DEPARTAMENTO DE
MATEMÁTICA

Train faster, generalize better: Stability of stochastic gradient descent

Luiz Fernando Bossa

25 de junho de 2025

MTM

Sumário

Estabilidade de algoritmos iterativos randomizados

Estabilidade do Método do Gradiente Estocástico

Operações que induzem estabilidade

Minimização de Risco Convexo

- Temos uma distribuição de probabilidade $\mathcal{D}$ sobre um espaço de dados Z.
- Temos uma amostra $S = (z_1, \ldots, z_n)$ de tamanho n extraída i.i.d. de $\mathcal{D}$.
- Ω o espaço de parâmetros do modelo.
- f é a função de perda, $f : \Omega \times Z \to \mathbb{R}$

- O risco populacional é definido como

$$R[w] := \mathbb{E}_{z\sim\mathcal{D}}[f(w;z)]$$

- O risco empírico é definido como a perda média sobre a amostra S:

$$R_S[w] := \frac{1}{n}\sum_{i=1}^{n} f(w;z_i)$$

- O erro de generalização é definido como

$$R_S[w] - R[w]$$

- Quando os parâmetros w são dados por um algoritmo A aplicado à amostra S, faz sentido definir

$$\epsilon_{gen} := |R_S[A(S)] - R[A(S)]|$$

Definition (2.1)

Um algoritmo randomizado A é ϵ-uniformemente estável se para todos os conjuntos de dados S, $S' \in Z^n$ tal que S e S' diferem em no máximo uma amostra, temos

$$\sup_z \mathbb{E}_A[f(A(S); z) - f(A(S'); z)] \leq \epsilon \tag{2.3}$$

- A esperança é tomada apenas sobre a aleatoriedade interna de A.
- Denotamos por $\epsilon_{stab}(A, n)$ o ínfimo sobre todos os ϵ para os quais (2.3) é válido.
- Omitiremos (A, n) quando o contexto for claro.

Theorem (2.2)

Seja A ϵ-uniformemente estável. Então,

$$|\mathbb{E}_{S,A}[R_S[A(S)] - R[A(S)]]| \leq \epsilon$$

$S = (z_1, \ldots, z_n)$ e $S' = (z'_1, \ldots, z'_n)$ duas amostras aleatórias independentes. Seja $S^{(i)} = (z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_n)$

$$\begin{aligned}
\mathbb{E}_S\mathbb{E}_A[R_S[A(S)]] &= \mathbb{E}_S\mathbb{E}_A[\frac{1}{n}\sum_{i=1}^{n} f(A(S); z_i)] \\
&= \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_A[\frac{1}{n}\sum_{i=1}^{n} f(A(S^{(i)}); z'_i)] \\
&= \mathbb{E}_S\mathbb{E}_{S'}\mathbb{E}_A[\frac{1}{n}\sum_{i=1}^{n} f(A(S); z'_i)] + \delta \\
&= \mathbb{E}_S\mathbb{E}_A[R[A(S)]] + \delta
\end{aligned}$$

onde podemos expressar δ como

$$\delta = \mathbb{E}_S \mathbb{E}_{S'} \mathbb{E}_A [\frac{1}{n} \sum_{i=1}^{n} f(A(S^{(i)}); z_i') - \frac{1}{n} \sum_{i=1}^{n} f(A(S); z_i')]$$

Além disso, tomando o supremo sobre quaisquer dois conjuntos de dados S, S' diferindo em apenas uma amostra, podemos limitar a diferença como

$$|\delta| \leq \sup_{S,S',z} \mathbb{E}_A [f(A(S); z) - f(A(S'); z)] \leq \epsilon,$$

pela nossa suposição sobre a estabilidade uniforme de A. A afirmação segue.

Definition (2.3)

Uma regra de atualização é η-expansiva se

$$\sup_{v,w\in\Omega} \frac{\|G(v) - G(w)\|}{\|v - w\|} \leq \eta.$$

Definition (2.4)

Uma regra de atualização é σ-limitada se

$$\sup_{w\in\Omega} \|w - G(w)\| \leq \sigma.$$

Lemma (2.5)

Fixe uma sequência arbitrária de atualizações $G_1, \ldots, G_T$ *e outra sequência* $G'_1, \ldots, G'_T$. *Seja* $w_0 = w'_0$ *um ponto de partida em* Ω *e defina* $\delta_t = \|w'_t - w_t\|$ *onde* w_t *e* w'_t *são definidos recursivamente através de*

$$w_{t+1} = G_{t+1}(w_t), \qquad w'_{t+1} = G'_{t+1}(w'_t), \quad t \geq 0$$

Então, temos a relação de recorrência:

$$\delta_0 = 0$$

$$\delta_{t+1} \leq \begin{cases} \eta\delta_t & G_t = G'_t \text{ é } \eta\text{-expansiva} \\ \min(\eta, 1)\delta_t + 2\sigma & G_t \text{ e } G'_t \text{ são } \sigma\text{-limitadas, } G_t \text{ é } \eta\text{-expan.} \end{cases}$$

O primeiro limite em δ_t segue diretamente da suposição de que $G_t = G'_t$ e da definição de η-expansividade. Para o segundo limite, vamos usar a desigualdade triangular e truque de soma zero:

$$\delta_{t+1} = \|G(w_t) - G'(w'_t)\| \leq \|G(w_t) - w_t + w'_t - G'(w'_t)\| + \|w_t - w'_t\|$$

$$\leq \delta_t + \|G(w_t) - w_t\| + \|G'(w'_t) - w'_t\| \leq \delta_t + 2\sigma$$

Alternativamente, podemos limitar δ_{t+1} como

$$\begin{aligned}
\delta_{t+1} &= \|G_t(w_t) - G'_t(w'_t)\| \\
&= \|G_t(w_t) - G_t(w'_t) + G_t(w'_t) - G'_t(w'_t)\| \\
&\leq \|G_t(w_t) - G_t(w'_t)\| + \|G_t(w'_t) - G'_t(w'_t)\| \\
&\leq \eta\delta_t + 2\sigma.
\end{aligned}$$

Estabilidade de algoritmos iterativos randomizados

Estabilidade do Método do Gradiente Estocástico

Operações que induzem estabilidade

Minimização de Risco Convexo

Definition (3.1)

Para um tamanho de passo não negativo $\alpha > 0$ e uma função $f : \Omega \to \mathbb{R}$, definimos a regra de atualização do gradiente $G_{f,\alpha}$ como

$$G_{f,\alpha}(w) = w - \alpha \nabla f(w).$$

Dada uma amostra $S = (z_1, \ldots, z_n)$, podemos fazer as atualizações da seguinte maneira:

- Escolher i de maneira uniforme em $[n]$ e calcular o gradiente em z_i
- Escolher uma permutação aleatória de de $[n]$, fixar essa permutação e fazer o gradiente de maneira sucessiva nessa ordem.

Definition (3.2)

Dizemos que f é L-Lipschitz se para todos os pontos u no domínio de f temos $\|\nabla f(x)\| \leq L$. Isso implica que

$$|f(u) - f(v)| \leq L\|u - v\|.$$

Lemma (3.3)

Assuma que f é L-Lipschitz. Então, a atualização de gradiente $G_{f,\alpha}$ é (αL)-limitada.

Definition (3.4)

Uma função $f : \Omega \to \mathbb{R}$ é convexa se para todo $u, v \in \Omega$ temos

$$f(u) \geq f(v) + \langle \nabla f(v), u - v \rangle.$$

Definition (3.5)

Uma função $f : \Omega \to \mathbb{R}$ é γ-fortemente convexa se para todo $u, v \in \Omega$ temos

$$f(u) \geq f(v) + \langle \nabla f(v), u - v \rangle + \frac{\gamma}{2}\|u - v\|^2.$$

Definition (3.6)

Uma função $f : \Omega \to \mathbb{R}$ é β-suave se para todo $u, v \in \Omega$ temos

$$\|\nabla f(u) - \nabla f(v)\| \leq \beta \|u - v\|.$$

Lemma (3.7)

Assuma que f é β-suave. Então, as seguintes propriedades são válidas:

1. *$G_{f,\alpha}$ é $(1+\alpha\beta)$-expansiva.*
2. *Assuma adicionalmente que f é convexa. Então, para qualquer $\alpha \leq 2/\beta$, a atualização de gradiente $G_{f,\alpha}$ é 1-expansiva.*
3. *Assuma adicionalmente que f é γ-fortemente convexa. Então, para $\alpha \leq \frac{2}{\beta+\gamma}$, $G_{f,\alpha}$ é $\left(1-\frac{\alpha\beta\gamma}{\beta+\gamma}\right)$-expansiva.*

Theorem (3.8)

Assuma que a função de perda $f(\cdot; z)$ é β-suave, convexa e L-Lipschitz para todo z. Suponha que executamos SGM com tamanhos de passo $\alpha_t \leq 2/\beta$ por T passos. Então, SGM satisfaz estabilidade uniforme com

$$\epsilon_{stab} \leq \frac{2L^2}{n} \sum_{t=1}^{T} \alpha_t.$$

Sejam S e S' duas amostras de tamanho n diferindo em uma única amostra, $\{G_i\}_{i=1}^T$ e $\{G'_i\}_{i=1}^T$ as atualizações de gradiente estocástico correspondentes, w_T e w'_T os parâmetros finais correspondentes.

- Aplica Lipschitz em $f(\cdot; z)$:

$$\mathbb{E}|f(w_T; z) - f(w'_T; z)| \leq L\mathbb{E}[\|w_T - w'_T\|] = L\mathbb{E}[\delta_T] \quad (3.3)$$

- Com probabilidade $1 - \frac{1}{n}$, temos que G_t e G'_t são idênticas, e nossas hipóteses permitem aplicar o Lema 3.7(2) para concluir que $G_{f,\alpha}$ é 1-expansiva.
- Com probabilidade $\frac{1}{n}$, a amostra escolhida é diferente, e usamos que G_t e G'_t são $\alpha_t L$-limitadas (Lema 3.3);

$$\begin{aligned}
\mathbb{E}[\delta_{t+1}] &\leq \left(1 - \frac{1}{n}\right) \mathbb{E}[\eta\delta_t] + \frac{1}{n}\mathbb{E}[\eta\delta_t + 2\alpha_t L] \qquad (3.4)\\
&= \left(1 - \frac{1}{n}\right) \mathbb{E}[\delta_t] + \frac{1}{n}\mathbb{E}[\delta_t] + \frac{2\alpha_t L}{n}\\
&= \mathbb{E}[\delta_t] + \frac{2\alpha_t L}{n}
\end{aligned}$$

- Desenrolando essa recursão, e lembrando que $\delta_0 = 0$, obtemos

$$\mathbb{E}[\delta_T] \leq \sum_{t=1}^{T} \frac{2\alpha_t L}{n} = \frac{2L}{n} \sum_{t=1}^{T} \alpha_t$$

- Voltando para (3.3), temos

$$\mathbb{E}|f(w_T; z) - f(w'_T; z)| \leq L\mathbb{E}[\delta_T] \leq \frac{2L^2}{n} \sum_{t=1}^{T} \alpha_t$$

- Como ϵ_{stab} é o mínimo dentre os valores possíveis, segue o resultado.

□

- Assuma que a função de perda $f(w; z)$ é fortemente convexa com respeito a w, para todo z.
- Suponha Ω compacto e convexo, e que podemos calcular a projeção sobre Ω, cuja qual denotaremos por

$$\Pi_\Omega(v) = \arg\min_{w \in \Omega} \|w - v\|^2$$

- Gradiente estocástico com projeção é definido como

$$w_{t+1} = \Pi_\Omega(w_t - \alpha_t \nabla f(w_t; z_t)) \tag{3.5}$$

- Risco empírico com termo de regularização:

$$R_{S,\mu}[w] := \frac{1}{n}\sum_{i=1}^{n} f(w; z_i) + \frac{\mu}{2}\|w\|^2 \tag{3.6}$$

- O minimizador de (3.6) ocorre dentro da bola de raio $r = \sqrt{2/\mu}$, então podemos tomar Ω como sendo essa bola e a projeção vira uma operação de escala.

- Vamos trocar a função de perda $f(w; z)$ por sua versão regularizada $f(w; z) + \frac{\mu}{2}\|w\|^2$.
- Também vamos tomar a constante de Lipschitz como

$$L = \sup_{w \in \Omega} \sup_{z} \|\nabla f(w; z)\|_2 \tag{3.7}$$

Theorem (3.9)

Assuma que a função de perda $f(\cdot\,; z)$ é γ-fortemente convexa e β-suave para todo z. Então, SGM satisfaz estabilidade uniforme com

$$\epsilon_{stab} \leq \frac{2L^2}{\gamma n}.$$

A demonstração é análoga à do Teorema 3.8, usaremos a mesma notação.

- Usando Lipschitz em $f(\cdot; z)$:

$$\mathbb{E}|f(w_T; z) - f(w'_T; z)| \leq L\mathbb{E}[\delta_T] \tag{3.8}$$

- Note que, como a projeção não aumenta distâncias,

$$\begin{aligned}
\delta_t &= \|w_t - w'_t\| \\
&= \|\Pi_\Omega(w_t - \alpha_t \nabla f(w_t; z_t)) - \Pi_\Omega(w'_t - \alpha_t \nabla f(w'_t; z'_t))\| \\
&\leq \|w_t - \alpha_t \nabla f(w_t; z_t) - w'_t - \alpha_t \nabla f(w'_t; z'_t)\|
\end{aligned}$$

Demonstração do Teorema 3.9

- Com probabilidade $1 - \frac{1}{n}$, temos que G_t e G'_t são idênticas, e nossas hipóteses permitem aplicar o Lema 3.7(3) para concluir que $G_{f,\alpha}$ é $(1 - \alpha\gamma)$-expansiva.
- Com probabilidade $\frac{1}{n}$, a amostra escolhida é diferente, e usamos que G_t e G'_t são (αL)-limitadas (Lema 3.3);

$$\begin{aligned}
\mathbb{E}[\delta_{t+1}] &\leq \left(1 - \frac{1}{n}\right)\mathbb{E}[(1-\alpha\gamma)\delta_t] + \frac{1}{n}\mathbb{E}[(1-\alpha\gamma)\delta_t + 2\alpha L] \quad (3.9)\\
&= (1 - \frac{1}{n})(1-\alpha\gamma)\mathbb{E}[\delta_t] + \frac{1}{n}(1-\alpha\gamma)\mathbb{E}[\delta_t] + \frac{2\alpha L}{n}\\
&= (1-\alpha\gamma)\mathbb{E}[\delta_t] + \frac{2\alpha L}{n}
\end{aligned}$$

Desenvolvendo essa recursão, obtemos

$$\begin{aligned}
\mathbb{E}[\delta_T] &\leq (1-\alpha\gamma)\mathbb{E}[\delta_{T-1}] + \frac{2\alpha L}{n} \\
&\leq (1-\alpha\gamma)\left((1-\alpha\gamma)\mathbb{E}[\delta_{T-2}] + \frac{2\alpha L}{n}\right) + \frac{2\alpha L}{n} \\
&= (1-\alpha\gamma)^2\mathbb{E}[\delta_{T-2}] + \frac{2\alpha L}{n}\left((1-\alpha\gamma)+1\right) \\
&\vdots \\
&\leq (1-\alpha\gamma)^T\mathbb{E}[\delta_0] + \frac{2\alpha L}{n}\sum_{t=0}^{T}(1-\alpha\gamma)^t
\end{aligned}$$

Como $\mathbb{E}[\delta_0] = 0$, e o somatório é uma soma parcial da série geométrica, temos

$$\begin{aligned}\mathbb{E}[\delta_T] &\leq \frac{2\alpha L}{n} \sum_{t=0}^{T} (1 - \alpha\gamma)^t \\ &< \frac{2\alpha L}{n} \cdot \frac{1}{\alpha\gamma} = \frac{2L}{\gamma n}\end{aligned}$$

Voltando para (3.3),

$$\epsilon_{stab} \leq L\mathbb{E}[\delta_T] \leq \frac{2L^2}{\gamma n}$$

Theorem (3.10)

Assuma que a função de perda $f(\cdot; z) \in [0, 1]$ é γ-fortemente convexa, tem gradientes limitados por L, e é β-suave para todo z. Suponha que executamos SGM com tamanhos de passo $\alpha_t = \frac{1}{\gamma t}$. Então, SGM tem estabilidade uniforme de

$$\epsilon_{stab} \leq \frac{2L^2 + \beta\rho}{\gamma n}$$

onde $\rho = \sup_{w \in \Omega} \sup_z f(w; z)$.

Lemma (3.11)

Assuma que a função de perda $f(\cdot; z)$ é não negativa e L-Lipschitz para todo z. Então, para todo $z \in Z$ e todo $t_0 \in [n]$, temos

$$\mathbb{E}|f(w_T; z) - f(w'_T; z)| \leq \frac{t_0}{n} \sup_{w,z} f(w; z) + L\mathbb{E}[\delta_T | \delta_{t_0} = 0].$$

Seja $\mathcal{E} = \mathbf{1}[\delta_{t_0} = 0]$ o evento que $\delta_{t_0} = 0$ – ie, até o passo t_0, as atualizações são idênticas. Temos:

$$\begin{aligned}
&\mathbb{E}|f(w_T; z) - f(w'_T; z)| = \\
&\mathbb{P}\{\mathcal{E}\}\mathbb{E}[|f(w_T; z) - f(w'_T; z)| \mid \mathcal{E}] + \mathbb{P}\{\mathcal{E}^c\}\mathbb{E}[|f(w_T; z) - f(w'_T; z)| \mid \mathcal{E}^c] \\
&\qquad \leq \mathbb{E}[|f(w_T; z) - f(w'_T; z)| \mid \mathcal{E}] + \mathbb{P}\{\mathcal{E}^c\} \cdot \sup_{w,z} f(w; z) \\
&\qquad\qquad \leq L\mathbb{E}[\|w_T - w'_T\| \mid \mathcal{E}] + \mathbb{P}\{\mathcal{E}^c\} \cdot \sup_{w,z} f(w; z)
\end{aligned}$$

Resta limitar $\mathbb{P}\{\mathcal{E}^c\}$.

- Seja $i^* \in [n]$ a posição onde as amostras diferem e I a variável aleatória que assume o índice da primeira iteração na qual a amostra z_{i^*} é escolhida.
- O evento $\mathcal{E}^c = \{\delta_{t_0} \neq 0\}$ está contido no evento $\{I \leq t_0\}$, pois se tomamos a amostra diferente, então nesse passo ela gera gradientes diferentes e as sequências de atualizações divergem.

Assim, temos

$$\mathbb{P}\{\mathcal{E}^c\} = \mathbb{P}\{\delta_{t_0} \neq 0\} \leq \mathcal{P}\{I \leq t_0\}$$

- Na regra da permutação aleatória, I é uniforme sobre $[n]$, então
$$\mathbb{P}\{I \leq t_0\} = \frac{t_0}{n}$$
- Na regra da seleção aleatória, I é uniforme sobre $[n]$ com probabilidade $1/n$ de escolher qualquer amostra. Assim, temos
$$\mathbb{P}\{I \leq t_0\} = \bigcup_{i=1}^{t_0} \mathbb{P}\{I = i\} \leq \sum_{i=1}^{t_0} \mathbb{P}\{I = i\} = \frac{t_0}{n}$$

Theorem (3.12)

Assuma que $f(\cdot; z) \in [0, 1]$ é uma função de perda L-Lipschitz e β-suave para todo z. Suponha que executamos SGM por T passos com tamanho de passos $\alpha_t \leq c/t$ monotonicamente não-crescente. Então, SGM tem estabilidade uniforme com

$$\epsilon_{stab} \leq \frac{1 + 1/\beta c}{n - 1} (2cL^2)^{\frac{1}{\beta c+1}} T^{\frac{\beta c}{\beta c+1}}$$

Em particular, omitindo fatores constantes, obtemos

$$\epsilon_{stab} \lessapprox \frac{T^{1-1/(\beta c+1)}}{n}$$

Definition (4.1)

Seja $f : \Omega \to \Omega$ uma função diferenciável. Definimos a atualização de gradiente com decaimento de peso na taxa μ como

$$G_{f,\mu,\alpha}(w) = (1 - \alpha\mu)w - \alpha\nabla f(w).$$

- Essa atualização é obtida de

$$\nabla\left(f(w) + \frac{\mu}{2}\|w\|^2\right) = \nabla f(w) + \mu w$$

Lemma (4.2)

Assuma que f é β-suave. Então, $G_{f,\mu,\alpha}$ é $(1+\alpha(\beta-\mu))$-expansiva.

Pela desigualdade triangular + suavidade,

$$\begin{aligned}\|G_{f,\mu,\alpha}(v) - G_{f,\mu,\alpha}(w)\| &\leq (1-\alpha\mu)\|v-w\| + \alpha\|\nabla f(w) - \nabla f(v)\| \\ &\leq (1-\alpha\mu)\|v-w\| + \alpha\beta\|w-v\| \\ &= (1-\alpha\mu+\alpha\beta)\|v-w\|.\end{aligned}$$

- Regularização nos adiciona suavidade.

Recorte do Gradiente

- Precisamos evitar valores de gradiente com a norma muito grande, pois isso pode levar a saltos grandes nas atualizações.
- Assim, podemos fazer o recorte do gradiente,

$$\nabla_C f(w) = \begin{cases} \nabla f(w) & \text{se } \|\nabla f(w)\| \leq C \\ C \frac{\nabla f(w)}{\|\nabla f(w)\|} & \text{se } \|\nabla f(w)\| > C \end{cases}$$

- O operador de dropout é uma técnica amplamente utilizada em redes neurais para prevenir overfitting. Ao invés de utilizamos todo o gradiente, zeramos algumas de suas componentes aleatoriamente.

Definition (4.3)

Dizemos que uma função aleatória $D : \Omega \to \Omega$ é um operador de dropout com taxa de dropout s se para cada $v \in D$ temos $\mathbb{E}[\|Dv\|] = s\|v\|$. Para $f : \Omega \to \Omega$ diferenciável, definimos a atualização de dropout $DG_{f,\alpha}$ como

$$DG_{f,\alpha}(v) = v - \alpha D\nabla f(v).$$

Lemma (4.4)

Assuma que f é L-Lipschitz. Então, a atualização de dropout $DG_{f,\alpha}$ com taxa de dropout s é $(\alpha s L)$-limitada.

Lipschitz + linearidade da expectativa,

$$\mathbb{E}\|G_{f,\alpha}(v) - v\| = \alpha\mathbb{E}\|D\nabla f(v)\| = \alpha s\mathbb{E}\|\nabla f(v)\| \leq \alpha s L.$$

Definition (4.5)

Para um tamanho de passo não negativo $\alpha \geq 0$ e uma função $f : \Omega \to \mathbb{R}$, definimos a regra de atualização proximal $P_{f,\alpha}$ como

$$P_{f,\alpha}(w) = \arg\min_{v} \frac{1}{2}\|w - v\|^2 + \alpha f(v) \tag{4.1}$$

Lemma (4.6)

Se f é convexa, a atualização proximal (4.1) é 1-expansiva.

- A ideia aqui é tomar a média dos parâmetros w_t.

Theorem (4.7)

Assuma que $f : \Omega \to [0, 1]$ é uma função convexa, L-Lipschitz e β-suave e que executamos SGD com tamanhos de passo $\alpha_t \leq \alpha \leq 2/\beta$ por T passos. Então, a média das T primeiras iterações do SGD tem estabilidade uniforme de

$$\epsilon_{stab} \leq \frac{\alpha T L^2}{n}$$

Estabilidade de algoritmos iterativos randomizados

Estabilidade do Método do Gradiente Estocástico

Operações que induzem estabilidade

Minimização de Risco Convexo

- Erro de otimização

$$\epsilon_{opt}(w) := \mathbb{E}[R_S[w] - R_S[w_\star^S]]$$

com $w_\star^S$ sendo o minimizador do risco empírico.

- Pelo Teorema 2.2

$$\mathbb{E}[R[w]] \leq \mathbb{E}[R_S[w]] + \epsilon_{stab} = \mathbb{E}[R_S[w_\star^S]] + \epsilon_{opt}(w) + \epsilon_{stab}$$

- Tentamos minimizar ϵ_{opt} sem aumentar muito ϵ_{stab}.

Lemma

Seja w_* o minimizador do risco da população e w_*^S o minimizador do risco empírico dado um conjunto de dados S. Então $\mathbb{E}[R_S[w_*^S]] \leq R[w_*]$.

$$
\begin{aligned}
R[w_*] &= \inf_w R[w] = \inf_w \mathbb{E}_z[f(w;z)] \\
&= \inf_w \mathbb{E}_S \left[\frac{1}{n} \sum_{i=1}^{n} f(w;z_i) \right] \\
&\geq \inf_w \mathbb{E}_S \left[\frac{1}{n} \sum_{i=1}^{n} f(w_*^S;z_i) \right] \\
&= \mathbb{E}_S \left[\frac{1}{n} \sum_{i=1}^{n} f(w_*^S;z_i) \right] = \mathbb{E}[R_S[w_*^S]].
\end{aligned}
$$

Theorem (5.2 - Nemirovski e Yudin)

Assuma que executamos o SDG com tamanho de passo constante α em uma função convexa $R[w] = \mathbb{E}_z[f(w; z)]$. Assuma que $\|\nabla f(w; z)\| \leq L$ e $\|w_0 - w\| \leq D$. Denote $\bar{w}_T$ a média das T iterações do algoritmo. Então temos

$$R[\bar{w}_T] \leq R[w_*] + \frac{1}{2}\frac{D^2}{T\alpha} + \frac{1}{2}L^2\alpha.$$

Corolário 5.3

Seja f uma função de perda convexa satisfazendo $\|\nabla f(w, z)\| \leq L$ e seja w_* um minimizador do risco da população $R[w] = \mathbb{E}_z[f(w; z)]$. Suponha que fazemos uma única passagem de SDG sobre $S = (z_1, \ldots, z_n)$ com um passo adequado e começando de w_0 próximo de w_* a menos de D. Então, a média $\bar{w}_n$ das iterações satisfaz

$$\mathbb{E}[R[\bar{w}_n]] \leq R[w_*] + \frac{DL}{\sqrt{n}}.$$

Proposição 5.4 - Resultado d'Os Cara

Seja S uma amostra de tamanho n. Seja f uma função de perda convexa β-suave satisfazendo $\|\nabla f(w, z)\| \leq L$ e seja w_*^S um minimizador do risco empírico. Suponha que rodamos T iterações de SDG com um passo adequado e começando de w_0 próximo de w_* a menos de D. Então, a média $\bar{w}_T$ sobre as iterações satisfaz

$$\mathbb{E}[R[\bar{w}_T]] \leq \mathbb{E}[R_S[w_*^S]] + \frac{DL}{\sqrt{n}}\sqrt{\frac{n+2T}{T}}.$$

Demonstração.

Aplicando o Teorema 5.2 ao risco empírico R_S, obtemos o erro de otimização $\epsilon_{opt}(\bar{w}_T) \leq \frac{1}{2}\frac{D^2}{T\alpha} + \frac{1}{2}L^2\alpha$. Combinando as duas desigualdades:

$$\mathbb{E}[R[\bar{w}_T]] \leq \mathbb{E}[R_S[w_*^S]] + \frac{1}{2}\frac{D^2}{T\alpha} + \frac{1}{2}L^2(1 + \frac{2T}{n})\alpha.$$

Escolhendo $\alpha = \frac{D\sqrt{n}}{L\sqrt{T(n+2T)}}$ resulta no limite fornecido na proposição. □

- Os resultados não são diretamente comparáveis, pois um usa o risco populacional e o outro o risco empírico.
- Se $T = n$, eles perdem por um fator $\sqrt{3}$
- A aproximação deles permite $T > n$, e quando T vai pro infinito, perdem apenas por $\sqrt{2}$.