

Convergence Analysis of SGD in Convex Optimization Problems

by Hossain Rahnama

Sep 27, 2026

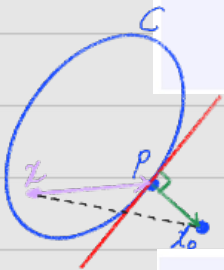
تصور کردن روی یک مجموعه‌ی محدب بسته

گزاره ۵.۴. فرض کنید $C \subseteq \mathbb{R}^n$ مجموعه‌ای ناتهی، بسته و محدب است. نقطه‌ی دلخواه $x_0 \in \mathbb{R}^n$ را در نظر بگیرید. در این صورت، یک تصویر یکتا از x_0 بر روی C وجود دارد، یعنی مساله‌ی بهینه‌سازی زیر یک جواب یکتا دارد:

$$\min_{x \in C} \|x - x_0\|, \quad (7.4)$$

که آن را به صورت $p := \Pi_C(x_0)$ نمایش می‌دهیم. همچنین، برای هر $x \in C$ داریم:

$$(x - p)^T (p - x_0) \geq 0. \quad (8.4)$$



• حالت خاص و معروف این گزاره؟

نتیجه ۶.۴. فرض کنید $C \subseteq \mathbb{R}^n$ مجموعه‌ای ناتهی، بسته و محدب است. نقطه‌ی دلخواه $x_0 \in \mathbb{R}^n$ و تصویر آن روی C ، یعنی $p := \Pi_C(x_0)$ را در نظر بگیرید. در این صورت، برای هر $x \in C$ داریم $\|x - p\| \leq \|x - x_0\|$.

برهان. کافی است که نشان دهیم $\|x - x_0\|^2 - \|x - p\|^2 \geq 0$.

$$\begin{aligned} \|x - x_0\|^2 - \|x - p\|^2 &= \|(x - p) + (p - x_0)\|^2 - \|x - p\|^2 \\ &= \|x - p\|^2 + \|p - x_0\|^2 + 2(x - p)^T(p - x_0) - \|x - p\|^2 \\ &= \|p - x_0\|^2 + \underbrace{2(x - p)^T(p - x_0)}_{\geq 0} \end{aligned}$$

بهینه‌سازی مقعدها با SGD، طول گام ثابت

$$\min_{x \in \mathcal{W}} f(x)$$

الگوریتم ۳.۴ گرادینان کاهشی تصادفی با طول گام ثابت بر روی مجموعه‌ی محدب بسته.

ورودی: تابع محدب $f: \mathbb{R}^d \rightarrow \mathbb{R}$ ، مجموعه‌ی محدب بسته $\mathcal{W}$.

خروجی: تقریبی از یک کمینه‌کننده‌ی سراسری f روی $\mathcal{W}$ و مقدار کمینه‌ی متناظر آن.

۱: روال STOCHASTIC-GRADIENT-DESCENT

۲: * مقداردهی اولیه

۳: نقطه‌ی آغازین w_1 ، تعداد تکرار T و طول گام ثابت $s > 0$ را انتخاب کن.

۴: * حلقه‌ی همگرایی

۵: برای $t = 1, \dots, T$ انجام بده

۶: جهت مناسب d_t را از توزیع $\mathbb{D}_t \in \text{Pb}(\mathbb{R}^d)$ به گونه‌ای انتخاب کن که $\mathbb{E}(d_t | w_t) \in \partial f(w_t)$.

۷: قرار بده $u_t = w_t - s d_t$

۸: قرار بده $w_{t+1} = \Pi_{\mathcal{W}}(u_t)$

۹: * محاسبه‌ی خروجی

۱۰: قرار بده $\bar{w} = \frac{1}{T} \sum_{t=1}^T w_t$

۱۱: برگردان $(\bar{w}, f(\bar{w}))$

• گام تصویرسازی طبق گزاره‌ی ۱.۳ خوش‌صرف است.

• تفاوت اصلی با حالت نامعید؟

• سبک‌های قبلی کار می‌کنند؟

همگرایی برای توابع محدب لیپشیتز در حالت معید

۱.۴. فرض کنید $f: \mathbb{R}^d \rightarrow \mathbb{R}$ تابعی محدب باشد و $w^* \in \arg \min(f, \mathbb{R}^d)$ دنباله‌ی نقاط تصادفی $(w_1, \dots, w_T)$ با $\bar{w} = \frac{1}{T} \sum_{t=1}^T w_t$ و جهت‌های تصادفی $(d_1, \dots, d_T)$ را در نظر بگیرید به طوری که برای هر $t \in [T]$ داشته باشیم $\mathbb{E}(d_t | w_t) \in \partial f(w_t)$ در این صورت، داریم:

$$\mathbb{E}(f(\bar{w}) - f(w^*)) \leq \frac{1}{T} \sum_{t=1}^T \mathbb{E}((w_t - w^*)^\top d_t).$$

این معنی دارد؟

• شرط ضعیف‌تری؟

۷.۴. فرض کنید $\mathcal{W} \subseteq \mathbb{R}^d$ یک مجموعه‌ی محدب بسته و $w^* \in \mathcal{W}$ نقطه‌ای دلخواه است. دنباله‌ی جهت‌های تصادفی $(d_1, \dots, d_T)$ و طول گام‌های $(s_1, \dots, s_T)$ را در نظر بگیرید که برای هر $t \in [T]$ داریم $s_t > 0$. از روی آن‌ها دنباله‌ی نقاط تصادفی $(w_1, \dots, w_T)$ را می‌سازیم به طوری که برای هر $t \in [T]$ داریم $u_t = w_t - s_t d_t$ و $w_{t+1} = \Pi_{\mathcal{W}}(u_t)$ در این صورت، نامساوی زیر برقرار است:

$$\mathbb{E}((w_t - w^*)^\top d_t) \leq \frac{1}{2s_t} \mathbb{E}(\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2) + \frac{s_t}{2} (\text{tr}(\text{cov}(d_t)) + \|\mathbb{E}(d_t)\|^2).$$

سبک‌زن. در لیم ۳.۴ به جای w_{t+1} قرار دهیم w_t بنابراین داریم

$$\mathbb{E}((w_t - w^*)^\top d_t) = \frac{1}{2s_t} \mathbb{E}(\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2) + \frac{s_t}{2} \mathbb{E}(\|d_t\|^2)$$

نتیجه‌ی ۶.۲: $\|w_{t+1}^* - w_{t+1}\| \leq \|w_t^* - w_t\|$

$$\stackrel{\text{۶.۲۰}}{=} \frac{1}{2s_t} \mathbb{E}(\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2) + \frac{s_t}{2} (\text{tr}(\text{cov}(d_t)) + \|\mathbb{E}(d_t)\|^2)$$

همگرایی برای توابع محدب لیپشیتز در حالت مقید

گزاره ۸.۴. فرض کنید $\mathcal{W} \subseteq \mathbb{R}^d$ یک مجموعه‌ی محدب بسته و $f: \mathbb{R}^d \rightarrow \mathbb{R}$ تابعی محدب و ρ -لیپشیتز است. همچنین، بگیرد $\mathbf{w}^* \in \arg \min(f, \mathcal{W})$ به طوری که $\|\mathbf{w}_1 - \mathbf{w}^*\| \leq c$. فرض کنید عدد $\sigma > 0$ وجود دارد به طوری که برای هر $t \in [T]$ داریم $\text{tr}(\text{cov}(\mathbf{d}_t)) \leq \sigma^2$. تعریف کنید $r := \sqrt{\rho^2 + \sigma^2}$. پس از T تکرار حلقه‌ی همگرایی در الگوریتم ۳.۴ با طول گام ثابت $s = c/(r\sqrt{T})$ داریم:

$$\mathbb{E}(f(\bar{\mathbf{w}}) - f(\mathbf{w}^*)) \leq \frac{rc}{\sqrt{T}}.$$

بدین ترتیب، برای هر $\epsilon > 0$ ، عدد $M = r^2 c^2 / \epsilon^2$ تضمین می‌کند که برای هر تعداد تکرار $T \geq M$ داریم:

$$\mathbb{E}(f(\bar{\mathbf{w}}) - f(\mathbf{w}^*)) \leq \epsilon.$$

• بنابراین اثبات کاملاً مشابه با حالت نامقید کار می‌کند.

• گزاینده پیش‌تر به دست آورده‌یم.

$$\frac{1}{2T} \|\mathbf{w}_1 - \mathbf{w}^*\|^2 + \frac{s}{2T} \sum_{t=1}^T \text{tr}(\text{cov}(\mathbf{d}_t)) + \frac{s}{2T} \sum_{t=1}^T \|\mathbf{E}(\mathbf{d}_t)\|^2$$

بدست صفر می‌آید

فرض کنسرتک وار بایس

لیپشیتز بودن تابع صفر

توابع محدب قوی

تعریف ۹.۳. فرض کنید $C \subseteq \mathbb{R}^n$ یک مجموعه‌ی محدب و $f: C \rightarrow \mathbb{R}$ یک تابع است. در این صورت، f را تابعی محدب قوی با ثابت $\gamma > 0$ بر روی C گویند اگر برای هر دو نقطه‌ی متمایز $\mathbf{x}, \mathbf{y} \in C$ و هر $\lambda \in (0, 1)$ داشته باشیم:

$$f(\lambda \mathbf{x} + (1 - \lambda) \mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}) - \frac{\gamma}{2} \lambda (1 - \lambda) \|\mathbf{x} - \mathbf{y}\|^2.$$

• تابع محدب قوی، محدب اکید تر هستند.

گزاره ۸.۳. فرض کنید $C \subseteq \mathbb{R}^n$ یک مجموعه‌ی محدب است و $\gamma > 0$. تابع $f: C \rightarrow \mathbb{R}$ بر روی C محدب قوی با ثابت γ است اگر و تنها اگر تابع $f(\mathbf{x}) - \frac{\gamma}{2} \|\mathbf{x}\|^2$ بر روی C محدب باشد.

• مثال. تابع $f(\mathbf{x}) = a \|\mathbf{x}\|^2$. تابع $h(\mathbf{x}) = |\mathbf{x}|$ آن محدب قوی است؟ محدب اکید چگونه؟

نتیجه ۹.۳. فرض کنید $C \subseteq \mathbb{R}^n$ یک مجموعه‌ی محدب است. اگر تابع $f: C \rightarrow \mathbb{R}$ محدب قوی با ثابت $\gamma > 0$ و تابع $g: C \rightarrow \mathbb{R}$ محدب باشد، آن‌گاه $f + g: C \rightarrow \mathbb{R}$ محدب قوی با ثابت γ است.

برهان.

$$(f+g)(\mathbf{x}) - \frac{\gamma}{2} \|\mathbf{x}\|^2 = \left(f(\mathbf{x}) - \frac{\gamma}{2} \|\mathbf{x}\|^2 \right) + g(\mathbf{x}) = \underbrace{h(\mathbf{x})}_{\text{محدب}} + g(\mathbf{x})$$

محدب $h(\mathbf{x})$

جمع دو تابع محدب

گزاره ۱۰.۳. فرض کنید $C \subseteq \mathbb{R}^n$ یک مجموعه‌ی محدب باز و $f: C \rightarrow \mathbb{R}$ تابعی محدب است. در این صورت، گزاره‌های زیر معادل‌اند.

• تابع f روی C محدب قوی با ثابت $\gamma > 0$ است.

• ثابت $\gamma > 0$ وجود دارد به طوری که برای هر $x, y \in C$ و هر $v \in \partial f(y)$ داریم:

$$f(x) \geq f(y) + v^T(x - y) + \frac{\gamma}{2} \|x - y\|^2.$$

• ثابت $\gamma > 0$ وجود دارد به طوری که برای هر $x, y \in C$ و هر $u \in \partial f(x)$ و هر $v \in \partial f(y)$ داریم:

$$(u - v)^T(x - y) \geq \gamma \|x - y\|^2.$$

• با گزاره‌ی ۱۱.۱ مقایسه کنید. اگر f محدب باشد؟

برهان. با توجه به تعریف محدب قوی بودن، برای هر $x, y \in C$ و هر $\lambda \in (0, 1)$ داریم

$$f(y + \lambda(x - y)) + \frac{1}{2} \lambda(1 - \lambda) \|x - y\|^2 \leq \lambda f(x) + (1 - \lambda) f(y) = f(y) + \lambda(f(x) - f(y))$$

از طرفی با توجه به تعریف زیردیفانسیب داریم

$$f(y + \lambda(x - y)) \geq f(y) + \lambda v^T(x - y)$$

ترکیب دونا مساوی بالا نتیجه می‌دهد

$$f(y) + \lambda v^T(x - y) + \frac{1}{2} \lambda(1 - \lambda) \|x - y\|^2 \leq f(y) + \lambda(f(x) - f(y))$$

با $\lambda \rightarrow 0$ نتیجه می‌شود

$$v^T(x - y) + \frac{1}{2} \lambda \|x - y\|^2 \leq f(x) - f(y) \Rightarrow f(x) \geq f(y) + v^T(x - y) + \frac{1}{2} \lambda \|x - y\|^2$$

بهینه‌سازی مقید با SGD، طول گام متغیر

الگوریتم ۴.۴ گرادینان کاهشی تصادفی با طول گام متغیر بر روی مجموعه‌ی محدب بسته.

ورودی: تابع محدب قوی $f: \mathbb{R}^d \rightarrow \mathbb{R}$ با ثابت γ ، مجموعه‌ی محدب بسته $\mathcal{W}$.
خروجی: تقریبی از کمینه‌کننده‌ی سراسری f روی $\mathcal{W}$ و مقدار کمینه‌ی متناظر آن.

۱: روال STOCHASTIC-GRADIENT-DESCENT

۲: * مقداردهی اولیه

۳: نقطه‌ی آغازین w_1 و تعداد تکرار T را انتخاب کن.

۴: * حلقه‌ی همگرایی

۵: برای $t = 1, \dots, T$ انجام بده

۶: جهت مناسب d_t را از توزیع $\mathbb{D}_t \in \text{Pb}(\mathbb{R}^d)$ به گونه‌ای انتخاب کن که $\mathbb{E}(d_t | w_t) \in \partial f(w_t)$.

۷: قرار بده $s_t = \frac{1}{\gamma_t}$.

۸: قرار بده $u_t = w_t - s_t d_t$.

۹: قرار بده $w_{t+1} = \Pi_{\mathcal{W}}(u_t)$.

۱۰: * محاسبه‌ی خروجی

۱۱: قرار بده $\bar{w} = \frac{1}{T} \sum_{t=1}^T w_t$.

۱۲: برگردان $(\bar{w}, f(\bar{w}))$.

• طول گام متغیر چه شکلی دارد؟

• چرا این تابع خاص برای طول گام انتخاب شده است؟

همگرایی برای توابع محدب قوی در حالت مقید

$$w^* \in \arg \min(f, W)$$

لم ۹.۴. فرض کنید $W \subseteq \mathbb{R}^d$ یک مجموعه‌ی محدب بسته و $f: \mathbb{R}^d \rightarrow \mathbb{R}$ تابعی محدب قوی با ثابت $\gamma > 0$ است. دنباله‌ی نقاط تصادفی $(w_1, \dots, w_T)$ با $\bar{w} = 1/T \sum_{t=1}^T w_t$ و جهت‌های تصادفی $(d_1, \dots, d_T)$ را در نظر بگیرید به طوری که برای هر $t \in [T]$ داشته باشیم $\mathbb{E}(d_t | w_t) \in \partial f(w_t)$ در این صورت، داریم:

$$\mathbb{E}(f(\bar{w}) - f(w^*)) \leq \frac{1}{T} \sum_{t=1}^T \left(\mathbb{E}((w_t - w^*)^\top d_t) - \frac{\gamma}{2} \mathbb{E}(\|w_t - w^*\|^2) \right).$$

برهان. گزاره ۱۰.۳

$$\begin{aligned} \mathbb{E}(f(\bar{w}) - f(w^*)) &\stackrel{\text{نیمه}}{\leq} \frac{1}{T} \sum_{t=1}^T \mathbb{E}(f(w_t) - f(w^*)) \\ &\stackrel{\text{نیمه}}{\leq} \frac{1}{T} \sum_{t=1}^T \left(\mathbb{E}(f(w_t) - f(w^*)) \right) \\ &\stackrel{\text{نیمه}}{\leq} \frac{1}{T} \sum_{t=1}^T \left(\mathbb{E}(\mathbb{E}(d_t | w_t)^\top (w_t - w^*)) - \frac{\gamma}{2} \mathbb{E}(\|w_t - w^*\|^2) \right) \\ &\stackrel{\text{نیمه}}{\leq} \frac{1}{T} \sum_{t=1}^T \left(\mathbb{E}(\mathbb{E}((w_t - w^*)^\top d_t | w_t)) - \frac{\gamma}{2} \mathbb{E}(\|w_t - w^*\|^2) \right) \\ &\stackrel{\text{نیمه}}{=} \frac{1}{T} \sum_{t=1}^T \left(\mathbb{E}((w_t - w^*)^\top d_t) - \frac{\gamma}{2} \mathbb{E}(\|w_t - w^*\|^2) \right) \end{aligned}$$

ایده ریاضی کل
Law of total expectation

همگرایی برای توابع محدب قوی در حالت مقید

گزاره ۱۰.۴. فرض کنید $W \subseteq \mathbb{R}^d$ یک مجموعه‌ی محدب بسته و $f: \mathbb{R}^d \rightarrow \mathbb{R}$ تابعی محدب قوی با ثابت $\gamma > 0$ باشد. همچنین، بگیرید $w^* \in \arg \min(f, W)$ فرض کنید عدد $r > 0$ وجود دارد به طوری که برای هر $t \in [T]$ داریم $\mathbb{E}(\|d_t\|^2) \leq r^2$. پس از T تکرار حلقه‌ی همگرایی در الگوریتم ۴.۴ داریم:

$$\mathbb{E}(f(\bar{w}) - f(w^*)) \leq \frac{r^2}{2\gamma} \left(\frac{\ln T}{T} + \frac{1}{T} \right).$$

برهان. لم ۹.۴ و لم ۱۰.۳

$$\begin{aligned} J(\bar{w}) &\stackrel{\text{نیمه}}{\leq} \frac{1}{2T} \sum_{t=1}^T \mathbb{E} \left(\left(\frac{1}{s_t} - \delta \right) \|w_t - w^*\|^2 - \frac{1}{s_t} \|w_{t+1} - w^*\|^2 \right) + \frac{1}{2T} \sum_{t=1}^T s_t \mathbb{E}(\|d_t\|^2) \\ &\stackrel{\text{نیمه}}{=} \frac{\delta}{2T} \mathbb{E} \left(\sum_{t=1}^T (t-1) \|w_t - w^*\|^2 - t \|w_{t+1} - w^*\|^2 \right) + \frac{1}{2T} \sum_{t=1}^T s_t \mathbb{E}(\|d_t\|^2) \\ &\stackrel{\text{نیمه}}{=} \frac{\delta}{2T} \mathbb{E}(\|w_{T+1} - w^*\|^2) + \frac{r^2}{2\delta T} \sum_{t=1}^T \frac{1}{t} \leq \frac{r^2}{2\delta T} \sum_{t=1}^T \frac{1}{t} = \frac{r^2}{2\delta T} \left(1 + \sum_{t=2}^T \frac{1}{t} \right) \\ &\stackrel{\text{نیمه}}{\leq} \frac{r^2}{2\delta T} (1 + \ln T) \end{aligned}$$

مجموع هارمونیک
۱۵
۱۵