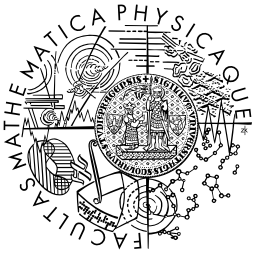


MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Marie Jelínková

Konvergence stochastického gradientu v úlohách strojového učení

Katedra pravděpodobnosti a matematické statistiky

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

Studijní program: Matematika

Studijní obor: 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 svému vedoucímu doc. RNDr. Martinu Brandovi, Ph.D. za čas věnovaný této bakalářské práci.

Název práce: Konvergence stochastického gradientu v úlohách strojového učení

Autor: Marie Jelínková

Katedra: Katedra pravděpodobnosti a matematické statistiky

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

Abstrakt: Tato práce se zabývá hledáním minima účelové funkce, která je součtem diferencovatelné (možno i nekonvexní) a obecné konvexní funkce. Zaměřili jsme se na metody stochastického a projektovaného gradientového sestupu ze strojového učení. Představujeme metodu kombinující oba přístupy. Postupně zavedeme potřebné pojmy a představíme RSPG algoritmus schopný řešit daný problém. Dokážeme jeho konvergenci pro konvexní i nekonvexní funkce. Součástí práce je i krátká numerická studie.

Klíčová slova: strojové, učení, optimalizace, stochastický, gradient, projekce

Title: Convergence of stochastic gradient descent in machine learning problems

Author: Marie Jelínková

Department: Department of Probability and Mathematical Statistics

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

Abstract: The aim of this thesis is solving minimization problems where the objective function is a sum of a differentiable (yet possibly non-convex) and general convex function. We focus on methods of stochastic and projected gradient descent from machine learning. By combining those two approaches we introduce an algorithm for solving such problems. The work is composed in a gradual manner where we firstly define necessary concepts needed for describing RSPG algorithm. Then we proceed to show the convergence of the algorithm for both convex and non-convex objective functions. A short numerical study is also included at the end.

Keywords: machine, learning, optimization, stochastic, gradient, projection

Obsah

Úvod

Strojové učení je obor zabývající se využitím velkých objemů dat s cílem pochopení principů reálného světa. K tomu využívá statistické modely, jejichž parametry odhaduje pomocí složitých funkcí mnoha proměnných. Hledání extrémů funkcí více proměnných vede často ke složitým optimalizačním úlohám. Přístupy matematické analýzy založené na hledání bodů s nulovým gradientem, mohou být obecně velmi těžko řešitelné. Z tohoto důvodu se ve strojovém učení používá metoda největšího spádu. Přesné spočítání gradientu v každém bodě je výpočetně náročné a proto se mnohé algoritmy spoléhají jen na stochastický odhad gradientu.

Samotná metoda největšího spádu nefunguje, pokud hledáme extrémy pouze na omezené konvexní množině, protože snadno můžeme vystoupit z dané množiny. Pro vynucení omezujících podmínek můžeme například po každém kroku algoritmu provést projekci zpět na množinu. Tento přístup v práci představujeme jako PG algoritmus (*angl. projected gradient*).

Hlavním cílem této práce je zaměřit se na případ funkcí, kdy je určení gradientu výpočetně náročné nebo nemožné, přestože víme, že funkce jsou diferencovatelné. Konkrétně na případ, kde účelová funkce je součtem potenciálně nekonvexní, ale diferencovatelné funkce a konvexní funkce. Pro řešení optimalizačních úloh kombinujeme odhad gradientu s metodou projekce gradientu v metodě nazvané RSPG algoritmus (*Randomized Stochastic Projected Gradient*).

V následujících kapitolách zformulujeme PG a RSPG algoritmy a dokážeme jejich vlastnosti týkající se konvergence. Na závěr provedeme numerickou studii na jednoduchém problému stochastického programování pro podpoření našich teoretických závěrů.

1. Optimalizační problém

V celé této bakalářské práci budeme primárně vycházet z práce Ghadimi a kol. (2016), kterou doplníme dalšími zdroji.

1.1 Základní definice

Normou $\|\cdot\|$ budeme rozumět obecnou euklidovskou normu, tedy pro $x \in \mathbb{R}^n$

$$\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Pro $x \in \mathbb{R}$ budeme $\lceil x \rceil$ značit horní celou část x a $\lfloor x \rfloor$ dolní celou část x.

Definice 1 (L-lipschitzovská derivace)**.** *Nechť $\Omega \subseteq \mathbb{R}^n$ je uzavřená konvexní množina a $f : \Omega \to \mathbb{R}$ je spojitě diferencovatelná funkce. Řekneme, že funkce má L-lipschitzovskou derivaci pro nějaké $L \in \mathbb{R},\ L > 0$, pokud platí*

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\| \qquad \forall x,y \in \Omega.$$

Lemma 1 (Bubeck, 2015)**.** *Buď $\Omega \subseteq \mathbb{R}^n$ je uzavřená konvexní množina a nechť $f : \Omega \to \mathbb{R}$ je spojitě diferencovatelná funkce s L-lipschitzovskou derivací pro nějaké $L \in \mathbb{R},\ L > 0$. Potom pro každé $x, y \in \Omega$ platí*

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|y - x\|^2.$$

Důkaz. Viz Bubeck, 2015, Lemma 3.4, str. 40.

□

Definice 2 (Subgradient a subdiferenciál)**.** *[Rockafellar, 1970] Nechť $h : X \to \mathbb{R}$ je konvexní funkce. Vektor g se nazývá subgradient konvexní funkce h v bodě x, pokud*

$$h(z) \geq h(x) + \langle g, z - x \rangle \qquad \forall z \in X. \tag{1.1}$$

Množinu všech subgradientů h v bodě x značíme $\partial f(x)$ a nazýváme subdiferenciál f v bodě x.

Subdiferenciál může být i prázdná množina, ale tímto případem se nebudeme zabývat, protože ho budeme uvažovat pouze pro konvexní funkce. $\partial f(x)$ je konvexní množina. Navíc x je globální minimum pouze pokud subdiferenciál obsahuje nulu. Pokud je konvexní funkce f součtem dvou konvexních funkcí h a g, potom

$$\partial f(x) = \{p + r, p \in \partial h(x), r \in \partial g(x)\}, \tag{1.2}$$

důkaz plyne z rozepsání definice pro h a r. Platí, že pokud je funkce konvexní a diferencovatelná v bodě x, tak její subdiferenciál obsahuje derivaci x. Pro h konvexní a diferencovatelnou v bodě x, tak platí

$$h(z) \geq h(x) + \langle \nabla h(x), z - x \rangle \qquad \forall z \in X. \tag{1.3}$$

Více vlastností lze například najít v Lachout (2021) str. 42 - 45, kde najdeme i důkazy předešlých tvrzení.

1.2 Definice problému

Předpoklad 1. *Předpokládejme, že $X \subseteq \mathbb{R}^n$ je uzavřená konvexní množina a $f : X \to \mathbb{R}$ je spojitě diferencovatelná funkce s L-lipschitzovskou derivací pro nějaké $L \in \mathbb{R}$, $L > 0$. Dále nechť $h : X \to \mathbb{R}$ je konvexní funkce se známým předpisem a pro každé $x \in X$ platí, že $f(x) + h(x)$ je konečné.*

V této práci budeme uvažovat problém

$$\psi^* := \min_{x \in X} \{\psi(x) := f(x) + h(x)\}. \tag{1.4}$$

za platnosti předpokladu 1.

Ačkoli známe tvar h, tak nemusí být hladká, například $h(x) = |x|$. Můžeme ale volit i $h(x) \equiv 0$.

Funkce f naopak nemusí být konvexní. Přestože gradient f existuje, tak ho nemusíme mít k dispozici, protože počítat ho může být výpočetně náročné. Máme k dispozici pouze stochastický odhad gradientu, který dostaneme pomocí borelovské funkce G, zavedené v následující podmínce.

Předpoklad 2 (Odhad gradientu)**.** *Mějme zvolenou konstantu $\sigma > 0$ a funkci f. Předpokládejme že, existuje pravděpodobnostní prostor $(\Xi, \mathcal{B}, \lambda)$, kde $\Xi \subset \mathbb{R}^d$ a $\mathcal{B}$ je borelovská $\sigma-$algebra na množině Ξ. Dále předpokládejme, že pro každý bod $x \in X$ existuje borelovská funkce $G : \mathbb{R}^n \times \Xi \to \mathbb{R}^n$, splňující pro dané $x \in X$*

$$a)\ \mathbb{E}_\xi[G(x,\xi)] = \nabla f(x) \tag{1.5}$$

$$b)\ \mathbb{E}_\xi[\|G(x,\xi) - \nabla f(x)\|^2] \leq \sigma^2. \tag{1.6}$$

Volbou měřitelné funkce G a pravděpodobnostního prostoru $(\Xi, \mathcal{B}, \lambda)$ se v této práci zabývat nebudeme.

1.3 Projekce a proximální zobrazení

Množinou $X \subset R^n$ budeme vždy rozumět uzavřenou konvexní množinu. V této sekci definujeme projekce a proximální zobrazení, které budeme dále používat v následujících algoritmech. Projekce budeme používat při hledání řešení k projektování zpět na množinu X. Proximální zobrazení nám umožňuje následně volit různé metriky.

Definice 3. *Konvexní funkce $h : X \to \mathbb{R}$ se nazývá silně konvexní funkce s parametrem $\alpha > 0$, pokud splňuje pro každé $x, y \in X$ a každé $p \in \partial h(x)$*

$$h(y) \geq h(x) + \langle p, y - x\rangle + \frac{\alpha}{2}\|y - x\|^2. \tag{1.7}$$

Tato definice je ekvivalentní s vlastností

$$\langle x - y, p - g\rangle \geq \alpha \parallel x - y \parallel^2 \qquad \forall x,y \in X, p \in \partial h(x), g \in \partial h(y), \tag{1.8}$$

důkaz plyne jen z rozepsání (1.7) podle x a y.

Lemma 2. *Nechť $h(x)$ je konvexní funkce a $g(x)$ je silně konvexní funkce s parametrem $\alpha > 0$. Potom funkce $f(x) = g(x) + h(x)$ je silně konvexní s parametrem $\alpha > 0$.*

Důkaz. Protože je $h(x)$ konvexní tak platí nerovnost (1.1) pro každé $p \in \partial h(x)$, použitím silné konvexity $g(x)$ a nerovnosti (1.7) z definice 3 dostáváme pro každé $r \in \partial g(x)$

$$g(y) + h(y) \geq g(x) + h(x) + \langle r, y - x\rangle + \langle p, y - x\rangle + \frac{\alpha}{2}\|y - x\|^2.$$

Potom podle (1.2) označíme $v = p + r$ a získáváme

$$f(y) \geq f(x) + \langle v, y - x\rangle + \frac{\alpha}{2}\|y - x\|^2.$$

Tato nerovnost platí pro každé $v \in \partial f(x)$, protože platí (1.2) a předchozí nerovnosti platí pro všechna $p \in \partial h(x)$ a $r \in \partial g(x)$.

□

Definice 4 (Funkce generující vzdálenost)**.** *Funkce $\omega : X \to \mathbb{R}$ se nazývá funkce generující vzdálenost s modulem spojitosti $\alpha > 0$, pokud je spojitě diferencovatelná a silně konvexní s parametrem $\alpha > 0$.*

Definice 5 (Proximální zobrazení)**.** *Nechť ω je funkce generující vzdálenost s modulem spojitosti $\alpha > 0$. Potom proximální zobrazení přidružené ω definujeme*

$$V(x,z) = \omega(x) - (\omega(z) + \langle \nabla\omega(z), x - z\rangle). \tag{1.9}$$

V jiné literatuře se proximální zobrazení nazývá také Bregmanova divergence nebo Bregmanova vzdálenost (Bregman (1967)). Protože je proximální zobrazení definované funkcí generující vzdálenost ω, která je silně konvexní, a tedy i konvexní, a spojitá, tak je proximální zobrazení podle (1.3) vždy nezáporné číslo. Navíc protože proximální zobrazení definujeme přidružené vůči silně konvexní funkci, tak $V(x,y) \geq \frac{\alpha}{2}\|x - y\|^2$ pro každé $x, y \in X$. Dále je proximální zobrazení silně konvexní v první proměnné, protože ω je silně konvexní funkce, nemusí ale být obecně silně konvexní v druhé proměnné.

Proximální zobrazení není metrika, protože obecně neplatí trojúhelníková nerovnost. Platí ale následující identita, kde důkaz jsme rozepsali sami.

Lemma 3. *Nechť proximální zobrazení $V(x,y)$ je z definice 5, potom pro každé $u, x, y \in X$ platí*

$$V(u, x) = V(u, y) + V(y, x) + \langle \nabla\omega(y) - \nabla\omega(x), u - y\rangle.$$

Důkaz. Zvolme libovolné $u, x, y \in X$, počítejme:

$$\begin{aligned}
V(u, x) =& \omega(u) - \omega(x) - \langle \nabla\omega(x), u - x\rangle \\
=& \omega(u) - \omega(y) - \langle \nabla\omega(y), u - y\rangle + \omega(y) - \omega(x) - \langle \nabla\omega(x), y - x\rangle \\
&- \langle \nabla\omega(x), u - x\rangle + \langle \nabla\omega(y), u - y\rangle + \langle \nabla\omega(x), y - x\rangle \\
=& V(u, y) + V(y, x) + \langle \nabla\omega(y), u - y\rangle - \langle \nabla\omega(x), u - y\rangle \\
=& V(u, y) + V(y, x) + \langle \nabla\omega(y) - \nabla\omega(x), u - y\rangle,
\end{aligned}$$

kde v jednotlivých rovnostech využíváme definici 5 a vlastností skalárního součinu.

□

Přestože proximální zobrazení není metrika, tak můžeme různou volbou ω volit určité metriky. Je výhodné postupovat s proximálním zobrazením, protože ve strojovém učení se používá více metrik. Pokud zvolíme za

$$\omega(x) = \frac{\|x\|_2^2}{2},$$

tak dosazením a snadnými úpravami skalárního součinu dostaneme euklidovskou metriku

$$V(x,z) = \frac{1}{2}\|x - z\|_2^2.$$

Nebo můžeme zvolit jinou, například kosinovou metriku. Další příklady lze najít například v Banerjee a kol. (2005), tabulka 1.

Definice 6 (Zobecněná projekce)**.** *Buď $h : X \to \mathbb{R}$ konvexní funkce, $\gamma > 0$, $g \in \mathbb{R}^n$ a $x \in X$ a , potom definujme zobecněnou projekci jako*

$$x^+ = \underset{u \in X}{\operatorname{argmin}}\{\langle g,u\rangle + \frac{1}{\gamma}V(u,x) + h(u)\}. \tag{1.10}$$

.

Pokud uvažujeme $V(x,z) = \frac{1}{2}\|x-z\|_2^2$ a $h \equiv 0$, tak x^+ je ortogonální projekce.

Povšimněme si ještě, že funkce $\langle g,u\rangle + \frac{1}{\gamma}V(u,x) + h(u)$ je pro fixní x a g silně konvexní funkce podle Lemma 2, protože skalární součin s fixní první proměnou je lineární funkce a tedy i konvexní.

V tomto článku předpokládáme, že proximální zobrazení je definované tak, že (1.10) je snadno řešitelný, pro jakékoli $\gamma > 0, g \in \mathbb{R}$ a $x \in X$.

Definujme dále pomocnou funkci P_X jako

$$P_X(x,g,\gamma) = \frac{1}{\gamma}(x - x^+), \tag{1.11}$$

kde x^+ je z (1.10) z definice 6.

Pokud za g dosadíme gradient funkce f v bodě x dostáváme zobecněný projektovaný gradient na množinu X, speciálně pokud za množinu X volíme $\mathbb{R}^n$, tak $P_X = \nabla f(x)$.

1.4 Vlastnosti generalizované projekce

V této sekci si ukážeme nějaké vlastnosti generalizované projekce, které následně budeme používat v důkazu konvergence dále zavedeného algoritmu.

Lemma 4 (Podmínka optimality prvého řádu)**.** *Nechť $X \subset \mathbb{R}^n$ je konvexní množina, $f : X \to \mathbb{R}$ je konvexní funkce. Potom x^+ je globální minimum funkce f na X, právě tehdy když existuje $p \in \partial f(x^+)$ takové, že pro každé $x \in X$ platí $\langle p, x - x^+\rangle \geq 0$.*

Důkaz. Viz Lachout (2021), věta 3.3., str 50.

□

Lemma 5. *Nechť x^+ je z definice generalizované projekce 6. Potom pro všechna $x \in X, g \in \mathbb{R}^n$ a $\gamma > 0$ platí*

$$\langle g, P_X(x,g,\gamma)\rangle \geq \alpha\|P_X(x,g,\gamma)\|^2 + \frac{1}{\gamma}[h(x^+) - h(x)]. \tag{1.12}$$

Důkaz. Důkaz přebíráme z Ghadimi a kol. (2016). Rozepsali jsme proč platí nerovnost 1.14. Nejprve vypočteme derivaci $V(x,z)$ podle x.

$$\frac{\partial V(x,z)}{\partial x} = \nabla\omega(x) - \nabla\omega(z). \tag{1.13}$$

Podle Lemmatu 2 je funkce $\langle g,u\rangle + \frac{1}{\gamma}V(u,x) + h(u)$ silně konvexní a tedy i konvexní. Z lemmatu 4 a toho, že x^+ je optimální, dostáváme $p \in \partial h(x^+)$ takové, že platí

$$\langle g + \frac{1}{\gamma}[\nabla\omega(x^+) - \nabla\omega(x)] + p, u - x^+\rangle \geq 0 \qquad \forall u \in X. \tag{1.14}$$

Protože daná nerovnost platí pro všechna $u \in X$, můžeme položit $u = x$ a jednoduchou úpravou skalárního součinu dostáváme

$$\begin{aligned}\langle g, x - x^+\rangle &\geq \frac{1}{\gamma}\langle\nabla\omega(x^+) - \nabla\omega(x), x^+ - x\rangle + \langle p, x^+ - x\rangle \\ &\geq \frac{\alpha}{\gamma}\|x^+ - x\|^2 + h(x^+) - h(x),\end{aligned}$$

kde v druhé nerovnosti používáme definici subdiferenciálu a vlastnosti silně konvexní funkce s derivací (1.8). Dosazením $x^+ - x = \gamma P_X(x,g,\gamma)$ z definice a vynásobením $\gamma > 0$ dostáváme požadovanou nerovnost.

□

Lemma 6. *Nechť x_1^+ respektive x_2^+ jsou dány vzorcem* (1.10)*, kde nahradíme g funkcemi g_1 respektive g_2. Potom platí*

$$\|x_2^+ - x_1^+\| \leq \frac{\gamma}{\alpha}\|g_2 - g_1\|. \tag{1.15}$$

Důkaz. Viz Ghadimi a kol. (2016) Lemma 2.

□

Z tohoto lemmatu plyne, že $P_X(x,\cdot,\gamma)$ je pro fixní γ a x $\frac{1}{\alpha}-$Lipschitzovská.

Důsledek 7. *Nechť $P_X(x,g,\gamma)$ je definovaná jako v* (1.11)*. Potom pro všechna g_1 a g_2 v $\mathbb{R}^n$ platí*

$$\|P_X(x,g_1,\gamma) - P_X(x,g_2,\gamma)\| \leq \frac{1}{\alpha}\|g_2 - g_1\|$$

Důkaz. Pouze dosadíme do předchozího lemmatu viz Ghadimi a kol. (2016) Tvrzení 1.

□

Lemma 8. *Nechť x^+ je dáno v definici 6, potom pro každé $u \in X$ platí*

$$\langle g, x^+\rangle + h(x^+) + \frac{1}{\gamma}V(x^+,x) \leq \langle g, u\rangle + h(u) + \frac{1}{\gamma}[V(u,x) - V(u,x^+)].$$

Důkaz. V tomto důkazu jsme použili ideu z důkazu lemmatu 1 v Lan (2011), avšak důkaz jsme vedli s jinou funkcí a rozepsali jsme některé kroky detailněji.

Součet silně konvexní funkce a konvexní funkce je silně konvexní funkce podle lemmatu 2. Tedy funkce $\langle g,\cdot\rangle + \frac{1}{\gamma}V(\cdot,x) + h(\cdot)$ je silně konvexní. Protože platí 1.13 a předpokládáme, že x^+ je minimum, tak podle Lemmatu 4 existuje $p \in \partial h(x^+)$ takové, že

$$\langle g + p + \frac{1}{\gamma}(\nabla\omega(x^+) - \nabla\omega(x)), u - x^+\rangle \geq 0. \tag{1.16}$$

Dále podle lemmatu 3 platí identita pro $V(u,x)$. Můžeme tedy psát $\forall u \in X$

$$\begin{aligned}
&\langle g, u\rangle + h(u) + \frac{1}{\gamma}V(u,x) \\
&\quad = \langle g, u\rangle + h(u) + \frac{1}{\gamma}[V(u,x^+) + V(x^+,x) + \langle\nabla\omega(x^+) - \nabla\omega(x), u - x^+\rangle] \\
&\quad \geq \langle g, x^+\rangle + h(x^+) + \langle g + p + \frac{1}{\gamma}(\nabla\omega(x^+) - \nabla\omega(x)), u - x^+\rangle \\
&\qquad + \frac{1}{\gamma}[V(u,x^+) + V(x^+,x)] \\
&\quad \geq \langle g, x^+\rangle + h(x^+) + \frac{1}{\gamma}[V(u,x^+) + V(x^+,x)].
\end{aligned}$$

První nerovnost platí, protože jsou $\langle g, u\rangle$ a $h(u)$ konvexní funkce (1.3). Druhá nerovnost platí z (1.16). Výslednou nerovnost dostaneme převedením členů v dokázané nerovnosti.

□

2. Algoritmus založený na projektovaném gradientu

V této kapitole budeme uvažovat optimalizační problém z první kapitoly (1.4) za předpokladu 1. Budeme předpokládat, že známe přesný gradient funkce f v každém bodě $x \in X$. Zavedeme PG algoritmus (*angl. projected gradient*) a dokážeme jeho konvergenci. Ghadimi a kol. (2016)

2.1 Algoritmus

Pro funkce z předpokladu 1 a za předpokladu, že známe gradient funkce f v každém bodě $x \in X$, zavádíme následující algoritmus.

Algoritmus 1 PG algoritmus

Ghadimi a kol. (2016)

Vstup: Počáteční bod $x_1 \in X$, počet iterací $N \in \mathbb{N}$ a délka kroků $\{\gamma_k\}_{k=1}^N, k \in \mathbb{N}$, kde $\gamma_k > 0$.

Pro $k = 1, ..., N$:

$x_{k+1} = \underset{u \in X}{\operatorname{argmin}}\{\langle \nabla f(x_k), u\rangle + \frac{1}{\gamma_k} V(u, x_k) + h(u)\}$

$g_{X,k} = P_X(x_k, \nabla f(x_k), \gamma_k)$

Výstup: $x_R \in \{x_1, \ldots, x_N\}$ takové, že $R = \underset{k \in \{1,\ldots,N\}}{\operatorname{argmin}} \|g_{X,k}\|$

V praxi se často používá jako výstup poslední x_{N+1}, nebo nejmenší hodnota $\psi(x_k)$ pro nějaké $k = 1, ..., N$. Protože ale f může být nekonvexní, tak bychom nemohli zaručit konvergenci algoritmu.

2.2 Důkaz konvergence

V algoritmu jsme zatím uvažovali pouze obecné délky kroků. Ukážeme, že pokud zvolíme správnou délku kroků, pak algoritmus konverguje. Následující větu přebíráme z Ghadimi a kol. (2016) věta 1, důkaz je proveden obdobně jako v daném článku, pouze v posledním kroku důkazu 2.2 jsme opravili nepřesnost.

Věta 9. *Předpokládejme že je splněn předpoklad 1 a funkce f má známý gradient v každém bodě $x \in X$. Zvolme počáteční bod x_1 a definujme si $D_\Psi := [\frac{\Psi(x_1) - \Psi^*}{L}]^{1/2}$. Pro výstup algoritmu x_R označme $g_{X,R} = P_X(x_R, \nabla f(x_R), \gamma_R)$ podle* (1.11).

Nechť délka kroku $\{\gamma_k\}$ v PG algoritmu 1 je zvolena tak, že pro každé k platí $0 < \gamma_k \leq 2\alpha/L$ a alespoň pro jedno k platí $\gamma_k < 2\alpha/L$. Potom platí

$$\|g_{X,R}\|^2 \leq \frac{LD_\Psi^2}{\sum_{k=1}^N (\alpha\gamma_k - \frac{L\gamma_k^2}{2})}$$

Důkaz. Zvolme $k \in \{1, 2, \ldots, N\}$. Protože má funkce f L-lipschitzovskou derivaci

a je spojitě diferencovatelná, tak podle Lemma 1 platí

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|y - x\|^2,$$

a tedy platí i

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2}\|y - x\|^2.$$

Volbou $y = x_{k+1}, x = x_k$ dostáváme

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= f(x_k) - \gamma_k \langle \nabla f(x_k), g_{X,k} \rangle + \frac{\gamma_k^2 L}{2}\|g_{X,k}\|^2, \end{aligned}$$

kde v rovnosti jsme použili definici $g_{X,k}$. Protože $x_{k+1} = x_k^+$, $\gamma_k > 0$ a $g_{X,k} = P_X(x_R, \nabla f(x_R), \gamma_R)$, tak prostřední člen odhadneme pomocí Lemma 5 a získáváme

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \gamma_k[\alpha\|g_{X,k}\|^2 + \frac{1}{\gamma_k}(h(x_{k+1}) - h(x_k))] + \frac{\gamma_k^2 L}{2}\|g_{X,k}\|^2 \\ f(x_{k+1}) + h(x_{k+1}) &\leq f(x_k) + h(x_k) - (\gamma_k\alpha - \frac{\gamma_k^2 L}{2})\|g_{X,k}\|^2. \end{aligned}$$

To je podle definice

$$\Psi(x_{k+1}) \leq \Psi(x_k) - (\gamma_k\alpha - \frac{\gamma_k^2 L}{2})\|g_{X,k}\|^2. \tag{2.1}$$

Protože jsme výpočty provedli pro libovolné $k \in \{1, 2, \ldots, N\}$, tak předchozí nerovnost platí pro každé $k \in \{1, 2, \ldots, N\}$. Můžeme tedy psát

$$\begin{aligned} \Psi(x_{N+1}) &\leq \Psi(x_N) - (\gamma_N\alpha - \frac{\gamma_N^2 L}{2})\|g_{X,N}\|^2 \\ &\leq \Psi(x_{N-1}) - (\gamma_{N-1}\alpha - \frac{\gamma_{N-1}^2 L}{2})\|g_{X,N-1}\|^2 - (\gamma_N\alpha - \frac{\gamma_N^2 L}{2})\|g_{X,N}\|^2 \\ &\leq \cdots \leq \Psi(x_1) - \sum_{k=1}^{N}((\gamma_k\alpha - \frac{\gamma_k^2 L}{2})\|g_{X,k}\|^2) \\ &\leq \Psi(x_1) - \|g_{X,R}\|^2 \sum_{k=1}^{N}(\gamma_k\alpha - \frac{\gamma_k^2 L}{2}), \end{aligned} \tag{2.2}$$

kde poslední nerovnost platí z definice $R = \underset{k\in\{1,\ldots,N\}}{\operatorname{argmin}} \|g_{X,k}\|$.

Z definice Ψ^* (1.5) a úpravou nerovnosti dostáváme

$$\|g_{X,R}\|^2 \sum_{k=1}^{N}(\gamma_k\alpha - \frac{\gamma_k^2 L}{2}) \leq \Psi(x_1) - \Psi(x_{N+1}) \leq \Psi(x_1) - \Psi^*.$$

Suma je kladná, protože $\exists k : \gamma_k < 2\alpha/L$ a $(\gamma_k\alpha - \frac{\gamma_k^2 L}{2}) \geq 0$ pro každé k z předpokladu věty. Vydělením obou stran výrazem $\sum_{k=1}^{N}(\gamma_k\alpha - \frac{\gamma_k^2 L}{2})$ dostáváme požadovanou nerovnost.

□

Tímto důkazem jsme dokázali omezit shora normu $\|g_{X,R}\|$. Protože L a D_ψ jsou pro danou volbu ψ a x_1 neměnná, tak horní odhad konverguje k nule s rostoucím N, za předpokladu volby délky kroků $0 \leq \gamma_k \leq 2\alpha/L$ pro všechna $k \in \mathbb{N}$ a $\gamma_k = 2\alpha/L$ pouze pro konečně mnoho $k \in \mathbb{N}$. Omezení je také zřejmé z následujícího důsledku.

Důsledek 10. *Ghadimi a kol. (2016) Předpokládejme, že platí všechny předpoklady věty 9 a že délka kroku v PG algoritmu je $\gamma_k = \alpha/L$ pro $\forall k \in \{1, \ldots, N\}$. Potom*

$$\|g_{X,R}\|^2 \leq \frac{2L^2 D_\Psi^2}{\alpha^2 N}.$$

Důkaz. Tento důkaz uvádíme pro úplnost, je převzaný z Lan (2011) (Důsledek 1). Protože $\forall k \in \{1, \ldots, N\}$ platí $\gamma_k = \alpha/L$, dostáváme

$$\sum_{k=1}^{N} (\gamma_k \alpha - \frac{\gamma_k^2 L}{2}) = \sum_{k=1}^{N} (\frac{\alpha^2}{L} - \frac{L\alpha^2}{2L^2}) = \frac{N\alpha^2}{2L}.$$

A z věty 9 plyne

$$\|g_{X,R}\|^2 \leq \frac{LD_\Psi^2}{\sum_{k=1}^{N} (\alpha\gamma_k - \frac{L\gamma_k^2}{2})} = \frac{2L^2 D_\Psi^2}{\alpha^2 N}.$$

□

Z předchozího důsledku vidíme, že pokud chceme omezit projektovaný gradient výstupu x_R nějakým $\epsilon \in \mathbb{R}, \epsilon \geq 0$, tak můžeme v omezení z důsledku měnit pouze počet iterací a počáteční bod x_1. D_Ψ ale bez znalosti optima nelze spočítat, proto se častěji využívá horní odhad. Ten závisí na parametrech úlohy a nelze ho obecně snadno určit.

3. Algoritmus založený na projektovaném stochastickém gradientu

V této kapitole představíme RSPG algoritmus, pro funkci ψ z (1.4) za předpokladů 1 a 2. Dokážeme jeho konvergenci ve střední hodnotě a pro ψ konvexní dokážeme, že algoritmus ve střední hodnotě hledá optimální řešení problému (1.4). Budeme vycházet z Ghadimi a kol. (2016).

3.1 Algoritmus

Pro funkce z předpokladu 1 zavádíme následující algoritmus.

Algoritmus 2 RSPG algoritmus

Ghadimi a kol. (2016)

Vstup: Počáteční bod $x_1 \in X$, maximální počet iterací $N \in \mathbb{N}$, délky kroků $\{\gamma_k\}_{k=1}^N$, kde $\gamma_k > 0$, velikost náhodného výběru $\{m_k\}_{k=1}^N$, pravděpodobnostní rozdělení P_R na množině $\{1, 2, \ldots, N\}$ a odhad gradientu G z předpokladu 2.

Krok: Nechť R je hodnota náhodné proměnné z pravděpodobnostního rozdělení P_R.

Pro $k = 1, ..., R$:

Odhadni m_k-krát gradient pomocí nezávislých realizací $G(x_k,\xi_{k,i})$, splňující podmínku 2, $i = 1, \cdots, m_k$.

Polož $G_k = \frac{1}{m_k} \sum_{i=1}^{m_k} G(x_k,\xi_{k,i})$.

Spočti $x_{k+1} = \underset{u \in X}{\operatorname{argmin}}\{\langle G_k, u\rangle + \frac{1}{\gamma_k} V(u, x_k) + h(u)\}$.

Výstup: x_{R+1}.

Algoritmus jsme oproti verzi RSPG algoritmu z Ghadimi a kol. (2016) drobně upravili, abychom mohli použít důkaz uvedený v článku.

Protože postupujeme podle odhadu gradientu G_k, který splňuje pouze předpoklad 2, tak v každém kroku se nutně nemusí zmenšit vzdálenost k optimu $\psi(x_k) - \psi^*$ (1.4) oproti předchozímu kroku. Podle podmínky 2 ale můžeme chybu omezit ve střední hodnotě a jsme tedy schopni dokázat konvergenci algoritmu k optimu pouze ve střední hodnotě.

3.2 Konvergence stochastického gradientu

Věta 11. *Ghadimi a kol. (2016) Nechť jsou splněny předpoklady 1 a 2.*

Předpokládejme že délky kroků $\{\gamma_k\}_{k=1}^N$ v RSPG algoritmu volíme tak, že pro každé $k \in \{1, 2, \ldots, N\}$ platí $0 < \gamma_k \leq \alpha/L$ a alespoň pro jedno k platí $\gamma_k < \alpha/L$. Nechť pravděpodobnostní rozdělení P_R je zvoleno jako

$$P_R(k) = P(R = k) = \frac{\alpha\gamma_k - L\gamma_k^2}{\sum_{k=1}^N (\alpha\gamma_k - L\gamma_k^2)}. \tag{3.1}$$

Zvolme počáteční bod $x_1 \in X$, označme $D_\Psi := [\frac{\Psi(x_1)-\Psi^*}{L}]^{1/2}$ a pro P_X z (1.11) označme pro každé $k \in \{1, 2, \ldots, N\}$

$$\hat{g}_{X,k} = P_X(x_k, G(x_k), \gamma_k),$$
$$g_{X,k} = P_X(x_k, \nabla f(x_k), \gamma_k).$$

Pro každé $k \in \{1, 2, \ldots, N\}$ označme

$$\xi_k := (\xi_{k,1}, \xi_{k,2}, \ldots, \xi_{k,m_k}),$$
$$\xi_{[k]} := (\xi_1, \xi_2, \ldots, \xi_k),$$

kde $\xi_{k,i} \in \Xi$ jsou z dané nezávislé realizace odhadu gradientu v bodě $f(x_k)$ v RSPG algoritmu 2 za předpokladu 2.

Potom platí

$$\mathbb{E}_{R,\xi_{[N]}}\left[\|\hat{g}_{X,R}\|^2\right] \leq \frac{LD_\Psi^2 + \frac{\sigma^2}{\alpha}\sum_{k=1}^N \frac{\gamma_k}{m_k}}{\sum_{k=1}^N(\alpha\gamma_k - L\gamma_k^2)}. \tag{3.2}$$

Důkaz.

Důkaz provedeme podobně jako ve větě 2 v Ghadimi a kol. (2016). Důkaz rozdělíme na 4 části. V první části si odvodíme klíčovou nerovnost. Tento krok je shodný s důkazem z článku, až na opravení nepřesnosti na konci. V druhé části doplníme důkaz nulové střední hodnoty druhého členu výrazu na pravé straně nerovnosti. Ve třetí části omezíme poslední člen nerovnosti, tuto část přejímáme a pouze jsme doplnili argumenty na počítání střední hodnoty. V poslední části shrneme dosažené výsledky z předchozího. Poslední část uvádíme pro úplnost a přejímáme jí z Ghadimi a kol. (2016).

1. část

Označme $\delta_k \equiv G_k - \nabla f(x_k)$ pro $k \in \{1, 2, \ldots, N\}$ a zvolme fixní k. Protože má funkce f L-lipschitzovskou derivaci a je spojitě diferencovatelná, tak podle Lemma 1 platí

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= f(x_k) - \gamma_k\langle \nabla f(x_k), \hat{g}_{X,k}\rangle + \frac{\gamma_k^2 L}{2}\|\hat{g}_{X,k}\|^2 \\ &= f(x_k) - \gamma_k\langle G(x_k), \hat{g}_{X,k}\rangle + \frac{\gamma_k^2 L}{2}\|\hat{g}_{X,k}\|^2 + \gamma_k\langle \delta_k, \hat{g}_{X,k}\rangle, \end{aligned}$$

kde první rovnost plyne z definice $\hat{g}_{X,k}$ a druhá z definice δ_k. Dosazením do lemmatu 5 pro $x = x_k, \gamma = \gamma_k > 0$, $g = G_k$ a $x_{k+1} = x_k^+$ získáme

$$f(x_{k+1}) \leq f(x_k) - \gamma_k\alpha\|\hat{g}_{X,k}\|^2 - [h(x_{k+1}) - h(x_k)] + \frac{\gamma_k^2 L}{2}\|\hat{g}_{X,k}\|^2 + \gamma_k\langle \delta_k, \hat{g}_{X,k}\rangle$$

a po převedení členu $h(x_{k+1})$ na druhou stranu obdržíme

$$f(x_{k+1}) + h(x_{k+1}) \leq f(x_k) + h(x_k) - \gamma_k\alpha\|\hat{g}_{X,k}\|^2 + \frac{\gamma_k^2 L}{2}\|\hat{g}_{X,k}\|^2 + \gamma_k\langle \delta_k, \hat{g}_{X,k}\rangle,$$

a tedy s použitím definice ψ (1.4) a rozšířením o $\gamma_k\langle\delta_k, g_{X,k}\rangle - \gamma_k\langle\delta_k, g_{X,k}\rangle$ získáváme

$$\begin{aligned}\psi(x_{k+1}) &\leq \psi(x_k) - \gamma_k\alpha\|\hat{g}_{X,k}\|^2 + \frac{\gamma_k^2 L}{2}\|\hat{g}_{X,k}\|^2 + \gamma_k\langle\delta_k, \hat{g}_{X,k}\rangle \\ &= \psi(x_k) - \Big(\gamma_k\alpha - \frac{\gamma_k^2 L}{2}\Big)\|\hat{g}_{X,k}\|^2 + \gamma_k\langle\delta_k, g_{X,k}\rangle + \gamma_k\langle\delta_k, \hat{g}_{X,k} - g_{X,k}\rangle.\end{aligned}$$

Z Cauchyovy-Schwarzovy nerovnosti dostáváme

$$\psi(x_{k+1}) \leq \psi(x_k) - \Big(\gamma_k\alpha - \frac{\gamma_k^2 L}{2}\Big)\|\hat{g}_{X,k}\|^2 + \gamma_k\langle\delta_k, g_{X,k}\rangle + \gamma_k\|\delta_k\|\|\hat{g}_{X,k} - g_{X,k}\|.$$

Důsledek 7 nám pro $x = x_k, g_1 = G_k, g_2 = \nabla f(x_k), \hat{g}_{X,k} = P_X(x_k, G_k,\gamma_k)$ a $g_{X,k} = P_X(x_k, \nabla f(x_k),\gamma_k)$ dává vztah

$$\|\hat{g}_{X,k} - g_{X,k}\| \leq \frac{1}{\alpha}\|G_k - \nabla f(x_k)\| = \frac{1}{\alpha}\|\delta_k\|.$$

Dosazením vztahu do předchozí nerovnosti získáváme

$$\psi(x_{k+1}) \leq \psi(x_k) - \Big(\gamma_k\alpha - \frac{\gamma_k^2 L}{2}\Big)\|\hat{g}_{X,k}\|^2 + \gamma_k\langle\delta_k, g_{X,k}\rangle + \frac{\gamma_k}{\alpha}\|\delta_k\|^2,$$

a tedy převedením některých členů

$$\Big(\gamma_k\alpha - \frac{\gamma_k^2 L}{2}\Big)\|\hat{g}_{X,k}\|^2 \leq \psi(x_k) - \psi(x_{k+1}) + \gamma_k\langle\delta_k, g_{X,k}\rangle + \frac{\gamma_k}{\alpha}\|\delta_k\|^2.$$

Poznamenejme ještě, že $\gamma_k\alpha - \gamma_k^2 L \leq \gamma_k\alpha - \frac{\gamma_k^2 L}{2}$ pro každé $k \in \{1, 2, \ldots, N\}$, protože $\gamma_k \leq \frac{\alpha}{L}$. Předchozí nerovnost platí pro každé $k \in \{1, 2, \ldots, N\}$, můžeme dané nerovnosti sečíst a dostáváme

$$\begin{aligned}\sum_{k=1}^{N}(\gamma_k\alpha - \gamma_k^2 L)\|\hat{g}_{X,k}\|^2 &\leq \sum_{k=1}^{N}\Big(\gamma_k\alpha - \frac{\gamma_k^2 L}{2}\Big)\|\hat{g}_{X,k}\|^2 \\ &\leq \sum_{k=1}^{N}(\psi(x_k) - \psi(x_{k+1})) + \sum_{k=1}^{N}[\gamma_k\langle\delta_k, g_{X,k}\rangle + \frac{\gamma_k}{\alpha}\|\delta_k\|^2] \\ &= \psi(x_1) - \psi(x_{N+1}) + \sum_{k=1}^{N}[\gamma_k\langle\delta_k, g_{X,k}\rangle + \frac{\gamma_k}{\alpha}\|\delta_k\|^2].\end{aligned}$$

Díky definici ψ^* (1.4) můžeme pracovat s nerovností

$$\sum_{k=1}^{N}(\gamma_k\alpha - \gamma_k^2 L)\|\hat{g}_{X,k}\|^2 \leq \psi(x_1) - \psi^* + \sum_{k=1}^{N}\gamma_k\langle\delta_k, g_{X,k}\rangle + \sum_{k=1}^{N}\frac{\gamma_k}{\alpha}\|\delta_k\|^2. \qquad (3.3)$$

2. část

Nyní spočteme střední hodnotu členu $\sum_{k=1}^{N}\langle\delta_k, g_{X,k}\rangle$ z předchozí nerovnosti. Budeme zkráceně psát, že funkce závisí na $\xi_{[k]}$, ale budeme tím myslet, že závisí na realizaci náhodné veličiny $G(x_i,\xi_{i,j}),\ i = 1, \ldots, k,\ j = 1, \ldots m_i$.

Poznamenejme, že x_k závisí jen na $\xi_{[k-1]}$ pro fixní x_1, $\{\gamma_k\}_{k=1}^{N}$ a funkci $\psi(x)$. Proto je x_k také náhodnou veličinou pro každé k. Podle první částí (1.5) předpokladu 2 pro odhad gradientu a nezávislosti platí

$$\mathbb{E}_{\xi_{[k]}}\Big[G_k - \nabla f(x_k)\Big|\xi_{[k-1]}\Big] = \frac{1}{m_k}\sum_{i=1}^{m_k}\mathbb{E}_{\xi_{[k]}}\Big[G(x_k,\xi_{k,i}) - \nabla f(x_k)\Big|\xi_{[k-1]}\Big] = \vec{0} \in \mathbb{R}^n.$$

Podle této nerovnosti můžeme počítat

$$\mathbb{E}_{\xi_{[k]}}\Big[\langle\delta_k, g_{X,k}\rangle\Big|\xi_{[k-1]}\Big]=\mathbb{E}_{\xi_{[k]}}\Big[\langle G_k-\nabla f(x_k), P_X(x_k,\nabla f(x_k),\gamma_k)\rangle\Big|\xi_{[k-1]}\Big]=0,$$

kde poslední nerovnost plyne z linearity skalárního součinu pro fixní druhý člen. Zároveň můžeme střední hodnotu vůči $\xi_{[k]}$ nahradit $\xi_{[N]}$, protože G_k a x_k nezávisí na $\xi_{k+1},\dots,\xi_N$.

Protože $\xi_{[N]}$ obsahuje všechny realizace $\xi_{[k]}$ tak i $\sigma-$algebra generovaná všemi realizacemi $G(x_i,\xi_{i,j}),\ i=1,\dots k\ j=1,\dots m_i$ bude podmnožinou $\sigma-$algebry generovanou všemi realizacemi $G(x_i,\xi_{i,j}),\ i=1,\dots N,\ j=1,\dots m_i$ pro každé k. Podle rovnosti $\mathbb{E}[X]=\mathbb{E}[\mathbb{E}[X|\mathcal{F}]]$ pro $X\in L_1(S,\mathcal{S},\mu)$ a $\mathcal{F}\subseteq\mathcal{S}$ píšeme

$$\mathbb{E}_{\xi_{[N]}}\Big[\sum_{k=1}^{N}\langle\delta_k, g_{X,k}\rangle\Big]=\sum_{k=1}^{N}\mathbb{E}_{\xi_{[N]}}\Big[\langle\delta_k, g_{X,k}\rangle\Big]=\sum_{k=1}^{N}\mathbb{E}_{\xi_{[N]}}\Big[\mathbb{E}_{\xi_{[N]}}\Big[\langle\delta_k, g_{X,k}\rangle\Big|\xi_{[k-1]}\Big]\Big]=0. \tag{3.4}$$

3. část

V této části odhadneme ve střední hodnotě poslední člen nerovnosti (3.3).

Označme $\delta_{k,i}\equiv G(x_k,\xi_{k,i})-\nabla f(x_k)$ a $S_{k,j}=\sum_{i=1}^{j}\delta_{k,i}$ pro $i,j=1,\dots,m_k$ a pro $k=1,\dots,N$. Dále $S_{k,0}=0\ \forall k=1,\dots,N$. Protože $\mathbb{E}_{\xi_{[N]}}[\delta_{k,i}]=\vec{0}\in\mathbb{R}^n$ analogickým důkazem jako v druhé části a za předpokladu 2, tak

$$\mathbb{E}_{\xi_{[N]}}\Big[\langle S_{k,i-1},\delta_{k,i}\rangle\Big]=\mathbb{E}_{\xi_{[N]}}\Big[\mathbb{E}_{\xi_{[N]}}\Big[\langle S_{k,i-1},\delta_{k,i}\rangle\Big|S_{k,i-1}\Big]\Big]=0\qquad\forall i=1,\dots,m_k.$$

Podle předchozího platí

$$\begin{aligned}\mathbb{E}_{\xi_{[N]}}\Big[\|S_{k,m_k}\|^2\Big]&=\mathbb{E}_{\xi_{[N]}}\Big[\|S_{k,m_k-1}+\delta_{k,m_k}\|^2\Big]\\&=\mathbb{E}_{\xi_{[N]}}\Big[\|S_{k,m_k-1}\|^2+\|\delta_{k,m_k}\|^2+2\langle S_{k,m_k-1},\delta_{k,m_k}\rangle\Big]\\&=\mathbb{E}_{\xi_{[N]}}\Big[\|S_{k,m_k-1}\|^2\Big]+\mathbb{E}_{\xi_{[N]}}\Big[\|\delta_{k,m_k}\|^2\Big]\\&=\dots\\&=\sum_{j=1}^{m_k}\mathbb{E}_{\xi_{[N]}}\Big[\|\delta_{k,j}\|^2\Big].\end{aligned}$$

Z tohoto výsledku a z druhé části (1.6) podmínky 2 dostáváme

$$\begin{aligned}\mathbb{E}_{\xi_{[N]}}\Big[\|\delta_k\|^2\Big]&=\mathbb{E}_{\xi_{[N]}}\Big[\|\frac{1}{m_k}\sum_{i=1}^{m_k}\delta_{k,i}\|^2\Big]=\frac{1}{m_k^2}\mathbb{E}_{\xi_{[N]}}\Big[\|S_{k,m_k}\|^2\Big]\\&=\frac{1}{m_k^2}\sum_{i=1}^{m_k}\mathbb{E}_{\xi_{[N]}}\Big[\|\delta_{k,i}\|^2\Big]\le\frac{\sigma^2}{m_k},\end{aligned} \tag{3.5}$$

protože platí stejně jako v předchozí části za předpokladu 2

$$\mathbb{E}_{\xi_{[N]}}\Big[\|\delta_{k,i}\|^2\Big]=\mathbb{E}_{\xi_{[N]}}\Big[\mathbb{E}_{\xi_{[N]}}\Big[\|\delta_{k,i}\|^2\Big|\xi_{[k-1]}\Big]\Big]\le\sigma^2.$$

Můžeme tedy omezit

$$\mathbb{E}_{\xi_{[N]}}\Big[\sum_{k=1}^{N}\frac{\gamma_k}{\alpha}\|\delta_k\|^2\Big]=\sum_{k=1}^{N}\frac{\gamma_k}{\alpha}\mathbb{E}_{\xi_{[N]}}\Big[\|\delta_k\|^2\Big]\le\sum_{k=1}^{N}\frac{\sigma^2\gamma_k}{\alpha m_k}. \tag{3.6}$$

4. část

Použitím vlastnosti střední hodnoty $X \leq Y s.j.$ implikuje $EX \leq EY s.j$ na nerovnost (3.3) z 1. části důkazu a použitím výsledků z druhé a třetí části získáváme

$$\mathbb{E}_{\xi_{[N]}}\Big[\sum_{k=1}^{N}(\gamma_k\alpha - \gamma_k^2 L)\|\hat{g}_{X,k}\|^2\Big] \leq \mathbb{E}_{\xi_{[N]}}\Big[\psi(x_1) - \psi^* + \sum_{k=1}^{N}[\gamma_k\langle\delta_k, g_{X,k}\rangle + \frac{\gamma_k}{\alpha}\|\delta_k\|^2\Big]$$
$$\sum_{k=1}^{N}(\gamma_k\alpha - \gamma_k^2 L)\mathbb{E}_{\xi_{[N]}}\Big[\|\hat{g}_{X,k}\|^2\Big] \leq \psi(x_1) - \psi^* + \frac{\sigma^2}{\alpha}\sum_{k=1}^{N}\frac{\gamma_k}{m_k}.$$

Podle podmínek na délky kroků $\{\gamma_k\}_{k=1}^N$ je $\sum_{k=1}^{N-1}(\gamma_k\alpha - \gamma_k^2 L) > 0$, můžeme tedy tímto výrazem vydělit a obdržíme

$$\frac{\sum_{k=1}^{N}(\gamma_k\alpha - \gamma_k^2 L)\mathbb{E}_{\xi_{[N]}}\Big[\|\hat{g}_{X,k}\|^2\Big]}{\sum_{k=1}^{N-1}(\gamma_k\alpha - \gamma_k^2 L)} \leq \frac{\psi(x_1) - \psi^* + \frac{\sigma^2}{\alpha}\sum_{k=1}^{N}\frac{\gamma_k}{m_k}}{\sum_{k=1}^{N-1}(\gamma_k\alpha - \gamma_k^2 L)}.$$

Podle počítání střední hodnoty pro diskrétní náhodnou veličinu a díky volbě P_R platí rovnost

$$\begin{aligned}\mathbb{E}_{R,\xi_{[N]}}\Big[\|\hat{g}_{X,k}\|^2\Big] &= \sum_{k=1}^{N}\mathbb{E}_{\xi_{[N]}}\Big[\|\hat{g}_{X,k}\|^2\Big]P_R(R=k) \\ &= \frac{\sum_{k=1}^{N}(\gamma_k\alpha - \gamma_k^2 L)\mathbb{E}_{\xi_{[N]}}\Big[\|\hat{g}_{X,k}\|^2\Big]}{\sum_{k=1}^{N}(\gamma_k\alpha - \gamma_k^2 L)}.\end{aligned}$$

Dosazením tohoto poznatku do předchozí nerovnosti již plyne dokazované tvrzení.

□

Tento samotný výsledek konvergenci nezaručuje. Jak ukázali autoři článku Ghadimi a kol. (2016), pro různé volby malých m_k můžeme omezit zdola omezující podmínku (3.2) z věty 11. Například pro volbu $m_k = 1$ pro každé $k \in \mathbb{N}$ dostáváme

$$\frac{LD_\Psi^2 + \frac{\sigma^2}{\alpha}\sum_{k=1}^{N}\gamma_k}{\sum_{k=1}^{N}(\alpha\gamma_k - L\gamma_k^2)} \geq \frac{\frac{\sigma^2}{\alpha}\sum_{k=1}^{N}\gamma_k}{\sum_{k=1}^{N}\gamma_k(\alpha - L\gamma_k)} \geq \frac{\frac{\sigma^2}{\alpha}\sum_{k=1}^{N}\gamma_k}{\alpha\sum_{k=1}^{N}\gamma_k} = \frac{\sigma^2}{\alpha^2}.$$

Pro konvergenci tedy potřebujeme i větší m_k.

Důsledek 12. *Nechť platí všechny předpoklady a označení jako ve větě 11. Dále nechť máme v RSPG algoritmu 2 délku kroků $\gamma_k = \frac{\alpha}{2L}$ pro každé $k \in \{1, 2, \ldots, N\}$ a velikost náhodného výběru $m_k = m \in N$ pro $k = 1,2,\ldots,N$. Potom platí následující dvě nerovnosti*

$$\mathbb{E}_{R,\xi_{[N]}}\Big[\|\hat{g}_{X,R}\|^2\Big] \leq \frac{4L^2D_\psi^2}{\alpha^2 N} + \frac{2\sigma^2}{\alpha^2 m}$$
$$\mathbb{E}_{R,\xi_{[N]}}\Big[\|g_{X,R}\|^2\Big] \leq \frac{8L^2D_\psi^2}{\alpha^2 N} + \frac{6\sigma^2}{\alpha^2 m}.$$

Důkaz. Důkaz přejímáme z Ghadimi a kol. (2016) důsledek 3, a uvádíme ho pro úplnost. Dovysvětlili jsme navíc proč platí nerovnost v odhadu $\mathbb{E}_{R,\xi_{[N]}}\Big[\|g_{X,R}\|^2\Big]$.

Dosazením předpokladu $m_k = m$ pro $k = 1, 2, \ldots, N$ a délky kroků $\gamma_k = \frac{\alpha}{2L}$ pro každé $k \in \{1, 2, \ldots, N\}$ do nerovnosti z věty 11 dostáváme

$$\begin{aligned}\mathbb{E}_{R,\xi_{[N]}}\left[\|\hat{g}_{X,R}\|^2\right] &\leq \frac{LD_\Psi^2 + \frac{\sigma^2}{\alpha}\sum_{k=1}^N \frac{\gamma_k}{m}}{\sum_{k=1}^N(\alpha\gamma_k - L\gamma_k^2)} = \frac{LD_\Psi^2 + \frac{\sigma^2}{m\alpha}\sum_{k=1}^N \frac{\alpha}{2L}}{\sum_{k=1}^N(\frac{\alpha^2}{2L} - L\frac{\alpha^2}{4L^2})} \\ &= \frac{LD_\Psi^2 + \frac{N\sigma^2}{2mL}}{\frac{N\alpha^2}{4L}} = \frac{4L^2D_\psi^2}{\alpha^2 N} + \frac{2\sigma^2}{\alpha^2 m},\end{aligned}$$

Tím je dokázaná první část tvrzení.

Pro zjednodušení zápisu budeme v tomto důkazu zkracovat zápis $\mathbb{E}_{R,\xi_{[N]}}$ záměnou za $\mathbb{E}$. Počítejme nejprve

$$\begin{aligned}\mathbb{E}\left[\|g_{X,R}\|^2\right] &\leq \mathbb{E}\left[(\|g_{X,R} - \hat{g}_{X,R}\| + \|\hat{g}_{X,R}\|)^2\right] \\ &\leq 2\mathbb{E}\left[\max\left\{\|g_{X,R} - \hat{g}_{X,R}\|^2, \|\hat{g}_{X,R}\|^2\right\}\right] \\ &\leq 2\mathbb{E}\left[\|g_{X,R} - \hat{g}_{X,R}\|^2 + \|\hat{g}_{X,R}\|^2\right] \\ &= 2\mathbb{E}\left[\|g_{X,R} - \hat{g}_{X,R}\|^2\right] + 2\mathbb{E}\left[\|\hat{g}_{X,R}\|^2\right].\end{aligned}$$

Druhý člen je omezen vztahem z první části této věty. Omezíme nyní druhý člen. Podle důsledku 7 máme pro $x = x_k, g_1 = G_k, g_2 = \nabla f(x_k), \hat{g}_{X,k} = P_X(x_k, G_k, \gamma_k)$ a $g_{X,k} = P_X(x_k, \nabla f(x_k), \gamma_k)$ nerovnost

$$\|\hat{g}_{X,k} - g_{X,k}\| \leq \frac{1}{\alpha}\|G_k - \nabla f(x_k)\|.$$

Podle třetí části (3.5) důkazu věty 11, dostáváme

$$2\mathbb{E}\left[\|g_{X,R} - \hat{g}_{X,R}\|^2\right] \leq \frac{2}{\alpha}\mathbb{E}\left[\|G_R - \nabla f(x_R)\|^2\right] \leq \frac{2\sigma^2}{m\alpha}.$$

Složením těchto výsledků dostáváme požadovanou nerovnost

$$\mathbb{E}\left[\|g_{X,R}\|^2\right] \leq 2(\frac{4L^2D_\psi^2}{\alpha^2 N} + \frac{2\sigma^2}{\alpha^2 m}) + \frac{2\sigma^2}{m\alpha} = \frac{8L^2D_\psi^2}{\alpha^2 N} + \frac{6\sigma^2}{m\alpha}.$$

□

Jak vidíme, pokud budeme brát pouze malé velikosti vzorků m, tak nemůžeme zaručit konvergenci, protože na pravé straně pořád budeme mít $\frac{6\sigma^2}{m\alpha}$, které dále už nelze omezit. V následujícím důsledku si zavedeme další podmínky, pro které už dokážeme zaručit konvergenci ve střední hodnotě.

Důsledek 13. *Nechť jsou splněny předpoklady 1 a 2 a voleny délky kroků jako v důsledku 12. Zvolme počáteční bod x_1 a označme D_ψ^2 jako ve větě 11. Předpokládejme, že odhad gradientu můžeme použít nejvýše $M-$krát, $M \in \mathbb{N}$. Dále předpokládejme, že odhad gradientu v RSPG algoritmu 2 v každém kroku použijeme $m-$krát, kde*

$$m = \left\lceil \min\left\{\max\left\{\frac{\sigma\sqrt{6M}}{4L\hat{D}}, 1\right\}, M\right\}\right\rceil, \tag{3.7}$$

pro nějaké $\hat{D} > 0$. Nechť maximální počet kroků je $N = \lfloor \frac{M}{m} \rfloor$ a pravděpodobnostní rozdělení P_R je voleno pro dané N podle vzorce 3.1. Potom

$$\mathbb{E}_{R,\xi_{[N]}}[\|g_{X,R}\|^2] \leq \frac{L}{\alpha^2}\Big[\frac{16LD_\psi^2}{M} + \frac{4\sqrt{6}\sigma}{\sqrt{M}}\Big(\frac{D_\psi^2}{\hat{D}} + \hat{D}\max\Big\{\frac{\sigma\sqrt{6}}{4L\hat{D}\sqrt{M}}, 1\Big\}\Big)\Big].$$

Důkaz uvádět nebudeme, lze najít v Ghadimi a kol. (2016) důsledek 4.

V této části kapitoly, jsme ukázali, že jeden běh RSPG algoritmu nezaručuje konvergenci k extrému. Konvergenci máme zaručenou jen ve střední hodnotě a pro reálné použití je proto nutné algoritmus pouštět opakovaně a následně zvolit nejlepší výstup.

3.3 Důkaz nalezení globálního minima

V celé této části budeme předpokládat, že funkce f je konvexní. Tedy i ψ je konvexní, protože je součtem dvou konvexních funkcí. Tedy funkce ψ má na konvexní množině X globální minimum.

Věta 14. *Nechť platí všechny předpoklady a označení jako ve větě 11. Dále nechť f je konvexní funkce a optimální řešení problému* (1.4) *označme x^*. Předpokládejme ještě, že posloupnost délek kroků $\{\gamma_k\}_{k=1}^N$ je neklesající, tedy*

$$0 < \gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_N \leq \frac{\alpha}{L}.$$

Pak platí

$$\mathbb{E}_{R,\xi_{[N]}}\Big[\psi(x_{R+1}) - \psi(x^*)\Big] \leq \frac{(\alpha - L\gamma_1)V(x^*,x_1) + \frac{\sigma^2}{2}\sum_{k=1}^N \frac{\gamma_k^2}{m_k}}{\sum_{k=1}^N(\alpha\gamma_k - L\gamma_k^2)},$$

kde $V(x^*,x_1)$ je z definice 5.

Důkaz lze nalézt v Ghadimi a kol. (2016) Věta 2, který jsme doplnili o důkaz Lemmatu 8.

Věta 15. *Nechť platí všechny předpoklady a označení jako ve větě 11. Dále nechť f je konvexní funkce a optimální řešení problému* (1.4) *označme x^*. Předpokládejme ještě, že posloupnost délek kroků $\{\gamma_k\}_{k=1}^N$ je nerostoucí, tedy*

$$0 < \gamma_N \leq \gamma_{N-1} \leq \cdots \leq \gamma_1 \leq \frac{\alpha}{L}.$$

Pak platí

$$\mathbb{E}_{R,\xi_{[N]}}\Big[\psi(x_{R+1}) - \psi(x^*)\Big] \leq \frac{\frac{\sigma^2}{2}\sum_{k=1}^N \frac{\gamma_k^2}{m_k} + (\alpha - L\gamma_1)\max_{u\in X}\Big\{V(x^*,u)\Big\}}{\sum_{k=1}^N(\alpha\gamma_k - L\gamma_k^2)},$$

kde $V(x^*,u)$ je z definice 5.

Důkaz viz Ghadimi a kol., 2016, věta 2.

Důsledek 16. *Nechť jsou splněny všechny předpoklady důsledku 12 a nechť f z předpokladu 1 je konvexní funkce. Položme x^* jako optimální hodnotu problému 1.4. Potom platí*

$$\mathbb{E}_{R,\xi_{[N]}}[\psi(x_{R+1}) - \psi(x^*)] \leq \frac{2LV(x^*, x_1)}{\alpha N} + \frac{\sigma^2}{2mL}.$$

Důkaz. Tento důkaz jsme analogicky rozepsali jako důkaz důsledku 12 z (Ghadimi a kol., 2016). Protože délka kroku $\gamma_k = \frac{\alpha}{2L}$ pro každé $k \in \{1, 2, \ldots, N\}$ je neklesající, můžeme použít odhad z věty 14. Dosazením délky kroků a velikosti náhodného výběru m dostáváme

$$\begin{aligned}
\mathbb{E}_{R,\xi_{[N]}}[\psi(x_{R+1}) - \psi(x^*)] &\leq \frac{(\alpha - L\gamma_1)V(x^*,x_1) + \frac{\sigma^2}{2}\sum_{k=1}^{N}\frac{\gamma_k^2}{m_k}}{\sum_{k=1}^{N}(\alpha\gamma_k - L\gamma_k^2)} \\
&= \frac{(\alpha - \frac{\alpha}{2})V(x^*,x_1) + \frac{\sigma^2}{2}\sum_{k=1}^{N}\frac{\alpha^2}{4mL^2}}{\sum_{k=1}^{N}(\frac{\alpha^2}{2L} - \frac{\alpha^2}{4L})} \\
&= \frac{\frac{\alpha}{2}V(x^*,x_1) + \frac{N\sigma^2\alpha^2}{8mL^2}}{\frac{N\alpha^2}{4L}} \\
&= \frac{2LV(x^*, x_1)}{\alpha N} + \frac{\sigma^2}{2mL}.
\end{aligned}$$

□

Podle předchozího důsledku je vzdálenost optimální hodnoty a výsledku algoritmu ve střední hodnotě omezená výrazem, který s malou velikostí náhodného výběru m může být velký. Musíme tedy stejně jako v sekci 3.2 vybrat speciální velikost náhodného výběru, abychom zaručili konvergenci k nule.

Důsledek 17. *Nechť jsou splněny všechny předpoklady důsledku 13 a nechť f z předpokladu 1 je konvexní funkce. Položme x^* jako optimální hodnotu problému 1.4. Potom*

$$\begin{aligned}
&\mathbb{E}_{R,\xi_{[N]}}\Big[\psi(x_{R+1}) - \psi(x^*)\Big] \\
&\qquad \leq \frac{4LV(x^*, x_1)}{\alpha M} + \frac{\sqrt{6}\sigma}{\alpha\sqrt{M}}\Big(\frac{V(x^*, x_1)}{\hat{D}} + \frac{\alpha\hat{D}}{3}\max\Big\{\frac{\sigma\sqrt{6}}{4L\hat{D}\sqrt{M}}, 1\Big\}\Big).
\end{aligned}$$

Důkaz. Důkaz jsme rozepsali podle vzoru důkazu důsledku 4 z Ghadimi a kol. (2016). Zřejmě platí nerovnost $N \geq \frac{M}{2m}$, kterou dosadíme do důsledku 16 a dostáváme

$$
\begin{aligned}
&\mathbb{E}_{R,\xi_{[N]}}\Big[\psi(x_{R+1}) - \psi(x^*)\Big] \\
&\quad\leq \frac{2LV(x^*,x_1)}{\alpha N} + \frac{\sigma^2}{2mL} \\
&\quad= \frac{4mLV(x^*,x_1)}{\alpha M} + \frac{\sigma^2}{2mL} \\
&\quad= \frac{4LV(x^*,x_1)}{\alpha M}\max\Big\{\frac{\sigma\sqrt{6M}}{4L\hat{D}},1\Big\} + \frac{\sigma^2}{2L\min\Big\{\frac{\sigma\sqrt{6M}}{4L\hat{D}},M\Big\}} \\
&\quad\leq \frac{4LV(x^*,x_1)}{\alpha M}\Big(\frac{\sigma\sqrt{6M}}{4L\hat{D}} + 1\Big) + \max\Big\{\frac{2\sigma\hat{D}}{\sqrt{6M}}, \frac{\sigma^2}{2LM}\Big\} \\
&\quad\leq \frac{4LV(x^*,x_1)}{\alpha M} + \frac{\sqrt{6}\sigma}{\alpha\sqrt{M}}\Big(\frac{V(x^*,x_1)}{\hat{D}} + \frac{\alpha\hat{D}}{3}\max\Big\{\frac{\sigma\sqrt{6}}{4L\hat{D}\sqrt{M}},1\Big\}\Big).
\end{aligned}
$$

□

Z předešlé úlohy vyplývá, že pokud zvětšujeme počet odhadů gradientu M, tak pravá strana konverguje k nule jako $\mathcal{O}(\frac{1}{\sqrt{M}})$. Máme tedy zaručeno, že střední hodnota výstupu x_{R+1} konverguje ve střední hodnotě k x^*.

4. Numerická studie

V celé této kapitole budeme uvažovat $\omega(x) = \|x\|_2^2/2$ a tedy proximální zobrazení

$$V(x, z) = \frac{1}{2}\|x - z\|^2$$

a $h(x) \equiv 0$. Z rovnosti (1.8) dostáváme silnou konvexitu s modulem spojitosti $\alpha = 1$.

Budeme zkoumat chování RSPG algoritmu pro řešení problému stochastického programovaní. Zaměříme se na problém, který lze reprezentovat jako lineární program. Pro tento problém je snadné gradient přímo spočítat, budeme ale předpokládat, že gradient neznáme. Volíme tento problém pro snadné ověření korektnosti výstupu RSPG algoritmu.

Stochastický program je optimalizační problém, ve kterém jsou některé parametry modelu nejisté. Tyto parametry však pochází ze známého pravděpodobnostního rozdělení. Stochastické programování se využívá v mnoha odvětvích, například ve financích, nebo v optimalizaci průmyslu. Pomáhá nám se rozhodovat optimálně vzhledem k nejistotě výsledku. Birge a Louveaux (2011)

4.1 Problém farmáře

Příklad čerpáme z úvodního příkladu v knize Birge a Louveaux (2011), tento problém jsme ale modifikovali, abychom mohli použít RSPG algoritmus 2.

Jedná se o problém kdy má farmář k dispozici 3 typy plodin a musí se rozhodnout, jakou plochu oseje každou z daných plodin. Farmář potřebuje určité množství surovin k vlastnímu užitku, má ale možnost suroviny dokoupit. Na konci roku vypěstované a nespotřebované suroviny prodá za nejistou cenu, která je při rozhodovacím procesu neznámá. Budeme předpokládat, že množství vypěstovaných surovin je fixní. Farmář se snaží maximalizovat svůj zisk neboli minimalizovat náklady ve střední hodnotě.

Farmář vlastní 500 hektarů půdy a může pěstovat pšenici, kukuřici a řepu. Budeme předpokládat, že prodejní ceny pšenice a kukuřice jsou z normálního rozdělení se známou střední hodnotou a rozptylem, které jsou zadány v tabulce 4.1. Prodejní cena řepy se liší od prodaného objemu do 6 000 tun a nad 6 000 tun, obě tyto kategorie jsou z normálního rozdělení s parametry definovanými v tabulce 4.1. Předpokládejme, že všechny tyto náhodná rozdělení jsou nezávislé.

	střední hodnota	rozptyl
pšenice	170	2500
kukuřice	150	2025
řepa do 6 000 tun	36	256
řepa nad 6 000 tun	10	25

Pozn: Všechny jednotky uvádíme v $.

Tabulka 4.1: Parametry normálního rozdělení prodejní ceny plodin.

Označme Z_1, Z_2 , Z_3, Z_4 náhodné reálné veličiny reprezentující prodejní ceny pšenice, kukuřice, řepy do 6 000 tun a řepy nad 6 000 tun respektive. Další parametry problému shrnujeme v tabulce 4.2.

	Pšenice	Kukuřice	Řepa
Cena osiva [$/$ha$]	150	230	260
Výnosnost pole [T/ha]	2,5	3	20
Nákupní cena [$/$T$]	238	210	–
Vlastní potřeba [T]	200	240	–

Pozn: T značí tunu, *ha* hektar

Tabulka 4.2: Parametry úlohy.

Označme dále $x = (x_1, x_2, x_3, y_1, y_2, w_1, w_2, w_3, w_4)^T$, kde
x_1 = počet hektarů oseté pšenicí,
x_2 = počet hektarů oseté kukuřicí,
x_3 = počet hektarů oseté řepou,
y_1 = počet nakoupených tun pšenice,
y_2 = počet nakoupených tun kukuřice.
w_1 = počet prodaných tun pšenice,
w_2 = počet prodaných tun kukuřice,
w_3 = počet prodaných tun řepy do 6000 tun,
w_4 = počet prodaných tun řepy nad 6000 tun.
Potom formulaci lineárního problému zapíšeme podle Birge a Louveaux (2011)

$$\begin{aligned} min \quad & \mathbb{E}_{Z_1,Z_2,Z_3,Z_4}\Big[150x_1 + 230x_2 + 260x_3 + 238y_1 + 210y_2 \\ & \qquad - Z_1w_1 - Z_2w_2 - Z_3w_3 - Z_4w_4\Big] \end{aligned}$$

$$\begin{aligned} za\ podmínek \qquad & x_1 + x_2 + x_3 \leq 500 \\ & -2{,}5x_1 - y_1 + w_1 \leq -200 \\ & -3x_2 - y_2 + w_2 \leq -240 \\ & -20x_3 + w_3 + w_4 \leq 0 \\ & w_3 \leq 6000 \\ & x_1, x_2, x_3, y_1, y_2, w_1, w_2, w_3, w_4 \geq 0. \end{aligned}$$

Omezující podmínky, které generují množinu X, nezávisí na hodnotě náhodných veličin a množina X je tedy stejná pro všechny realizace. Množina X je konvexní, protože je generovaná pouze lineárními funkcemi, navíc je zřejmě omezená. Označme $f(x)$ účelovou funkci a $h(x) \equiv 0$ pro všechna $x \in X$.

Funkce h je zřejmě konvexní. Protože operátor střední hodnoty je lineární, tak je funkce f na X lineární, a proto má na množině X konstantní gradient. Funkce f má tedy na množině X L-lipschitzovskou derivaci pro jakékoli $L > 0$. Součet $f(x) + h(x)$ je na omezené konvexní množině konečný. Máme tedy splněn předpoklad 1 a můžeme označit podle 1.4 $\psi(x) = f(x)$.

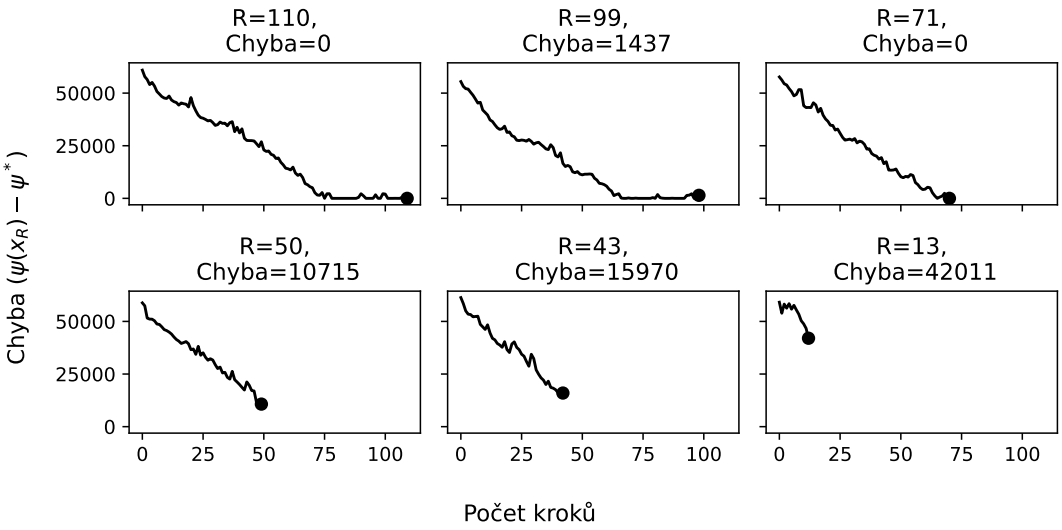
Ačkoli funkce f je na množině X definovaná tak, že bychom mohli použít vlastnost linearity střední hodnoty a gradient spočítat, budeme gradient funkce

f odhadovat podle předpokladu 2. Protože jsou Z_1, Z_2, Z_3 a Z_4 náhodné veličiny z normálního rozdělení, tak díky nezávislosti náhodných veličin je splněn předpoklad 2 pro součet jejich rozptylů $\sigma^2 = 4\,381$.

Pokud spočítáme nejdříve střední hodnotu, tak z Birge a Louveaux (2011) máme optimální řešení $(120, 80, 300, 0, 0, 100, 0, 6\,000, 0)^T$ s optimální hodnotou $\psi^* = -118\,000$, kterou použijeme k porovnání chyby algoritmu.

Pro náš případ si zvolíme $L = 0{,}05$ a stejné délky kroků $\gamma = \frac{\alpha}{2L} = 10$. Pro x_1 zvolíme scénář, že nic nezasadí. Položíme $\hat{D} = 6\,000$. Zvolme $M = 500$ a $m = 4$ podle vzorce 3.7. Budeme uvažovat podle důkazu důsledku 17 maximální počet kroků $N = \lfloor \frac{M}{m} \rfloor = 125$. Pravděpodobnostní rozdělení P_R podle (3.1) je pro konstantní délku kroků rovnoměrné na množině $\{1, 2, \ldots, 125\}$.

Algoritmus jsme naimplementovali podle 2 a pustili 6−krát. V následujícím grafu jsme porovnali výstupy algoritmu pro různé realizace náhodné veličiny R. Na vodorovnou osu jsme zanesli počet kroků R daného běhu, na osu y jsme zanesli absolutní chybu výstupu, tedy $\psi(x_R) - \psi^*$.



Obrázek 4.1: Výsledky RSPG algoritmu na problému farmáře.

Zvolíme výstup s nejnižší chybou, tedy pro $R = 71$, nebo $R = 110$. Dostáváme optimální hodnotu $-118\,600$ \$ s optimálním řešením zasadit 170 hektarů pšenice, 80 hektarů kukuřice a 250 hektarů řepy, které je shodné s optimálním řešením problému.

Jak vidíme z těchto běhů, tak není optimální se spolehnout na jeden běh algoritmu, protože nejhorší běh je s vysokou chybou 42 011 \$.

Závěr

V této práci jsme nejdříve zavedli optimalizační úlohu a definovali její předpoklady. Následně jsme zadefinovali proximální zobrazení a zobecněnou projekci a dokázali některé jejich vlastnosti. Poté jsme představili PG algoritmus a dokázali jeho konvergenci i pro nekonvexní funkce při splnění určitých předpokladů. Tento algoritmus jsme dále rozšířili na RSPG algoritmus, který nevyžaduje počítání gradientu funkce. Omezením velikosti projektovaného gradientu ve střední hodnotě jsme dokázali konvergenci RSPG algoritmu pro nekonvexní funkce. Pro konvexní funkce jsme dokázali, že algoritmus ve střední hodnotě nalezne globální optimum účelové funkce na konvexní množině X.

V praktické části jsme aplikovali dosažené teoretické výsledky na jednoduchý problém lineárního stochastického programování. Algoritmus se choval dle očekávání a nalezl optimální řešení problému ve dvou z šesti běhů.

Seznam použité literatury

Banerjee, A., Merugu, S., Dhillon, I. S. a Ghosh, J. (2005). Clustering with bregman divergences. *Journal of Machine Learning Research*, **6**, 1705–1749. ISSN 1532-4435.

Birge, J. a Louveaux, F. (2011). *Introduction to Stochastic Programming*. Springer Series in Operations Research and Financial Engineering. Springer New York. ISBN 9781461402374. URL https://books.google.cz/books?id=Vp0Bp8kjPxUC.

Bregman, L. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, **7** (3), 200–217. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(67)90040-7. URL https://www.sciencedirect.com/science/article/pii/0041555367900407.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, **8**(3-4), 231–357. ISSN 1935-8237. doi: 10.1561/2200000050. URL http://dx.doi.org/10.1561/2200000050.

Ghadimi, S., Lan, G. a Zhang, H. (2016). Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, **155**, 267–305. doi: 10.1007/s10107-014-0846-1.

Lachout, P. (2021). Optimization Theory - direct approach. https://www2.karlin.mff.cuni.cz/~lachout/Vyuka/T-Optima/210929-TO-OptI-text.pdf. [Online; accessed 30-June-2022].

Lan, G. (2011). An optimal method for stochastic composite optimization. *Mathematical Programming*, **133**. doi: 10.1007/s10107-010-0434-y. URL https://doi.org/10.1007/s10107-010-0434-y.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press. ISBN 978-1-4008-7317-3.

Seznam použitých zkratek

$\mathbb{R}$ reálná čísla
$\mathbb{N}$ přirozená čísla
$\lfloor x \rfloor$ dolní celá část
$\lceil x \rceil$ horní celá část
$\langle x,y \rangle$ skalární součin
$\|c\|$ euklidovská norma
$\|c\|_p$ p-norma
$\nabla f(x)$ gradient funkce f v bodě x
$\partial f(x)$ množina subgradientů funkce f v bodě x
s.j. skoro jistě