

Convergence Analysis of SGD in Convex Optimization Problems

by Hossain Rahnama

Oct 11, 2026

• مساله‌های بهینه‌سازی در یادگیری

یادگیری نظارتی

$$\min_{h \in \mathcal{H}} R(h) = \mathbb{E}(\ell(h, \frac{x, y}{z}))$$

($\subseteq \mathcal{F}_2(X, Y)$)

$$\min_{h \in \mathcal{H}} \Phi_S(h) = R_S(h) + \lambda \Omega(h)$$

$$\mathcal{H} = \{f_{w, \theta} \in \mathcal{F}_2(X, Y) : w \in \mathbb{R}^N\}$$

$$\min_{w \in \mathbb{R}^N} R(w) = \mathbb{E}(\ell(w, (x, y)))$$

($\subseteq \mathbb{R}^N$)

$$\min_{w \in \mathbb{R}^N} \Phi_S(w)$$

($\subseteq \mathbb{R}^N$)

$$\min_{w \in \mathbb{R}^N} \frac{1}{N} \sum_{i=1}^N \ell(w, (x_i, y_i)) + \lambda \Omega(w)$$

Regularizer
Tikhonov $\Omega(w) = \|w\|^2$
Regularized ERM or RLM

Empirical Risk Minimization (ERM)

مساله‌های یادگیری محب

۳.۳. فرض کنید $\mathcal{W}$ محدب است و برای هر $z \in \mathcal{S}$ تابع $\ell(\cdot, z) : \mathbb{R}^k \rightarrow \mathbb{R}$ محدب است. در این صورت، تابع ریسک $R : \mathbb{R}^k \rightarrow \mathbb{R}$ و تابع ریسک تجربی $R_S : \mathbb{R}^k \rightarrow \mathbb{R}$ محدب هستند. بنابراین، مساله‌ی یادگیری میانی ۲.۲ و مساله‌ی یادگیری جایگزین ۲.۲ محدب هستند.

برهان. برای هر $\lambda \in [0, 1]$ و هر $w_1, w_2 \in \mathbb{R}^k$ داریم

$$\ell(\lambda w_1 + (1-\lambda)w_2, z) \leq \lambda \ell(w_1, z) + (1-\lambda) \ell(w_2, z)$$

$$\mathbb{E}(\ell(\lambda w_1 + (1-\lambda)w_2, z)) \leq \lambda \mathbb{E}(\ell(w_1, z)) + (1-\lambda) \mathbb{E}(\ell(w_2, z))$$

$$R(\lambda w_1 + (1-\lambda)w_2) \leq \lambda R(w_1) + (1-\lambda) R(w_2)$$

همچنین جمع توابع محدب، یک تابع محدب است، پس R_S محدب است.

مثال. تابع هدف در (۱۵.۴) محدب قوی با ثابت λ است. نتیجه ۳.۳ را ببینید.

تعریف ۳.۳ (یادگیری محدب). مساله‌ی یادگیری نظارتی با رده‌ی فرضیات $\mathcal{W}$ را یک مساله‌ی یادگیری محدب گوئیم اگر شرایط زیر برآورده شوند.

- رده‌ی فرضیات $\mathcal{W}$ محدب باشد.
- خانواده‌ی توابع $\mathcal{F} = \{\ell(\cdot, z) : \mathbb{R}^k \rightarrow \mathbb{R} \mid z \in \mathcal{S}\}$ را در نظر بگیرید. هر $l \in \mathcal{F}$ تابعی محدب روی $\mathcal{W}$ است.

مسئله‌های یادگیری محذب

مثال. مسئله‌ی رگرسیون با توابع آفین.

$$\mathcal{X} = \mathbb{R}^d, \quad \mathcal{Y} = \mathbb{R}, \quad \mathcal{W} = \mathbb{R}^{d+1}, \quad \ell(w, b; (x, y)) = \frac{1}{2} (w^T x + b - y)^2$$

$$\mathcal{H} = \mathcal{H}_n(\mathbb{R}^d, \mathbb{R}) := \{h \in \mathcal{F}_n(\mathbb{R}^d, \mathbb{R}) : h(x) = w^T x + b \text{ } \wedge \text{ } w \in \mathbb{R}^d \wedge b \in \mathbb{R}\}$$

مثال. مسئله‌ی دسته‌بندی دودوره‌ای با توابع آفین.

$$\mathcal{X} = \mathbb{R}^d, \quad \mathcal{Y} = \{-1, 1\}, \quad \mathcal{W} = \mathbb{R}^{d+1}, \quad \ell(w, b; (x, y)) = \frac{1}{2} (y \neq \text{sgn}(w^T x + b))$$

$$\mathcal{H} = \text{sgn}(\mathcal{H}_n(\mathbb{R}^d, \mathbb{R}))$$

می‌توان نشان داد که

$$\dim(\mathcal{H}) = d+1$$

و این که حل مسئله‌ی جایگزین NP-Hard است.

گرایان کاهش تصادفی برای کمینه‌سازی ریسک تجربی منظم شده

$$\min_{w \in \mathbb{R}^d} \Phi(w) = \frac{1}{N} \sum_{i=1}^N \ell(w, (x_i, y_i)) + \frac{\lambda}{2} \|w\|^2$$

الگوریتم ۵.۴ گرایان کاهش تصادفی برای کمینه‌سازی ریسک تجربی منظم شده.

ورودی: تابع زبان $\ell: \mathbb{R}^d \times \mathcal{S} \rightarrow \mathbb{R}$ ، ثابت $\lambda > 0$ و نمونه‌ی S .

خروجی: تقریبی از کمینه‌کننده‌ی سراسری Φ_S روی $\mathbb{R}^d$ 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

معمولاً با انتخاب اندیس تصادفی از $[m]$

۶: مثال ζ_t را با توزیع یکنواخت $\mathcal{U}$ به صورت تصادفی از $S = (z_1, \dots, z_m)$ انتخاب کن.

۷: قرار بده $d_t = \frac{1}{\lambda} \nabla \ell(\cdot, \zeta_t)(w_t)$ به طوری که $g(w_t, \zeta_t) \in \partial \ell(\cdot, \zeta_t)(w_t)$.

۸: قرار بده $s_t = \frac{1}{\lambda} d_t$

گرمشقی پذیرا باشد

۹: قرار بده $w_{t+1} = w_t - s_t d_t$

$$\partial \ell(\cdot, \zeta_t)(w_t) = \{\nabla \ell(\cdot, \zeta_t)(w_t)\} = \{\nabla_{w_t} \ell(w_t, \zeta_t)\}$$

۱۰: * محاسبه‌ی خروجی

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

۱۲: برگردان $(\bar{w}, \Phi_S(\bar{w}))$

• گام تصورسازی را حذف کردیم، چون مسئله‌ی کمینه‌سازی نامعینه است و در واقع $\nabla \ell(w_t) = d_t$.

فرض ۱.۴. در الگوریتم ۵.۴ متغیرهای تصادفی $\zeta_{[T]}$ مستقل و دارای توزیع یکنواخت $\mathcal{U}$ روی نمونه‌ی S هستند.

گزاره ۱۱.۴. مسئله‌ی یادگیری محذب با رده‌ی فرضیات $\mathcal{W} = \mathbb{R}^d$ را در نظر بگیرید که مساله‌ی جایگزین متناظر آن (۱۰.۴) است. با فرض ۱.۴، در هر تکرار $t \in [T]$ از الگوریتم ۵.۴ شرط $\mathbb{E}(d_t | w_t) \in \partial \Phi_S(w_t)$ برقرار است.

گزارش کار تحقیق اقتصادی برای کمینه سازی ریسک تجربی منظم شده

برهان. دو گام زیر را به ترتیب جلو می بریم.

• برای هر $t \in [T]$ ، توابع مانند ψ_t و ϕ_t وجود دارند که

$$\phi_t = \psi_t(\xi_1, \dots, \xi_t) = \psi_t(\xi_{[t]}) \quad \psi_{t+1} = \psi_t(\xi_1, \dots, \xi_t, \xi_{t+1}) = \psi_t(\xi_{[t+1]})$$

• کمیت $\psi_t(\phi_t, d_t)$ را می نویسیم.

$$w_{t+1} = w_t - \gamma_t d_t = w_t - \frac{1}{2\lambda t} (2\lambda w_t + g_t) = \left(\frac{t-1}{t}\right) w_t - \frac{1}{2\lambda t} g_t$$

$$\frac{t w_{t+1}}{\cancel{t_{t+1}}} = \frac{(t-1) w_t}{\cancel{t_t}} - \frac{1}{2\lambda} g_t \implies \psi_{t+1} = \psi_t - \frac{1}{2\lambda} g_t$$

$$= \psi_{t-1} - \frac{1}{2\lambda} (g_{t-1} + g_t) = \dots = \psi_1 - \frac{1}{2\lambda} \sum_{i=1}^t g_i$$

$$(*) \quad w_{t+1} = -\frac{1}{2\lambda t} \sum_{i=1}^t g_i(w_i, \xi_i) := \phi_t(w_{[t]}, \xi_{[t]})$$

$$t=1, \quad w_2 = -\frac{1}{2\lambda} g_1(w_1, \xi_1) = \phi_1(w_{[1]}, \xi_{[1]}) \longrightarrow \psi_1 = \phi_1$$

فرض کنید برای $t < k$ حجم مورد نظر تکرار باشد. برای $t=k$ داریم

$$w_{k+1} = \phi_k(w_{[k]}, \xi_{[k]}) = \phi_k(w_1, \psi_1(\xi_{[1]}), \dots, \psi_k(\xi_{[k-1]}), \xi_{[k]}) := \psi_k(\xi_{[k]})$$

نتیجه می دهیم که ψ_t از ϕ_t مشتق است.

$$\begin{aligned} E(d_t | w_t) &= 2\lambda E(w_t | w_t) + E(g(w_t, \xi_t) | w_t) = 2\lambda w_t + \frac{1}{N} E(g(w_t, \xi_t)) \\ &= 2\lambda w_t + \sum_{i=1}^N g(w_t, z_i) \cdot \frac{1}{N} = 2\lambda w_t + \frac{1}{N} \sum_{i=1}^N g(w_t, z_i) \end{aligned}$$

حال نشان می دهیم که $\partial E(d_t | w_t) \subseteq \partial \Phi_t(w_t)$

گزاره ۳۱.۱. فرض کنید $A \subseteq \mathbb{R}^n$ و $f: A \rightarrow \mathbb{R}$ و $g: A \rightarrow \mathbb{R}$ توابعی دلخواه باشند و $x \in A$.

• برای هر عدد حقیقی $\lambda > 0$ داریم $\partial(\lambda f)(x) = \lambda \partial f(x)$

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

$$2\lambda w_t = \nabla H(w_t) \in \partial H(w_t)$$

$$g(w_t, z_i) \in \partial \ell(\cdot, z_i)(w_t) \longrightarrow \sum_{i=1}^N g(w_t, z_i) \in \sum_{i=1}^N \partial \ell(\cdot, z_i)(w_t) \subseteq \partial \left(\sum_{i=1}^N \ell(\cdot, z_i)(w_t) \right)$$

گزاره ۳۱.۱

$$\longrightarrow \frac{1}{N} \sum_{i=1}^N g(w_t, z_i) \in \partial \left(\frac{1}{N} \sum_{i=1}^N \ell(w_t, z_i) \right)$$

بنابراین نتیجه می شود که

$$E(d_t | w_t) = 2\lambda w_t + \frac{1}{N} \sum_{i=1}^N g(w_t, z_i) \in \partial H(w_t) + \partial \left(\frac{1}{N} \sum_{i=1}^N \ell(w_t, z_i) \right)$$

$$\subseteq \partial \left(H(w_t) + \frac{1}{N} \sum_{i=1}^N \ell(w_t, z_i) \right) = \partial \Phi_t(w_t)$$

گزاره ۳۱.۱

نگاهی به کنترل خطا، تقسیم پذیری و بهینه سازی

• مساله های زیر مورد توجه ما هستند.

$$\min_{w \in \mathcal{W}} R(w) \quad \min_{w \in \mathcal{W}} \tilde{\Phi}(w) = R(w) + \frac{\gamma}{2} \Omega(w) \quad \gamma \geq 0$$

$w^* \in \arg \min(R, \mathcal{W})$ $w^* \in \arg \min(\tilde{\Phi}, \mathcal{W})$ جواب الگوریتم بهینه سازی: w^*

به γ وابسته است

• خطای نه در طی این فرایند داریم

$$0 \leq R(w^*) - R(w^*) = (R(w^*) - \tilde{\Phi}(w^*)) + (\tilde{\Phi}(w^*) - \tilde{\Phi}(w^*)) + (\tilde{\Phi}(w^*) - R(w^*))$$

$$\leq (R(w^*) - R(w^*)) + (\tilde{\Phi}(w^*) - \tilde{\Phi}(w^*)) + (\tilde{\Phi}(w^*) - R(w^*))$$

$$\leq \sup_{w \in \mathcal{W}} |R(w) - \tilde{\Phi}(w)| + (\tilde{\Phi}(w^*) - \tilde{\Phi}(w^*)) + (\tilde{\Phi}(w^*) - R(w^*))$$

خطای تقریب تابع هدف خطای الگوریتم بهینه سازی خطای تقسیم پذیری

نامساوی های با احتمال بالا (تقریب)

$$\mathbb{E}(\tilde{\Phi}(w^*) - R(w^*)) \leq \frac{\gamma}{2} \Omega(w^*) = \frac{\gamma}{2} \|w^*\|^2$$

• خطای مساله های محب یا گزینی خطای تقسیم پذیری قابل کنترل ندارند!

گزاره ۴.۳. مساله های یادگیری ۱.۲ متناظر با رگرسیون را در نظر بگیرید که تابع زبان آن مربعی باشد و $\mathcal{X} = \mathcal{Y} = \mathbb{R}$. رده های فرضیات $\mathcal{H}$ را توابع خطی $\text{Lin}(\mathbb{R}, \mathbb{R})$ فرض کنید که با $\mathcal{W} = \mathbb{R}$ در تناظر یک به یک است. این مساله یادگیری پذیر PAC با فرض $\mathcal{W}$ نیست.

توابع هموار

تعریف ۱.۳ (تابع هموار). مجموعه ای باز $A \subseteq \mathbb{R}^n$ و تابع $f: A \rightarrow \mathbb{R}$ را در نظر بگیرید. تابع f روی A هموار است اگر روی آن مشتق پذیر باشد و یک عدد حقیقی $\beta > 0$ وجود داشته باشد به طوری که برای هر $x, y \in A$ داشته باشیم $\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$. به عدد β ثابت همواری گوئیم. گرادیان آن لایپشیتز باشد

گزاره ۱.۳. مجموعه ای محدب باز $C \subseteq \mathbb{R}^n$ و تابع محدب $f: C \rightarrow \mathbb{R}$ را در نظر بگیرید که روی C مشتق پذیر است. در این صورت، موارد زیر معادل اند.

- f تابعی β -هموار روی C است.
- ثابت $\beta > 0$ وجود دارد که برای هر دو نقطه ای دلخواه $x, y \in C$ داریم:

$$f(x) \leq f(y) + \nabla f(y)^T (x - y) + \frac{\beta}{2} \|x - y\|^2. \quad (1.3)$$

برهان. تعريف كنيد $u := x - y$ و $\phi(t) := f(y + tu)$. چرا ϕ فوش تعريف است؟

$$f(x) - f(y) = \phi(1) - \phi(0) = \int_0^1 \phi'(t) dt = \int_0^1 \nabla f(y + tu)^T u dt = \nabla f(y)^T u + \int_0^1 (\nabla f(y + tu) - \nabla f(y))^T u dt$$

$$\leq \nabla f(y)^T u + \int_0^1 \beta \|tu\| \|u\| dt = \nabla f(y)^T u + \frac{\beta}{2} \|u\|^2$$

لم ۲.۳ (خود کران داری). تابع محدب $f: \mathbb{R}^n \rightarrow \mathbb{R}$ را در نظر بگیرید. فرض کنید تابع f روی $\mathbb{R}^n$ مثبت و β -هموار است. در این صورت، برای هر $y \in \mathbb{R}^n$ داریم $\|\nabla f(y)\|^2 \leq 2\beta f(y)$.

برهان. بگیرد $x = y - \frac{1}{\beta} \nabla f(y)$ و از گزاره ۱.۳ استفاده کنید.

$$f(x) \leq f(y) - \frac{1}{\beta} \nabla f(y)^T \nabla f(y) + \frac{\beta}{2} \left\| \frac{1}{\beta} \nabla f(y) \right\|^2 \Rightarrow \frac{1}{2\beta} \|\nabla f(y)\|^2 \leq f(y) - f(x) \leq f(y)$$

≥ 0

یادگیری محدب لیپشیتز و هموار

فرض ۱.۳. اگر $\mathbf{w}^* \in \arg \min(R, \mathcal{W})$ ، آنگاه ثابت $c > 0$ را به گونه‌ای انتخاب کنید که $\|\mathbf{w}^*\| \leq c$.

تعریف ۴.۳ (یادگیری محدب لیپشیتز). مساله‌ی یادگیری نظارتی با رده‌ی فرضیات $\mathcal{W} \subseteq \mathbb{R}^k$ را در نظر بگیرید. این مساله را محدب لیپشیتز با پارامترهای (ρ, c) می‌گوییم اگر یک مساله‌ی یادگیری محدب با شرایط زیر باشد.

- خانواده‌ی توابع $\mathcal{F} = \{\ell(\cdot, z) : \mathbb{R}^k \rightarrow \mathbb{R} \mid z \in \mathcal{S}\}$ را در نظر بگیرید. ثابت $\rho > 0$ وجود دارد به طوری که هر $l \in \mathcal{F}$ تابعی ρ -لیپشیتز روی $\mathbb{R}^k$ است.
- فرض ۱.۳ با ثابت $c > 0$ برقرار است.

فرض ۲.۳. اگر $\mathbf{w}^* \in \arg \min(R, \mathcal{W})$ ، آنگاه ثابت $a > 0$ را به گونه‌ای انتخاب کنید که $|R(\mathbf{w}^*)| \leq a$.

تعریف ۵.۳ (یادگیری محدب هموار). مساله‌ی یادگیری نظارتی با رده‌ی فرضیات $\mathcal{W} \subseteq \mathbb{R}^k$ را در نظر بگیرید. این مساله را محدب هموار با پارامترهای (β, c, a) می‌گوییم اگر یک مساله‌ی یادگیری محدب با شرایط زیر باشد.

- خانواده‌ی توابع $\mathcal{F} = \{\ell(\cdot, z) : \mathbb{R}^k \rightarrow \mathbb{R} \mid z \in \mathcal{S}\}$ را در نظر بگیرید. ثابت $\beta > 0$ وجود دارد به طوری که هر $l \in \mathcal{F}$ تابعی β -هموار روی $\mathbb{R}^k$ است. همچنین، هر $l \in \mathcal{F}$ تابعی مثبت روی $\mathbb{R}^k$ است.
- فرض ۱.۳ با ثابت $c > 0$ و فرض ۲.۳ با ثابت $a > 0$ برقرارند.

$$k=d, \ell(\mathbf{w}, (x, y)) =$$

نام مساله	نوع مساله	$\mathcal{W}$	$\mathcal{X}$	r	$\mathcal{Y}$	$\varphi(\mathbf{w}^\top \mathbf{x}, y)$
ماشین بردار پشتیبان ^۱	محدب لیپشیتز	$\mathbb{R}^d$	$B_r(0)$	ρ	$\{-1, 1\}$	$\max\{0, 1 - y(\mathbf{w}^\top \mathbf{x})\}$
رگرسیون لجستیک ^۲	محدب هموار	$\mathbb{R}^d$	$B_r(0)$	$2\sqrt{\beta}$	$\{-1, 1\}$	$\ln(1 + \exp(-y(\mathbf{w}^\top \mathbf{x})))$
رگرسیون قدرمطلق	محدب لیپشیتز	$\mathbb{R}^d$	$B_r(0)$	ρ	$[-1, 1]$	$ \mathbf{w}^\top \mathbf{x} - y $
رگرسیون ریب	محدب هموار	$\mathbb{R}^d$	$B_r(0)$	$\sqrt{\beta/2}$	$[-1, 1]$	$(\mathbf{w}^\top \mathbf{x} - y)^2$

همگرایی SGD برای مساله‌ی یادگیری محدب لیپشیتز

نتیجه ۱۲.۴. مساله‌ی یادگیری محدب لیپشیتز با پارامترهای (ρ, c) و رده‌ی فرضیات $\mathcal{W} = \mathbb{R}^d$ را در نظر بگیرید که مساله‌ی جایگزین متناظر آن (۱۰.۴) است. با فرض ۱.۴، در هر تکرار $t \in [T]$ از حلقه‌ی همگرایی در الگوریتم ۵.۴ داریم $\mathbb{E}(\mathbf{d}_t | \mathbf{w}_t) \in \partial \Phi_S(\mathbf{w}_t)$. همچنین، نرم جهت تصادفی کران‌دار است، یعنی $\|\mathbf{d}_t\| \leq 2\rho$.

$$\|\mathbf{d}_t\| = \|2\lambda \mathbf{w}_t + \mathbf{g}_t\| \leq 2\lambda \|\mathbf{w}_t\| + \|\mathbf{g}_t\| \stackrel{(*)}{=} 2\lambda \left\| \frac{1}{2\lambda} \sum_{i=1}^t \mathbf{g}_i \right\| + \|\mathbf{g}_t\| \leq \frac{1}{\lambda} \sum_{i=1}^t \|\mathbf{g}_i\| + \|\mathbf{g}_t\| \quad (**)$$

$$\rho + \rho = 2\rho$$

نرم جهت تصادفی کران‌دار

گزاره ۱۳.۴. مساله‌ی یادگیری محدب لیپشیتز با پارامترهای (ρ, c) و رده‌ی فرضیات $\mathcal{W} = \mathbb{R}^d$ را در نظر بگیرید که مساله‌ی جایگزین متناظر آن (۱۰.۴) است. با فرض ۱.۴، پس از T تکرار حلقه‌ی همگرایی در الگوریتم ۵.۴ داریم:

$$\mathbb{E}(\Phi_S(\bar{\mathbf{w}}) - \Phi_S(\mathbf{w}^*)) \leq \frac{\rho^2}{\lambda} \left(\frac{\ln T}{T} + \frac{1}{T} \right).$$

در نوسان صفدی قبل

برهان. با توجه به گزاره ۱۰.۴ بگیریم

$$r = 2\rho, \quad \mathcal{X} = B_r, \quad \mathcal{W} = \mathbb{R}^d$$

نتیجه ۱۴.۴. مساله ی یادگیری محدب هموار با تابع زیان کران دار و پارامترهای (β, c, a) و رده ی فرضیات $\mathcal{W} = \mathbb{R}^d$ را در نظر بگیرید که مساله ی جایگزین متناظر آن (۱۰.۴) است. با فرض ۱.۴، در هر تکرار $t \in [T]$ از حلقه ی همگرایی در الگوریتم ۵.۴ داریم $\mathbb{E}(\mathbf{d}_t | \mathbf{w}_t) \in \partial \Phi_S(\mathbf{w}_t)$. همچنین، نرم جهت تصادفی کران دار است، یعنی $\|\mathbf{d}_t\| \leq 2\sqrt{2\beta a}$.

سپهرخان

$$\|\mathbf{d}_t\| \leq \frac{1}{T} \sum_{i=1}^T \|\mathbf{g}_i\| + \|\mathbf{g}_t\| \leq \frac{1}{T} \sum_{i=1}^T \sqrt{2\beta c(\mathbf{w}_i, \mathbf{z}_i)} + \sqrt{2\beta c(\mathbf{w}_t, \mathbf{z}_t)}$$

$$\leq \sqrt{2\beta a} + \sqrt{2\beta a} = 2\sqrt{2\beta a}$$

معادله (**)

کران داری تابع زیان

گزاره ۱۵.۴. مساله ی یادگیری محدب هموار با تابع زیان کران دار و پارامترهای (β, c, a) و رده ی فرضیات $\mathcal{W} = \mathbb{R}^d$ را در نظر بگیرید که مساله ی جایگزین متناظر آن (۱۰.۴) است. با فرض ۱.۴، پس از T تکرار حلقه ی همگرایی در الگوریتم ۵.۴ داریم:

$$\mathbb{E}(\Phi_S(\bar{\mathbf{w}}) - \Phi_S(\mathbf{w}^*)) \leq \frac{2\beta a}{\lambda} \left(\frac{\ln T}{T} + \frac{1}{T} \right).$$

سپهرخان. با توجه به گزاره ۱۵.۴ بگیریم

$$r = 2\sqrt{2\beta a}, \quad \delta = 2\lambda, \quad \mathcal{W} = \mathbb{R}^d$$

مراجع

- [1] S. Shalev-Shwartz, S. Ben-David. *Understanding Machine Learning: from Theory to Algorithms*. Cambridge University Press, 2014.
- [2] M. Mohri, A. Rostamzadeh, A. Talwalkar. *Foundations of Machine Learning*, MIT Press, 2018.
- [3] F. Bach. *Learning Theory from First Principles*, MIT Press, 2025.
- [4] A. Beck. *First-Order Methods in Optimization*, SIAM, 2017.
- [5] D. Bertsekas, *Convex Optimization Theory*, Athena Scientific, 2009.
- [6] A. Rakhlin, O. Shamir, K. Sridharan. *Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization*, ICML, 2011.
- [7] O. Shamir, T. Zhang. *Stochastic Gradient Descent for Non-Smooth Optimization*. ICML, 2013.
- [8] E. Hazan, A. Agarwal, S. Kale. *Logarithmic Regret Algorithms for Online Convex Optimization*. *Machine Learning*. vol.69, no.2, pp.169-192, 2007.
- [9] H. Rahnama, *Stochastic Algorithms for Optimization in Machine Learning*. Master's Thesis, Sharif University of Technology, 2025.