| |w_1 - w_2| \leq
 \end{aligned}$$

Якщо f - β -м., то

$$\leq \langle \nabla f(w_2), w_1 - w_2 \rangle + \beta \underbrace{|w_2 - w_2|}_{\parallel} |w_1 - w_2| = \langle \nabla f(w_2), w_1 - w_2 \rangle + \alpha \beta |w_1 - w_2|^2 \quad (*)$$

Якщо $\alpha \leq \frac{1}{2}$, то це $(*) \leq \langle \nabla f(w_2), w_1 - w_2 \rangle + \frac{\beta}{2} |w_1 - w_2|^2$. Якщо ж $\alpha > \frac{1}{2}$, то повторимо ці ж виваженки з w_1 замість w_2 :

$$f(w_2) - f(w_1) = \langle \nabla f(w_2), w_2 - w_1 \rangle = \langle \nabla f(w_1), w_2 - w_1 \rangle + \langle \nabla f(w_2) - \nabla f(w_1), w_2 - w_1 \rangle \leq \langle \nabla f(w_1), w_2 - w_1 \rangle + \frac{\beta}{2} |w_2 - w_1|^2, \text{ нічого змінюючи } w_1 \text{ і } w_2 \text{ місцями.}$$

$$\forall \text{ пари, } f(v) \leq f(w) + \langle \nabla f(w), v - w \rangle + \frac{\beta}{2} |v - w|^2 \quad \forall v, w \quad (*)$$

$$\forall \beta\text{-м. } f. \text{ існує деяка, } f(v) \geq f(w) + \langle \nabla f(w), v - w \rangle,$$

якщо f опукла.

$$\text{Якщо ж } (*) \text{ з } v = w - \frac{1}{\beta} \nabla f(w), \text{ то}$$

$$f(v) \leq f(w) - \frac{1}{\beta} |\nabla f(w)|^2 + \frac{1}{2\beta} |\nabla f(w)|^2,$$

$$\frac{1}{2\beta} |\nabla f(w)|^2 \leq f(w) - f(v)$$

Якщо при цьому $f(v) \geq 0$ (зокр., при $f \geq 0$),

$$|\nabla f(w)|^2 \leq 2\beta f(w),$$

тобто $f \in$ самобмеженню (self-bounded).

Ex. 1. Оск. для $f(x) = x^2$ $f'(x) = 2x$ ор. 2-линейная,
 $f \in 2$ -магнор.

2. Для $f(x) = \ln(1+e^x)$ $f'(x) = \frac{e^x}{1+e^x} = \frac{1}{e^{-x}+1}$, отнсе

$$|f''(x)| = \left| \frac{e^{-x}}{(e^{-x}+1)^2} \right| = \left| \frac{1}{(1+e^x)(1+e^{-x})} \right| = \left| \frac{1}{2+2\cosh x} \right| \leq \frac{1}{4}, \text{ ману}$$

$f' \frac{1}{4}$ -линей. $\Rightarrow f \frac{1}{4}$ -м.

Р. Если $f: \mathbb{R} \rightarrow \mathbb{R}$ β -магнор, то $g(w) = f(\langle v, w \rangle + b) \in \beta|v|^2$ -магнор ($\forall v \in \mathbb{R}^d, b \in \mathbb{R}$).

▷ Дифференцируем: $\nabla g(w) = f'(\langle v, w \rangle + b) v$. За условия, f
 β -м. $\Rightarrow f'$ β -линей. $\Rightarrow w \mapsto f'(\langle v, w \rangle + b) \beta|v|$ -линей. За Р. 1.1
будет. Ману $\forall w_1, w_2$ $|\nabla g(w_1) - \nabla g(w_2)| = |f'(\langle v, w_1 \rangle + b) - f'(\langle v, w_2 \rangle + b)| |v| \leq \beta|v|^2 |w_1 - w_2|$, что и необходимо. $\triangle$

Ex. 1. Квадратичная функция впрам $w \mapsto (\langle w, x \rangle - y)^2 \forall x \in \mathbb{R}^d$,
 $y \in \mathbb{R} \in 2|x|^2$ -магнор за Ex. 1. будет и Р.

2. Логистическая ф-ция впрам $w \mapsto \ln(1+e^{-y\langle w, x \rangle}) \forall x \in \mathbb{R}^d$,
 $y = \pm 1 \in \frac{|x|^2}{4}$ -м, за Ex. 2. будет и Р. (а также $|x|$ -линей, жк)

показано раніше).

Ці приклади мотивують наступні означення:

def. Класична задача машинного навчання визначена на мн. Z класів іпотез $\mathcal{H}$ і ф-цією втрат $\ell: \mathcal{H} \times Z \rightarrow \mathbb{R}_+$, причому $\mathcal{H}$ отождествляється з підмножиною $\mathbb{R}^d: \mathcal{H} \subset \mathbb{R}^d$.
Тоді ця задача зветься

- опуклою, якщо $\mathcal{H}$ опукла і $\forall z \in Z$ ф-ція $\ell(\cdot, z): \mathcal{H} \rightarrow \mathbb{R}_+$ опукла;
- ρ -ліпшицевою для $\rho > 0$, якщо $\forall z \in Z$ $\ell(\cdot, z)$ ρ -ліпш.;
- β -ладною для $\beta > 0$, якщо $\forall z \in Z$ $\ell(\cdot, z)$ β -ладна.;
- обмеженою (bounded), якщо $\mathcal{H}$ – обмежена множина!

$\exists B > 0: \forall w \in \mathcal{H} \quad |w| \leq B$ (тобто $\mathcal{H} \subset \mathcal{D}_B(0)$ – замкн. кулі радіуса B з центром у 0).

Ex. 1. Для задачі лінійної регресії на $\mathbb{R}^d$ з обмеженими іпотезами маємо $Z = \mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R} = \mathbb{R}^{d+1}$,

і $\mathcal{H}$ скл. з ф-цій вигляду $x \mapsto \langle w, x \rangle$, що можна
ототожнити з векторами коефіцієнтів $w \in \mathbb{R}^d$. Тл.ч.,
 $\mathcal{H} = \mathbb{R}^d$ - окрла. Поді, як показати вище, квадратична
ф-ція втрам $\ell_{sq}(w, (x, y)) = (\langle w, x \rangle - y)^2 \in \text{окрлого}$.

Тому це опукла задача, Оск., крім того, $\ell_{sq} \in 2|x|^2$ -м.
Якщо об'єксити X , взавши замість $\mathbb{R}^d$ евкл. замкнену
кулю $\mathbb{D}_\beta(0) = \{x \mid |x| \leq \beta\}$, то задача стане $2\beta^2$ -макмою.
Якщо ж об'єксити $\bar{u}$ $K := \mathbb{D}_\beta(0) = \{w \mid |w| \leq \beta\}$, то задача
стане об'єксною (при цьому вона залишиться опуклою,
бо $\mathbb{D}_\beta(0)$ опукла - Впр.).

З. Для n -ї не задані φ -ція абсолютної втрапи
 $\ell_{abs}(w, (x, y)) = |\langle w, x \rangle - y|$ є опуклою і $|x|$ -ліній. (До
 $| \cdot |$ опукла і 1-ліній.). Понам ліній. перспекія з аБС.
 Втрапи опукла і ρ -ліній. на $X = \mathbb{D}_\rho(0)$. Якщо
 взяти $K = \mathbb{D}_\rho(0)$ вона знову не буде обмеженою.

З. В задані логістичної перспекії $Z = X \times Y$, де
 $X \subset \mathbb{R}^d$, а $Y = \{\pm 1\}$. Осн. логістична φ -ція втр.
 $\ell_{\log}(w, (x, y)) = \ln(1 + e^{-y \langle w, x \rangle})$, як ми показали,
 опукла, $|x|$ -ліній. і $\frac{|x|^2}{4}$ - магна, для $X = \mathbb{D}_\rho(0)$
 ця задача опукла, ρ -ліній. і $\frac{\rho^2}{4}$ - магна (і знову не
 обмежена, якщо взяти $K = \mathbb{D}_\rho(0)$).

Реш. У опуклих задачах ERM означає знаходження
 w з опуклої $K \subset \mathbb{R}^d$, що мінімізує ліній. ризик
 $L_S(w) = \frac{1}{m} \sum_{i=1}^m \ell(w, z_i)$ для вибірки $S = \{z_i\}_{i=1}^m$.
 Осн. усі $\ell(\cdot, z_i)$ опуклі, L_S теж є опуклою

ф-ції на функції $\mathcal{H}$ за означ з Рч. вище. Т.ч., ERM є
задачею опуклої оптимізації. Для таких задач відомі
ефективні алгоритми (за год. умов, гув. галі).

З іншого боку, опуклість ще не гарантує навчальності
у статистичному сенсі (і разом з обмеженістю такою):
Рч. Задача лінійної регресії на $\mathbb{R}$ (тобто при $d=1$)
з квадратичною втрапою не є адекватно PAC навч.
для $\mathcal{H} = \mathbb{R}$ або $\mathcal{H} = \mathbb{D}_1(0) = [-1, 1]$ (тобто ці класи не
є адм. PAC навч.).

⇒ Тут ми припускаємо, що алгоритми A детермініс-
тські (тобто $A(S)$ однозначно визначена S).

Покладено $\epsilon = \frac{1}{8}$ $\delta = \frac{1}{2}$. Якщо δ навчальність мала
місце, то існували δ алгоритми A і $m_{\mathcal{H}} : (0,1)^2 \rightarrow \mathbb{N}$

maxi, что $\forall m > m_{\mathcal{K}}(\varepsilon, \delta) \forall$ разбиения $\mathcal{Q}$ на $Z = X \times Y =$
 $= \mathbb{R}^2$ zi им. $\geq 1 - \delta = \frac{1}{2}$ для $S \sim \mathcal{Q}^m$

$$L_{\mathcal{Q}}(A(S)) \leq \min_{w \in \mathcal{K}} L_{\mathcal{Q}}(w) + \varepsilon.$$

Для данного m выберем $\mu := \frac{1}{2m \ln \frac{3}{2}}$ разбиено два разбиения $\mathcal{Q}_1, \mathcal{Q}_2$ на $\mathbb{R}^2$:

- при $\mathcal{K} = \mathbb{R}$ пусть $\mathcal{Q}_1$ сосредоточенный у $z_1 = (1, 0)$ zi им. μ и $z_2 = (\mu, -1)$ zi им. $1 - \mu$, $\mathcal{Q}_2$ - тоже у z_2 (zi им. 1).

Поскольку zi им. $\geq (1 - \mu)^m > e^{-2\mu m} = \frac{2}{3} > \frac{1}{2}$ S складывается лишь

из z_2 . Пусть A повертает за максимум S $A(S) = w$. Если

$w < -\frac{1}{2\mu}$, то для $\mathcal{Q} = \mathcal{Q}_1$

$$L_{\mathcal{Q}_1}(w) = \mu (1 \cdot w - 0)^2 + (1 - \mu)(\mu \cdot w + 1)^2 \geq \mu w^2 \geq \frac{1}{4\mu},$$

$$\text{а } \min_{w^*} L_{\mathcal{Q}_1}(w^*) \leq L_{\mathcal{Q}_1}(0) = \mu (1 \cdot 0 - 0)^2 + (1 - \mu)(\mu \cdot 0 + 1)^2 =$$

$$= 1 - \mu, \text{ тогда } L_{\mathcal{Q}_1}(w) - \min_{w^*} L_{\mathcal{Q}_1}(w^*) \geq \frac{1}{4\mu} - (1 - \mu) = \frac{\mu}{2} \frac{1}{\ln \frac{3}{2}} - 1 +$$

$$+ \mu > \frac{1}{2 \ln \frac{3}{2}} - 1 > \frac{1}{8} = \varepsilon.$$

Якщо $w > -\frac{1}{2\mu}$, то для $\mathcal{D} = \mathcal{D}_2$

$$L_{\mathcal{D}_2}(w) = (\mu \cdot w + 1)^2 > \frac{1}{4} > \frac{1}{8} = \varepsilon, \text{ а } \min_{w^*} L_{\mathcal{D}_2}(w^*) = 0 = L_{\mathcal{D}_2}\left(-\frac{1}{\mu}\right).$$

Піч., отримали протиріччя. $\downarrow$

— при $\mathcal{K} = [-1, 1]$ діємо так само для $z_1 = (\frac{1}{\mu}, 0)$ і $z_2 = (1, -1)$. Тенер ми не обидва отримуюмо для $w < -\frac{1}{2}$ і $w > -\frac{1}{2}$ sign. (де у другому випад. $L_{\mathcal{D}_2} = 0$ для $w^* = -1$). $\triangle$

Якщо не до опуклості та обмеженості додати лінійності або квадратності, то наврядісь все матиме місце (див. далі).

Рем. Якщо $\ell(\cdot, z)$ не є опуклим, можна спробувати їх обмежити згори опуклим — т.зв. сурогатною ф-цією втрач (англ. surrogate loss function). Піч. це повинна бути $\ell^s : \mathcal{K} \times \mathcal{Z} \rightarrow \mathbb{R}_+$ така, що

$$- \ell(w, z) \leq \ell^s(w, z) \quad \forall w, z.$$

$$- \ell^s(\cdot, z) \text{ опукла } \forall z.$$

Потім $\bar{L}_{\mathcal{D}}$, що обчислена за ℓ , не перевищуватиме

$L_{\mathcal{D}}^3$, що об'є. за l^3 (за монотонністю мат. оч.). Тому, якщо (напр., в умові півності)

$$L_{\mathcal{D}}^3(A(S)) \leq \min_{w \in \mathcal{H}} L_{\mathcal{D}}^3(w) + \epsilon$$

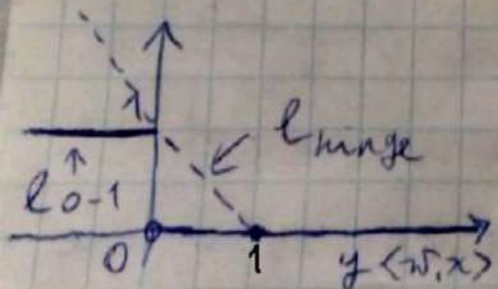
то $L_{\mathcal{D}}(A(S)) \leq \min_{w'} L_{\mathcal{D}}^3(w') + \epsilon$. Також що нерівність переписується у вигляді

$$L_{\mathcal{D}}(A(S)) \leq \min_{w'} L_{\mathcal{D}}(w') + \left(\min_{w'} L_{\mathcal{D}}^3(w') - \min_{w'} L_{\mathcal{D}}(w') \right) + \epsilon,$$

де поряг з початковою апроксимацією $\min_{w'} L_{\mathcal{D}}(w')$ і деяким ϵ (що тут обчислюється за допомогою l^3) викликає помилку оптимізації, що залежить від різниці між ϕ -цілою витрат.

Ек. 0-1 витрата l_{0-1} $(w, (x, y)) = \begin{cases} 1, & y \langle w, x \rangle < 0, \\ 0, & y \langle w, x \rangle \geq 0 \end{cases}$

(де y задані для клас. $Y = \{\pm 1\}$) не є опуклою (див. рис. для



$d=1$, але обмежена зверху петлевою витратою

(hinge loss) $l_{\text{hinge}}(w, (x, y)) = \max\{0, 1 - y \langle w, x \rangle\}$

(див. рис.), що є опуклою за Рн. Вище як max опуклих.

Застосованна опуклості: регуляризація

Пит ми продовжемо розглядати задачу ML на $\mathcal{Z}$ з класом гіпотез $\mathcal{H} \subset \mathbb{R}^d$.

def. Алгоритм ML здійснює регуляризовану мінімізацію втрат (regularized loss minimization, RLM) на класі $\mathcal{H}$ з функцією регуляризації $R: \mathbb{R}^d \rightarrow \mathbb{R}$, якщо за вибіркою S він повертає гіпотезу з найменшою сумою L_S і R :

$$A(S) = h_S \in \argmin_{w \in \mathcal{H}} (L_S(w) + R(w)).$$

Рем. Пит застосовуємо той же підхід, що у SRM (цей клас можна розм. як узгальнена SRM для такої задачі): виходить переважність в $\mathcal{H}$ гіпотез з меншою $R(w)$, що інтуїтивно збільшується зі "складністю" гіпотези; тоді RLM-алгоритм може обрати гіпотезу w з меншим $R(w)$ навіть якщо

у неї сильний емп. ризик $L_S(w)$. Найпростіший
 видір такої міри складності - гевісала, подібно
 евал. норма, вектора w . Тоді $\mathcal{H} = \mathbb{R}^d$ можна
 представити (ан-по до SRM) як $\bigcup_{i=1}^{\infty} \mathcal{H}_i$, де $\mathcal{H}_i =$
 $= \mathcal{D}_i(0) = \{w \mid |w| \leq i\}$. Більш точно,

def. Пеналізаційною функцією регуляризації зветься
 $R: \mathbb{R}^d \rightarrow \mathbb{R}: w \mapsto \lambda |w|^2$, де $\lambda > 0$. Задача лінійної
 регресії на $\mathbb{R}^d$ з квадратичного φ -гією втрам і
 такою R , точніше, $\forall$ алгоритм, що згідноє з цією
 задачею RLM з такою R , зветься хребтовою регресією
 (ridge regression): для $S = \{(x_i, y_i)\}_{i=1}^m \in (\mathcal{X} \times \mathcal{Y})^m \subset \mathbb{R}^{(d+1)m}$

$$h_S \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \left(\frac{1}{m} \sum_{i=1}^m \frac{1}{2} (\langle w, x_i \rangle - y_i)^2 + \lambda |w|^2 \right) \quad (**)$$

Далі ми покажемо, що такий алгоритм забезпечує
 адекватну PAC навчальність, якщо $\mathcal{X} = \mathcal{D}_R(0) = \{x \mid |x| \leq R\}$,
 $\mathcal{Y}$ - $\mathbb{R}$ -визірзок, а $\mathcal{H} = \mathcal{D}_B(0) = \{w \mid |w| \leq R\}$ (у цьому випадку

ця задача не тільки обмежена, опукла і ^{як показано вище,} гладка, але і лінійцева - Впр.). Сам алгоритм тут аналогічний до звичайного методу найменших квадратів: щоб мінімізувати цю ф-цію^(**), знайдемо градієнт за w і прирівняємо до 0:

$$2\lambda w + \frac{1}{m} \sum_{i=1}^m (\langle w, x_i \rangle - y_i) x_i = 0$$

Ан-но до методу найм. квадр. позначаючи (для векторів-стовпчиків x_i, y, w) м-цю $A := \sum_{i=1}^m x_i x_i^T$ і вектор $b := \sum_{i=1}^m y_i x_i$, маємо

$$2\lambda w + \frac{1}{m} (A w - b) = 0,$$

$$(2\lambda m I + A) w = b,$$

де I - одинична m -ця. Оск. m -ця A невід'ємно знако-
визначена (тобто усі її вл. значення ≥ 0 , див. вище до-
слідження у методі найм. квадратів), m -ця $2\lambda m I + A$
тодісно визначена (усі її вл. зн. $\geq 2\lambda m > 0$), отже
(Впр.)

невироджена : тут завжди можна знайти розв'язок як $w = (2\lambda m I + A)^{-1} b$.

Впр. Показати, що цей розв'язок завжди буде локальним мінімумом (зокр. при $\lambda=0$, тобто для "звич." пайн. квадратів) - вивести з опуклості.
причому якщо A невироджена

Стійкість

Лем. PAC навчаєсть алгоритма, що описаний вище, виконуватиме з дією загального поняття (властивості) стійкості : невелика зміна у S дає невелику зміну у $A(S)$.
Дано відповідні точки означення.

Рем. Лема $S = \{z_i\}_{i=1}^m$ - навч. вибірка, $i, z' \in \mathbb{Z}$. $\forall i = \overline{1, m}$
позн. $S^{(i)} := \{z_1, \dots, z_{i-1}, z', z_{i+1}, \dots, z_m\}$. Тоді для "гарного" алг. A $\ell(A(S^{(i)}), z_i) > \ell(A(S), z_i)$, бо вибірка у першому випадку не включає z_i . Але якщо різниця

мін цими витратами надто велика, це означає, що A схильний до перенавчання і не є стійким: додавання z до вибірки суттєво змінює поведінку на новому.

Лем. У позн. як вище A зветься у середньому стійким при заміні одного аргумента (on-average-one-stable), якщо $\exists$ така ф-ція $\epsilon: \mathbb{N} \rightarrow \mathbb{R}$ така, що $\forall$ розподіл $\mathcal{D}$ на $\mathbb{Z}$

$$\mathbb{E}_{(S, z') \sim \mathcal{D}^{m+1}, i \sim U(m)} \left(\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \right) \leq \epsilon(m),$$

де $U(m) = \left(\frac{1}{m}, \dots, \frac{1}{m} \right)$ - однор. розподіл на $\{1, \dots, m\}$. Далі будемо говорити просто, що A стабільний (стійкий) з параметром (чате) ϵ .

Рез. Для $\forall m$, розподіл $\mathcal{D}$ на $\mathbb{Z}$, вибірки S з m ел-тів, що i.i.d. відн. до $\mathcal{D}$ і все одною z' , що має i.i.d. відн. до $\mathcal{D}$ (відн. S), $\forall$ алгоритма A

$$\mathbb{E}_{S \sim \mathcal{D}^m} \left(L_{\mathcal{D}}(A(S)) - L_S(A(S)) \right) = \mathbb{E}_{(S, z') \sim \mathcal{D}^{m+1}, i \sim U(m)} \left(\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \right).$$

Зокрема, це $\leq \varepsilon(m)$ для стійкого A , тобто цей алгоритм не схильний до переобчислення.

Діємо, за означеннями L_D і L_S :

$$L_D(A(S)) = \mathbb{E}_{z' \sim D} \ell(A(S), z') \Rightarrow \left[\begin{array}{l} \text{власн. } E, \\ z' \text{ незал. від } S \end{array} \right] \Rightarrow \mathbb{E}_{S \sim D^m} L_D(A(S)) =$$

$$= \mathbb{E}_{(S, z') \sim D^{m+1}} \ell(A(S), z') = \left[\begin{array}{l} \forall i (S^{(i)})_{z_i} \text{ max} \\ \text{само } \sim D^{m+1} \end{array} \right] = \mathbb{E}_{(S, z') \sim D^{m+1}} \ell(A(S^{(i)}) / z_i) =$$

$$= \sum_{(S, z') \sim D^{m+1}, i \sim U(m)} \ell(A(S^{(i)}) / z_i).$$

$$L_S(A(S)) = \mathbb{E}_{i \sim U(m)} \ell(A(S), z_i) \xrightarrow{\text{власн. } E} \mathbb{E}_{S \sim D^m} L_S(A(S)) = \mathbb{E}_{S \sim D^m, i \sim U(m)} \ell(A(S), z_i),$$

звідки маємо потрібне. $\triangle$

(нум не хі саме, що $(S, z') \sim D^{m+1}$ до z' не нале)

def. Ф-ція $f: C \rightarrow \mathbb{R}$, де $C \subset \mathbb{R}^d$ опукла, зветься λ -строю (сильно)

опуклою (λ -strongly convex) для $\lambda > 0$, якщо $\forall w, u \in C$,

$\alpha \in [0, 1]$

$$f(\alpha w + (1-\alpha)u) \leq \alpha f(w) + (1-\alpha)f(u) - \frac{\lambda}{2} \alpha(1-\alpha) |w-u|^2$$

Рем. Тобто ми додаємо додатковий ≤ 0 член у праву.

частину функції опуклості, підсилюючи її. Зокрема, $\forall$ опукла φ -функція $\in O$ -с.он.

Рл. 1. $f(w) = \lambda |w|^2 \in \mathbb{Z}\lambda$ -с.он.

2. Якщо f - λ -с.он. а g - он. то $f+g$ - λ -с.он.

3. Якщо f - λ -с.он. ^{диференційовна в u} , а u - м. екстремума f (зокрема, м. мінімуму), то $\forall w$ $f(w) - f(u) \geq \frac{\lambda}{2} |w-u|^2$.

Д. 1. $\forall w, u$ за власт. скал. добутку

$$\begin{aligned} \lambda |\alpha w + (1-\alpha)u|^2 &= \lambda (\alpha^2 |w|^2 + (1-\alpha)^2 |u|^2 + 2\alpha(1-\alpha) \langle w, u \rangle) = \\ &= \lambda (\alpha(\alpha-1+1) |w|^2 + (1-\alpha)(- \alpha + 1) |u|^2 + 2\alpha(1-\alpha) \langle w, u \rangle) = \\ &= \alpha \cdot \lambda |w|^2 + (1-\alpha) \cdot \lambda |u|^2 + \lambda \alpha(1-\alpha) (-|w|^2 - |u|^2 + 2 \langle w, u \rangle) = \\ &= \alpha \cdot \lambda |w|^2 + (1-\alpha) \cdot \lambda |u|^2 - \lambda \alpha(1-\alpha) |w-u|^2, \text{ маємо у } \underline{\text{def.}} \text{ маємо} \\ &\text{рівність (як керує } \mathbb{Z}\lambda \text{).} \end{aligned}$$

2. Діючи, $\forall w, u, \alpha$

$$\begin{aligned} (f+g)(\alpha w + (1-\alpha)u) &= f(\alpha w + (1-\alpha)u) + g(\alpha w + (1-\alpha)u) \leq [\underline{\text{def}}] \leq \\ &\leq \alpha f(w) + (1-\alpha) f(u) - \frac{\lambda}{2} \alpha(1-\alpha) |w-u|^2 + \alpha g(w) + (1-\alpha) g(u) = \end{aligned}$$

$$= \alpha(f+g)(w) + (1-\alpha)(f+g)(u) - \frac{\lambda}{2} \alpha(1-\alpha) |w-u|^2$$

3. Для $\alpha > 0$ з доб., позитивна на α , маємо

$$\frac{f(\alpha w + (1-\alpha)u) - f(u)}{\alpha} \leq f(w) - f(u) - \frac{\lambda}{2} (1-\alpha) |w-u|^2$$

Ліва іє $\frac{f(u + \alpha(w-u)) - f(u)}{\alpha} \xrightarrow{\alpha \rightarrow 0} g'(0)$, де $g(\alpha) = f(u + \alpha(w-u))$.

Оскільки u -м.екстр. f , 0 -м.екстр. $g \Rightarrow g'(0) = 0$. Справа не має $f(w) - f(u) - \frac{\lambda}{2} |w-u|^2$ при $\alpha \rightarrow 0$, що м.ч. > 0 , що й потрібне. Δ

Ром. Якщо задана ML функція, тодімо має функції $\ell(\cdot, z)$

$\forall z$, то й ємн. вираз $L_S(w) = \frac{1}{m} \sum_{i=1}^m \ell(w, z_i)$ (де $S = \{z_i\}_{i=1}^m$) задає функцію оп-ційо $\text{big } w$ за даним z набор.

Рн. Тоді за м.1. і 2. Рн. Внаслідок, $f_S(w) := L_S(w) + \frac{\lambda}{2} |w|^2$

2.2-с.он. $\forall \lambda > 0$. $\forall$ RLM-адзоримна A його рез-т по-
том $A(S)$ - м. (модальною за доб. RLM) мінімум

f_S , тому за н.3. $\forall v$

$$f_S(v) - f_S(A(S)) \geq \frac{\lambda}{2} |v - A(S)|^2$$

Ми будемо перевіряти стійкість A , тому нехай z' - год.
 приклад i $S^{(i)} = (z_1, \dots, z_i', \dots, z_m)$ $\forall i = \overline{1, m}$ як раніше. За
 дов. L_S , маємо $\forall v, u, i$

$$f_S(v) - f_S(u) = L_S(v) + \lambda |v|^2 - (L_S(u) + \lambda |u|^2) = \underbrace{L_{S^{(i)}}(v) + \lambda |v|^2 - (L_{S^{(i)}}(u) + \lambda |u|^2)}_K + \underbrace{\ell(v, z_i) - \ell(u, z_i)}_m + \underbrace{\ell(u, z_i') - \ell(v, z_i')}_m.$$

Подаємо тут $v = A(S^{(i)})$, $u = A(S)$. Оскільки v - м. мінімуму
 $L_{S^{(i)}}(\cdot) + \lambda |\cdot|^2$, то нерівності перші чотири члени правої частини
 в сумі ≤ 0 , отже

$$f_S(A(S^{(i)})) - f_S(A(S)) \leq \frac{1}{m} (\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)) + \frac{1}{m} (\ell(A(S^{(i)}), z_i') - \ell(A(S), z_i'))$$

, тому маємо

$$\lambda |A(S^{(i)}) - A(S)|^2 \leq \frac{1}{m} (\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)) + \frac{1}{m} (\ell(A(S^{(i)}), z_i') - \ell(A(S), z_i')) \quad (***)$$

Якщо $\forall z$ $\ell(\cdot, z)$ ρ -ліній. (крім того, що опукла), то в правій
 частині (***) обидва доданки $\leq \frac{\rho}{m} |A(S^{(i)}) - A(S)|$:

$$\lambda |A(S^{(i)}) - A(S)|^2 \leq \frac{2\rho}{m} |A(S^{(i)}) - A(S)|$$

Т.ч., з лінійності $\ell(\cdot, z_i)$ маємо $\forall S, z', i$

$$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \leq \beta |A(S^{(i)}) - A(S)| \leq \frac{2\beta^2}{\lambda m},$$

а отже і мат. очікування за $(S, z') \sim \mathcal{D}^{m+1}, i \sim U(m) \leq \frac{2\beta^2}{\lambda m}$:

Сок. 1. Якщо задача ML опукла і β -ліпнищева, то $\forall$ RLM-алгоритм A з φ -цією регуляризації $\lambda |w|^2$ ($\forall \lambda > 0$) є стійким з параметром $\varepsilon(m) = \frac{2\beta^2}{\lambda m}$. Зокрема, $\forall m, \mathcal{D}$ та $\mathcal{Z}$

$$\mathbb{E}_{S \sim \mathcal{D}^m} (L_{\mathcal{D}}(A(S)) - L_S(A(S))) \leq \frac{2\beta^2}{\lambda m}$$

(тут остання частина була з Рн. 6 вище).