# Clase 17: Descenso en Gradiente Estocástico

Responsable: Jesús Bernardo Solórzano Flores

*EST-25134, Primavera 2021*
*Dr. Alfredo Garbuno Iñigo*
*Marzo 16, 2021*

## 1. Introducción

Recordemos que el objetivo del aprendizaje es minimizar una función de pérdida $L_{\mathcal{D}}(h) = \mathbb{E}_{S\sim\mathcal{D}}[l(h,z)]$. Hasta ahora se han discutido métodos de aprendizaje que dependen del riesgo empírico. Esto es, partimos de una muestra de entrenamiento $S$ y definimos, en particular, la función de pérdida empírica

$$L_S(h) = \frac{1}{m}\sum_{i=1}^m l\Big(h, z^{(i)}\Big) \quad \text{con } z^{(i)} \sim \mathcal{D},\ S \sim \mathcal{D}^m$$

En Minimización de Pérdida Regularizada (RLM) se elegía una hipótesis que minimizara conjuntamente $L_S(h)$ y la función de regularización sobre $h$, esto es

$$\text{mín}\, L_s(h) + R(h) = \text{mín}\, J_S(h)$$

El método de descenso en gradiente para minimizar una función $f(w)$ convexa y diferenciable utiliza iteraciones dadas por

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \eta\, \boldsymbol{p}_t$$

con la dirección de descenso $\boldsymbol{p}_t = -\nabla_{\boldsymbol{w}} J_S(\boldsymbol{w})$. El método de descenso en gradiente estocástico permite cambiar esta dirección $\boldsymbol{p}_t$ por un estimador insesgado aleatorio $\hat{\boldsymbol{p}}_t$, esto es, un estimador tal que $\mathbb{E}[\hat{\boldsymbol{p}}_t] = \boldsymbol{p}_t$. Utilizaremos SGD para funciones convexas y Lipschitz. Además, GD tendrá la misma tasa de convergencia, en promedio, que SDG.

## 2. Descenso en Gradiente

El gradiente de una función diferenciable $f : \mathbb{R}^d \leftarrow \mathbb{R}$ en $\boldsymbol{w} \in \mathbb{R}^d$, denotado por $\nabla f(\boldsymbol{w})$, es el vector de derivadas parciales de $f$, esto es, $\nabla f(\boldsymbol{w}) = \Big(\frac{\partial f}{\partial w_1}, \frac{\partial f}{\partial w_2}, \ldots, \frac{\partial f}{\partial w_d}\Big)$. Descenso en gradiente es un algoritmo iterativo que en cada iteración, empezando con un valor inicial $\boldsymbol{w}^{(1)} = \boldsymbol{0}$, genera una actualización en la dirección de máximo descenso, esto es

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta\, \nabla f(\boldsymbol{w}^{(t)}),$$

donde $\eta > 0$. Después de $T$ iteraciones, el algoritmo regresa como respuesta

i) el vector de medias, $\overline{\boldsymbol{w}}_T = \dfrac{1}{T}\displaystyle\sum_{t=1}^{T} \boldsymbol{w}^{(t)}$,

ii) la última iteración, $\boldsymbol{w}^{(T)}$,

iii) o el punto con la menor pérdida posible $\boldsymbol{w}(T) = \text{argmin}_{t\in\{1,\ldots,T\}} f(\boldsymbol{w}^{(t)})$.

Si utilizamos la aproximación de Taylor de $f$ alrededor de $\boldsymbol{w}$ dada por $f(\boldsymbol{u}) \approx f(\boldsymbol{w}) + \langle \boldsymbol{u} - \boldsymbol{w}, \nabla f(\boldsymbol{w})\rangle$ y utilizando la propiedad de convexidad, evaluando la función en cualquier punto, se tiene que

$$f(\boldsymbol{u}) \geq f(\boldsymbol{w}) + \langle \boldsymbol{u} - \boldsymbol{w}, \nabla f(\boldsymbol{w}) \rangle.$$

Entonces, para un $\boldsymbol{w}$ relativament cercano a $\boldsymbol{w}^{(t)}$ se satisface que

$$f(\boldsymbol{w}) \approx f(\boldsymbol{w}^{(t)}) + \langle \boldsymbol{w} - \boldsymbol{w}^{(t)}, \nabla f(\boldsymbol{w}^{(t)}) \rangle,$$

es decir, $f(\boldsymbol{w})$ será aproximadamente la cota inferior. Luego, podemos minimizar la aproximación de $f(\boldsymbol{w})$ con la mejor actualización

$$\boldsymbol{w}^{(t+1)} = \text{argmin}_{\boldsymbol{w}} \, \|\boldsymbol{w} - \boldsymbol{w}^{(t)}\|^2 + \eta\Big(f(\boldsymbol{w}^{(t)}) + \langle \boldsymbol{w} - \boldsymbol{w}^{(t)}, \nabla f(\boldsymbol{w}^{(t)}) \rangle\Big)$$

El parámetro $\eta$ me permite enontrar el compromiso entre la aproximación a la recta y buscar puntos más cercanos al punto en el que estamos. Si derivamos respecto a $\boldsymbol{w}$ e igualamos a cero, la solución coincide con el método iterativo de la ecuación (1).

### 2.1. Análisis de Descenso en Gradiente

Supongamos que la función objetivo es Lipschitz-convexa-acotada. Sea $\boldsymbol{w}^*$ el minimizador y $\|\boldsymbol{w}^*\| \leq B$, para alguna $B$. Consideremos $\overline{\boldsymbol{w}}_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{w}^{(t)}$.Queremos medir $f(\boldsymbol{w}^{(t)}) - f(\boldsymbol{w}^*)$, es decir, medir el desempeño del candidato respecto al mímimo de la función. Por la definición de $\overline{\boldsymbol{w}}_T$ y la desigualdad de Jensen, tenemos que

$$\begin{aligned} f(\boldsymbol{w}^{(t)}) - f(\boldsymbol{w}^*) &= f\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{w}^{(t)}\right) - f(\boldsymbol{w}^*) \\ &\leq \frac{1}{T}\sum_{t=1}^{T} \Big(f(\boldsymbol{w}^{(t)})\Big) - f(\boldsymbol{w}^*) \\ &= \frac{1}{T}\sum_{t=1}^{T} \Big(f(\boldsymbol{w}^{(t)}) - f(\boldsymbol{w}^*)\Big). \end{aligned}$$

Por la convexidad de $f$, para todo $i \in \{1, \ldots, T\}$ se satisface que

$$f(\boldsymbol{w}^{(i)}) - f(\boldsymbol{w}^*) \leq \langle \boldsymbol{w}^{(i)} - \boldsymbol{w}^*, \nabla f(\boldsymbol{w}^{(i)}) \rangle.$$

Combinando estas ecuaciones obtenemos que

$$f(\overline{\boldsymbol{w}}_T) - f(\boldsymbol{w}^*) \leq \frac{1}{T}\sum_{t=1}^{T} \langle \boldsymbol{w}^{(t)} - \boldsymbol{w}^*, \nabla f(\boldsymbol{w}^{(t)}) \rangle.$$

Para acotar el lado derecho de la desigualdad usamos el siguiente lemma.

**Lemma 2.1.** *Sea una sucesión $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_T$. Con un proceso iterativo que empieza en el vector $\boldsymbol{w}^{(1)} = \boldsymbol{0}$ con regla de actualización de la forma $\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta\, \boldsymbol{v}_t$, se satisface que*

$$\tfrac{1}{T}\sum_{t=1}^{T} \langle \boldsymbol{w}^{(t)} - \boldsymbol{w}^*, \boldsymbol{v}_t \rangle \leq \frac{\|\boldsymbol{w}^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \|\boldsymbol{v}_t\|^2.$$

En particular, si consideramos $B, \rho > 0$, si para todo $t$ tenemos que $\|\boldsymbol{v}_t\| \leq \rho$, y fijando $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, entonces para todo $\boldsymbol{w}^*$ tal que $\|\boldsymbol{w}^*\| \leq B$ tenemos que

$$\frac{1}{T}\sum_{t=1}^{T}\langle \boldsymbol{w}^{(t)} - \boldsymbol{w}^*, \boldsymbol{v}_t\rangle \leq \frac{B\rho}{\sqrt{T}}$$

Esto quiere decir que si queremos un nivel de precisión $f(\overline{\boldsymbol{w}}_T) - f(\boldsymbol{w}^*) \leq \varepsilon$, entonces lo que necesitamos es aplicar el método de descenso en gradiente con $T \geq \frac{B^2\rho^2}{\varepsilon^2}$ iteraciones.

## 3. Descenso en Gradiente Estocástico

En descenso en gradiente estocásitco no necesitamos que la actualización en la dirección sea exactamente el gradiente. Lo que utilizaremos es un estimador insesgado del gradiente.

**Teorema 3.1.** *Sean $B, \rho > 0$. Sean $f$ una función convexa y $\boldsymbol{w}^* \in \operatorname{argmin}_{\{\boldsymbol{w}:\|\boldsymbol{w}\|\leq B\}} f(\boldsymbol{w})$. Entonces, después de $T$ iteraciones en SGD con $\eta = \frac{B^2}{\rho^2 T}$ y suponiendo que para todo paso $t$ se cumple que $\|\boldsymbol{v}_t\| \leq \rho$ con probabilidad 1, se tiene que*

$$\mathbb{E}[f(\overline{\boldsymbol{w}}_T)] - f(\boldsymbol{w}^*) \leq \frac{B\rho}{\sqrt{T}}.$$

Notemos que podemos hacer modificaciones que pueden funcionar bajo las siguientes consideraciones:

1. Hasta ahora hemos considerado que el espacio de las actualizaciones sea acotado, $\|\boldsymbol{w}\| \leq B$. No hay nada dentro de la actualización usando el gradiente que garantice que $\|\boldsymbol{w}^{(t+1)}\| \leq B$. Esto puede ocasionar que la dirección de descenso proponga un candidato tal que $\|\boldsymbol{w}\|^{(t+1)} > B$, lo que implicaría que $\boldsymbol{w}^{(t+1)} \notin H$. Para ello, podemos usar proyecciones

   $$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta\, \boldsymbol{v}_t$$

   y proyectar sobre $\boldsymbol{w}^{(t+1)} = \operatorname{argmin}_{\{\boldsymbol{w}:\|\boldsymbol{w}\leq B\}}\|\boldsymbol{w} - \boldsymbol{w}^{\left(t+\frac{1}{2}\right)}\|$. Así, garantizamos que $\boldsymbol{w}^{(t)}, \overline{\boldsymbol{w}}_T \in H$.
2. Podemos hacer un tamaño de paso variable tomando

   $$\eta_t = \frac{B}{\rho\sqrt{t}}.$$

   Conforme avancemos y tengamos más observaciones, el proceso iterativo de búsqueda se mantiene controlado dentro de una vecindad de búsqueda.

### 3.1. SDG con Error de Generalización

¿Cómo conectamos SDG con el error de generalización? Recordemos que en aprendizaje nos enfrentamos con el problema de la minimización del error de generalización

$$L_{\mathcal{D}}(\boldsymbol{w}) = \mathbb{E}_{z\sim\mathcal{D}}[l(\boldsymbol{w}, z)]$$

Tomemos $z \sim \mathcal{D}$ y evaluemos el gradiente con esa observación, entoces

$$\nabla\, l(\boldsymbol{w}^{(t)}, z) = \tilde{\nabla}\, L_{\mathcal{D}}(\boldsymbol{w}^{(t)})$$

Luego, por el Teorema 3.1,

$$\mathbb{E}[L_{\mathcal{D}}(\overline{\boldsymbol{w}}_T) - L_{\mathcal{D}}(\boldsymbol{w}^*)] \leq \varepsilon,$$

de donde se sigue que

$$\mathbb{E}[L_{\mathcal{D}}(\overline{\boldsymbol{w}}_T)] \leq L_{\mathcal{D}}(\boldsymbol{w}^*) + \varepsilon$$

y, por definición,

$$\mathbb{E}[L_{\mathcal{D}}(\overline{\boldsymbol{w}}_T)] \leq \min_{\boldsymbol{w} \in H} L_{\mathcal{D}}(\boldsymbol{w}) + \varepsilon.$$

Por lo tanto tenemos aprendizaje siempre y cuando $T \geq \frac{B^2\rho^2}{\varepsilon^2}$, $T$ iteraciones del SGD con muestras aleatorias de $\mathcal{D}$.

### 3.2. SGD con RLM

Recordemos que la función objetivo es $L_S(\boldsymbol{w}) + R(\boldsymbol{w})$, donde $L_S(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} l(\boldsymbol{w}, z^{(i)})$. Por propiedades del operador gradiente, tenemos que

$$\nabla\, L_{\mathcal{D}}(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m} \nabla\, l(\boldsymbol{w}, z^{(i)})$$

¿Cómo podemos construir un estimador de la dirección de descenso? Tomemos un índice $i \sim U\{1, \ldots, m\}$ y evaluamos en ese índice

$$\nabla\, l(\boldsymbol{w}, z^{(i)}) = \tilde{\nabla}\, L_S(\boldsymbol{w}).$$

Si calculamos el valor esperado

$$\mathbb{E}_i\Big[\nabla\, l(\boldsymbol{w}, z^{(i)})\Big] = \sum_{i=1}^{m} \nabla\, l(\boldsymbol{w}, z^{(i)}) \cdot \frac{1}{m} = \nabla\, L_S(\boldsymbol{w}).$$

Por lo tanto, hemos construido un estimador insesgado de la dirección de descenso.

Notemos que en cada iteración del método los índices se escogieron de manera independiente y sobre el mismo conjunto $\{1, \ldots, m\}$, es decir, los puntos se escogieron a través de un muestreo aleatorio con reemplazo. Este mecanismo puede acarrear algunas desventajas. Supongamos que tenemos $m$ observaciones y obtenemos una muestra con reemplazo de índices de tamaño $m$. Entonces la probabilidad de escoger una observación al menos una vez es

$$1 - \mathbb{P}(\text{ no escoger } i) = 1 - \Big(1 - \tfrac{1}{m}\Big)^m \approx 1 - e^{-1} \approx 63\,\%$$

En consecuencia, puede ser que estemos desperdiciando en promedio el 37 % de los datos para evaluar el gradiente. En otras palabras, puede ser que, al no observar estos datos, lo que aprendió el algoritmo no sea suficientemente generalizable.

En la práctica, lo que podemos hacer es proceder de la siguiente manera:

1. Permutamos nuestras observaciones.
2. Aplicamos SGD con $m$ iteraciones utilizando el orden en que aparecieron las muestras.
3. Una vez que pasemos por las $m$ iteraciones, volvemos a permutar (cambiamos el orden), aplicamos SGD y repetimos hasta convergencia.

Replicamos los pasos hasta alcanzar un número deseado de tolerancia, $Tol$, tal que $\|\nabla f(\boldsymbol{w}^{(T)})\| \leq Tol$. De esta manera, nos aseguramos que pasamos por todas las observaciones y que, en promedio, nuestras direcciones de descenso aleatorio sean direcciones insesgadas.