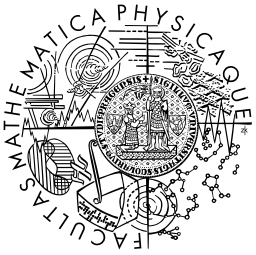


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Ondřej Komora

Řídká řešení v optimalizačních úlohách klasifikace

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Martin Branda, Ph.D.

Studijní program: Obecná matematika

Praha 2022

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Děkuji panu doc. RNDr. Martinovi Brandovi, Ph.D., za jeho vedení a za jeho cenné připomínky při vypracovávání této práce.

Dále bych chtěl poděkovat svým rodičům, kteří mě v průběhu studia neustále podporovali.

Název práce: Řídká řešení v optimalizačních úlohách klasifikace

Autor: Ondřej Komora

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Martin Branda, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Hlavním cílem této práce je podat ucelený popis metod proximálního a obyčejného stochastického subgradientového sestupu, které se používají při hledání řídkých řešení v optimalizačních úlohách klasifikace. Zavedeme a interpretujeme pojmy vedoucí k definici těchto metod a podrobně diskutujeme předpoklady, za nichž dokážeme jejich konvergenci ke kritickému bodu. Na závěr v numerické ukázce na konkrétní úloze demonstrujeme, jak jsme za pomoci těchto metod a vhodné volby tzv. regularizace ovlivnili řídkost řešení této úlohy.

Klíčová slova: optimalizace, řídkost, subgradient, proximální operátor

Title: Sparse solutions in labeling optimization problems

Author: Ondřej Komora

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Martin Branda, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The main goal of this work is to give self-contained description of proximal and ordinary stochastic subgradient descent methods, which are used in finding sparse solutions of labeling optimization problems. We will define and interpret necessary concepts leading to the definition of those methods and we will discuss in detail conditions, under which we show convergence of these methods to critical point. At the end, we will present a numerical experiment on concrete optimization task where we demonstrate use of these methods. In this experiment we will also show how a suitable choice of so called regularization can influence sparsity of solution of this particular task.

Keywords: optimization, sparsity, subgradient, proximal operator

Obsah

Úvod

Při úlohách klasifikace chceme roztřídit pozorování do předem známých kategorií. Často přitom pracujeme s rozsáhlými datasety s mnoha vysvětlujícími proměnnými. Některé z těchto proměnných však mohou mít malý či zanedbatelný vliv na to, do jaké kategorie dané pozorování spadá. Přesto však tyto proměnné mohou mít negativní vliv na náš model, do kterého vnášejí zbytečnou chybu a zhoršují jeho predikční schopnosti. Cílem hledání řídkého řešení je odstranit tyto málo významné proměnné z našeho modelu a docílit tím toho, že náš model bude více stabilní při predikci kategorií.

Řešení klasifikačních úloh většinou probíhá ve dvou fázích. V první se optimalizuje model, který je v našem kontextu vyjádřen *ztrátovou funkcí*. Tato ztrátová funkce je závislá na parametrech, které optimalizujeme tak, aby hodnota ztrátové funkce byla co nejmenší. Ztrátová funkce je často interpretována tak, že odhaduje pravděpodobnosti příslušnosti pozorování do jednotlivých kategorií. Ve druhé fázi dochází k zařazování pozorování do kategorií na základě pravděpodobností spočtených v první fázi. Rozdělení na tyto fáze nám umožňuje převést klasifikační úlohu z diskrétní optimalizace na spojitou, která se zpravidla snadněji provádí, obzvlášť v kontextu strojového učení.

Ve strojovém učení existuje několik způsobů jak ovlivnit řídkost řešení klasifikační úlohy. Používají se různé metody pro redukci dimenzionality (např. analýza hlavních faktorů pomocí singulárního rozkladu), iterovanou eliminaci proměnných, hodnoty SHAP, apod. V této práci se zaměříme na hledání řídkého řešení pomocí tzv. *regularizace*. Regularizace je funkce, která se přičítá ke ztrátové funkci a má specifické vlastnosti ovlivňující predikční schopnosti ztrátové funkce. O tom, jak volba regularizace může ovlivnit řídkost řešení, budeme mluvit v sekci 5.1.

Uvažujme však nejprve ztrátovou funkci $f\colon \mathbb{R}^n \to \mathbb{R}$ a řešme úlohu

$$\min_{x\in\mathbb{R}^n} f(x).$$

Klasickou metodou řešení těchto úloh je metoda největšího spádu. Pokud je funkce f diferencovatelná, pak lze dosáhnout kritického bodu, tj. bodu kde je gradient nulový, iterací

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),\ k \in \mathbb{N}, \tag{1}$$

pro vhodně zvolenou posloupnost reálných čísel $\{\alpha_k\}$ a startovací bod $x_1 \in \mathbb{R}^n$. Iteraci provádíme dokud je gradient $\nabla f(x_k)$ nenulový. Tento přístup má však dvě nevýhody. Jednak určit gradient ztrátové funkce může být velmi náročné, a také ztrátová funkce nemusí být diferencovatelná. První případ vyvstává z faktu, že ztrátové funkce jsou často definovány pomocí pozorování z rozsáhlých datasetů, které mohou obsahovat desítky tisíc a více pozorování. Určit gradient ztrátové funkce pak vyžaduje průchod celého datasetu, což je výpočetně náročné. Druhý případ nastává poměrně často. V neuronových sítích se běžně používají nediferencovatelné aktivace, např. ReLU či Leaky ReLU. Dále i některé běžně používané ztrátové funkce jsou nediferencovatelné, např. takzvaný *hinge loss* jenž je základem modelu *soft-margin SVM*[1], či *mean absolute error*[2].

[1] v angličtině zkratka SVM znamená *support vector machine*

[2] v češtině *střední absolutní chyba*

Problém složitosti výpočtu gradientu se dá obejít tím, že gradient budeme v jistém smyslu odhadovat. Tak vznikl algoritmus *stochastického gradientového sestupu*[3]. Předpis toho algoritmu je podobný jako v (1), jen místo přesného gradientu používáme jeho odhad.

V případě, že ztrátová funkce je nediferencovatelná, gradient nahradíme *subgradientem*. Subgradient v této práci formálně definujeme a ukážeme některé jeho vlastnosti a proč je důležitý v optimalizačních úlohách.

Mějme ztrátovou funkci $f\colon \mathbb{R}^n \to \mathbb{R}$ a regularizaci $g\colon \mathbb{R}^n \to \mathbb{R}$. Hlavní náplní této práce bude diskutovat některé optimalizační metody řešící úlohu

$$\min_{x\in\mathbb{R}^n} f(x) + g(x), \tag{2}$$

se zaměřením na řídkost řešení této úlohy. Řídkost řešení formálně definujeme jako podíl nulových parametrů ku celkovému počtu. V této práci navíc budeme uvažovat, že máme oba výše zmíněné problémy spojené s metodou gradientového sestupu. Tedy funkce f a g jsou nediferencovatelné a je těžké spočítat jejich gradient, resp. subgradient. V práci se omezíme na funkce, které jsou tzv. *lokálně lipschitzovské*.

Hlavní metody řešící úlohu (2), které budeme v této práci diskutovat, jsou metody obyčejného a proximálního stochastického subgradientového sestupu[4]. Obě metody využívají subgradient a v jistém smyslu jeho odhad.

Obsah práce je strukturován následovně. V první kapitole této práce zavedeme základní pojmy a uvedeme některé věty, které dají do souvislosti tyto pojmy a jejich význam pro optimalizaci.

V druhé kapitole se zaměříme na tzv. *diskrétní aproximace trajektorie*. Jedná se o specifický typ algoritmu, který asymptoticky sleduje *trajektorii diferenciální inkluze*. Tyto pojmy čtenáři objasníme.

Obsah třetí kapitoly je zaměřen na samotné metody SSGD a PSSGD. Nejdříve je zavedeme a formulujeme podmínky konvergence k tzv. *kritickému bodu*. Poté ukážeme souvislost těchto algoritmů s kapitolou 2, neboli jak souvisí s diskrétními aproximacemi trajektorie. Na závěr provedeme důkaz jejich konvergence ke kritickému bodu.

Čtvrtá kapitola je věnovaná interpretaci podmínek konvergence algoritmů SSGD a PSSGD, neboť tyto podmínky nejsou přímočaré a vyžadují podrobnější komentář.

Pátá kapitola obsahuje numerickou ukázku. Popíšeme zde detaily této ukázky, včetně volby datasetů, ztrátové funkce a regularizace. Vysvětlíme, jak regularizace dokáže ovlivnit řídkost řešení. Ukázku provedeme použitím obou metod SSGD a PSSGD a porovnáme model ztrátové funkce a regularizace oproti samotné ztrátové funkci.

Na konci shrneme výsledky této práce a také přiložíme přílohu s důkazy některých použitých tvrzení.

[3] v angličtině *stochastic gradient descent*, odtud budeme používat zkratku SGD

[4] v anglické literatuře se používají názvy *stochastic subgradient descent* a *proximal stochastic subgradient descent*, proto tyto metody budeme zkracovat jako SSGD, resp. PSSGD

Značení

V celé práci budeme pracovat s euklidovskou normou $\|\cdot\|$ definovanou pomocí standardního skalárního součinu, který budeme značit $\langle\cdot,\cdot\rangle$. Otevřenou kouli okolo bodu $x \in \mathbb{R}^n$ o poloměru $r > 0$ označíme $B_r(x)$. Definujeme $\operatorname{dist}(x, A) = \inf\{\|x - a\| : a \in A\}$, kde $x \in \mathbb{R}^n$ a $A \subseteq \mathbb{R}^n$ je neprázdná množina. Pokud bude z kontextu zřejmé o jaký index se jedná, budeme někdy zapisovat posloupnost jako $\{x_k\}$ místo $\{x_k\}_{k=1}^{\infty}$.

V práci často budeme pracovat se součty množin. Ten definujeme v klasickém Minkowského smyslu: pro $A, B \subseteq \mathbb{R}^n$ definujeme jejich součet jako

$$A + B := \{a + b \colon a \in A, b \in B\}.$$

Operací $\operatorname{conv} A$ myslíme konvexní obal množiny A. Uzávěr množiny A značíme $\overline{A}$. Symbolem $\mathcal{P}(A)$ značíme potenční množinu množiny A.

Pokud nebude řečeno jinak, funkce f a g budou vždy zobrazení z $\mathbb{R}^n$ do $\mathbb{R}$, kde $n \in \mathbb{N}$. Zápisem $H \colon \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$ budeme definovat množinová zobrazení, tj. zobrazení jejichž hodnoty jsou množiny. Někdy však budeme množinovým zobrazením říkat pouze zobrazení, z kontextu však bude vždy jasné, o jaký typ zobrazení se jedná. Nechť $x \in \mathbb{R}^n$, inverz množinového zobrazení definujeme jako

$$H^{-1}(x) := \{y \in \mathbb{R}^n \colon x \in H(y)\}.$$

Součet množinových zobrazení $H, G \colon \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$ definujeme pro libovolné $x \in \mathbb{R}^n$ jako $(H + G)(x) := H(x) + G(x)$, kde $+$ je v tomto případě Minkowského součet množin.

Nechť $\mathbb{R}_+$ značí kladná reálná čísla. Křivku $\varphi \colon \mathbb{R}_+ \to \mathbb{R}^n$ nazvěme absolutně spojitou jestliže je její derivace lebesgueovsky integrovatelná a pro skoro všechna $t \geq 0$ platí

$$\varphi(t) = \varphi(0) + \int_0^t \dot{\varphi}(\tau)\mathrm{d}\tau,$$

kde $\dot{\varphi}$ značí derivaci podle proměnné t.

Prostor k-krát diferencovatelných funkcí mezi prostory X a Y budeme značit $\mathcal{C}^k(X, Y)$, přičemž prostor spojitých funkcí budeme značit pouze $\mathcal{C}(X, Y)$.

1. Základní pojmy

Uvažujme ztrátovou funkci $f\colon \mathbb{R}^n \to \mathbb{R}$ a optimalizační úlohu

$$\min_{x\in\mathbb{R}^n} f(x).$$

V této práci budeme často uvažovat takzvaně *lokálně lipschitzovské* funkce, tzn. pro každé $x \in \mathbb{R}^n$ existuje $\varepsilon > 0$ a konstanta $K > 0$ takové, že pro každé $y, z \in B_\varepsilon(x)$ platí

$$|f(y) - f(z)| \leq K \|y - z\|. \tag{1.1}$$

Pokud řekneme, že funkce f je v bodě x lokálně lipschitzovská s konstantou $K > 0$, máme tím na mysli to, že existuje $\eta > 0$ takové, že pro každé $y, z \in B_\eta(x)$ platí (1.1).

Jak jsme již předeslali v úvodní kapitole, v této práci se budeme zabývat obecně nediferencovatelnými funkcemi. Přistoupíme tak k jisté aproximaci gradientu. Aproximace, se kterou budeme pracovat, se nazývá *subgradient.*

Definice 1.1 (Clarkův subdiferenciál a subgradient)**.** *Clarkův subdiferenciál funkce f v bodě $x \in \mathbb{R}^n$ definujeme jako množinu*

$$\partial f(x) := \operatorname{conv}\left\{\lim_{n\to\infty} \nabla f(y_n)\colon y_n \to x, y_n \in \mathbb{R}^n, \nabla f(y_n) \text{ existuje } \forall n \in \mathbb{N}\right\}.$$

Každý prvek této množiny nazýváme subgradient.

Jedná se o poměrně komplikovaný pojem, který uvažuje limitu gradientů ze všech možných směrů vedoucích k bodu x. V případě, že f je konvexní, Clarkův subdiferenciál přechází přirozeně v pojem subdiferenciálu z konvexní analýzy. Tedy v případě konvexní funkce f je $\partial f(x)$ množina vektorů v splňující

$$f(y) - f(x) \geq \langle v, y - x\rangle,\ \forall y \in B_\varepsilon(x),$$

pro nějaké $\varepsilon > 0$. Chování subgradientových metod je dobře známé pro konvexní funkce, a proto se v této práci zabýváme obecnějším případem.

Důležité je si uvědomit, že pokud má funkce f v bodě x gradient, pak zřejmě platí $\partial f(x) = \{\nabla f(x)\}$ a Clarkův subdiferenciál se redukuje na prostý gradient.

Na první pohled není zcela zřejmé, jaké má Clarkův subdiferenciál vlastnosti. Víme zatím jenom to, že je to konvexní množina. V následujících tvrzeních si ukážeme nejdůležitější vlastnosti. Důkaz následujícího tvrzení čtenář najde v Rockafellar and Wets [2009, Tvrzení 8.7].

Tvrzení 1.2. *Nechť f je funkce a $x \in \mathbb{R}^n$. Potom $\partial f(x)$ je uzavřená množina, která je tzv. z vnějšku spojitá*[1]*. To znamená, že pro každé dvě posloupnosti $\{x_k\}$ a $\{y_k\}$ takové, že $x_k \to x^*$ a $y_k \in \partial f(x_k)$, platí* $\operatorname{dist}(y_k, \partial f(x^*)) \to 0$.

Nyní se podívejme na vlastnosti Clarkova subdiferenciálu v případě lokálně lipschitzovských funkcí. Ukazuje se, že Clarkův subdiferenciál je v tomto případě kompaktní a neprázdná množina.

[1] v anglické literatuře se používá pojem *outer continuous*

Tvrzení 1.3. *Nechť f je v bodě $x \in \mathbb{R}^n$ lokálně lipschitzovská s konstantou $K > 0$. Potom*

$$\partial f(x) \neq \emptyset \text{ a } \partial f(x) \subseteq \overline{B_K(0)}.$$

Je-li navíc g další lokálně lipschitzovská funkce v bodě x, pak

$$\partial(f+g)(x) \subseteq \partial f(x) + \partial g(x).$$

Důkaz. Viz příloha A, tvrzení A.1.

□

Subdiferenciál lokálně lipschitzovské funkce je tedy vždy omezená množina. V kombinaci s jeho uzavřeností dostáváme, že je i kompaktní. Dále budeme Clarkův subdiferenciál nazývat pouze subdiferenciál.

Význam subdiferenciálu pro optimalizaci je ihned zřejmý z následující věty, která byla dokázana např. v Clarke et al. [2008, Věta 1.5]. Tato věta říká za jakých podmínek funkce f klesá ve směru subgradientu.

Věta 1.4 (O sestupu ve směru subgradientu)**.** *Nechť $0 \notin \partial f(x)$ a dále nechť $d \in \partial f(x)$ je prvek s minimální normou. Potom pro dostatečně malé $\alpha > 0$ platí*

$$f(x - \alpha d) < f(x).$$

Zaveďme si nyní pojem tzv. *kritického bodu*. Tento pojem je analogický pojmu kritického bodu z matematické analýzy a v případě diferencovatelných funkcí v něj přechází. My ho však budeme definovat v obecnější formě pro množinová zobrazení.

Definice 1.5 (Kritický bod)**.** *Mějme množinové zobrazení $L\colon \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$. Řekneme, že $x^* \in \mathbb{R}^n$ je kritickým bodem zobrazení L, jestliže platí $0 \in L(x^*)$. Dále řekneme, že x^* je kritickým bodem funkce f, pokud je x^* kritickým bodem zobrazení ∂f, tj. $0 \in \partial f(x^*)$.*

Poznámka. Zobrazení L v předchozí větě může být například subdiferenciál nějaké funkce f. Nejvíce zajímavý případ tak pro nás bude množinové zobrazení ∂f.

Význam kritického bodu spočívá v následující větě.

Věta 1.6 (Optimalita a kritický bod)**.** *Pokud má funkce f v bodě $x^* \in \mathbb{R}^n$ lokální minimum, pak musí být x^* kritickým bodem funkce f, neboli $0 \in \partial f(x^*)$.*

Důkaz. Viz Rockafellar and Wets [2009, Věta 10.1.].

□

Opačná implikace v předchozí větě zjevně neplatí. Problém, že funkce v kritickém bodě nemusí mít lokální minimum, je známým problém pro optimalizační metody využívající gradient či subgradient. Nelze se mu vyhnout a nevyhneme se mu ani my. Přesto se gradientové a subgradientové metody osvědčily v mnoha oblastech využívající teorii optimalizace, např. ve strojovém a hlubokém učení. Proto má smysl se těmto metodám věnovat, i přes jejich nedostatky.

V úvodní kapitole jsme mluvili o tom, že spočítat přesně gradient, či subgradient ztrátové funkce může být velmi náročné. Proto směřujeme k tomu, že budeme subgradient odhadovat. Pro tento účel předpokládáme existenci vhodného výběrového mechanismu.

Předpoklad 1.7 (Odhadovací mechanismus pro subgradient)**.** *Mějme pravděpodobnostní prostor* $(\Omega, \mathcal{A}, P)$*. Pro funkci* f *předpokládáme, že existuje měřitelné zobrazení* $\zeta : \mathbb{R}^n \times \Omega \to \mathbb{R}^n$ *takové, že pro každé* $x \in \mathbb{R}^n$ *platí*

$$\mathbb{E}_\omega \left[\zeta(x, \omega)\right] \in \partial f(x).$$

Operace $\mathbb{E}_\omega$ *značí střední hodnotu v proměnné* $\omega \in \Omega$*.*

V prostředí strojového učení a této práce nám takto definovaný pravděpodobnostní prostor slouží k náhodnému výběru z dat a odhadu subgradientu ztrátové funkce. Výše zmíněný předpoklad nám pak zaručuje, že střední hodnota tohoto odhadu je skutečný subgradient funkce.

Kdykoliv budeme pracovat s tímto odhadovacím mechanismem, budeme používat pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$ a budeme mít na paměti, že je to pravděpodobnostní prostor, který nám umožnil zavést tento mechanismus.

Nyní jsme probrali důležité pojmy k optimalizaci nekonvexních a nediferencovatelných funkcí a jsme připraveni pracovat s algoritmy využívající subgradient funkce. V další kapitole však budeme diskutovat tzv. *diskrétní aproximace trajektorie*. Smysl tohoto názvu čtenáři objasníme větou 2.4 o aproximaci trajektorie. V kapitole 3 ukážeme, jak spolu souvisí tyto diskrétní aproximace a metody SSGD a PSSGD.

2. Diskrétní aproximace trajektorie

Mějme množinové zobrazení $H\colon \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$. Dále nechť $\{\eta_k\}$ je posloupnost prvků prostoru $\mathbb{R}^n$ a $\{\alpha_k\}$ posloupnost kladných reálných čísel. Zvolme $x_1 \in \mathbb{R}^n$. Pro každé $k \in \mathbb{N}$ budeme iterovat

$$x_{k+1} = x_k + \alpha_k(y_k + \eta_k), \tag{2.1}$$

kde $y_k \in H(x_k), k \in \mathbb{N}$, jsou libovolně zvolené. Takto zavedené značení budeme používat po celou kapitolu. Poznamenejme, že tato iterace je deterministická a nepracujeme zde s náhodou.

Za zobrazení H můžeme například zvolit subdiferenciál nějaké funkce f, tedy $H = \partial f$. Prvky y_k pak lze interpretovat jako subgradient této funkce v bodě x_k. Součet $y_k + \eta_k$ zase můžeme chápat jako subgradient v bodě x_k, přičemž jsme při určování toho subgradientu udělali chybu η_k. V této kapitole se však zaměříme na obecnější případ kdy H je libovolné množinové zobrazení.

Zaveďme si nyní podmínky za kterých platí konvergence posloupnosti $\{x_k\}$, vygenerované předpisem (2.1), ke kritickému bodu. Podmínky jsou převzaté z Davis et al. [2020], ovšem my je zde z důvodu přehlednosti uvádíme a přikládáme k nim podrobnější komentář. Předpoklady budeme používat po celý zbytek kapitoly.

Předpoklad 2.1. *Předpokládejme, že je pro iteraci* (2.1) *splněno následující:*

1. $\sup_{k\in\mathbb{N}} \|x_k\| < \infty$ *a* $\sup_{k\in\mathbb{N}} \|y_k\| < \infty$.
2. *Posloupnost* $\{\alpha_k\}$ *má vlastnosti*

$$\alpha_k > 0,\ \sum_{k=1}^{\infty} \alpha_k = \infty,\ \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

3. *Řada* $\sum_{k=1}^{\infty} \alpha_k \eta_k$ *konverguje k nějakému vektoru* $v \in \mathbb{R}^n$.
4. *Pro každou rostoucí posloupnost přirozených čísel* $\{n_k\}_{k=1}^{\infty}$ *takovou, že* $\{x_{n_k}\}_{k=1}^{\infty}$ *konverguje k nějakému bodu* $x^* \in \mathbb{R}^n$, *platí*

$$\lim_{n\to\infty} \operatorname{dist}\left(\frac{1}{n}\sum_{k=1}^{n} y_{n_k}, H(x^*)\right) = 0.$$

Předpoklad 2.1.1 asi nikoho nepřekvapí. Pokud jsou x_k či y_k neomezené, tak vzhledem k předpisu iterace (2.1) lze stěží očekávat konvergenci hodnot x_k.

Vlastnosti iteračních kroků uvedené v bodě 2.1.2 jsou standardní v literatuře pojednávající o stochastickém gradientovém sestupu (SGD). Robbins and Monro [1951] ukázali, že tyto vlastnosti iteračních kroků jsou nutné ke konvergenci metody SGD ke kritickému bodu. Vzhledem k tomu, že SSGD ve speciálním případě přechází v SGD, je rozumné vlastnosti iteračních kroků požadovat i po iteraci (2.1). Jedná se o mírný předpoklad a lze jej jednoduše splnit například volbou kroků $\alpha_k = 1/k$.

Bod 2.1.3 pojednává o omezenosti chyb vážených kroky α_k. Intuitivně bychom očekávali, že pro konvergenci algoritmu nesmí být chyby příliš velké. To nám právě zaručí tento předpoklad.

Předpoklad 2.1.4 nám říká, že aritmetický průměr $\frac{1}{n}\sum_{k=1}^{n} y_{n_k}$ dobře aproximuje množinu $H(x^*)$ pro velká n. Tento předpoklad trochu připomíná spojitost z vnějšku zmíněnou v tvrzení 1.2, jen zde uvažujeme, že se aritmetický průměr prvků $y_{n_k} \in H(x_{n_k}), k \in \mathbb{N}$, blíží k $H(x^*)$, a ne samotná y_{n_k}. Vskutku, z tvrzení 3.6 vyplyne, že pokud jsou hodnoty zobrazení H kompaktní, konvexní a z vnějšku spojité množiny, pak je splněn předpoklad 2.1.4.

Zbývá zavést ještě jednu sadu předpokladů. Budeme předpokládat existenci funkce splňující určité podmínky. Těmto podmínkám budeme říkat Ljapunovy.

Předpoklad 2.2. *Předpokládejme, že existuje spojitá, zdola omezená funkce* $L : \mathbb{R}^n \to \mathbb{R}$ *taková, že*

1. *(Sardova vlastnost) Množina* $L(\mathbb{R}^n \setminus H^{-1}(0))$ *je hustá v* $\mathbb{R}$.
2. *(Sestupová vlastnost) Pro každou absolutně spojitou křivku* $\varphi\colon \mathbb{R}_+ \to \mathbb{R}^n$ *takovou, že* $\varphi(0)$ *není kritickým bodem* H *a*

$$\dot{\varphi}(t) \in H(\varphi(t)) \text{ pro s.v. } t \in \mathbb{R}_+ \tag{2.2}$$

platí, že existuje $T > 0$ *takové, že*

$$L(\varphi(T)) < \sup_{t\in[0,T]} L(\varphi(t)) \leq L(\varphi(0)).$$

Poznámka. Vztah (2.2) nazýváme *diferenciální inkluze*. Absolutně spojitou křivku splňující vlastnost diferenciální inkluze (2.2) nazýváme *trajektorie zobrazení* H.

Sardova vlastnost nám říká, že zobrazíme-li nekritické body zobrazení H pomocí funkce L, dostaneme hustou množinu v $\mathbb{R}$.

Tyto podmínky můžou působit poměrně komplikovaně. Jejich interpretaci, včetně objasnění názvu Sardova podmínka a sestupová vlastnost, proto věnujeme celou kapitolu 4.

Vyslovme nyní větu o konvergenci posloupnosti $\{x_k\}$ ke kritickému bodu. Důkaz zde nebudeme uvádět, čtenáře odkážeme na Davis et al. [2020, Sekce 3.3], kde je důkaz přehledně a dobře sepsán.

Věta 2.3 (Konvergence ke kritickému bodu)**.** *Nechť jsou splněny předpoklady 2.1 a 2.2. Potom pro libovolný limitní bod* x^* *posloupnosti* $\{x_k\}$ *platí* $0 \in H(x^*)$ *a posloupnost* $\{L(x_k)\}$ *konverguje.*

2.1 Věta o aproximaci trajektorie

Následující sekci přebíráme z Davis et al. [2020] a Duchi and Ruan [2018]. Sekci zde uvádíme z ilustrativních a interpretačních důvodů.

Definujme $t_k = \sum_{i=1}^{k} \alpha_i$. Pro posloupnost $\{x_k\}$ vytvořenou iterací (2.1) definujme funkci $x\colon \mathbb{R}_+ \to \mathbb{R}^n$ jako $x(t) = x_1, t \in [0, t_1)$, a

$$x(t) = x_k + \frac{t - t_k}{t_{k+1} - t_k}(x_{k+1} - x_k),\ \forall t \in [t_k, t_{k+1}),\ \forall k \in \mathbb{N}.$$

Podmínka $\sum_{k=1}^{\infty} \alpha_k = \infty$ zaručuje, že tato funkce je definovaná pro každé $t \geq 0$. Funkce x není nic jiného než spojitá lineární interpolace posloupnosti $\{x_k\}$, platí totiž $x(t_k) = x_k$ a mezi body x_{k+1} a x_k je x lineární funkce. Dále definujme posunuté křivky $x^\tau(t) := x(\tau + t),\ t \in \mathbb{R}_+, \tau \geq 0$.

Následující věta je výsledkem Duchi and Ruan [2018, Věta 3.7].

Věta 2.4. *(O aproximaci trajektorie) Nechť je splněna sada předpokladů 2.1. Je-li* $\{\tau_k\}_{k=1}^{\infty}$ *taková, že* $\tau_k \to \infty$, *potom všechny limitní body* $\varphi(\cdot)$ *posloupnosti funkcí* $\{x^{\tau_k}(\cdot)\}_{k=1}^{\infty}$ *v prostoru* $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^n)$ *jsou trajektoriemi diferenciální inkluze*

$$\dot{\varphi}(t) \in H(\varphi(t)) \text{ pro s.v. } t \geq 0.$$

Konvergence je stejnoměrná v následujícím smyslu. Pokud je funkce $\varphi(\cdot)$ *limitou nějaké vybrané podposloupnosti* $\{x^{\tau_{k_i}}(\cdot)\}_{i=1}^{\infty}$, *potom*

$$\lim_{i \to \infty} \sup_{t \in [0,T]} \|x^{\tau_{k_i}}(t) - \varphi(t)\| = 0,\ \forall T > 0.$$

Věta o aproximaci trajektorie nám říká, že lineární interpolace $x(t)$ posloupnosti $\{x_k\}$ pro dostatečně velké t dobře aproximuje trajektorii zobrazení H. Tento fakt plyne ze zvětšujících se posunutí τ_k. Touto větou jsme zároveň obhájili název této kapitoly, neboť hodnoty x_k lze pro dostatečně velká $k \in \mathbb{N}$ pomocí této věty interpretovat jako diskrétní aproximace trajektorie zobrazení H. Tato věta je důležitým interpretačním nástrojem a v kapitole 4 se o ní zmíníme v souvislosti s metodami SSGD a PSSGD.

3. Optimalizační metody pro hledání řídkých řešení

V této kapitole se zaměříme na metody SSGD a PSSGD a jejich konvergenci ke kritickému bodu. Obě metody používají odhad subgradientu funkce, budeme tedy potřebovat výběrový mechanismus z předpokladu 1.7 i s pravděpodobnostním prostorem na kterém jsme ho zaváděli. Začneme s metodou SSGD.

3.1 Metoda SSGD

Mějme lokálně lipschitzovskou funkci f. Metoda SSGD je určena pro optimalizační úlohy typu

$$\min_{x\in\mathbb{R}^n} f(x).$$

Algoritmus SSGD pro tuto úlohu vypadá následovně.

Algoritmus 1: Metoda SSGD

Vstup: *Kladná posloupnost* $\{\alpha_k\}_{k=1}^{\infty}, x_1 \in \mathbb{R}^n$
Výstup: x_k *je kritickým bodem funkce* f
Začátek:
  $k \leftarrow 1$
  Opakuj:
    Vyber $\omega_k \in \Omega$ a spočítej $\zeta(x_k, \omega_k)$
    $x_{k+1} \leftarrow x_k - \alpha_k \zeta(x_k, \omega_k)$
    $k \leftarrow k + 1$
  Dokud: $0 \notin \partial f(x_k)$
  Vrať: x_k
Konec

Metoda připomíná algoritmus SGD, akorát zde nepředpokládáme diferencovatelnost funkce a pracujeme se subgradienty. Iteraci $x_{k+1} = x_k - \alpha_k \zeta(x_k, \omega_k)$ budeme nazývat *subgradientová aktualizace*. Vektor x_1 slouží jako startovací bod.

Diskutujme nyní předpoklady konvergence algoritmu SSGD ke kritickému bodu.

3.1.1 Předpoklady konvergence SSGD

Předpoklady jsou převzaty z Davis et al. [2020] a my je zde pro přehlednost uvádíme v upravené formě. Zavedené podmínky budeme používat při důkazu konvergence ke kritickému bodu v sekci 3.1.2.

Předpoklad 3.1. *Nechť je pro metodu SSGD definovanou algoritmem 1 splněno následující:*

1. *Posloupnost* $\{x_k\}$ *je omezená, neboli* $\sup_{k\in\mathbb{N}} \|x_k\| < \infty$.
2. *Posloupnost iteračních kroků* $\{\alpha_k\}$ *má vlastnosti*

$$\alpha_k > 0,\ \sum_{k=1}^{\infty} \alpha_k = \infty,\ \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

3. *Nechť pro výběrový mechanismus z předpokladu 1.7 platí, že existuje funkce* $b\colon \mathbb{R}^n \to \mathbb{R}$ *omezená na všech omezených množinách taková, že*

$$\mathbb{E}_\omega \left[\|\zeta(x,\omega)\|^2 \right] \leq b(x).$$

Body 3.1.1 a 3.1.2 známe již z kapitoly 2. Přidáváme však předpoklad 3.1.3, který zaručuje konečný rozptyl normy odhadu subgradientu. Platí totiž

$$\operatorname{var}_\omega \left(\|\zeta(x,\omega)\| \right) = \mathbb{E}_\omega \left[\|\zeta(x,\omega)\|^2 \right] - \left(\mathbb{E}_\omega \left[\|\zeta(x,\omega)\| \right] \right)^2 \leq b(x). \tag{3.1}$$

Podobně jako v kapitole 2 požadujeme existenci funkce splňující Ljapunovy podmínky. Zde však předpokládáme, že tyto podmínky splňuje samotná funkce f.

Předpoklad 3.2. *Nechť f je navíc zdola omezená funkce splňující*

1. *(Sardova vlastnost) Množina nekritických bodů funkce f je hustá v $\mathbb{R}^n$.*
2. *(Sestupová vlastnost) Pro každou trajektorii φ diferenciální inkluze*

$$\dot{\varphi}(t) \in -\partial f(\varphi(t)) \text{ pro s.v. } t \in \mathbb{R}_+$$

splňující $0 \notin \partial f(\varphi(0))$ platí, že existuje $T > 0$ s vlastností

$$f(\varphi(T)) < \sup_{t \in [0,T]} f(\varphi(t)) \leq f(\varphi(0)).$$

Poznámka. Předpoklad omezenosti zdola funkce f není příliš omezující. V praxi se téměř výhradně pracuje se zdola omezenými ztrátovými funkcemi.

O funkci f splňující obě podmínky 3.2.1 a 3.2.2 budeme říkat, že splňuje Ljapunovy podmínky.

Za těchto předpokladů můžeme dokázat konvergenci metody SSGD ke kritickému bodu.

3.1.2 Důkaz konvergence SSGD

Struktura důkazu je podobná jako v článku Davis et al. [2020, Kapitola 4], ovšem my ho zde upravujeme pro účely použití odhadového mechanismu z předpokladu 1.7. Navíc přikládáme důkazy některých tvrzení, které považujeme za vlastní. Zužitkujeme přitom předpoklady zavedené v předchozí sekci.

Nechť $y_k := -\mathbb{E}_\omega \left[\zeta(x_k,\omega) \right]$ a $\eta_k := \mathbb{E}_\omega \left[\zeta(x_k,\omega) \right] - \zeta(x_k,\omega_k), \omega_k \in \Omega$. Vektory η_k lze chápat jako náhodné veličiny závisející na konkrétní realizaci náhody $\omega_k \in \Omega$ a v našem případě slouží jako posloupnost náhodných chyb, které děláme při odhadu subgradientu. Prvky y_k jsou omezené, neboť $y_k \in -\partial f(x_k)$. To plyne z předpokladu 1.7 na odhadový mechanismus a z omezenosti subdiferenciálu, viz tvrzení 1.3. Platí tak $y_k + \eta_k = -\zeta(x_k,\omega_k)$ a zřejmě

$$x_{k+1} = x_k - \alpha_k \zeta(x_k,\omega_k) = x_k + \alpha_k (y_k + \eta_k).$$

Dostáváme tak souvislost s diskrétními aproximacemi trajektorie, které byly diskutovány v kapitole 2. Dokážeme-li tedy předpoklady konvergence těchto aproximací ke kritickému bodu, přesněji předpoklady 2.1 a 2.2, aplikací věty 2.3 na

posloupnost $\{x_k\}$ dostaneme konvergenci této posloupnosti ke kritickému bodu zobrazení $-\partial f$. Tím dokážeme také konvergenci metody SSGD ke kritickému bodu funkce f.

Všimněme si, že předpoklad 3.2 zaručuje splnění předpokladu 2.2, neboť tyto podmínky splňuje samotná funkce f. Dále si uvědomme, že stačí ukázat jen $\sup_{k\in\mathbb{N}} \|y_k\| < \infty$ a platnost předpokladů 2.1.3 a 2.1.4. Začneme omezeností posloupnosti $\{y_k\}$. Nejdříve si však ukážeme jednu charakterizaci vlastnosti lokální lipschitzovskosti a její užitečnou vlastnost. Platnost toho lemmatu ukazujeme v příloze A, tvrzení A.2.

Lemma 3.3. *Vlastnost lokální lipschitzovskosti funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ je ekvivalentní s vlastností*

$$\forall x \in \mathbb{R}^n\colon L_f(x) := \limsup_{\substack{y\to x, z\to x \\ y,z\neq x}} \frac{|f(y)-f(z)|}{\|y-z\|} < \infty.$$

Dále platí-li výše zmíněná vlastnost, jsme schopni pro každé $x \in \mathbb{R}^n$ a libovolné $\varepsilon > 0$ najít $\xi > 0$ takové, že pro libovolné $y, z \in B_\xi(x), y \neq z$, platí

$$\frac{|f(y)-f(z)|}{\|y-z\|} \leq L_f(x) + \varepsilon.$$

Omezenost posloupnosti $\{y_k\}$ snadno vyplyne z následujícího tvrzení.

Tvrzení 3.4. *Nechť je posloupnost $\{x_k\}$ omezená. Potom existuje $M > 0$ takové, že platí $\bigcup_{k\in\mathbb{N}} \partial f(x_k) \subseteq \overline{B_M(0)}$. Jinými slovy, množina $\bigcup_{k\in\mathbb{N}} \partial f(x_k)$ je omezená.*

Důkaz. Jelikož je funkce f lokálně lipschitzovská, užitím lemmatu 3.3 můžeme pro každé $k \in \mathbb{N}$ najít $\delta_k > 0$ takové, že pro libovolné $y, z \in B_{\delta_k}(x_k), y \neq z$, platí

$$\frac{|f(y)-f(z)|}{\|y-z\|} \leq L_f(x_k) + 1.$$

Z tvrzení 1.3 víme, že

$$\bigcup_{k\in\mathbb{N}} \partial f(x_k) \subseteq \bigcup_{k\in\mathbb{N}} \overline{B_{L_f(x_k)+1}(0)} \subseteq \overline{B_M(0)},$$

kde $M := 1 + \sup_{k\in\mathbb{N}} L_f(x_k)$. Stačí tak ukázat $\sup_{k\in\mathbb{N}} L_f(x_k) < \infty$. Potom totiž bude uvažovaná množina obsažena v kouli o poloměru M, což je omezená množina.

Pro spor předpokládejme $\sup_{k\in\mathbb{N}} L_f(x_k) = \infty$. Bez újmy na obecnosti předpokládejme, že $L_f(x_k) \to \infty$ a $x_k \to x^*$, jinak si přeuspořádáme a přeindexujeme prvky posloupnosti $\{x_k\}$. Pro každé $k \in \mathbb{N}$ nalezneme posloupnosti $\{y_k^i\}_{i=1}^\infty$ a $\{z_k^i\}_{i=1}^\infty$ realizující hodnoty $L_f(x_k)$ ve smyslu

$$L_f(x_k) = \lim_{i\to\infty} \frac{|f(y_k^i) - f(z_k^i)|}{\|y_k^i - z_k^i\|},$$

přičemž $\lim_{i\to\infty} y_k^i = \lim_{i\to\infty} z_k^i = x_k$ pro každé $k \in \mathbb{N}$ a zároveň $y_k^i, z_k^i \neq x_k, i \in \mathbb{N}$.

Dalším použitím lemmatu 3.3 můžeme nalézt $\eta > 0$ takové, že pro každé $y, z \in B_\eta(x^*), y \neq z$, platí

$$\frac{|f(y)-f(z)|}{\|y-z\|} \leq L_f(x^*) + 1.$$

Najdeme $k_0 \in \mathbb{N}$ takové, že pro každé $k \geq k_0$ platí $x_k \in B_{\eta/2}(x^*)$. Potom pro dostatečně velká $i \in \mathbb{N}$ je také $y_k^i, z_k^i \in B_\eta(x^*)$, neboť $y_k^i, z_k^i \to x_k, i \to \infty$. Pro libovolné $k \geq k_0$ dostáváme

$$L_f(x_k) = \lim_{i\to\infty} \frac{|f(y_k^i) - f(z_k^i)|}{\|y_k^i - z_k^i\|} \leq L_f(x^*) + 1.$$

Limitním přechodem $k \to \infty$ dostáváme $\infty \leq L_f(x^*) + 1$, a to je spor, neboť $L_f(x^*) < \infty$.

□

Poznámka. Vezměme konstantu $M > 0$ z předchozího tvrzení. Potom pro každé $k \in \mathbb{N}$ platí $y_k \in \overline{B_M(0)}$ a zřejmě $\sup_{k\in\mathbb{N}} \|y_k\| \leq M$. Předpoklad 2.1.1 je tedy splněn.

Nyní dokážeme předpoklad 2.1.3. Důležité je správně interpretovat následující tvrzení. To nám říká, že s pravděpodobností jedna, tj. skoro jistě, je splněn předpoklad 3.1.3. S pravděpodobností nula tak můžeme narazit na realizaci náhodných veličin η_k takovou, že řada $\sum_{k=1}^\infty \alpha_k\eta_k$ diverguje a nebudou splněny předpoklady věty 2.3. Prezentujeme zde vlastní argument pro platnost následujícího tvrzení. Důkaz je však podobný Davis et al. [2020, Lemma 4.1], kde se používají martingaly.

Tvrzení 3.5. *Řada $\sum_{k=1}^\infty \alpha_k\eta_k$ konverguje skoro jistě.*

Důkaz. Nechť $S_n = \sum_{k=1}^n \alpha_k\eta_k$. Ukážeme, že $S_n \in \mathbb{L}_2(\Omega, \mathcal{A}, P)$[1], $n \in \mathbb{N}$, a posloupnost těchto veličin je $\mathbb{L}_2$-cauchyovská v následujícím smyslu. Pro každé $\varepsilon > 0$ existuje $n_0 \in \mathbb{N}$ takové, že pro libovolné $n, m \in \mathbb{N}, m \geq n_0, n \geq n_0$, platí

$$\mathbb{E}\left[\|S_m - S_n\|^2\right] < \varepsilon.$$

To nám zaručí $S_n \xrightarrow[n\to\infty]{\mathbb{L}_2} S$, kde S je $\mathbb{L}_2$ náhodná reálná veličina. Tento fakt plyne z Lachout [2004, Věta 6.14]. Konvergence řady náhodných veličin v $\mathbb{L}_2$ implikuje konvergenci v pravděpodobnosti, a tedy i konvergenci skoro jistě. Viz Lachout [2004, Věty 6.10 a 11.3].

Uvědomíme si, že pro každou náhodnou veličinu X platí odhad rozptylu

$$\mathbb{E}\left[\|X - \mathbb{E}[X]\|^2\right] \leq \mathbb{E}\left[\|X\|^2\right]. \tag{3.2}$$

Snadno pak odhadneme

$$\mathbb{E}\left[\|\eta_k\|^2\right] \leq \mathbb{E}_\omega\left[\|\zeta(x_k, \omega)\|^2\right] \leq b(x_k),$$

kde b je funkce z předpokladu 3.1.3. Množina $\{x_1, x_2, \ldots\}$ je omezená, a tedy i pravá strana nerovnosti je omezená, neboť b je omezená na všech omezených množinách. Zřejmě tak platí

$$\sum_{k=1}^\infty \alpha_k^2\, \mathbb{E}\left[\|\eta_k\|^2\right] \leq \sum_{k=1}^\infty \alpha_k^2 b(x_k) < \infty,$$

[1] dále budeme zkracovat pouze na $\mathbb{L}_2$

neboť $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$. Tím jsme ukázali $S_n \in \mathbb{L}_2, n \in \mathbb{N}$. Dále pro každé $n, m \in \mathbb{N}, m > n$, máme

$$\mathbb{E}\left[\|S_m - S_n\|^2\right] \leq \sum_{k=n+1}^{m} \alpha_k^2 \, \mathbb{E}\left[\|\eta_k\|^2\right] \leq \sum_{k=n+1}^{\infty} \alpha_k^2 \, \mathbb{E}\left[\|\eta_k\|^2\right] \overset{n\to\infty}{\longrightarrow} 0,$$

což je přesně $\mathbb{L}_2$-cauchyovskost.

□

Zbývá jen dokázat předpoklad 2.1.4. Tvrzení ukážeme v trochu obecnější variantě. Důkaz tohoto tvrzení není v článku Davis et al. [2020] uveden, níže uvedený důkaz tedy prezentujeme jako svůj vlastní.

Tvrzení 3.6. *Nechť $G\colon \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$ je množinové zobrazení, jehož hodnoty jsou konvexní, kompaktní a z vnějšku spojité množiny. Nechť $\{z_k\}$ a $\{y_k\}$ jsou posloupnosti splňující $z_k \to z^*$ a $y_k \in G(z_k)$, potom*

$$\lim_{n\to\infty} \operatorname{dist}\left(\frac{1}{n}\sum_{k=1}^{n} y_k, G(z^*)\right) = 0.$$

Důkaz. Z toho, že G je z vnějšku spojité zobrazení, máme

$$\lim_{n\to\infty} \operatorname{dist}\left(y_n, G(z^*)\right) = 0.$$

Zvolme $\varepsilon > 0$. Definujme ε-obal množiny $G(z^*)$ jako

$$F_\varepsilon := \left\{x \in \mathbb{R}^n \colon \operatorname{dist}\left(x, G(z^*)\right) \leq \varepsilon\right\}.$$

Množina F_ε je zřejmě kompaktní a konvexní, neboť $G(z^*)$ je kompaktní a konvexní. Existuje $n_0 \in \mathbb{N}$ takové, že pro každé $n \geq n_0$ je splněno $\operatorname{dist}\left(y_n, G(z^*)\right) \leq \varepsilon$, tedy $y_n \in F_\varepsilon$. Pro $n \geq n_0$ platí $\frac{1}{n+1-n_0}\sum_{j=n_0}^{n} y_j \in F_\varepsilon$ z konvexity této množiny. Označme $K := \sup_{v\in F_\varepsilon}\|v\|$ a $L := \max\left\{\|y_1\|, \ldots, \|y_{n_0-1}\|\right\}$. Obě jsou to konečné konstanty. Pro $n \geq n_0$ odhadneme

$$\begin{aligned}\left\|\frac{1}{n}\sum_{j=1}^{n} y_j - \frac{1}{n+1-n_0}\sum_{j=n_0}^{n} y_j\right\| &\leq \frac{n_0-1}{n(n+1-n_0)}\sum_{j=n_0}^{n}\|y_j\| + \frac{1}{n}\sum_{j=1}^{n_0-1}\|y_j\| \\ &\leq (K+L)\frac{n_0-1}{n},\end{aligned}$$

což je pro dostatečně velké n menší než ε. Dostáváme proto

$$\operatorname{dist}\left(\frac{1}{n}\sum_{j=1}^{n} y_j, F_\varepsilon\right) \leq \varepsilon,$$

a z toho plyne také

$$\operatorname{dist}\left(\frac{1}{n}\sum_{j=1}^{n} y_j, G(z^*)\right) \leq 2\varepsilon.$$

Vzhledem k tomu, že ε bylo zvoleno libovolně, dokázali jsme tím toto tvrzení.

□

Poznámka. V tvrzeních 1.2 a 1.3 jsme uvedli, že hodnoty subdiferenciálu lokálně lipschitzovské funkce jsou kompaktní, konvexní a z vnějšku spojité množiny. Volbou $G = -\partial f$ splníme předpoklad 2.1.4. Máme-li totiž nějakou rostoucí posloupnost přirozených čísel $\{n_k\}_{k=1}^{\infty}$ takovou, že $\lim_{k\to\infty} x_{n_k} = x^*$, pak v předchozím tvrzení můžeme uvažovat posloupnost $\{x_{n_k}\}_{k=1}^{\infty}$ místo $\{z_k\}$. Platí pak $y_{n_k} \in -\partial f(x_{n_k})$ a jednoduše lze použít předchozí tvrzení na splnění předpokladu 2.1.4.

Dovršili jsme tak důkaz věty o konvergenci algoritmu SSGD ke kritickému bodu, viz Davis et al. [2020, Věta 4.2].

Věta 3.7. *Každý limitní bod posloupnosti $\{x_k\}$ vygenerované algoritmem 1 je kritickým bodem funkce f a hodnoty $\{f(x_k)\}$ konvergují.*

3.2 Metoda PSSGD

Zaměřme se nyní na metodu PSSGD. Mějme dvě lokálně lipschitzovské funkce f a g. Metoda PSSGD je určena pro úlohu

$$\min_{x\in\mathbb{R}^n} f(x) + g(x). \tag{3.3}$$

V tomto případě nám funkce g slouží jako regularizace. Všimněme si, že součet lokálně lipschitzovských funkcí je zřejmě lokálně lipschitzovská funkce. Součet funkcí lze chápat jako jednu funkci a můžeme tak aplikovat metodu SSGD, pokud jsme schopni odhadnout subgradient tohoto součtu. Metoda PSSGD se liší tím, že odhadujeme subgradient funkce f a poté aplikujeme tzv. *proximální operátor*. To nám umožňuje vyhnout se odhadu subgradientu funkce g. Tento operátor si nyní zavedeme.

Definice 3.8 (Proximální operátor)**.** *Buď $g\colon \mathbb{R}^n \to \mathbb{R}$ zdola omezená, lokálně lipschitzovská funkce. Proximální operátor definujeme jako měřitelné zobrazení $\mathcal{P}_{(\cdot)}g(\cdot)\colon (0,\infty)\times\mathbb{R}^n \to \mathbb{R}^n$ realizující pro výběr*

$$\mathcal{P}_\lambda g(x) \in \operatorname*{argmin}_{y\in\mathbb{R}^n} \left[g(y) + \frac{1}{2\lambda}\|y-x\|^2 \right].$$

Poznámka. Existenci takového zobrazení nám zaručuje [Rockafellar and Wets, 2009, Důsledek 14.6]. Důležitá je vlastnost omezenosti zdola funkce g, která je nutná k tomu, aby byl tento pojem dobře definován. Odteď tedy budeme uvažovat pouze zdola omezenou funkci g, což je obvyklá vlastnost regularizačních funkcí.

Proximální operátor zaručuje jakýsi kompromis mezi vzdáleností od určitého bodu a optimalizací funkce g. Tento kompromis je řízen parametrem λ. Vysoké hodnoty λ znamenají zanedbatelný vliv vzdálenosti od určitého bodu a více se tak zaměřujeme na minimalizaci g.

Nyní jsme připraveni zavést metodu PSSGD pro úlohu (3.3). Ta vypadá ná-

sledovně:

Algoritmus 2: Metoda PSSGD

Vstup: *Kladná posloupnost* $\{\alpha_k\}_{k=1}^{\infty}, x_1 \in \mathbb{R}^n$
Výstup: x_k *je kritickým bodem zobrazení* $\partial f + \partial g$
Začátek:
  $k \leftarrow 1$
  Opakuj:
    Vyber $\omega_k \in \Omega$ a spočítej $\zeta(x_k, \omega_k)$
    $x_k^+ \leftarrow x_k - \alpha_k \zeta(x_k, \omega_k)$
    $x_{k+1} \leftarrow \mathcal{P}_{\alpha_k} g(x_k^+)$
    $k \leftarrow k + 1$
  Dokud: $0 \notin (\partial f + \partial g)(x_k)$
  Vrať: x_k
Konec

Poznámka. Z tvrzení 1.3 a 1.6 víme, že pokud je nějaký bod x^* kritickým bodem funkce $f+g$, musí být $0 \in \partial(f+g)(x^*) \subseteq \partial f(x^*) + \partial g(x^*)$. Pracujeme tak s množinovým zobrazením $-\partial f - \partial g$ místo $-\partial(f+g)$. Pokud se nám podaří splnit předpoklady věty 2.3 o konvergenci diskrétních aproximací trajektorie, dokážeme tím, že hodnoty x_k konvergují ke kritickému bodu zobrazení $-\partial f - \partial g$. Nemůžeme sice zaručit to, aby kritický bod tohoto zobrazení byl i kritickým bodem zobrazení $-\partial(f+g)$, ale nalezení kritického bodu zobrazení $-\partial f - \partial g$ je nutnou podmínkou[2] nalezení kritického bodu zobrazení $-\partial(f+g)$. To nás opravňuje k práci se zobrazením $-\partial f - \partial g$.

První krok algoritmu se neliší od metody SSGD. Jedná se o subgradientovou aktualizaci. Poté však aplikujeme proximální operátor na regularizaci g, tento krok budeme nazývat *proximální aktualizace.*

Diskutujme nyní předpoklady za kterých ukážeme konvergenci algoritmu 2 ke kritickému bodu.

3.2.1 Předpoklady konvergence PSSGD

Všechny předpoklady nyní přehledně uveďme. Předpoklady jsou převzaty z Davis et al. [2020]. Předpoklady, které si nyní zavedeme, si přeneseme do sekce 3.2.2, kde dokážeme konvergenci metody PSSGD ke kritickému bodu.

Předpoklad 3.9. *Předpokládejme, že pro metodu PSSGD definovanou algoritmem 2 je splněno*

1. $\sup_{k\in\mathbb{N}} \|x_k\| < \infty$.
2. *Pro posloupnost iteračních kroků* $\{\alpha_k\}$ *předpokládejme*

$$\alpha_k > 0, \ \sum_{k=1}^{\infty} \alpha_k = \infty, \ \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

[2] existují podmínky, kdy nastává rovnost $\partial(f+g)(x) = \partial f(x) + \partial g(x)$, ovšem tím bychom kladli příliš omezující podmínky na funkce f a g

3. *Pro výběrový mechanismus z předpokladu 1.7 požadujeme, aby existovala funkce* $b\colon \mathbb{R}^n \to \mathbb{R}$ *omezená na všech omezených množinách taková, že*

$$\mathbb{E}_\omega \left[\|\zeta(x,\omega)\|^2 \right] \leq b(x). \tag{3.4}$$

Navíc chtějme, aby pro každou konvergentní posloupnost $\{z_k\}$ *platilo*

$$\mathbb{E}_\omega \left[\sup_{k\in\mathbb{N}} \|\zeta(z_k,\omega)\| \right] < \infty. \tag{3.5}$$

4. *Nechť existuje funkce* $r\colon \mathbb{R}^n \to \mathbb{R}$, *která je omezená na všech omezených množinách, splňující pro každé* $y \in \mathbb{R}^n$ *nerovnost*

$$g(x) - g(y) \leq r(x)\,\|x - y\|\,.$$

Předpoklady 3.9.1 a 3.9.2 již známe z kapitoly 2 a z předpokladů konvergence metody SSGD.

Vlastnost (3.4) slouží k odhadu rozptylu náhodné veličiny $\zeta(x,\omega)$, viz (3.1). Vztah (3.5) zase požaduje, aby střední hodnota odhadu subgradientů byla omezená pro konvergentní posloupnosti. Tento předpoklad nevyplývá z odhadu rozptylu (3.4) a je třeba tuto vlastnost předpokládat.

Předpoklad 3.9.4 poněkud připomíná lokální lipschitzovskost, ovšem jedná se o silnější vlastnost. Naštěstí je splněna pro široké spektrum regularizací, jak ukazuje následující tvrzení. Důkaz je uveden v Davis et al. [2020, Lemma 6.1].

Tvrzení 3.10. *Je-li funkce* g *zdola omezená a lokálně lipschitzovská, která je buď konvexní, globálně lipschitzovská, nebo splňuje vlastnost* $\lim_{\|x\|\to\infty} g(x) = \infty$, *je předpoklad 3.9.4 splněn.*

Zbývá ještě uvést předpoklad existence funkce splňující Ljapunovy podmínky.

Předpoklad 3.11. *Po funkci* $f + g$ *navíc požadujme, aby byla zdola omezená a splňovala*

1. *(Sardova vlastnost) Množina nekritických bodů funkce* $f + g$ *je hustá v* $\mathbb{R}^n$.

2. *(Sestupová vlastnost) Pro každou trajektorii* φ *diferenciální inkluze*

$$\dot{\varphi}(t) \in -(\partial f + \partial g)(\varphi(t)) \text{ pro s.v. } t \in \mathbb{R}_+,$$

splňující $0 \notin (\partial f + \partial g)(\varphi(0))$ *platí, že existuje* $T > 0$ *takové, že*

$$(f+g)(\varphi(T)) < \sup_{t\in[0,T]} (f+g)(\varphi(t)) \leq (f+g)(\varphi(0)).$$

Nyní jsme probrali základní předpoklady pro práci s algoritmem PSSGD a jsme připraveni dokázat jeho konvergenci ke kritickému bodu.

3.2.2 Důkaz konvergence PSSGD

Budeme postupovat podobně jako při důkazu věty 3.7. Nejprve si zavedeme vektory y_k a η_k a ukážeme, že jsou dobře definované. Poté ukážeme, že je splněna formule (2.1) definující diskrétní aproximace trajektorie. Nakonec ověříme předpoklady věty 2.3 konvergenci těchto aproximací ke kritickému bodu, a tím dokážeme i konvergenci algoritmu PSSGD ke kritickému bodu.

Definujme množinové zobrazení $F := -\partial f - \partial g$. Zaveďme si

$$F_k(x) := -\partial f(x) + \alpha_k^{-1}\,\mathbb{E}_\omega\left[\mathcal{P}_{\alpha_k} g(x - \alpha_k\zeta(x,\omega)) - (x - \alpha_k\zeta(x,\omega))\right].$$

Množiny $F_k(x)$ lze interpretovat jako subdiferenciály funkce f, jen zatížené o průměrnou chybu kterou uděláme, když aplikujeme proximální operátor po subgradientové aktualizaci $x - \alpha_k\zeta(x,\omega)$. Zobrazení F_k jsou deterministická a liší se pouze konstantami α_k.

Dále si definujme η_k jako náhodné veličiny

$$\eta_k := \alpha_k^{-1}\left[\mathcal{P}_{\alpha_k} g(x_k - \alpha_k\zeta(x_k,\omega_k)) - x_k\right] \\ - \alpha_k^{-1}\,\mathbb{E}_\omega\left[\mathcal{P}_{\alpha_k} g(x_k - \alpha_k\zeta(x_k,\omega)) - x_k\right],$$

kde η_k závisí na realizaci náhody $\omega_k \in \Omega$. Nabízí se přirozená interpretace těchto veličin. Pro konkrétní realizaci náhody se jedná o vychýlení iteračního kroku algoritmu 2 od očekávané hodnoty x_{k+1}, neboť $x_{k+1} = \mathcal{P}_{\alpha_k} g(x_k - \alpha_k\zeta(x_k,\omega_k)) - x_k$.

Na první pohled není zřejmé, jaké mají η_k a F_k vlastnosti. To, že chyby η_k nejsou v jistém smyslu příliš velké, si ukážeme později v tvrzení 3.16. Následující tvrzení ukazuje, že proximální aktualizace je v jistém smyslu omezená. Přímým důsledkem je omezenost hodnot zobrazení F_k, jak ukážeme v důsledku tohoto tvrzení. Důkaz tvrzení je převzat z Davis et al. [2020, Lemma A.1], my ho zde podrobněji rozvedeme.

Tvrzení 3.12. *Pro libovolné $x, u \in \mathbb{R}^n$ a $\alpha > 0$ platí*

$$\|\mathcal{P}_\alpha g(x - \alpha u) - x\| \le 2\alpha(r(x) + \|u\|),$$

kde r je funkce z předpokladu 3.9.4.

Důkaz. Označme $x^+ = \mathcal{P}_\alpha g(x - \alpha u)$. Z definice proximálního operátoru dostaneme

$$g(x^+) + \frac{1}{2\alpha}\|x^+ - (x - \alpha u)\|^2 \le g(x) + \frac{1}{2\alpha}\|\alpha u\|^2.$$

Rozepsáním zjistíme, že levá strana je rovna

$$g(x^+) + \frac{1}{2\alpha}\|x^+ - x\|^2 + \frac{1}{2\alpha}\|\alpha u\|^2 + \langle x^+ - x, u\rangle.$$

Úpravami a pomocí předpokladu 3.9.4 můžeme odhadnout

$$\frac{1}{2\alpha}\|x^+ - x\|^2 \le g(x) - g(x^+) - \langle x^+ - x, u\rangle \le r(x)\|x^+ - x\| + \|u\|\,\|x^+ - x\|.$$

Pokud $\frac{1}{2\alpha}\|x^+ - x\| \neq 0$, dostaneme tvrzení vydělením nerovnosti tímto výrazem. Opačný případ je triviální.

□

Důsledek 3.13. *Množiny $F_k(x)$ jsou omezené pro každé $x \in \mathbb{R}^n$.*

Důkaz. Vzhledem k tomu, že můžeme zapsat

$$F_k(x) = -\partial f(x) + \mathbb{E}_\omega\left[\zeta(x,\omega)\right] + \alpha_k^{-1}\,\mathbb{E}_\omega\left[\mathcal{P}_{\alpha_k} g(x - \alpha_k\zeta(x,\omega)) - x\right], \tag{3.6}$$

zbývá dokázat pouze

$$\left\|\alpha^{-1}\,\mathbb{E}_\omega\left[\mathcal{P}_\alpha g(x - \alpha\zeta(x,\omega)) - x\right]\right\| < \infty$$

pro každé $x \in \mathbb{R}^n$ a $\alpha > 0$. Subdiferenciál lokálně lipschitzovské funkce je totiž omezený a $\mathbb{E}_\omega\left[\zeta(x,\omega)\right] \in \partial f(x)$. Aplikací Jensenovy nerovnosti na konkávní odmocninu dostaneme

$$\mathbb{E}_\omega\left[\|\zeta(x,\omega)\|\right] \le \sqrt{\mathbb{E}_\omega\left[\|\zeta(x,\omega)\|^2\right]} \le \sqrt{b(x)}, \tag{3.7}$$

dle předpokladu 3.9.3. Jakákoliv norma je konvexní zobrazení, a tak pomocí další aplikace Jensenovy nerovnosti dostaneme

$$\begin{aligned}\left\|\alpha^{-1}\,\mathbb{E}_\omega\left[\mathcal{P}_\alpha g(x - \alpha\zeta(x,\omega)) - x\right]\right\| &\le \alpha^{-1}\,\mathbb{E}_\omega\left[\|\mathcal{P}_\alpha g(x - \alpha\zeta(x,\omega)) - x\|\right] \\ &\le 2r(x) + 2\,\mathbb{E}_\omega\left[\|\zeta(x,\omega)\|\right] \le 2\left(r(x) + \sqrt{b(x)}\right),\end{aligned} \tag{3.8}$$

přičemž jsme využili předchozího tvrzení 3.12 a (3.7). Dosazením α_k dostaneme omezenost množin $F_k(x)$.

□

Pomocí následujícího lemmatu ukážeme souvislost s kapitolou 2 o diskrétních aproximacích trajektorie.

Lemma 3.14. *Existují $y_k \in F_k(x_k)$ takové, že platí*

$$x_{k+1} = x_k + \alpha_k(y_k + \eta_k),\ k \in \mathbb{N}. \tag{3.9}$$

Důkaz. Jelikož musí $y_k \in F_k(x_k)$, můžeme volbou $\mathbb{E}_\omega\left[\zeta(x_k,\omega)\right] \in \partial f(x_k)$ vzhledem k alternativnímu zápisu $F_k(x_k)$ jako v (3.6) vzít

$$y_k = \alpha_k^{-1}\,\mathbb{E}_\omega\left[\mathcal{P}_{\alpha_k} g(x_k - \zeta(x_k,\omega)) - x_k\right].$$

Pak dostaneme

$$\begin{aligned}\alpha_k(y_k + \eta_k) &= \mathbb{E}_\omega\left[\mathcal{P}_{\alpha_k} g(x_k - \zeta(x_k,\omega)) - x_k\right] \\ &\quad + \left[\mathcal{P}_{\alpha_k} g(x_k - \alpha_k\zeta(x_k,\omega_k)) - x_k\right] - \mathbb{E}_\omega\left[\mathcal{P}_{\alpha_k} g(x_k - \alpha_k\zeta(x_k,\omega)) - x_k\right] \\ &\quad = \mathcal{P}_{\alpha_k} g(x_k - \alpha_k\zeta(x_k,\omega_k)) - x_k = x_{k+1} - x_k.\end{aligned}$$

□

V kapitole 2 jsme předpokládali, že $y_k \in H(x_k)$ pro pevně dané množinové zobrazení H. Zde však pracujeme s $y_k \in F_k(x_k)$, kde F_k jsou různá množinová zobrazení. My však ukážeme, že to nevadí a že jsou splněny předpoklady věty 2.3

o konvergenci posloupnosti $\{x_k\}$ ke kritickému bodu zobrazení F. Konvergence metody PSSGD ke kritickému bodu pak bude zaručena.

Všimněme si, že předpoklad 3.11 je stejný s předpokladem 2.2, jen zde Ljapunovy podmínky splňuje samotná funkce $f + g$. Navíc předpoklady na iterační kroky $\{\alpha_k\}$ jsou totožné a stejně tak předpokládáme i omezenost posloupnosti $\{x_k\}$. Stačí tak dokázat $\sup_{k\in\mathbb{N}} \|y_k\| < \infty$ a předpoklady 2.1.3 a 2.1.4. Tyto předpoklady postupně dokážeme. Začneme s omezeností posloupnosti $\{y_k\}$. Důkaz je upravenou a rozvedenou verzí Davis et al. [2020, Lemma A.6].

Tvrzení 3.15. *Posloupnost $\{y_k\}$ je omezená.*

Důkaz. Jelikož $y_k \in F_k(x_k)$, s odkazem na (3.6) stačí ukázat

$$\sup_{k\in\mathbb{N}} \left\| \alpha_k^{-1} \mathbb{E}_\omega \left[\mathcal{P}_{\alpha_k} g(x_k - \alpha_k \zeta(x_k, \omega)) - x_k \right] \right\| < \infty,$$

neboť z tvrzení 3.4 víme, že množina $\bigcup_{k\in\mathbb{N}} \partial f(x_k)$ je omezená.

Z důsledku 3.13, přesněji z odhadu (3.8), víme

$$\sup_{k\in\mathbb{N}} \left\| \alpha_k^{-1} \mathbb{E}_\omega \left[\mathcal{P}_{\alpha_k} g(x_k - \alpha_k \zeta(x_k, \omega)) - x \right] \right\| \leq \sup_{k\in\mathbb{N}} \left[2r(x_k) + 2\sqrt{b(x_k)} \right].$$

Množina $\{x_1, x_2, \ldots\}$ je omezená díky předpokladu 3.9.1, takže i pravá strana této nerovnosti je omezená dle vlastností funkcí r a b.

□

Nyní se zaměříme na předpoklad 2.1.3. Prezentujeme zde podobný argument jako v důkazu tvrzení 3.5, přičemž interpretace nadcházejícího tvrzení je stejná jako u tvrzení 3.5. S pravděpodobností jedna realizace náhodných veličin $\{\eta_k\}$ splňuje předpoklad 2.1.3.

Tvrzení 3.16. *Řada $\sum_{k=1}^{\infty} \alpha_k \eta_k$ konverguje skoro jistě.*

Důkaz. Definujme náhodnou veličinu částečných součtů $S_n := \sum_{k=1}^{n} \alpha_k \eta_k$. Jako v tvrzení 3.5 ukážeme, že posloupnost těchto veličin je $\mathbb{L}_2$-cauchyovská.

Využijeme odhad rozptylu (3.2). Odhadneme

$$\begin{aligned}
\mathbb{E}\left[\|\eta_k\|^2 \right] &\leq \alpha_k^{-2} \mathbb{E}_\omega \left[\|\mathcal{P}_{\alpha_k} g(x_k - \zeta(x_k, \omega)) - x_k\|^2 \right] \\
&\leq \mathbb{E}_\omega \left[4 \left(r(x_k) + \|\zeta(x_k, \omega)\| \right)^2 \right] \\
&= 4\, \mathbb{E}_\omega \left[r^2(x_k) + 2 r_k(x_k) \|\zeta(x_k, \omega)\| + \|\zeta(x_k, \omega)\|^2 \right] \\
&\leq 4 \left(r^2(x_k) + 2 r(x_k) \sqrt{b(x_k)} + b(k_k) \right) = 4 \left(r(x_k) + \sqrt{b(x_k)} \right)^2.
\end{aligned}$$

V druhé nerovnosti jsme použili tvrzení 3.12 a ve třetí nerovnosti jsme použili odhady (3.4) a (3.7). Z omezenosti množiny $\{x_1, x_2, \ldots\}$ plyne omezenost pravé strany. Získáváme tak

$$\sum_{k=1}^{\infty} \alpha_k^2\, \mathbb{E}\left[\|\eta_k\|^2 \right] \leq \sum_{k=1}^{\infty} 4\alpha_k^2 \left(r(x_k) + \sqrt{b(x_k)} \right)^2 < \infty,$$

z čehož $S_n \in \mathbb{L}_2, n \in \mathbb{N}$. Dále pro každé $n, m \in \mathbb{N}, m > n$, platí

$$\mathbb{E}\left[\|S_m - S_n\|^2\right] \leq \sum_{k=n+1}^{m} \alpha_k^2 \, \mathbb{E}\left[\|\eta_k\|^2\right] \leq \sum_{k=n+1}^{\infty} \alpha_k^2 \, \mathbb{E}\left[\|\eta_k\|^2\right] \overset{n\to\infty}{\longrightarrow} 0,$$

což jsme chtěli dokázat.

□

Předtím, než dokážeme platnost posledního předpokladu si uvedeme dvě lemmata.

Lemma 3.17. *Nechť $\{z_k\}$ je omezená posloupnost. Dále nechť $\{\beta_k\}$ je libovolná posloupnost splňující $\beta_k > 0$ a $\sum_{k=1}^{\infty} \beta_k^2 < \infty$. Potom $\beta_k \|\zeta(z_k, \omega)\| \to 0$ skoro jistě a $\beta_k \zeta(z_k, \omega) \to 0$ skoro jistě.*

Důkaz. Důkaz je podobný důkazům tvrzení 3.5 a 3.16. Nechť $X_k(\omega) := \|\zeta(z_k, \omega)\|$. Ukážeme, že $S_n := \sum_{k=1}^{n} \beta_k X_k$ je $\mathbb{L}^2$-cauchyovská posloupnost. Použijeme předpoklad 3.9.3 a odhadneme

$$\sum_{k=1}^{\infty} \beta_k^2 \, \mathbb{E}\left[\|X_k\|^2\right] \leq \sum_{k=1}^{\infty} \beta_k^2 b(z_k) < \infty,$$

neboť $\{z_k\}$ je omezená, a tedy i $\{b(z_k)\}$ je omezená. Dostaneme $S_n \in \mathbb{L}_2, n \in \mathbb{N}$ a pro každé $n, m \in \mathbb{N}, m > n$, máme

$$\mathbb{E}\left[\|S_m - S_n\|^2\right] \leq \sum_{k=n+1}^{m} \beta_k^2 \, \mathbb{E}\left[\|X_k\|^2\right] \leq \sum_{k=n+1}^{\infty} \beta_k^2 b(z_k) \overset{n\to\infty}{\longrightarrow} 0.$$

Tedy S_n konvergují skoro jistě k nějaké reálné náhodné veličině S, viz důkaz tvrzení 3.4. Proto musí $\beta_k \|\zeta(z_k, \omega)\| \to 0$ skoro jistě a také $\beta_k \zeta(z_k, \omega) \to 0$ skoro jistě.

□

Bez důkazu si uvedeme následující lemma. Platnost tohoto lemmatu snadno plyne z Rockafellar and Wets [2009, Příklad 10.2].

Lemma 3.18. *Pro každou funkci $g\colon \mathbb{R}^n \to \mathbb{R}$ a $x \in \mathbb{R}^n$ platí*

$$\alpha^{-1}(\mathcal{P}_\alpha g(x) - x) \in -\partial g(\mathcal{P}_\alpha g(x)).$$

Splnění posledního předpokladu 2.1.4 věty 2.3 vyplyne z následujícího tvrzení. Toto tvrzení je mírným zobecněním podmínky 2.1.4. Důkaz přebíráme z Davis et al. [2020, Lemma A.7] a přikládáme k němu podrobnější komentáře.

Tvrzení 3.19. *Uvažujme libovolnou konvergentní posloupnost $z_k \to z^*$. Potom pro každou rostoucí posloupnost přirozených čísel $\{n_k\}_{k=1}^{\infty}$ a libovolné $y_k \in F_{n_k}(z_k), k \in \mathbb{N}$, platí*

$$\lim_{n\to\infty} \operatorname{dist}\left(\frac{1}{n}\sum_{k=1}^{n} y_k, F(z^*)\right) = 0.$$

Důkaz. Zaveďme $\beta_k := \alpha_{n_k}$. Dále si definujeme náhodné veličiny $z_k^+(\omega) := z_k - \beta_k \zeta(z_k, \omega)$. Vyberme si libovolnou posloupnost $\{w_k^f\}$ takovou, že $w_k^f \in \partial f(z_k)$ a označme si

$$w_k^g(\omega) = \beta_k^{-1} \left[\mathcal{P}_{\beta_k} g(z_k^+(\omega)) - z_k^+(\omega) \right].$$

Za tohoto značení pak

$$y_k := -w_k^f + \mathbb{E}_\omega[w_k^g(\omega)] = -w_k^f + \mathbb{E}_\omega \left[\mathcal{P}_{\beta_k} g(z_k - \beta_k \zeta(z_k, \omega)) - (z_k - \beta_k \zeta(z_k, \omega)) \right]$$

může být libovolný prvek množiny $F_{n_k}(z_k)$ za vhodné volby w_k^f. Uvědomíme si, že pro neprázdnou konvexní množinu $K \subseteq \mathbb{R}^n$ je funkce $x \mapsto \operatorname{dist}(x, K)$ konvexní. Množina $F(z^*)$ je neprázdná a konvexní. Platí totiž $F(z^*) = -\partial f(z^*) - \partial g(z^*)$ a subdiferenciály lokálně lipschitzovských funkcí jsou neprázdné, navíc Minkowského součet konvexních množin je konvexní množina. Viz vlastnosti subdiferenciálu 1.2 a 1.3. Pomocí tohoto pozorování dostaneme

$$\begin{aligned} \operatorname{dist} \left(\frac{1}{n} \sum_{k=1}^{n} \left(-w_k^f + \mathbb{E}_\omega \left[w_k^g(\omega) \right] \right), F(z^*) \right) & \\ \leq \frac{1}{n} \sum_{k=1}^{n} \operatorname{dist} \left(-w_k^f + \mathbb{E}_\omega \left[w_k^g(\omega) \right], F(z^*) \right) & \\ \leq \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}_\omega \left[\operatorname{dist} \left(-w_k^f + w_k^g(\omega), F(z^*) \right) \right]. & \quad (3.10) \end{aligned}$$

V druhé nerovnosti jsme použili Jensenovu nerovnost.

Naším cílem bude ukázat, že sčítance na pravé straně konvergují k nule nezávisle na volbě posloupnosti $\{w_k^f\}$. To provedeme ve 2 krocích:

1. Ukážeme, že skoro jistě platí

$$\operatorname{dist} \left(-w_k^f + w_k^g(\omega), F(z^*) \right) \overset{k \to \infty}{\longrightarrow} 0.$$

2. Ukážeme, že existuje integrovatelná majoranta posloupnosti

$$\left\{ \operatorname{dist} \left(-w_k^f + w_k^g(\omega), F(z^*) \right) \right\}_{k=1}^{\infty}.$$

Následně použijeme Lebesgueovu větu o konvergentní majorantě a dostaneme

$$\mathbb{E}_\omega \left[\operatorname{dist} \left(-w_k^f + w_k^g(\omega), F(z^*) \right) \right] \overset{k \to \infty}{\longrightarrow} 0.$$

Potom si uvědomíme jednoduchý fakt: pokud máme posloupnost $\{a_k\}$ konvergující k nule, pak zřejmě i $\frac{1}{n} \sum_{k=1}^{n} a_k$ konverguje k nule. Aplikací tohoto pozorování snadno dostaneme důkaz tvrzení, neboť pravá strana nerovnosti (3.10) zkonverguje k nule.

Část 1. Užitím lemmatu 3.18 získáme

$$w_k^g(\omega) \in -\partial g(\mathcal{P}_{\beta_k} g(z_k^+(\omega))). \quad (3.11)$$

Zřejmě platí $\beta_k > 0, \beta_k \to 0$ a $\sum_{k=1}^{\infty} \beta_k^2 < \infty$. Použijeme tvrzení 3.12 a lemma 3.17 na odhad, který konverguje k nule skoro jistě:

$$\| \mathcal{P}_{\beta_k} g(z_k^+(\omega)) - z_k \| \leq 2\beta_k r(z_k) + 2\beta_k \| \zeta(z_k, \omega) \| \overset{k \to \infty}{\longrightarrow} 0.$$

Dostáváme tak $\lim_{k\to\infty} \mathcal{P}_{\beta_k} g(z_k^+(\omega)) = \lim_{k\to\infty} z_k = z^*$ skoro jistě. Posloupnost $\{\mathcal{P}_{\beta_k} g(z_k^+(\omega))\}$ je omezená skoro jistě, neboť skoro jistě konverguje. Vzhledem k (3.11) můžeme říci, že je posloupnost $\{w_k^g(\omega)\}$ skoro jistě omezená, neboť je g lokálně lipschitzovská. To je důsledek tvrzení 3.4. Připomeňme zde spojitost z vnějšku subdiferenciálu. Jelikož $w_k^f \in \partial f(z_k)$ a dále $\lim_{k\to\infty} \mathcal{P}_{\beta_k} g(z_k^+(\omega)) = z^*$ a $w_k^g(\omega) \in -\partial g(\mathcal{P}_{\beta_k} g(z_k^+(\omega)))$, dostáváme

$$\operatorname{dist}(w_k^f, \partial f(z^*)) \to 0 \text{ a } \operatorname{dist}(w_k^g(\omega), -\partial g(z^*)) \to 0$$

za $k \to \infty$, přičemž druhá limita platí skoro jistě. Použijeme trojúhelníkovou nerovnost a $F(z^*) = -\partial f(z^*) - \partial g(z^*)$ na odhad

$$\operatorname{dist}(-w_k^f + w_k^g(\omega), F(z^*)) \le \operatorname{dist}(-w_k^f, -\partial f(z^*)) + \operatorname{dist}(w_k^g(\omega), -\partial g(z^*)) \overset{k\to\infty}{\longrightarrow} 0.$$

Tím jsme dokázali konvergenci k nule skoro jistě.

Část 2. Nyní ukážeme, že existuje integrovatelná majoranta. Odhadneme

$$\sup_{k\in\mathbb{N}} \left[\operatorname{dist}(-w_k^f + w_k^g(\omega), F(z^*))\right] \le \sup_{k\in\mathbb{N}} \|w_k^f\| + \sup_{k\in\mathbb{N}} \|w_k^g(\omega)\| + \operatorname{dist}(0, F(z^*)).$$

Ukážeme, že pravá strana má konečnou střední hodnotu. Platí $\sup_{k\in\mathbb{N}} \|w_k^f\| < \infty$, neboť $w_k^f \in \bigcup_{k\in\mathbb{N}} \partial f(z_k)$, a to je omezená množina dle tvrzení 3.4. Dále z tvrzení 3.12 platí, že

$$\|w_k^g(\omega)\| \le \beta_k^{-1} \|\mathcal{P}_{\beta_k} g(z_k^+(\omega)) - z_k\| + \|\zeta(z_k, \omega)\| \le 2r(z_k) + 3\|\zeta(z_k, \omega)\|,$$

tedy

$$\sup_{k\in\mathbb{N}} \|w_k^g(\omega)\| \le 2 \sup_{k\in\mathbb{N}} r(z_k) + 3 \sup_{k\in\mathbb{N}} \|\zeta(z_k, \omega)\|.$$

Protože je $\{x_1, \ldots\}$ omezená množina a předpokládáme 3.9.3, přesněji vlastnost (3.5) pro konvergentní posloupnosti, má pravá strana konečnou střední hodnotu. Omezenost $\operatorname{dist}(0, F(z^*))$ plyne snadno z lokální lipschitzovskosti funkcí f a g. Celkově jsme tedy našli integrovatelnou funkci v proměnné ω a důkaz je proveden. □

Poznámka. Vraťme se k posloupnosti $\{x_k\}$. Nechť nějaká její podposloupnost $\{x_{n_k}\}_{k=1}^{\infty}$ konverguje k $x^* \in \mathbb{R}^n$. Aplikací předchozí věty můžeme splnit předpoklad 2.1.4 pro libovolnou volbu $y_{n_k} \in F_{n_k}(x_{n_k})$, neboť vždy platí

$$\lim_{k\to\infty} \operatorname{dist}\left(\frac{1}{n}\sum_{k=1}^{n} y_{n_k}, F(x^*)\right) = 0.$$

Posloupnost $\{x_{n_k}\}_{k=1}^{\infty}$ tak hraje roli posloupnosti $\{z_k\}$ v důkazu tohoto tvrzení.

Dokázali jsme tak následující větu o konvergenci algoritmu PSSGD. Věta je převzata z Davis et al. [2020, Věta 6.2]

Věta 3.20 (Konvergence PSSGD)**.** *Nechť jsou splněny předpoklady 3.9 a 3.11. Pak pro každý limitní bod x^* posloupnosti $\{x_k\}$ vygenerované algoritmem 2 platí, že $0 \in (-\partial f - \partial g)(x^*)$ a posloupnost $\{(f+g)(x_k)\}$ konverguje.*

4. Interpretace podmínek konvergence

Přesuňme se nyní na interpretaci podmínek, za kterých jsme dokazovali konvergence algoritmů SSGD a PSSGD ke kritickému bodu. Předpoklady 3.1 a 3.9 jsme již okomentovali. V této kapitole se proto zaměříme na interpretaci trajektorie diferenciální inkluze, Ljapunových podmínek a zmíníme se o větě o aproximaci trajektorie. Pokud v průběhu této kapitoly zavedeme nějaké pojmy, či slovní spojení, budeme je používat po zbytek kapitoly.

K interpretaci podmínek konvergence nám pomohly články Murray et al. [1993] a Davis et al. [2020], ze kterých jsme studovali.

4.1 Trajektorie diferenciální inkluze

Připomeňme, že trajektorie diferenciální inkluze je absolutně spojitá křivka $\varphi\colon \mathbb{R}_+ \to \mathbb{R}^n$ splňující

$$\dot{\varphi}(t) \in H(\varphi(t)) \text{ pro s.v. } t > 0, \tag{4.1}$$

kde $H\colon \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$ je nějaké množinové zobrazení. Křivku φ splňující tuto vlastnost jsme nazvali *trajektorií zobrazení* H.

Vezměme libovolnou trajektorii zobrazení H a označme ji φ. Pak lze říci, že časová derivace $\dot{\varphi}$, neboli směr křivky φ, dobře popisuje chování zobrazení H v bodech křivky φ. Jelikož nejdůležitějším případem zobrazení H v této práci je subdiferenciál lokálně lipschitzovské funkce, budeme diskutovat případ $H := -\partial f$ pro f lokálně lipschitzovskou. Dále budeme říkat, že funkce f má nějakou vlastnost po své trajektorii, jestliže tuto vlastnost má zobrazení $t \mapsto f(\varphi(t))$, kde φ je libovolná trajektorie zobrazení $-\partial f$.

Pokud je f diferencovatelná, vztah (4.1) přejde v

$$\dot{\varphi}(t) = -\nabla f(\varphi(t)), \tag{4.2}$$

neboť v tomto případě platí $\partial f(x) = \{\nabla f(x)\}, \forall x \in \mathbb{R}^n$. Připomeňme, že má-li funkce f v bodě $x \in \mathbb{R}^n$ gradient, pak vektor $-\nabla f(x)$ je směr největšího spádu v následujícím smyslu. Pro $d \in \mathbb{R}^n$ definujme $g_d(\alpha) := f(x + \alpha d)$, potom výraz $\frac{\partial g_d}{\partial \alpha}(0)$ je minimalizován právě když $d = -\nabla f(x)$. Dá se tak říct, že na okolí bodu x funkce f klesá nejvíce právě ve směru $-\nabla f(x)$.

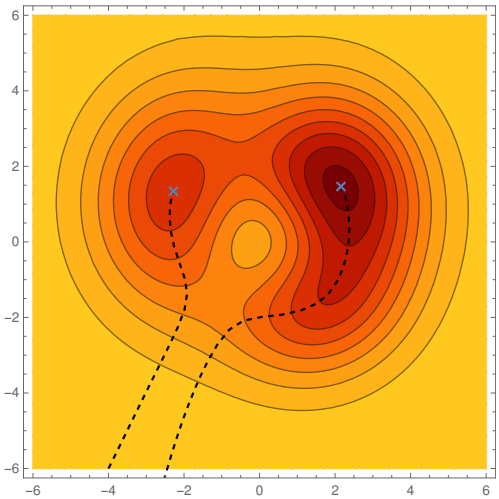
Z tohoto důvodu můžeme interpretovat, že trajektorie zobrazení $-\nabla f$ sleduje směr největšího spádu funkce f. Tj. směr trajektorie, neboli její časová derivace, je totožný se směrem největšího spádu funkce f v bodech této trajektorie. V případě diferencovatelných funkcí tedy očekáváme, že f bude nerostoucí po své trajektorii. Význam pro optimalizaci je proto zřejmý. Ilustraci diferenciální inkluze (4.2) lze nalézt v obrázku 4.1a.

Uvažujme nyní, že f není diferencovatelná všude, ale je lokálně lipschitzovská. Pak přecházíme k diferenciální inkluzi

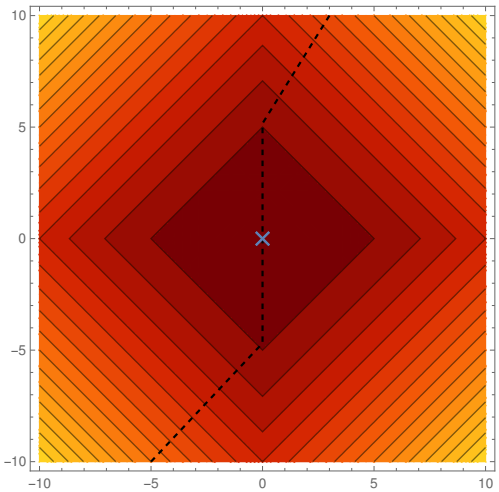
$$\dot{\varphi}(t) \in -\partial f(\varphi(t)) \text{ pro s.v. } t > 0. \tag{4.3}$$

Věta 1.4 o sestupu ve směru subgradientu nám v kombinaci s možností volby ze subdiferenciálu, reprezentovanou vztahem (4.3), dává možnost vybrat směr trajektorie takový, že f bude nerostoucí po své trajektorii. Ukázku diferenciální inkluze (4.3) lze nalézt v obrázku 4.1b.

Obecně se však může stát, že existuje trajektorie, po níž je funkce f neklesající, ale přitom funkce f po této trajektorii nedosáhla kritického bodu. Tyto případy pak musíme ošetřit dalšími předpoklady. Přecházíme tak postupně k Ljapunovým podmínkám.



(a) Trajektorie diferenciální inkluze (4.2), kde f je diferencovatelná.



(b) Trajektorie diferenciální inkluze (4.3), kde f je nediferencovatelná.

Obrázek 4.1: Ukázka trajektorií diferenciální inkluze pro případ diferencovatelné a nediferencovatelné funkce. Na obrázcích jsou barevně vyznačeny úrovňové množiny funkce f. Tmavší odstíny červené indikují nižší hodnoty funkce f. Přerušované čáry značí trajektorie diferenciální inkluze a křížky lokální minima funkce f.

4.2 Ljapunovy podmínky

Připomeňme zde, co jsou Ljapunovy podmínky. Jedná se o dvě podmínky, které jsme nazvali Sardova podmínka a sestupová vlastnost.

Sardova podmínka pro nějaké zobrazení $L\colon \mathbb{R}^n \to \mathbb{R}^m$ říká, že množina nekritických bodů L je hustá v prostoru $\mathbb{R}^n$. Jedná se o podmínku technického rázu a používá se při důkazu věty 2.3 o konvergenci diskrétní aproximace trajektorie. Název Sardova podmínka vychází ze Sardovy věty, viz de Pascale [2001], která říká, že pokud je L alespoň třídy $\mathcal{C}^k(\mathbb{R}^n, \mathbb{R}^m)$, kde $k \geq \max\{n-m+1, 1\}$, pak množina kritických bodů funkce L má nulovou lebesgueovu míru na $\mathbb{R}^n$.

Zaměřme se nyní na sestupovou vlastnost. Ta říká, že pro každou trajektorii φ zobrazení $H\colon \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$ takovou, že $\varphi(0)$ není kritickým bodem zobrazení H, existuje $T > 0$ splňující

$$f(\varphi(T)) < \sup_{t\in[0,T]} f(\varphi(t)) \leq f(\varphi(0)). \tag{4.4}$$

O funkci f splňující tuto podmínku říkáme, že splňuje sestupovou vlastnost pro zobrazení H. V kapitole 2 jsme předpokládali existenci funkce splňující sestupovou vlastnost pro obecné zobrazení H. V kapitole 3 jsme zase pro lokálně lipschitzovské funkce f a g předpokládali, že samotné f a $f+g$ splňují sestupovou vlastnost pro zobrazení $-\partial f$, resp. $-\partial f - \partial g$.

V předchozí sekci jsme zmiňovali, že funkce může být neklesající po nějaké své trajektorii, ale přitom po této trajektorii nedosáhla kritického bodu. Předpoklad sestupové vlastnosti tuto možnost odstraňuje. Znamená totiž to, že funkce f za dostatečně dlouhou dobu vždy ostře klesá po své trajektorii, dokud se neocitne v kritickém bodě. Demonstrujme nyní proč tomu tak je.

Předpokládejme, že f splňuje sestupovou vlastnost pro zobrazení $-\partial f$ a nechť φ je nějaká trajektorie zobrazení $-\partial f$. Označme si $T > 0$ konstantu splňující (4.4), ale zároveň předpokládejme, že $\varphi(T)$ není kritickým bodem $-\partial f$ a $f(\varphi(\tau)) \geq f(\varphi(T)),\ \forall \tau > T$. Pak si můžeme definovat časově posunutou křivku $\psi(t) := \varphi(t+T),\ t \geq 0$. To je zřejmě trajektorie zobrazení $-\partial f$, neboť

$$\dot{\psi}(t) = \dot{\varphi}(t+T) \in -\partial f(\varphi(t+T)) = -\partial f(\psi(t)) \text{ pro s.v. } t \geq 0.$$

Předpokládáme, že $\psi(0)$ není kritickým bodem $-\partial f$, a tak najděme $S > 0$ takové, že

$$f(\psi(S)) < \sup_{s \in [0,S]} f(\psi(s)) \leq f(\psi(0)).$$

Potom však platí

$$f(\varphi(S+T)) = f(\psi(S)) < f(\psi(0)) = f(\varphi(T)),$$

což je spor. Sestupová vlastnost, společně s ostatními předpoklady diskutovanými v kapitole 3, nám tak zaručí, že funkce f po své trajektorii dosáhne kritického bodu.

4.2.1 Funkce splňující Ljapunovy podmínky

Ověřovat Ljapunovy podmínky zmíněné v předchozí sekci není jednoduché. Proto v této sekci probereme některé důležité třídy funkcí, které tyto podmínky splňují. Výsledky uvedené v této sekci jsou převzaty z Davis et al. [2020].

Význam následující definice bude zřejmý z tvrzení, které bude následovat.

Definice 4.1 (Řetízkové pravidlo)**.** *Nechť $f\colon \mathbb{R}^n \to \mathbb{R}$ a $G\colon \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$. Řekneme, že f splňuje řetízkové pravidlo pro zobrazení G, pokud pro každou absolutně spojitou křivku $\varphi\colon \mathbb{R}_+ \to \mathbb{R}^n$ a skoro všechna $t \geq 0$ platí*

$$(f \circ \varphi)'(t) = \langle v, \dot{\varphi}(t) \rangle,\ \forall v \in G(\varphi(t)).$$

Tento fakt budeme zapisovat zkráceně jako $(f \circ \varphi)'(t) = \langle G(\varphi(t)), \dot{\varphi}(t) \rangle$.

Tvrzení 4.2. *Nechť funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ splňuje řetízkové pravidlo zobrazení $G\colon \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$. Potom funkce f splňuje sestupovou vlastnost pro zobrazení $-G$.*

Důkaz. Důkaz tohoto tvrzení je kombinací důkazů lemmat Davis et al. [2020, Lemma 5.2 a 6.3].

□

Přirozeně nás nejvíce zajímá případ $G = \partial f$. Zvolili jsme takhle formulovanou definici 4.1 a tvrzení 4.2 právě proto, že nám umožňuje přechod k součtu funkcí. Máme-li totiž dvě funkce f a g, které splňují řetízkové pravidlo pro zobrazení ∂f, resp. ∂g, pak zřejmě pro skoro všechna $t \geq 0$ platí

$$\begin{aligned}((f+g)\circ\varphi)'(t) &= (f\circ\varphi)'(t) + (g\circ\varphi)'(t) \\ &= \langle \partial f(\varphi(t)), \dot{\varphi}(t)\rangle + \langle \partial g(\varphi(t)), \dot{\varphi}(t)\rangle = \langle F(\varphi(t)), \dot{\varphi}(t)\rangle,\end{aligned}$$

kde $F := \partial f + \partial g$. Tedy $f + g$ splňuje řetízkové pravidlo pro zobrazení $\partial f + \partial g$ a použitím tvrzení 4.2 tak splníme předpoklad 3.11.2 sestupové vlastnosti funkce $f+g$ pro zobrazení $-\partial f - \partial g$. Tvrzení tak lze snadno aplikovat na metodu SSGD i PSSGD.

Dalším výsledkem Davis et al. [2020, Kapitola 5] je, že funkce, které jsou tzv. *definovatelné*, splňují Ljapunovy podmínky.

Tvrzení 4.3. *Definovatelné funkce splňují Ljapunovy podmínky.*

Definovatelné funkce jsou funkce, jejichž grafy jsou tzv. *definovatelné v o-minimální struktuře*. Diskuze pojmu o-minimální struktury je ovšem mimo rozsah této práce. Pro více informací čtenáře odkážeme na Davis et al. [2020]. Nám bude stačit, že definovatelnými funkcemi jsou například polynomy, exponenciální funkce a maximum a minimum ze dvou definovatelných funkcí. Navíc součet, násobení a složení definovatelných funkcí je definovatelná funkce a inverz definovatelné funkce je také definovatelný.

4.3 Souvislost s větou o aproximaci trajektorie

V kapitole 3 jsme pro algoritmy SSGD a PSSGD ukázali za diskutovaných předpokladů splnění sady předpokladů 2.1, které jsou zároveň předpoklady pro větu o aproximaci trajektorie 2.4. Vezmeme-li posloupnost $\{x_k\}$ vygenerovanou algoritmem SSGD nebo PSSGD, pak věta o aproximaci trajektorie říká, že pro dostatečně velké $k \in \mathbb{N}$ posloupnost $\{x_k\}$ dobře aproximuje trajektorii zobrazení $-\partial f$ (případ SSGD), resp. $-\partial f - \partial g$ (případ PSSGD). V kombinaci se sestupovou vlastností z Ljapunových podmínek pro funkce f, resp. $f + g$, pak dostáváme, že posloupnost $\{x_k\}$ pro dostatečně velká $k \in \mathbb{N}$ ostře klesá po odpovídající trajektorii diferenciální inkluze. Proto posloupnost $\{x_k\}$ konverguje ke kritickému bodu, neboť tam konvergují i trajektorie zobrazení $-\partial f$, resp. $-\partial f - \partial g$, jak jsme ukázali v předchozí sekci.

5. Numerická ukázka

V této kapitole provedeme praktickou ukázku algoritmů SSGD a PSSGD. Studie je inspirována článkem Zhang et al. [2020], ze kterého jsme také převzali některé regularizace uvedené v sekci 5.1. Dále jsme čerpali z Wang et al. [2018], konkrétně jsme převzali explicitní řešení proximálního operátoru na námi vybranou regularizaci.

K ukázce jsme si vybrali 2 datasety se kterými budeme pracovat, a to sice Digit MNIST a Fashion MNIST. Digit MNIST je dataset ručně psaných číslic reprezentovaný obrázky 28x28 pixelů. Tyto obrázky jsou černobílé a každá hodnota pixelu udává jeho světlost v rozmezí hodnot od 0 do 1, přičemž hodnota 0 značí černou barvu a hodnota 1 bílou. Celkově je v tomto datasetu 10 kategorií: číslice 0 až 9. Fashion MNIST je podobný dataset. Prvky tohoto datasetu jsou reprezentovány podobně jako Digit MNIST, tedy obrázky 28x28 pixelů v odstínech černé barvy. Rozdíl je ten, že ve Fashion MNIST datasetu jsou obrázky různých typů oblečení místo číslic. Kategorií je v tomto datasetu 10. Každé pozorování z obou datasetů má k sobě přiřazené tzv. *označení*, tedy kategorie do které spadá. Přehled těchto datasetů je uveden v tabulce 5.1.

Dataset	Počet pozorování	Počet sloupců (bez označení)
Digit MNIST	10330	784
Fashion MNIST	15000	784

Tabulka 5.1: Přehled rozsahu datasetů Digit MNIST a Fashion MNIST.

Každý z těchto datasetů má několik kategorií, do kterých se dají data klasifikovat. My si však vybereme pouze jednu kategorii a provedeme binární klasifikaci. Tedy budeme rozlišovat pouze zda dané pozorování patří do oné kategorie, či nikoliv.

Nechť je počet prvků daného datasetu $m \in \mathbb{N}$ a nechť jsou naše data reprezentována reálnými vektory $x^i \in \mathbb{R}^n, i \in \{1, \dots, m\}$. Součástí datasetu jsou i označení $y^i \in \{-1,1\}, i \in \{1, \dots, m\}$, která určují do jaké kategorie x^i spadá. Pokud $y^i = 1$ řekneme, že x^i spadá do vybrané kategorie a pokud naopak $y^i = -1$, pak do oné kategorie nespadá. Pro pohodlnost si na konec každého pozorování přidáme hodnotu -1. To nám umožní kompaktně zapsat ztrátovou funkci. Pracujeme tak s daty o $n + 1$ hodnotách.

Představme nyní náš model. Použijeme model tzv. *soft-margin SVM*[1] s regularizací MCP, jejíž definice je uvedená v sekci 5.1. Ztrátová funkce toho modelu je definována pro $w \in \mathbb{R}^{n+1}$, jako

$$\mathrm{SVM}(w) := \frac{1}{m} \sum_{i=1}^{m} \max\left\{0, 1 - y^i \langle w, x^i \rangle\right\}.$$

Ke ztrátové funkci SVM se také přičítá L_2 norma, která má zajistit, aby parametry modelu nebyly příliš velké. Po přičtení regularizace MCP tak dostaneme model

$$f(w; \lambda) = \mathrm{SVM}(w; \lambda) + \frac{\lambda}{2} \|w\|^2 + \lambda \, \mathrm{MCP}(w),$$

[1] zkratka SVM znamená v angličtině *support vector machine*

kde $\mathrm{MCP}(w) := \sum_{k=1}^{n+1} \mathrm{MCP}(w_k), w = (w_1, \ldots, w_{n+1}) \in \mathbb{R}^{n+1}$.

Parametr $\lambda \geq 0$ ovlivňuje kompromis mezi predikční schopností modelu a vlivu regularizace, v našem případě řídkosti řešení. Vysoké hodnoty λ více zvýrazňují vliv regularizace, ale zpravidla na úkor zhoršení predikční schopnosti modelu.

Odhad subgradientu ztrátové funkce realizujeme pomocí tzv. *mini-batchů*. Jedná se o náhodný výběr menšího množství pozorování z datasetu o předem daném rozsahu. Na těchto výběrech se pak spočte subgradient ztrátové funkce a ten se použije v iteraci algoritmů SSGD a PSSGD.

Studii provedeme tak, že nejdříve optimalizujeme parametry modelu bez regularizace pomocí metody SSGD. Poté přičteme vybranou regularizaci a tento model optimalizujeme pomocí metody SSGD i PSSGD.

V následující sekci čtenáři ukážeme, jak může volba regularizace ovlivnit řídkost řešení optimalizační úlohy a v sekci 5.2 prezentujeme výsledky této numerické ukázky.

5.1 Vliv regularizace na řídkost řešení

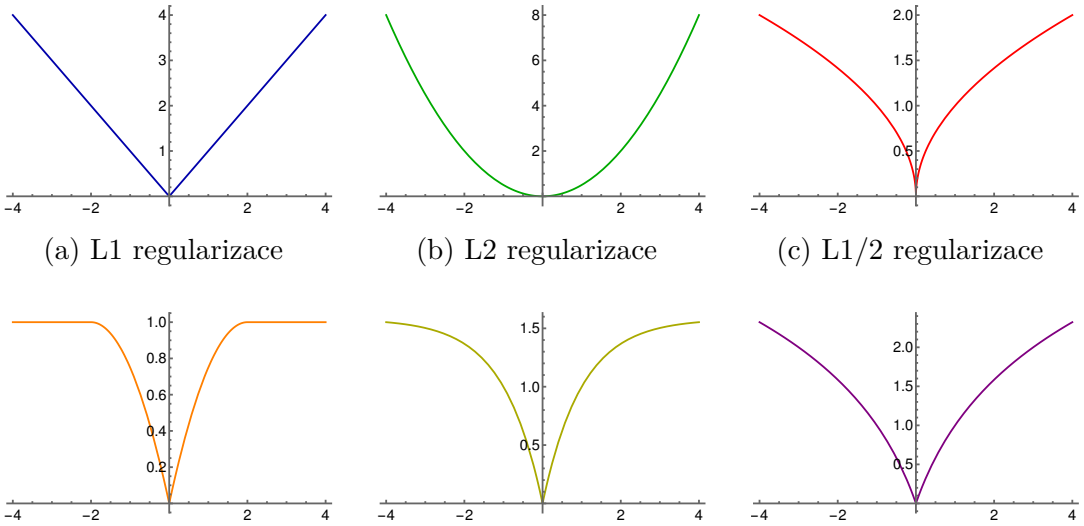
Regularizace, které mají ovlivnit řídkost řešení, jsou definovány tak, aby penalizovaly přílíš vysoké hodnoty parametrů, a nebo aby silně zmenšovaly hodnotu parametrů na nějaké blízkém okolí nuly. Zpravidla se aplikují po složkách vektoru parametrů.

V tabulce 5.2 si uvedeme několik příkladů regularizací, které ovlivňují řídkost řešení. V obrázku 5.1 jsou tyto regularizace graficky znázorněny.

Název regularizace	Definice	Podmínky
L1	$\lvert x\rvert$	$x \in \mathbb{R}^n$
L2	$\frac{1}{2}x^2$	$x \in \mathbb{R}^n$
L1/2	$\sqrt{\lvert x\rvert}$	$x \in \mathbb{R}^n$
MCP	$\begin{cases} \alpha \lvert x\rvert - \frac{x^2}{2\beta} & \lvert x\rvert \leq \alpha\beta \\ \frac{1}{2}\beta\alpha^2 & \lvert x\rvert > \alpha\beta \end{cases}$	$x \in \mathbb{R}^n, \alpha > 0, \beta > 1$
ETF	$\frac{\alpha(1-e^{-\beta\lvert x\rvert})}{1-e^{-\beta}}$	$x \in \mathbb{R}^n, \alpha > 0, \beta > 0$
LTF	$\frac{\alpha \log(\beta\lvert x\rvert+1)}{\log(\beta+1)}$	$x \in \mathbb{R}^n, \alpha > 0, \beta > 0$

Tabulka 5.2: Příklady některých regularizací ovlivňující řídkost řešení.

Vidíme, že tyto regularizace mají dvě společné vlastnosti. Předně mají minimum v bodě nula. To dává smysl, neboť chceme, aby hodnoty parametrů byly co nejmenší. Dále na okolí nuly ostře klesají směrem k minimu. Jelikož v metodách SSGD a PSSGD dochází i k minimalizaci regularizace (v obou případech v jiném smyslu), zajistíme tím to, že regularizace zvýhodňuje celkovou ztrátovou funkci pro malé hodnoty parametrů, přičemž největší zvýhodnění nastává pro nulové hodnoty parametrů. Regularizace nám tedy dělá jakýsi „kompromis“ mezi predikční schopností modelu a řídkosti řešení.



(a) L1 regularizace (b) L2 regularizace (c) L1/2 regularizace

(d) MCP regularizace s parametry $\alpha = 1$ a $\beta = 2$ (e) ETF regularizace s parametry $\alpha = 1$ a $\beta = 1$ (f) LTF regularizace s parametry $\alpha = 1$ a $\beta = 1$

Obrázek 5.1: Grafické znázornění některých regularizací

5.2 Výsledky numerické ukázky

Výsledky numerické ukázky jsou uvedeny v tabulkách 5.3 a 5.4. Přesnost definujeme jako podíl správně zařazených pozorování k celkovému počtu pozorování v datasetu.

Model	Metoda	Přesnost	Řídkost	Rozdíl přesností	Rozdíl řídkostí
SVM	SSGD	95.44%	0.00%	-	-
SVM+MCP	SSGD	94.68%	88.41%	-0.76%	88.41%
SVM+MCP	PSSGD	95.15%	38.73%	-0.29%	38.73%

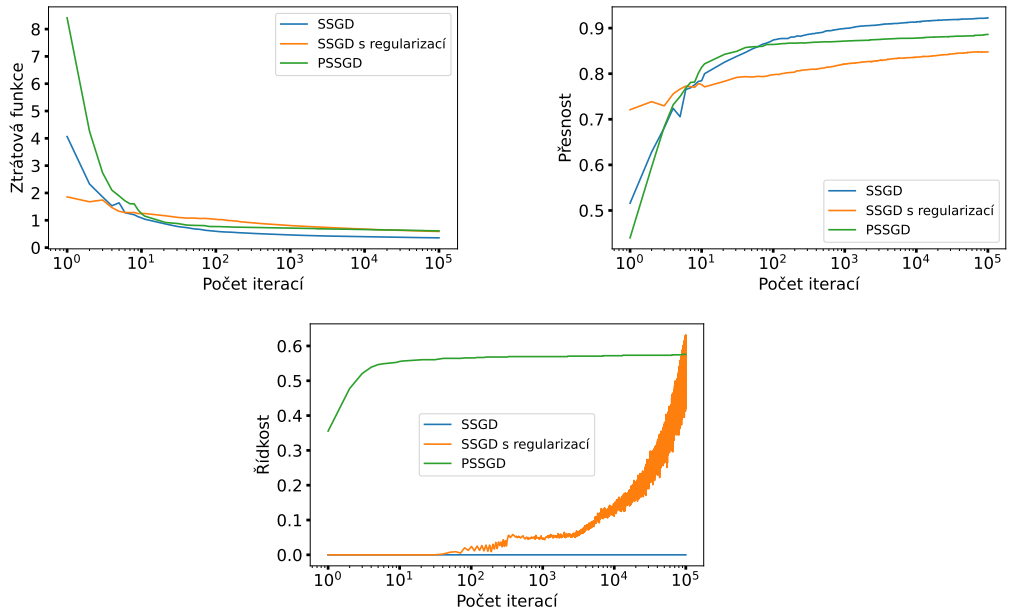
Tabulka 5.3: Výsledky numerické ukázky pro dataset Digit MNIST.

Model	Metoda	Přesnost	Řídkost	Rozdíl přesností	Rozdíl řídkostí
SVM	SSGD	91.16%	0.00%	-	-
SVM+MCP	SSGD	90.43%	89.68%	-0.73%	89.68%
SVM+MCP	PSSGD	91.06%	34.90%	-0.10%	34.90%

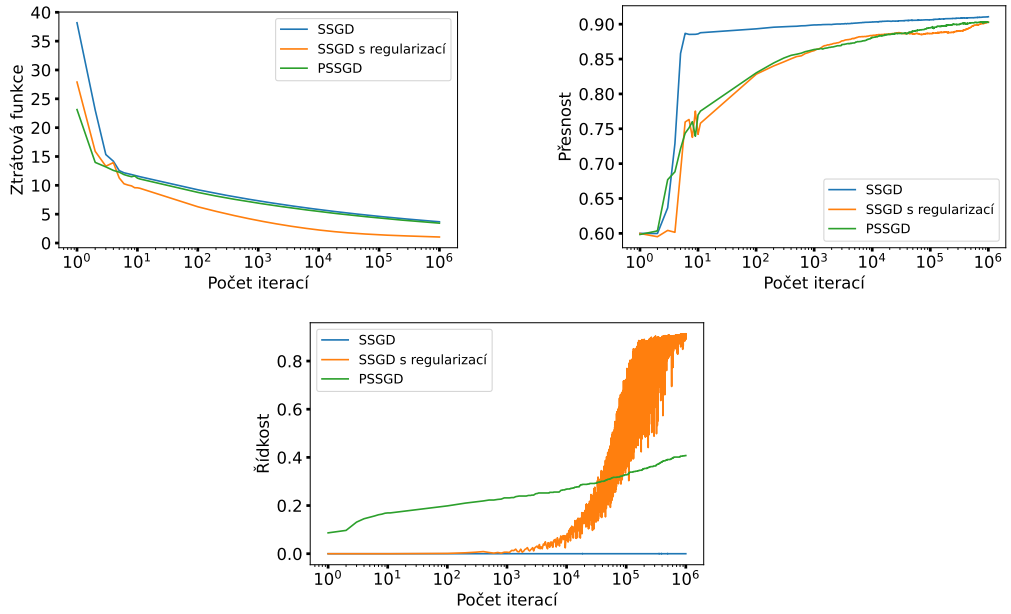
Tabulka 5.4: Výsledky numerické ukázky pro dataset Fashion MNIST.

Vidíme, že po přičtení regularizace skutečně došlo ke zvýšení řídkosti řešení, přičemž přesnost se změnila jen minimálně. V této konkrétní ukázce má metoda SSGD výrazně větší vliv na řídkost řešení, než metoda PSSGD. Ovšem metoda PSSGD má také výrazný vliv na řídkost řešení.

V obrázcích 5.2 a 5.3 jsme zaznamenali vývoj hodnot ztrátové funkce, přesnosti a řídkosti klasifikátoru v závislosti na počtu iterací algoritmu. Vidíme, že nejvýraznější optimalizace probíhá v prvních stovkách iterací, poté hodnoty ztrátové funkce pozvolna klesají. Výhoda metody PSSGD spočívá v tom, že po méně iteracích nachází poměrně řídké řešení a je stabilnější než metoda SSGD, což je



Obrázek 5.2: Grafické znázornění vývoje přesnosti a řídkosti v průběhu optimalizace na datasetu Digit MNIST.



Obrázek 5.3: Grafické znázornění vývoje přesnosti a řídkosti v průběhu optimalizace na datasetu Fashion MNIST.

vidět z kolísání řídkosti této metody. Samotný model SVM má sice nejlepší predikční schopnosti, ovšem nedochází v něm k žádnému ovlivňování řídkosti řešení.

Na závěr této ukázky poznamenejme, že modely tohoto typu jsou citlivé na počáteční volby tzv. *hyperparametrů*, v našem případě se jedná o α, β a λ. V závislosti na jejich volbě můžeme dostat jiné výsledky. Navíc jsou svou podstatou metody SSGD a PSSGD náhodné, a tedy lze pro jiné realizace odhadu subgradientu získat odlišné výsledky.

Závěr

V této práci jsme čtenáře seznámili s motivací hledání řídkých řešení v klasifikačních úlohách a se základními pojmy optimalizačních metod využívajících subgradient obecně nediferencovatelné funkce. Stručně jsme popsali co znamenají diskrétní aproximace trajektorie a formulovali jsme podmínky konvergence těchto aproximací ke kritickému bodu. Dále jsme čtenáři představili metody SSGD a PSSGD, také jsme ukázali, jaká je souvislost mezi těmito metodami a diskrétními aproximacemi trajektorie. Prezentovali jsme podmínky, za nichž jsme dokázali konvergenci těchto metod ke kritickému bodu, přičemž jsme ukázali několik vlastních důkazů dílčích tvrzení. Poté jsme interpretovali podmínky konvergence metod SSGD a PSSGD ke kritickému bodu a na závěr jsme demonstrovali použití těchto metod na konkrétní klasifikační úloze, přičemž jsme ukázali pozitivní vliv na řídkost řešení.

Metody SSGD a PSSGD mají mnoho možných rozšíření. Dá se například pracovat s různými normami nebo můžeme omezit metodu na nějakou podmnožinu $\mathbb{R}^n$. Tuto bakalářskou práci by šlo rozšířit například právě o tyto případy. Dále se lze více věnovat funkcím splňující Ljapunovy podmínky a podrobněji popsat o-minimální struktury a definovatelné funkce.

Vzhledem k důležitosti metod SSGD a PSSGD pro optimalizační úlohy ve strojovém a hlubokém učení, se rozhodně vyplatí zkoumat vlastnosti těchto metod i pro jejich rozšíření a funkce, které nemusí být lokálně lipschitzovské.

Seznam použité literatury

Francis H Clarke, Yuri S Ledyaev, Ronald J Stern, and Peter R Wolenski. *Nonsmooth analysis and control theory*, volume 178. Springer Science & Business Media, 2008.

Frank H Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.

Luigi de Pascale. The morse–sard theorem in sobolev spaces. *Indiana University Mathematics Journal*, 50(3):1371–1386, 2001.

John C. Duchi and Feng Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.

Juha Heinonen. *Lectures on Lipschitz analysis*. Number 100. University of Jyväskylä, 2005.

Petr Lachout. *Teorie pravděpodobnosti*, volume 2. Univerzita Karlova v Praze, 2004.

RM Murray, Z Li, SS Sastry, and SS Shankara. Lyapunov stability theory. *Caltech*, 1993.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Hao Wang, Zhanglei Shi, Chi-Sing Leung, and Hing Cheung So. Admm-mcp framework for sparse recovery with global convergence. 2018.

Yuqi Zhang, Haibin Zhang, and Yingjie Tian. Sparse multiple instance learning with non-convex penalty. *Neurocomputing*, 391:142–156, 2020.

A. Důkazy použitých tvrzení

Všechny důkazy v této příloze prezentujeme jako naše vlastní.

Tvrzení A.1. *Nechť f je v bodě $x \in \mathbb{R}^n$ lokálně lipschitzovská s konstantou $K > 0$. Potom*

$$\partial f(x) \neq \emptyset \text{ a } \partial f(x) \subseteq \overline{B_K(0)}.$$

Je-li navíc g další lokálně lipschitzovská funkce v bodě x, pak

$$\partial(f+g)(x) \subseteq \partial f(x) + \partial g(x).$$

Důkaz. Nechť $\eta > 0$ je takové, že pro každé $y, z \in B_\eta(x)$ platí

$$|f(y) - f(z)| \leq K \|y - z\|.$$

Podle Rademacherovy věty, viz např. Heinonen [2005, Věta 3.1], je lokálně lipschitzovská funkce diferencovatelná skoro všude. Uvažujme $n_0 \in \mathbb{N}$ takové, že $1/n_0 < \eta$. Pro $n \geq n_0, n \in \mathbb{N}$, pak můžeme najít na množině $B_{1/n}(x)$ bod y_n, ve kterém existuje gradient. Tímto způsobem jsme schopni najít nekonečnou posloupnost $\{\nabla f(y_n)\}_{n=n_0}^{\infty}$, přičemž $y_n \to x$. Zvolme $\varepsilon > 0$. Z definice gradientu dostaneme pro $u \in \mathbb{R}^n$ s dostatečně malou normou

$$|f(y_n + u) - f(y_n) - \langle \nabla f(y_n), u \rangle| \leq \varepsilon \|u\|.$$

Zvolme $u = \nabla f(y_n)/M$, kde $M > 0$ je dostatečně velké na to, aby $y_n + u \in B_\eta(x)$ a vektor u měl zároveň dostatečně malou normu. Pak

$$\begin{aligned}\frac{1}{M}\|\nabla f(y_u)\|^2 &= |\langle \nabla f(y_n), u \rangle| \\ &\leq |f(y_n+u) - f(y_n) - \langle \nabla f(y_n), u\rangle| + |f(y_n+u) - f(y_n)| \\ &\leq \frac{\varepsilon}{M}\|\nabla f(y_n)\| + \frac{K}{M}\|\nabla f(y_n)\|.\end{aligned}$$

Případ $\|\nabla f(y_n)\| = 0$ je triviální. V opačném případě nerovnost vydělíme výrazem $\|\nabla f(y_n)\|/M$. Vzhledem k tomu, že ε bylo libovolně malé, dostaneme

$$\|\nabla f(y_n)\| \leq K. \tag{A.1}$$

Posloupnost $\{\nabla f(y_n)\}_{n=n_0}^{\infty}$ je proto obsažena v kompaktní množině $\overline{B_K(0)}$ a můžeme z ní vybrat konvergentní podposloupnost. Výsledná limita je z definice prvkem Clarkova diferenciálu. Tím jsme dokázali jeho neprázdnost. Existuje-li v nějakém $z \in B_\eta(x)$ gradient, musí dle nerovnosti (A.1) platit $\nabla f(z) \in \overline{B_K(0)}$. Všechny konvexní kombinace gradientů na libovolně malém okolí x jsou prvky $\overline{B_K(0)}$ a z definice Clarkova subdiferenciálu vyvodíme $\partial f(x) \subseteq \overline{B_K(0)}$.

Druhá část tvrzení plyne z Clarke [1975, Tvrzení 1.12].

□

Tvrzení A.2. *Vlastnost lokální lipschitzovskosti funkce $f\colon \mathbb{R}^n \to \mathbb{R}$ je ekvivalentní s vlastností*

$$\forall x \in \mathbb{R}^n\colon L_f(x) := \limsup_{\substack{y \to x, z \to x \\ y, z \neq x}} \frac{|f(y) - f(z)|}{\|y - z\|} < \infty. \tag{A.2}$$

Dále platí-li výše zmíněná vlastnost, jsme schopni pro každé $x \in \mathbb{R}^n$ a libovolné $\varepsilon > 0$ najít $\xi > 0$ takové, že pro libovolné $y, z \in B_\xi(x), y \neq z$, platí

$$\frac{|f(y) - f(z)|}{\|y - z\|} \leq L_f(x) + \varepsilon.$$

Důkaz. "⇒". Nechť f je na nějakém okolí $B_\eta(x), \eta > 0$, lokálně lipschitzovská s konstantou $K > 0$. Pak pro $\forall y, z \in B_\eta(x)$ platí $|f(y) - f(z)| \leq K \|y - z\|$. Vzetím libovolných $y \to x$ a $z \to x$ takových, že $y, z \neq x$, dostáváme

$$\limsup_{\substack{y \to x, z \to x \\ y, z \neq x}} \frac{|f(y) - f(z)|}{\|y - z\|} \leq K,$$

neboť za dostatečně dlouhou dobu je $y, z \in B_\eta(x)$.

"⇐". Ukážeme rovnou druhou část tvrzení, neboť z něho lokální lipschitzovskost funkce f snadno vyplyne. Postupujme sporem. Nechť existuje $x \in \mathbb{R}^n$ a $\varepsilon > 0$ takové, že pro libovolné $\eta > 0$ jsme schopni najít $y, z \in B_\eta(x), y \neq z$, splňující

$$\frac{|f(y) - f(z)|}{\|y - z\|} > L_f(x) + \varepsilon.$$

Postupně pro libovolné $n \in \mathbb{N}$ najdeme $y_n, z_n \in B_{1/n}(x), y_n \neq z_n$, takové, že

$$\frac{|f(y_n) - f(z_n)|}{\|y_n - z_n\|} > L_f(x) + \varepsilon.$$

Zřejmě pak platí $y_n, z_n \to x$ a $y_n, z_n \neq x, n \in \mathbb{N}$, a dostáváme tak

$$\lim_{n \to \infty} \frac{|f(y_n) - f(z_n)|}{\|y_n - z_n\|} \geq L_f(x) + \varepsilon,$$

což je ale spor s vlastností (A.2). Dokázaná vlastnost je identická tomu, že f je lokálně lipschitzovská na okolí bodu x s konstantou $L_f(x) + \varepsilon$, kde $\varepsilon > 0$ je libovolně zvolené číslo.

□