باسمه تعالی

پروژه‌ی درس بهینه‌سازی محدّب ۱

مقدّمه‌ای بر بهینه‌سازی تصادفی

نام و نام خانوادگی: آیلار خرسندنیا، خشایار غفاری
شماره دانشجویی: ۹۸۱۰۰۴۱۸، ۹۸۱۰۰۲۱۵

استاد درس
دکتر محمّدحسین یاسائی میبدی

دانشکده‌ی مهندسی برق
دانشگاه صنعتی شریف

خرداد ۱۴۰۲

فهرست مطالب

۱ مقدّمه

در این پروژه می‌خواهیم مسائل بهینه‌سازی تصادفی را بررسی کنیم. در مسائل بهینه‌سازی تصادفی، بخشی از تابع هدف مسئله یک متغیّر تصادفی است. در نتیجه تابع هدف را می‌توان به صورت $F(\mathbf{x}, \mathbf{Z})$ نوشت که $\mathbf{Z} \in \mathbb{R}^m$ برداری تصادفی است. فضای نمونه‌ی این بردار تصادفی را $\mathcal{Z}$ و تابع چگالی/جرم احتمال آن را $p_{\mathbf{Z}}(\mathbf{z})$ می‌نامیم. همچنین فرض می‌کنیم تابع $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ نسبت به ورودی اوّل خود (یعنی $\mathbf{x}$) محدّب است.

حال تعریف می‌کنیم:

$$f(\mathbf{x}) = \mathbb{E}_{\mathbf{Z}}\left[F(\mathbf{x}, \mathbf{Z})\right] = \int_{\mathbf{z} \in \mathcal{Z}} F(\mathbf{x}, \mathbf{z}) p_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z}. \qquad (۱)$$

هدف مسئله‌ی بهینه‌سازی محدّب تصادفی حلّ مسئله‌ی زیر است:

$$\begin{aligned} \text{Minimize} \quad & f(\mathbf{x}) = \mathbb{E}_{\mathbf{Z}}\left[F(\mathbf{x}, \mathbf{Z})\right] \\ \text{subject to:} \quad & \mathbf{x} \in \mathcal{C}. \end{aligned} \qquad (۲)$$

که $\mathcal{C}$ یک مجموعه‌ی محدّب است.

۲ هادی اگر تویی که کسی گُم نمی‌شود![1]

می‌خواهیم مسئله‌ی (۲) را حل کنیم. ابتدا مسئله را ساده می‌کنیم. فرض کنید $\mathcal{C} = \mathbb{R}^n$ و تابع f در همه‌جا پیوسته و مشتق‌پذیر و با مشتق پیوسته است. به طور شهودی می‌دانیم که برای یافتن نقطه‌ی بهینه، باید قدم به قدم در خلاف جهت گرادیان حرکت کنیم. در نتیجه به الگوریتم زیر می‌رسیم:

Algorithm 1 Gradient Descent (GD)

parameters: Scalar $\eta > 0$, integer $T > 0$, Initial Point $\mathbf{x}_0$
Initialize $\mathbf{x}^{(1)} = \mathbf{x}_0$
for $t = 1, 2, \cdots, T$ **do**
  update $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta \nabla f(\mathbf{x}^{(t)})$
end
return $\mathbf{x}^{(T+1)}$

پرسش تئوری ۱. فرض کنید تابع $f : \mathbb{R}^n \to \mathbb{R}$ به صورت زیر تعریف شود:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x} + c$$

که $\mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{b} \in \mathbb{R}^n, c \in \mathbb{R}$. همچنین فرض کنید $\mathbf{A} \succ 0$. الگوریتم ۱ را با نقطه‌ی شروع $\mathbf{x}_0 = \mathbf{0}$ روی تابع f اجرا کنید. اگر تعریف کنیم $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{R}^n}\{f(\mathbf{x})\}$، رابطه‌ای میان $\mathbf{x}^{(t+1)} - \mathbf{x}^*$ و $\mathbf{x}^{(t)} - \mathbf{x}^*$ بیابید. تحت چه شرایطی خواهیم داشت $\lim_{T\to\infty}\|\mathbf{x}^{(T)} - \mathbf{x}^*\|_2 = 0$؟

پاسخ پرسش تئوری ۱. به وضوح با مشتق‌گیری داریم،

$$\mathbf{x}^* = A^{-1}b.$$

$$\mathbf{x}^{(t=1)} - \mathbf{x}^* = \mathbf{x}^{(t=1)} - A^{-1}b = (\mathbf{x}^{(t)} - \eta(A\mathbf{x}^{(t)} - b)) - A^{-1}b$$
$$= (\mathbf{x}^{(t)} - A^{-1}b)(I - \eta A) = (\mathbf{x}^{(t)} - \mathbf{x}^*)(I - \eta A)$$

پس با استقرا به راحتی می‌توان دید که

$$\mathbf{x}^{(T)} - \mathbf{x}^* = (\mathbf{x}^{(0)} - \mathbf{x}^*)(I - \eta A)^T$$

پس برای اینکه عبارت مد نظر به صفر میل کند باید مقادیر ویژه‌ی ماتریس $I - \eta A$ کمتر از ۱ و بزرگ تر از ۱− باشند یا معادلا مقادیر ویژه‌ی A بین ۰ و $\frac{2}{\eta}$ باشند.

[1] بی‌تو بهار قسمت مردم نمی‌شود/ هادی اگر تویی که کسی گُم نمی‌شود [محسن کاویانی]

پرسش تئوری ۲. نشان دهید رابطه‌ی به‌روز کردن الگوریتم ۱ یعنی $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta\nabla f(\mathbf{x}^{(t)})$ معادل با رابطه‌ی زیر است:

$$\mathbf{x}^{(t+1)} = \underset{\mathbf{x}\in\mathbb{R}^n}{\arg\min}\left\{f(\mathbf{x}^{(t)}) + \left\langle \nabla f(\mathbf{x}^{(t)}), \mathbf{x}-\mathbf{x}^{(t)}\right\rangle + \frac{1}{2\eta}\|\mathbf{x}-\mathbf{x}^{(t)}\|_2^2\right\}.$$

تعبیر شهودی این رابطه چیست؟

پاسخ پرسش تئوری ۲. به وضوح نقطه‌ی بهینه‌ی عبارت $f(\mathbf{x}^{(t)})+\left\langle \nabla f(\mathbf{x}^{(t)}), \mathbf{x}-\mathbf{x}^{(t)}\right\rangle+\frac{1}{2\eta}\|\mathbf{x}-\mathbf{x}^{(t)}\|_2^2$ وقتی رخ می‌دهد که مشتق در آن نقطه برابر صفر باشد زیرا یک تابع محدب بر حسب $\mathbf{x}$ است.

$$\frac{\partial}{\partial \mathbf{x}}(f(\mathbf{x}^{(t)})+\left\langle \nabla f(\mathbf{x}^{(t)}), \mathbf{x}-\mathbf{x}^{(t)}\right\rangle+\frac{1}{2\eta}\|\mathbf{x}-\mathbf{x}^{(t)}\|_2^2) = \nabla f(\mathbf{x}^{(t)})-\frac{1}{\eta}(\mathbf{x}-\mathbf{x}^{(t)})$$

پس در نقطه‌ی بهینه خواهیم داشت:

$$\nabla f(\mathbf{x}^{(t)}) - \frac{1}{\eta}(\mathbf{x}_{\text{opt}} - \mathbf{x}^{(t)}) = 0 \implies \mathbf{x}_{\text{opt}} = \mathbf{x}^{(t)} - \eta\nabla f(\mathbf{x}^{(t)}) = \mathbf{x}^{(t+1)}$$

پس تساوی مورد نظر اثبات شد. تعبیر شهودی این رابطه نیز این است که $\mathbf{x}^{(t+1)}$ درواقع نقطه‌ای در نزدیکی $\mathbf{x}^{(t)}$ است (بسته به پارامتر η) که تقریب خطی f در آن کمینه شود.

۳ از گوشه‌ای برون آی ای کوکبِ هدایت![۲]

حال یک قدم در راستای سخت‌کردن مسئله برمی‌داریم. در حالتی که تابع f در همه‌ی نقاط مشتق‌پذیر نباشد، باید از مفهوم زیرگرادیان استفاده کنیم. در درس با مفهوم زیرگرادیان آشنا شده‌ایم.

تعریف ۳-۱. تابع $f : \mathbb{R}^n \to \mathbb{R}$ را در نظر بگیرید، بردار $\mathbf{v}$ را یک زیرگرادیان برای تابع f در نقطه‌ی $\mathbf{x} \in \mathbb{R}^n$ می‌نامیم، اگر به ازای هر $\mathbf{y} \in \mathbb{R}^n$ داشته باشیم:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{y} - \mathbf{x}, \mathbf{v} \rangle \tag{۳}$$

مجموعه‌ی همه‌ی زیرگرادیان‌های f در نقطه‌ی $\mathbf{x}$ را با $\partial f(\mathbf{x})$ نشان می‌دهیم.

تعمیم الگوریتم ۱ به حالتی که به جای گرادیان، به زیرگرادیان دسترسی داشته باشیم، سرراست است:

Algorithm 2 Sub-Gradient Descent

parameters: Set of Scalars $\{\eta_t\}_{t=1}^{T}$ where for all $t \in [T]$, $\eta_t > 0$, integer $T > 0$, Initial Point $\mathbf{x}_0$
Initialize $\mathbf{x}^{(1)} = \mathbf{x}_0$
for $t = 1, 2, \cdots, T$ **do**
  choose $\mathbf{v}^{(t)}$ such that $\mathbf{v}^{(t)} \in \partial f(\mathbf{x}^{(t)})$
  update $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta_t \mathbf{v}^{(t)}$
end
return $\mathbf{x}^{(T+1)}$

پرسشی که وجود دارد، آنست که آیا همواره حرکت در خلاف جهت زیرگرادیان تابع را کاهش می‌دهد یا خیر؟

پرسش تئوری ۳. فرض کنید که $\mathbf{0} \notin \partial f(\mathbf{x}^{(t)})$ و تعریف کنید $\mathbf{x}^* = \arg\min_{\mathbf{x} \in \mathbb{R}^n} \{f(\mathbf{x})\}$. نشان دهید به ازای هر بردار $\mathbf{v}^{(t)}$ که در مورد آن داشته باشیم $\mathbf{v}^{(t)} \in \partial f(\mathbf{x}^{(t)})$، می‌توان یک $\eta_t > 0$ یافت که به ازای آن داشته باشیم:

$$\|\mathbf{x}^{(t+1)} - \mathbf{x}^*\|_2^2 = \|\mathbf{x}^{(t)} - \eta_t \mathbf{v}^{(t)} - \mathbf{x}^*\|_2^2 < \|\mathbf{x}^{(t)} - \mathbf{x}^*\|_2^2.$$

پاسخ پرسش تئوری ۳. به وضوح داریم $f(\mathbf{x}^{(t)}) > f(\mathbf{x}^*)$ چرا که در غیر این صورت باید $0 \in \partial f(\mathbf{x}^{(t)})$ باشد. و همچنین داریم،

$$f(\mathbf{x}^*) \geq f(\mathbf{x}^{(t)}) + \langle v^{(t)},\ \mathbf{x}^* - \mathbf{x}^{(t)} \rangle$$

پس داریم،

$$f(\mathbf{x}^{(t)}) > f(\mathbf{x}^{(t)}) + \langle v^{(t)},\ \mathbf{x}^* - \mathbf{x}^{(t)} \rangle \implies \langle v^{(t)},\ \mathbf{x}^* - \mathbf{x}^{(t)} \rangle < 0$$

[۲] در این شبِ سیاهم گم گشت راه مقصود/ از گوشه‌ای برون آی ای کوکب هدایت
از هر طرف که رفتم جز وحشتم نَیَفزود/ زِنهار از این بیابان وین راه بی‌نهایت [حافظ]

$$
\begin{aligned}
&\|\mathbf{x}^{(t)}-\mathbf{x}^*\|_2^2-\|\mathbf{x}^{(t)}-\eta_t\mathbf{v}^{(t)}-\mathbf{x}^*\|_2^2\\
&=(\mathbf{x}^{(t)}-\mathbf{x}^*)^T(\mathbf{x}^{(t)}-\mathbf{x}^*)-(\mathbf{x}^{(t)}-\mathbf{x}^*)^T(\mathbf{x}^{(t)}-\mathbf{x}^*)\\
&+2{\mathbf{x}^{(t)}}^T\eta_t\mathbf{v}^{(t)}-2{\mathbf{x}^*}^T\eta_t\mathbf{v}^{(t)}-\eta_t{\mathbf{v}^{(t)}}^T\eta_t\mathbf{v}^{(t)}\\
&=\eta_t((2\mathbf{x}^{(t)}-\mathbf{x}^*)^T\mathbf{v}^{(t)}-\eta_t\|\mathbf{v}^{(t)}\|_2^2)\\
&=\eta_t(2\langle v^{(t)},\ \mathbf{x}^*-\mathbf{x}^{(t)}\rangle-\eta_t\|\mathbf{v}^{(t)}\|_2^2)
\end{aligned}
$$

پس اگر η_t را به نحوی انتخاب کنیم که

$$0<\eta_t<2\frac{\langle v^{(t)},\ \mathbf{x}^*-\mathbf{x}^{(t)}\rangle}{\|\mathbf{v}^{(t)}\|_2^2}$$

آنگاه عبارت مد نظر، مثبت خواهد شد و حکم مساله ثابت می‌شود.

حالا آماده هستیم که همگرایی الگوریتم ۲ را بررسی کنیم. ابتدا دو فرض به مسئله اضافه می‌کنیم:

فرض ۳-۱. مقدار بهینه‌ی تابع f در بی‌نهایت رخ نمی‌دهد.
به تعبیر دیگر اگر تعریف کنیم: $\mathbf{x}^*=\arg\min_{\mathbf{x}\in\mathbb{R}^n}\{f(\mathbf{x})\}$، خواهیم داشت: $\|\mathbf{x}^*\|_2\le B$ که $B<+\infty$.

فرض ۳-۲. تابع f، یک تابع ρ-لیپشیتز است. یعنی به ازای هر $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$:

$$|f(\mathbf{x})-f(\mathbf{y})|\le\rho\|\mathbf{x}-\mathbf{y}\|_2.$$

پرسش تئوری ۴. برای هر تابع ρ-لیپشیتز مانند f نشان دهید که درباره‌ی هر بردار زیرگرادیان مانند $\mathbf{v}$ داریم:

$$\|\mathbf{v}\|_2\le\rho.$$

پاسخ پرسش تئوری ۴. با برهان خلف فرض کنید یک بردار در زیرمشتق مانند $\mathbf{v}$ وجود داشته باشد که

$$\|\mathbf{v}\|_2>\rho$$

$$\mathbf{v}\in\partial f(\mathbf{x})\implies f(\mathbf{x}+\mathbf{v})\ge f(\mathbf{x})+\mathbf{v}^T(\mathbf{x}+\mathbf{v}-\mathbf{x})=f(\mathbf{x})+\|\mathbf{v}\|_2^2$$

$$\frac{f(\mathbf{x}+\mathbf{v})-f(\mathbf{x})}{\|\mathbf{v}\|_2}\ge\|\mathbf{v}\|_2>\rho$$

پس فرض خلف غلط بوده و لذا حکم درست می‌باشد.

پرسش تئوری ۵. تعریف می‌کنیم: $\mathbf{x}^*=\arg\min_{\mathbf{x}\in\mathbb{R}^n}\{f(\mathbf{x})\}$، با توجّه به الگوریتم ۲ نشان دهید که به ازای هر η_t داریم:

$$\frac{1}{2}\|\mathbf{x}^{(t+1)}-\mathbf{x}^*\|_2^2\le\frac{1}{2}\|\mathbf{x}^{(t)}-\mathbf{x}^*\|_2^2-\eta_t\big(f(\mathbf{x}^{(t)})-f(\mathbf{x}^*)\big)+\frac{\eta_t^2}{2}\|\mathbf{v}^{(t)}\|_2^2.$$

پاسخ پرسش تئوری ۵. حکم نامساوی را ساده می‌کنیم تا به احکام معادل برسیم.

$$\frac{1}{2}\|\mathbf{x}^{(t+1)}-\mathbf{x}^*\|_2^2 \le \frac{1}{2}\|\mathbf{x}^{(t)}-\mathbf{x}^*\|_2^2 - \eta_t\big(f(\mathbf{x}^{(t)})-f(\mathbf{x}^*)\big) + \frac{\eta_t^2}{2}\|\mathbf{v}^{(t)}\|_2^2$$
$$\Leftrightarrow \frac{1}{2}\|\mathbf{x}^{(t+1)}-\mathbf{x}^*\|_2^2 - \frac{1}{2}\|\mathbf{x}^{(t)}-\mathbf{x}^*\|_2^2 - \frac{\eta_t^2}{2}\|\mathbf{v}^{(t)}\|_2^2 \le -\eta_t\big(f(\mathbf{x}^{(t)})-f(\mathbf{x}^*)\big) \Leftrightarrow$$
$$\frac{1}{2}\|\mathbf{x}^{(t+1)}-\mathbf{x}^{(t)}+\mathbf{x}^{(t)}-\mathbf{x}^*\|_2^2 - \frac{1}{2}\|\mathbf{x}^{(t)}-\mathbf{x}^*\|_2^2 - \frac{\eta_t^2}{2}\|\mathbf{v}^{(t)}\|_2^2 \le -\eta_t\big(f(\mathbf{x}^{(t)})-f(\mathbf{x}^*)\big)$$
$$\Leftrightarrow \langle \mathbf{x}^{(t+1)}-\mathbf{x}^{(t)},\ \mathbf{x}^{(t)}-\mathbf{x}^*\rangle \le -\eta_t(f(\mathbf{x}^{(t)})-f(\mathbf{x}^*)) \Leftrightarrow$$
$$\langle -\eta_t\mathbf{v}^{(t)},\ \mathbf{x}^{(t)}-\mathbf{x}^*\rangle \le -\eta_t(f(\mathbf{x}^{(t)})-f(\mathbf{x}^*)) \Leftrightarrow$$
$$\langle \mathbf{v}^{(t)},\ \mathbf{x}^{(t)}-\mathbf{x}^*\rangle \ge (f(\mathbf{x}^{(t)})-f(\mathbf{x}^*))$$

که چون $\mathbf{v}^{(t)} \in \partial f(\mathbf{x}^{(t)})$، نامساوی آخر واضح است.

پرسش تئوری ۶. فرض کنید $\{\eta_t\}_{t=1}^T$ هر دنباله‌ی دلخواهی از اعداد نامنفی باشد. همچنین فرض کنید که مفروضات ۳-۱ و ۳-۲ برقرار باشند. نشان دهید اگر در الگوریتم ۲ قرار دهیم: $\mathbf{x}_0 = \mathbf{0}$، آن‌گاه:

$$\sum_{t=1}^T \eta_t\big(f(\mathbf{x}^{(t)})-f(\mathbf{x}^*)\big) \le \frac{1}{2}B^2 + \frac{1}{2}\rho^2\sum_{t=1}^T \eta_t^2.$$

پاسخ پرسش تئوری ۶. از سوال قبل داریم:

$$\eta_t\big(f(\mathbf{x}^{(t)})-f(\mathbf{x}^*)\big) \le \frac{1}{2}\|\mathbf{x}^{(t)}-\mathbf{x}^*\|_2^2 + \frac{\eta_t^2}{2}\|\mathbf{v}^{(t)}\|_2^2 - \frac{1}{2}\|\mathbf{x}^{(t+1)}-\mathbf{x}^*\|_2^2$$

$$\Rightarrow \sum_{t=1}^T(\eta_t\big(f(\mathbf{x}^{(t)})-f(\mathbf{x}^*)\big)) \le \sum_{t=1}^T(\frac{1}{2}\|\mathbf{x}^{(t)}-\mathbf{x}^*\|_2^2 + \frac{\eta_t^2}{2}\|\mathbf{v}^{(t)}\|_2^2 - \frac{1}{2}\|\mathbf{x}^{(t+1)}-\mathbf{x}^*\|_2^2)$$
$$= \frac{1}{2}\sum_{t=1}^T(\eta_t^2\|\mathbf{v}^{(t)}\|_2^2) - \frac{1}{2}(\|\mathbf{x}^{(T)}-\mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(1)}-\mathbf{x}^*\|_2^2)$$
$$\le \frac{1}{2}\sum_{t=1}^T(\eta_t^2\rho^2) + \frac{1}{2}\|\mathbf{x}^{(1)}-\mathbf{x}^*\|_2^2 = \frac{1}{2}\rho^2\sum_{t=1}^T(\eta_t^2) + \frac{1}{2}\|\mathbf{x}^*\|_2^2$$
$$\le \frac{1}{2}\rho^2\sum_{t=1}^T(\eta_t^2) + \frac{1}{2}B^2 \quad \blacksquare$$

پرسش تئوری ۷. تعریف می‌کنیم:

$$\bar{\mathbf{x}}_T = \frac{\sum_{t=1}^T \eta_t\mathbf{x}^{(t)}}{\sum_{t=1}^T \eta_t}.$$

نشان دهید:

$$f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) \leq \frac{B^2 + \rho^2 \sum_{t=1}^T \eta_t^2}{2\sum_{t=1}^T \eta_t}.$$

پاسخ پرسش تئوری ۷. بنا به تعریف داریم،

$$f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) = f(\frac{\sum_{t=1}^T \eta_t \mathbf{x}^{(t)}}{\sum_{t=1}^T \eta_t}) - f(\mathbf{x}^*)$$

حال چون تابع f یک تابع محدب است، بنابر نامساوی ینسن داریم،

$$\begin{aligned} f(\frac{\sum_{t=1}^T \eta_t \mathbf{x}^{(t)}}{\sum_{t=1}^T \eta_t}) - f(\mathbf{x}^*) &\leq \frac{\sum_{t=1}^T \eta_t f(\mathbf{x}^{(t)})}{\sum_{t=1}^T \eta_t} - f(\mathbf{x}^*) \\ = \frac{\sum_{t=1}^T \eta_t (f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*))}{\sum_{t=1}^T \eta_t} &\leq \frac{\frac{1}{2}B^2 + \frac{1}{2}\rho^2 \sum_{t=1}^T \eta_t^2}{\sum_{t=1}^T \eta_t} \quad \blacksquare \end{aligned}$$

که نامساوی آخر را از پرسش قبل می‌دانیم.

پرسش تئوری ۸. قرار دهید $\eta_t = \eta$. کرانی که در پرسش تئوری ۷ اثبات کردید را بازنویسی کنید و مقدار بهینه‌ی η را برای به دست آوردن بهترین کران محاسبه کنید. در نهایت بهترین کران را گزارش کنید. پاسخ نهایی شما باید برحسب B, ρ, T باشد.

پاسخ پرسش تئوری ۸. بنابر قسمت قبل داریم،

$$f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) \leq \frac{B^2 + \rho^2 T \eta^2}{2\eta}$$

حال برای کمینه کردن قسمت راست نامساوی از آن مشتق می‌گیریم و برابر صفر می‌گذاریم.

$$\frac{\partial}{\partial \eta}(\frac{B^2 + \rho^2 T \eta^2}{2\eta}) = -\frac{B^2}{2T\eta^2} + \frac{\rho^2}{2} \Rightarrow \eta^* = \frac{B}{\rho\sqrt{T}}$$

پرسش تئوری ۹. به طور شهودی توجیه کنید که چرا کران نهایی برحسب $f(\bar{\mathbf{x}}_T)$ به دست آمد، نه برحسب $f(\mathbf{x}^{(T)})$ یا $f(\mathbf{x}^{(T+1)})$.

پاسخ پرسش تئوری ۹. چون در هر مرحله لزوما بهترین η ممکن را انتخاب نمی‌کنیم و داریم عدد ثابت انتخاب می‌کنیم، بنابراین در نهایت لزوما آخرین نقطه‌ای که به آن می‌رسیم نقطه بهینه نیست.
اگر تعداد زیادی نقطه حول نقطه بهینه در نقاطمان داشته باشیم درواقع به نوعی به طور میانگین داریم تابع را به میزان بهینه‌اش نزدیک می‌کنیم.

پرسش تئوری ۱۰. تعریف می‌کنیم: $t_T^* = \arg\min_{t\in[T]} f(\mathbf{x}^{(t)})$. نشان دهید:

$$f(\mathbf{x}^{(t_T^*)}) - f(\mathbf{x}^*) \leq \frac{B^2 + \rho^2 \sum_{t=1}^T \eta_t^2}{2\sum_{t=1}^T \eta_t}.$$

پاسخ پرسش تئوری ۱۰. بنابر پرسش تئوری ۷ کافی است نشان دهیم،

$$f(\mathbf{x}^{(t_T^*)}) \leq f(\bar{\mathbf{x}}_T)$$

$$f(\mathbf{x}^{(t_T^*)}) = \frac{\sum_{t=1}^T \eta_t f(\mathbf{x}^{(t_T^*)})}{\sum_{t=1}^T \eta_t} \leq \frac{\sum_{t=1}^T \eta_t f(\mathbf{x}^{(t)})}{\sum_{t=1}^T \eta_t} = f(\bar{\mathbf{x}}_T) \quad \blacksquare$$

پرسش شبیه‌سازی ۱. قرار دهید $m = 100, n = 10$. بردارهای تصادفی $\{\mathbf{Z}_i\}_{i=1}^m$ را به صورت مستقل و هم‌توزیع از توزیع $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ بسازید. همچنین متغیّرهای تصادفی $\{B_i\}_{i=1}^m$ را به صورت مستقل و هم‌توزیع از توزیع $\mathcal{N}(0, 1)$ بسازید. این متغیّرها در طول شبیه‌سازی تغییر نمی‌کنند. تابع هدف زیر را تشکیل دهید:

$$f(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^m |\mathbf{Z}_i^\top \mathbf{x} - B_i|.$$

زیرگرادیان این تابع را به صورت تئوری محاسبه کنید. سپس الگوریتم ۲ را با $\mathbf{x}_0 = \mathbf{0}, T = 4000$ و $\eta_t = \eta = \{0.01, 0.1, 1, 10\}$ اجرا کنید و کمیت $f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*)$ را در طول اجرای الگوریتم محاسبه و نمودار آن را رسم کنید.

پرسش شبیه‌سازی ۲. به ازای $\{\mathbf{Z}_i\}_{i=1}^m, \{B_i\}_{i=1}^m$ای که در پرسش شبیه‌سازی ۱ ساختید و پارامترهای داده‌شده در آن پرسش، الگوریتم ۲ را اجرا کنید و کمیت $f(\mathbf{x}^{(t_t^*)}) - f(\mathbf{x}^*)$ را در طول اجرای الگوریتم محاسبه و نمودار آن را رسم کنید.

---

۴ مرا زین قید ممکن نیست جستن![۳]

حال فرض کنید بخواهیم یک مسئله‌ی بهینه‌سازی محدّب و دارای قید مانند (۲) را با کمک الگوریتم ۲ حل کنیم. موقّتاً فرض می‌کنیم تابع f را در اختیار داریم و خاصیت تصادفی مسئله را فراموش می‌کنیم.
وقتی می‌خواهیم یک مسئله‌ی بهینه‌سازی مقیّد را حل کنیم، با پیمودن گامِ موجود در الگوریتم ۲، یعنی $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta_t \mathbf{v}^{(t)}$ ممکن است از مجموعه‌ی $\mathcal{C}$ خارج شویم! در نتیجه بعد از پیمودن هر گام، باید نقطه‌ی به دست آمده را روی مجموعه‌ی $\mathcal{C}$ تصویر کنیم. تابع زیر را به این منظور در نظر می‌گیریم:

$$\Pi_{\mathcal{C}}(\mathbf{x}) = \arg\min_{\mathbf{y}\in\mathcal{C}} \left\{ \|\mathbf{y} - \mathbf{x}\|_2 \right\}. \qquad (۴)$$

پرسش تئوری ۱۱. فرض کنید $\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{b}\}$ که $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$. فرم بسته‌ی تابع $\Pi_{\mathcal{C}}(\mathbf{x})$ را به ازای هر $\mathbf{x} \in \mathbb{R}^n$ بیابید.

پاسخ پرسش تئوری ۱۱. از جبرخطی می‌دانیم که معادله‌ی $A\mathbf{x} = b$، جواب دارد اگر و تنها اگر فضای پوچ ماتریسمان ناتهی باشد.
فرض کنید ماتریس F ماتریسی با ستون‌هایی که باشد که یک پایه‌ی متعامد یکه برای فضای پوچ ماتریس A باشند.
در درس دیدیم که هر جواب این معادله را می‌توان به شکل $Fw + x_0$ نوشت که x_0 یک جواب معادله است.
پس تابع هدف مساله به فرم $\|Fw + x_0 - y\|_2$ می‌شود که همان مساله کمترین مربعات است.
جواب این مساله نیز می‌دانیم که بر حسب $\mathbf{x}$ به این شکل می‌شود:

$$FF^Ty + \mathbf{x}_0 - FF^T\mathbf{x}_0$$

پرسش تئوری ۱۲. فرض کنید $\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$ که $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$. فرم بسته‌ی تابع $\Pi_{\mathcal{C}}(\mathbf{x})$ را به ازای هر $\mathbf{x} \in \mathbb{R}^n$ بیابید.

پاسخ پرسش تئوری ۱۲.

پرسش تئوری ۱۳. فرض کنید $\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 \leq b\}$ که $b \in \mathbb{R}^{\geq 0}$. فرم بسته‌ی تابع $\Pi_{\mathcal{C}}(\mathbf{x})$ را به ازای هر $\mathbf{x} \in \mathbb{R}^n$ بیابید.

پاسخ پرسش تئوری ۱۳. از آنجایی که $\|\mathbf{y} - \mathbf{x}\|_2$ یک عبارت مثبت است پس کمینه کردنش تفاوتی با کمینه کردن $\|\mathbf{y} - \mathbf{x}\|_2^2$ ندارد.

[۳] همی‌گویم بگریم در غمت زار/ دگر گویم بخندی بر گرستن
گر آزادم کنی ور بنده خوانی/ مرا زین قید ممکن نیست جستن
گرم دشمن شوی ور دوست گیری/ نخواهم دستت از دامن گسستن [سعدی]

حال برای مساله‌ی دوم که در بالا گفتیم، تابع لاگرانژ محاسبه می‌کنیم.

$$\mathcal{L}(\mathbf{y},\ \lambda) = \|\mathbf{y}-\mathbf{x}\|_2^2 + \lambda(\|\mathbf{y}\|_2^2 - b^2)$$
$$\frac{\partial}{\partial \mathbf{y}}\mathcal{L}(\mathbf{y},\ \lambda) = 2(\mathbf{y}-\mathbf{x}) + \lambda\mathbf{y} \implies \mathbf{y}^* = \frac{x}{2+\lambda}$$

پس $\mathbf{y}$ در واقع ضریبی از $\mathbf{x}$ می‌شود.
حال به وضوح اگر $\mathbf{x}$ در بازه‌ی مجاز باشد که جواب حاصل خود $\mathbf{x}$ است و اگر نباشد، چون باید ضریبی از خودش باشد، نقطه‌ی بهینه همان $\frac{b\mathbf{x}}{\|\mathbf{x}\|_2}$ می‌شود. (یا قرینه‌اش)

با این ویرایش، الگوریتم کاهش زیرگرادیان در حالت مقیّد به این صورت درمی‌آید:

Algorithm 3 Constrained Sub-Gradient Descent

parameters: Set of Scalars $\{\eta_t\}_{t=1}^T$ where for all $t \in [T]$, $\eta_t > 0$, integer $T > 0$, Initial Point $\mathbf{x}_0$
Initialize $\mathbf{x}^{(1)} = \mathbf{x}_0$
for $t = 1, 2, \cdots, T$ **do**
  choose $\mathbf{v}^{(t)}$ such that $\mathbf{v}^{(t)} \in \partial f(\mathbf{x}^{(t)})$
  update $\mathbf{x}^{(t+1)} = \Pi_{\mathcal{C}}\left(\mathbf{x}^{(t)} - \eta_t \mathbf{v}^{(t)}\right)$
end
return $\bar{\mathbf{x}}_T = \frac{\sum_{t=1}^T \eta_t \mathbf{x}^{(t)}}{\sum_{t=1}^T \eta_t}$

پرسش تئوری ۱۴. نشان دهید رابطه‌ی به‌روز کردن الگوریتم ۳ یعنی $\mathbf{x}^{(t+1)} = \Pi_{\mathcal{C}}\left(\mathbf{x}^{(t)} - \eta_t \mathbf{v}^{(t)}\right)$ معادل با رابطه‌ی زیر است:

$$\mathbf{x}^{(t+1)} = \arg\min_{\mathbf{x}\in\mathcal{C}} \left\{ f(\mathbf{x}^{(t)}) + \left\langle \mathbf{v}^{(t)}, \mathbf{x} - \mathbf{x}^{(t)} \right\rangle + \frac{1}{2\eta_t}\|\mathbf{x} - \mathbf{x}^{(t)}\|_2^2 \right\}.$$

تعبیر شهودی این رابطه چیست؟

پاسخ پرسش تئوری ۱۴.

$$\mathbf{x}^{(t+1)} = \Pi_{\mathcal{C}}\left(\mathbf{x}^{(t)} - \eta_t \mathbf{v}^{(t)}\right) = \arg\min_{\mathbf{x}\in\mathcal{C}} \left\{ \frac{1}{2\eta_t}\|\mathbf{x} - \mathbf{x}^{(t)} + \eta_t \mathbf{v}^{(t)}\|_2^2 \right\}$$

در حالت مقیّد نیاز به دو فرض جدید برای تحلیل تئوری الگوریتم ۳ داریم:

فرض ۴-۱. مقدار بهینه‌ی تابع f در بی‌نهایت رخ نمی‌دهد و مجموعه‌ی $\mathcal{C} \subset \mathbb{R}^n$ یک مجموعه‌ی محدّب و فشرده است. به تعبیر دیگر اگر تعریف کنیم: $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{C}}\{f(\mathbf{x})\}$، به ازای هر $\mathbf{x} \in \mathcal{C}$ خواهیم داشت: $\|\mathbf{x} - \mathbf{x}^*\|_2 \le B$ که $B < +\infty$.

فرض ۴-۲. تابع f، یک تابع ρ–لیپشیتز است. یعنی به ازای هر $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ داریم:

$$|f(\mathbf{x}) - f(\mathbf{y})| \le \rho\|\mathbf{x} - \mathbf{y}\|_2.$$

پرسش تئوری ۱۵. فرض کنید دنباله‌ی $\{\eta_t\}_{t=1}^{T}$ دنباله‌ای ناصعودی باشد و همچنین فرض کنید مفروضات ۴-۱ و ۴-۲ برقرار باشند. از روند اثبات قضایای مربوط به الگوریتم ۲ کمک بگیرید و نشان دهید که در این حالت، اگر الگوریتم ۳ را با هر $\mathbf{x}_0 \in \mathcal{C}$ اجرا کنیم، داریم:

$$\sum_{t=1}^{T}\left(f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*)\right) \leq \frac{1}{2\eta_T}B^2 + \frac{1}{2}\rho^2\sum_{t=1}^{T}\eta_t.$$

پاسخ پرسش تئوری ۱۵.

پرسش تئوری ۱۶. قرار دهید: $\eta_t = \frac{\alpha}{\sqrt{t}}$ و فرض کنید مفروضات ۴-۱ و ۴-۲ برقرار باشند. نشان دهید که در این حالت، اگر الگوریتم ۳ را با هر $\mathbf{x}_0 \in \mathcal{C}$ اجرا کنیم و تعریف کنیم $\bar{\mathbf{x}}_T = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}^{(t)}$، داریم:

$$f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) \leq \frac{B^2}{2\alpha\sqrt{T}} + \frac{\rho^2\alpha}{\sqrt{T}}.$$

همچنین مقدار بهینه‌ی α برای محاسبه‌ی بهترین کران را بیابید.

پاسخ پرسش تئوری ۱۶. مشابه اثباتی که در پرسش تئوری ۷ انجام دادیم پیش می‌رویم و از نامساوی ینسن و همچنین نامساوی پرسش قبل استفاده می‌کنیم.

$$f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) = f(\frac{1}{T}\sum_{t=1}^{T}f(\mathbf{x}^{(t)})) - f(\mathbf{x}^*) \leq \frac{1}{T}\sum_{t=1}^{T}(f(\mathbf{x}^{(t)})) - f(\mathbf{x}^*)$$

$$= \frac{1}{T}(\sum_{t=1}^{T}(f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*))) \leq \frac{1}{T}(\frac{1}{2\eta_T}B^2 + \frac{1}{2}\rho^2\sum_{t=1}^{T}\eta_t)$$

$$= \frac{1}{T}(\frac{\sqrt{T}}{2\alpha}B^2 + \frac{1}{2}\rho^2\sum_{t=1}^{T}\frac{\alpha}{\sqrt{t}})$$

پس الان کافی است ثابت کنیم که،

$$\frac{\rho^2\alpha}{2T}\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq \frac{\rho^2\alpha}{\sqrt{T}} \Leftrightarrow \sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leq 2\sqrt{T}$$

که این نامساوی به راحتی با استقرا و به توان ۲ رساندن ثابت می‌شود.

۵ مرا امیدِ وصالِ تو زنده می‌دارد![۴]

حالا می‌توانیم مسئله‌ای مانند مسئله‌ی (۲) را حل کنیم. برای حلّ این مسئله ابتدا مفهوم زیرگرادیان تصادفی را تعریف می‌کنیم:

تعریف ۵-۱. تابع $f : \mathbb{R}^n \to \mathbb{R}$ را در نظر بگیرید. بردار تصادفی $\mathbf{V} \in \mathbb{R}^n$ را یک زیرگرادیان تصادفی برای تابع f در نقطه‌ی $\mathbf{x} \in \mathbb{R}^n$ می‌نامیم، اگر به ازای هر $\mathbf{y} \in \mathbb{R}^n$ داشته باشیم $\mathbb{E}\left[\mathbf{V}\right] \in \partial f(\mathbf{x})$، یا به تعبیر دیگر:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{y} - \mathbf{x}, \mathbb{E}\left[\mathbf{V}\right] \rangle . \tag{۵}$$

و در نتیجه به الگوریتم کاهش گرادیان تصادفی در حالت مقیّد می‌رسیم:

Algorithm 4 Stochastic Gradient Descent (SGD)

parameters: Set of Scalars $\{\eta_t\}_{t=1}^{T}$ where for all $t \in [T]$, $\eta_t > 0$, integer $T > 0$, Initial Point $\mathbf{x}_0$
Initialize $\mathbf{x}^{(1)} = \mathbf{x}_0$
for $t = 1, 2, \cdots, T$ **do**
    choose $\mathbf{V}^{(t)}$ at random from a distribution such that $\mathbb{E}\left[\mathbf{V}^{(t)} | \mathbf{X}^{(t)}\right] \in \partial f(\mathbf{X}^{(t)})$
    update $\mathbf{X}^{(t+1)} = \Pi_{\mathcal{C}}\left(\mathbf{X}^{(t)} - \eta \mathbf{V}^{(t)}\right)$
end
return $\bar{\mathbf{X}}_T = \frac{1}{T} \sum_{t=1}^{T} \mathbf{X}^{(t)}$

پرسش تئوری ۱۷. در الگوریتم ۴ بردارهای $\mathbf{X}^{(t)}$ با حروف بزرگ نوشته شده‌اند به این دلیل که این بردارها تصادفی هستند. منشأ این خاصیت تصادفی چیست؟

پاسخ پرسش تئوری ۱۷. همانطور که در الگوریتم مشخص است، بردار $\mathbf{V}^{(t)}$ها به طور تصادفی انتخاب می‌شوند؛ پس بدیهتا $\mathbf{X}^{(t)}$ها نیز تصادفی هستند چرا که از روی $\mathbf{V}^{(t)}$ها ساخته می‌شوند.

پرسش تئوری ۱۸. چرا در انتخاب بردار زیرگرادیان تصادفی امید ریاضی شرطی دیده می‌شود؟ تعبیر شهودیِ این عبارت چیست؟

پاسخ پرسش تئوری ۱۸. چون به وضوح توزیع $\mathbf{V}^{(t)}$ با داشتن $\mathbf{X}^{(t)}$ مشخص می‌شود جون باید زیرگرادیان در آن نقطه باشد. پس باید از امید ریاضی شرطی استفاده کنیم.
تعبیر شهودی آن نیز این است که متوسط تقریب خطی $f(\mathbf{y})$ حول نقطه‌ی $f(\mathbf{x})$ زیر نمودار $f(\mathbf{y}) - f(\mathbf{x})$ باشد.

برای تحلیل تئوری الگوریتم ۴ نیاز به فرض ۴-۱ داریم. همچنین باید فرض ۴-۲ را به صورت زیر تغییر دهیم:

[۴] هزار دشمنم اَر می‌کنند قصد هلاک/ گَرَم تو دوستی از دشمنان ندارم باک
مرا امیدِ وصالِ تو زنده می‌دارد/ و گر نه هر دَمَم از هجرِ توست بیمِ هلاک
نَفَس نَفَس اگر از باد نَشنوم بویش/ زمان زمان چو گل از غم کُنَم گریبان چاک [حافظ]

فرض ۵-۱. درباره‌ی بردار گرادیان تصادفی می‌دانیم:

$$\mathbb{E}\left[\|\mathbf{V}\|_2^2\right] \leq \rho^2.$$

پرسش تئوری ۱۹. تعریف می‌کنیم $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{C}}\{f(\mathbf{x})\}$. با توجّه به الگوریتم ۴، اگر تعریف کنیم $\boldsymbol{\xi}_t = \mathbf{V}^{(t)} - \mathbb{E}\left[\mathbf{V}^{(t)} \mid \mathbf{X}^{(t)}\right]$، نشان دهید:

$$\frac{1}{2}\|\mathbf{X}^{(t+1)} - \mathbf{x}^*\|_2^2 \leq \frac{1}{2}\|\mathbf{X}^{(t)} - \mathbf{x}^*\|_2^2 - \eta_t\big(f(\mathbf{X}^{(t)}) - f(\mathbf{x}^*)\big) + \frac{\eta_t^2}{2}\|\mathbf{V}^{(t)}\|_2^2 - \eta_t\left\langle \boldsymbol{\xi}_t, \mathbf{X}^{(t)} - \mathbf{x}^*\right\rangle$$

پاسخ پرسش تئوری ۱۹.

پرسش تئوری ۲۰. نشان دهید:

$$\mathbb{E}\left[\left\langle \boldsymbol{\xi}_t, \mathbf{X}^{(t)} - \mathbf{x}^*\right\rangle\right] = 0.$$

پاسخ پرسش تئوری ۲۰.

$$\begin{aligned}\mathbb{E}\left[\left\langle \boldsymbol{\xi}_t, \mathbf{X}^{(t)} - \mathbf{x}^*\right\rangle\right] &= \mathbb{E}_{\mathbf{X}^{(t)},\mathbf{V}^{(t)}}\left[\left\langle \boldsymbol{\xi}_t, \mathbf{X}^{(t)} - \mathbf{x}^*\right\rangle\right] \\ &= \left\langle \mathbb{E}_{\mathbf{X}^{(t)}}\left[\mathbb{E}_{\mathbf{V}^{(t)}|\mathbf{X}^{(t)}}\left[\boldsymbol{\xi}_t\right]\right],\ \mathbf{X}^{(t)} - \mathbf{x}^*\right\rangle\end{aligned}$$

و چون $\boldsymbol{\xi}_t = \mathbf{V}^{(t)} - \mathbb{E}\left[\mathbf{V}^{(t)} \mid \mathbf{X}^{(t)}\right]$ امید ریاضی داخلی صفر می‌شود و کل عبارت برابر صفر می‌شود.

پرسش تئوری ۲۱. فرض کنید $\{\eta_t\}_{t=1}^T$ دنباله‌ای ناصعودی باشد. نشان دهید:

$$\mathbb{E}\left[f(\bar{\mathbf{X}}_T)\right] - f(\mathbf{x}^*) \leq \frac{B^2}{2T\eta_T} + \frac{1}{2T}\rho^2\sum_{t=1}^T \eta_t.$$

پاسخ پرسش تئوری ۲۱.

پرسش تئوری ۲۲. قرار دهید $\eta_t = \frac{B}{\rho\sqrt{t}}$. نشان دهید:

$$\mathbb{E}\left[f(\bar{\mathbf{X}}_T)\right] - f(\mathbf{x}^*) \leq \frac{3B\rho}{2\sqrt{T}}.$$

پاسخ پرسش تئوری ۲۲. از قسمت قبل داریم،

$$\mathbb{E}\left[f(\bar{\mathbf{X}}_T)\right] - f(\mathbf{x}^*) \leq \frac{B^2}{2T\eta_T} + \frac{1}{2T}\rho^2\sum_{t=1}^T \eta_t = \frac{\rho\sqrt{T}B^2}{2TB} + \frac{1}{2T}\rho^2\sum_{t=1}^T \frac{B}{\rho\sqrt{t}}$$

پس داریم،

$$\begin{aligned}\mathbb{E}\left[f(\bar{\mathbf{X}}_T)\right] - f(\mathbf{x}^*) &\leq \frac{\rho B}{2\sqrt{T}} + \frac{B}{2T}\rho\sum_{t=1}^T \frac{1}{\sqrt{t}} \leq \frac{\rho B}{2\sqrt{T}} + \frac{B}{2T}\rho 2\sqrt{T} \\ &= \frac{3B\rho}{2\sqrt{T}} \quad \blacksquare\end{aligned}$$

پرسش تئوری ۲۳. فرض کنید تابع f، قویاً محدّب با پارامتر $\lambda \geq 0$ باشد، یعنی به ازای هر $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ و هر بردار مانند $\mathbf{v}$ که $\mathbf{v} \in \partial f(\mathbf{x})$، داشته باشیم:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{v}, \mathbf{y} - \mathbf{x} \rangle + \frac{\lambda}{2} \|\mathbf{y} - \mathbf{x}\|_2^2.$$

همچنین فرض کنید مفروضات ۱-۴ و ۱-۵ برقرار باشند. نشان دهید در این حالت با انتخاب $\eta_t = \frac{1}{\lambda t}$ خواهیم داشت:

$$\mathbb{E}\left[f(\bar{\mathbf{X}}_T)\right] - f(\mathbf{x}^*) \leq \frac{B^2}{2\lambda} \frac{1 + \ln(T)}{T}.$$

پاسخ پرسش تئوری ۲۳.

حال برای حلّ مسئله‌ی (۲) می‌توانیم به ازای هر تحقّق بردار تصادفی $\mathbf{Z}$، مانند $\mathbf{z}$ و به ازای هر $\mathbf{x} \in \mathcal{C}$ بردار $\mathbf{v}$ را به گونه‌ای تعریف کنیم که $\mathbf{v} \in \partial_{\mathbf{x}} F(\mathbf{x}, \mathbf{z})$ باشد.

پرسش تئوری ۲۴. نشان دهید در این حالت $\mathbf{V}$ یک گرادیان تصادفی برای تابع $f(\mathbf{x}) = \mathbb{E}_{\mathbf{Z}}\left[F(\mathbf{x}, \mathbf{Z})\right]$ است.

پاسخ پرسش تئوری ۲۴.

$$F(\mathbf{y}, \mathbf{z}) \geq F(\mathbf{x}, \mathbf{z}) + \langle \mathbf{V}_{\mathbf{x},\mathbf{z}},\ \mathbf{y} - \mathbf{x} \rangle$$

پس با امید ریاضی گرفتن از دو طرف داریم،

$$\mathbb{E}_{\mathbf{Z}}\left[F(\mathbf{y}, \mathbf{Z}) - F(\mathbf{x}, \mathbf{Z}) - \langle \mathbf{V}, \mathbf{y} - \mathbf{x} \rangle\right] \geq 0$$

و با باز کردن امید ریاضی به حکم می‌رسیم.

پرسش شبیه‌سازی ۳. در پرسش شبیه‌سازی ۱ الگوریتم ۴ را اجرا کنید و کمیت $f(\mathbf{X}^{(t)}) - f(\mathbf{x}^*)$ را در طول الگوریتم محاسبه و نمودار آن را رسم کنید. نتیجه را با پرسش شبیه‌سازی ۱ مقایسه کنید.

پرسش شبیه‌سازی ۴. حجم محاسبات انجام‌شده در هر تکرار از الگوریتم ۲ و الگوریتم ۴، هنگامی که برای بهینه‌کردن تابع هدف تعریف‌شده در پرسش شبیه‌سازی ۱ استفاده می‌شوند را تقریب بزنید. حالا نموداری که در پرسش قبل برای مقایسه‌ی کمیت $f(\mathbf{X}^{(t)}) - f(\mathbf{x}^*)$ در طول دو الگوریتم ۲ و ۴ رسم کرده بودید را بر حسب حجم محاسبات بازترسیم کنید.

---

۶ نَبُوَد خیر در آن خانه که عصمت نَبُوَد! [۵]

فرض کنید شما یک تحلیلگر داده در یک شرکت املاک هستید. از شما خواسته شده است که قیمت فروش خانه‌ها را براساس ویژگی‌های آن‌ها پیش‌بینی کنید. همچنین از طرف شرکت به شما تعدادی خانه با قیمت مشخّص داده شده است تا مدل را براساس آن‌ها آموزش دهید. ویژگی‌هایی که از این خانه‌ها به شما داده شده است، شامل مساحت خانه برحسب متر مربع، تعداد اتاق‌خواب‌ها، تعداد سرویس‌های بهداشتی، عمر خانه برحسب سال و مکان خانه است. می‌خواهیم از مدل رگرسیون خطی برای مدل پیش‌بینی استفاده کنیم و برای آموزش این مدل نیز از الگوریتم SGD استفاده خواهیم کرد.

تعریف ۶-۱. فرض کنید مجموعه دادگانی از n مشاهده دارید که هر کدام دارای m ویژگی و یک مقدار خروجی هستند. $\mathbf{X} \in \mathbb{R}^{n\times m}$ ماتریس ویژگی است که در آن هر سطر یک بردار ویژگی برای یک مشاهده است. همچنین $\mathbf{y} \in \mathbb{R}^n$ بردار هدف است. در این بردار هر عنصر نماینده‌ی مقدار خروجی مشاهده‌شده است. $\mathbf{w} \in \mathbb{R}^m$ بردار وزن است که در آن هر عنصر وزنی که به هر ویژگی نسبت می‌دهیم را تعیین می‌کند. همچنین یک پارامتر اسکالر مانند b برای مدل در نظر می‌گیریم. در این صورت مقدار پیش‌بینی شده با مدل رگرسیون خطّی به صورت زیر است:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w} + b\boldsymbol{1}.$$

که $\boldsymbol{1} \in \mathbb{R}^n$ بردار تمام‌یک است.

هدف از یک مسئله رگرسیون خطی تخمین پارامتر های $\mathbf{w}$ و b است.

تعریف ۶-۲. خطای میانگین مربعات یا همان MSE به صورت زیر تعریف می‌شود:

$$l(\mathbf{w}, b) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2.$$

که y_i برابر با مقدار واقعی و $\hat{y}_i$ برابر با مقدار پیش بینی شده است.

پرسش شبیه‌سازی ۵. دادگان `houseData.csv` را فراخوانی کنید و نمودار هایی برای ارزیابی هر ویژگی آن رسم کنید. برای مثال قیمت خانه را بر اساس تعداد اتاق‌خواب‌ها، تعداد سرویس‌های بهداشتی، مکان و سنّ خانه رسم کنید و تحلیل خود را از دادگان و تاثیر هر ویژگی بنویسید.

پرسش شبیه‌سازی ۶. مجموعه دادگان را به دو دسته آموزش و آزمون تقسیم کنید. همچنین با کم کردن میانگین و تقسیم کرنم بر انحراف معیار دادگان را بهنجار [۶] کنید.

پرسش شبیه‌سازی ۷. تابعی برای محاسبه میانگین مربعات خطا (MSE) بین قیمت‌های پیش بینی شده و واقعی تعریف کنید.

---

[۵] گر مدد خواستم از پیرِ مُغان عیب مکن/ شیخ ما گفت که در صومعه همّت نبود
چون طهارت نَبُوَد کعبه و بتخانه یکیست/ نَبُوَد خیر در آن خانه که عصمت نبود
حافظا علم و ادب ورز که در مجلسِ شاه/ هر که را نیست ادب لایقِ صحبت نبود [حافظ]

[6] Normalize

پرسش شبیه‌سازی ۸. تابعی برای اجرای یک مرحله از الگوریتم SGD تعریف کنید. این تابع باید در ورودی پارامترهای فعلی که شامل $\mathbf{x}, b$ هستند و همچنین دسته‌ای از دادگان را بگیرد و در خروجی نیز پس از انجام الگوریتم پارامترهای به‌روزشده را برگرداند.

پرسش شبیه‌سازی ۹. در یک حلقه ترتیب داده‌های آموزش را به طور تصادفی عوض کنید. همچنین دادگان بخش آموزش را به تعدادی دسته[۷] تقسیم کنید و برای هر دسته الگوریتم SGD را اجرا کنید. نهایتا با پارامترهای بدست آمده قیمت خانه‌های بخش آموزش و آزمون را پیش بینی کنید و با توجه به مقدار واقعی قیمت‌ها، مقدار خطا برای هر یک از آن ها به دست آورید. این حلقه باید چند دوره[۸] اجرا شود. مقدار خطا برای هر یک از دوره‌ها برای دادگان آموزش و آزمون را در یک نمودار رسم کنید.

تعداد دوره‌ها، اندازه‌ی هر دسته و طول گام الگوریتم SGD را به طور دلخواه انتخاب کنید.

۷ نصیحتی کُنَمَت بشنو و بهانه مَگیر![۹]

لطفاً به نکات زیر دقّت کنید:

۱. این پروژه بخشی از نمره‌ی شما در این درس را تشکیل خواهد داد.

۲. می‌توانید پروژه را در قالب گروه‌های ۲ نفره انجام دهید. فرمی برای ثبت گروه‌ها در اختیار شما قرار خواهد گرفت. دقّت داشته باشید که در هنگام تحویل پروژه باید تمامی اعضای گروه به تمامی بخش‌ها مسلّط باشند و در نهایت همه‌ی اعضای یک گروه نمره‌ی واحدی را دریافت خواهند کرد.

۳. عنوان بخش‌های مختلف پروژه از آثار شعرا و بزرگان ادبیات فارسی انتخاب شده است. این اشعار بی‌ربط به مفاهیمی که در هر بخش با آن‌ها برخورد می‌کنید نیستند.

۴. تمامی شبیه‌سازی‌ها باید با کمک زبان Python انجام شود. شما تنها مجاز به استفاده از کتاب‌خانه‌های `numpy`، `scipy`، `random`، `plotly` و `matplotlib` هستید. اگر روی عنوان هر کتاب‌خانه کلیک کنید، به راهنمای آن کتاب‌خانه هدایت می‌شوید.

۵. مجموعه‌داده‌ی مورداستفاده در پرسش‌های شبیه‌سازی در CW بارگذاری شده است.

۶. تحویل پروژه به صورت گزارش و کدهای نوشته‌شده است. گزارش باید شامل پاسخ پرسش‌ها، تصاویر و نمودارها و نتیجه‌گیری‌های لازم باشد. توجه کنید که قسمت عمده بارم شبیه سازی را گزارش شما و نتیجه‌ای که از خروجی کد میگیرید دارد. همچنین تمیزی گزارش بسیار مهم است. کدها و گزارش را در یک فایل فشرده‌شده در سامانه‌ی درس‌افزار آپلود کنید.

۷. اگر برای پاسخ به پرسش‌ها، از منبعی (کتاب، مقاله، سایت و...) کمک گرفته‌اید، حتماً به آن ارجاع دهید.

[7] batch
[8] batch

[۹] نصیحتی کُنَمَت بشنو و بهانه مَگیر/ هر آنچه ناصِحِ مُشْفِق بگویَدَت بپذیر [حافظ]

۸. نوشتن گزارش کار با $\LaTeX$ نمره‌ی امتیازی دارد.

۹. پرسش‌های شبیه‌سازی با رنگ سبز و پرسش‌های تئوری با رنگ آبی مشخص شده‌اند.

۱۰. بخش‌های تئوری گزارش که در قالب پرسش‌ها طرح شده‌اند را می‌توانید روی کاغذ بنویسید و تصویر آن‌ها را در گزارش خود بیاورید، ولی توصیه‌ی برادرانه می‌کنم که این کار را نکنید!

۱۱. درصورت مشاهده‌ی تقلّب، نمره‌ی هردو فرد صفر منظور خواهد شد.

موفّق باشید!