

Métodos de primer orden

Axel Sirota

Facultad de Ciencias Exactas y Naturales

Departamento de Matemática

Hoja de ruta

- 1 Introducción
- 2 Convergencia de algoritmos de tipo batch
 - Convergencia Puntual
 - Convergencia c.t.p.
 - Complejidad exponencial
- 3 Algoritmos estocásticos
 - Convergencia en L1
 - Convergencia en c.t.p.
- 4 Conclusiones y continuación
- 5 Referencias

Marco teórico del problema

Consideremos una muestra aleatoria $\{x_i, y_i\}_{i \leq N} \subset \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ tomada bajo una distribución $\mathbb{P}(x, y)$.

El objetivo del *Machine Learning* es encontrar $h^* \in \mathcal{H} = \{h : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_y}\}$ tal que $R(h) = \mathbb{E}[1[h(x) \neq y]]$ sea mínima.

Dicho contexto es *variacional y estocástico*.

La práctica usual consiste en tomar una dada $\tilde{h} : \mathbb{R}^{d_x} \times \mathbb{R}^d \rightarrow \mathbb{R}^{d_y}$ que surge del conocimiento a priori del problema y tomar:

$$\mathcal{H}_{\tilde{h}} := \{\tilde{h}(\cdot; w) : w \in \mathbb{R}^d\}$$

Marco teórico del problema

Consideremos una muestra aleatoria $\{x_i, y_i\}_{i \leq N} \subset \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ tomada bajo una distribución $\mathbb{P}(x, y)$.

El objetivo del *Machine Learning* es encontrar $h^* \in \mathcal{H} = \{h : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_y}\}$ tal que $R(h) = \mathbb{E}[1[h(x) \neq y]]$ sea mínima.

Dicho contexto es *variacional y estocástico*.

La práctica usual consiste en tomar una dada $\tilde{h} : \mathbb{R}^{d_x} \times \mathbb{R}^d \rightarrow \mathbb{R}^{d_y}$ que surge del conocimiento a priori del problema y tomar:

$$\mathcal{H}_{\tilde{h}} := \{\tilde{h}(\cdot; w) : w \in \mathbb{R}^d\}$$

Marco teórico del problema

Consideremos una muestra aleatoria $\{x_i, y_i\}_{i \leq N} \subset \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ tomada bajo una distribución $\mathbb{P}(x, y)$.

El objetivo del *Machine Learning* es encontrar $h^* \in \mathcal{H} = \{h : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_y}\}$ tal que $R(h) = \mathbb{E}[1[h(x) \neq y]]$ sea mínima.

Dicho contexto es *variacional y estocástico*.

La práctica usual consiste en tomar una dada $\tilde{h} : \mathbb{R}^{d_x} \times \mathbb{R}^d \rightarrow \mathbb{R}^{d_y}$ que surge del conocimiento a priori del problema y tomar:

$$\mathcal{H}_{\tilde{h}} := \{\tilde{h}(\cdot; w) : w \in \mathbb{R}^d\}$$

Marco teórico del problema

Dada $\ell : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \rightarrow \mathbb{R}$ una métrica en \mathbb{R}^{d_y} , el objetivo se reduce a minimizar $R(w)$ donde:

$$R(w) = \int_{\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}} \ell(h(x; w), y) d\mathbb{P}(x, y) = \mathbb{E}[\ell(h(x; w), y)]$$

Al no conocer \mathbb{P} se optimiza $R_n(w)$:

$$R_n(w) = \frac{1}{n} \sum_{i=1}^n \ell(h(x_i; w), y_i)$$

Que resulta un estimador insesgado de R para cada $w \in \mathbb{R}^d$ por la desigualdad Hoeffding

No obstante, este subconjunto \mathcal{H} parametrizado además debe minimizar $L(h^*, w) = |R(h^*) - R_n(h_w)|$ donde h^* es la función óptima objetivo y h_w es la minimizante de R_n sobre el subconjunto \mathcal{H} .

Marco teórico del problema

Dada $\ell : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \rightarrow \mathbb{R}$ una métrica en \mathbb{R}^{d_y} , el objetivo se reduce a minimizar $R(w)$ donde:

$$R(w) = \int_{\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}} \ell(h(x; w), y) d\mathbb{P}(x, y) = \mathbb{E}[\ell(h(x; w), y)]$$

Al no conocer \mathbb{P} se optimiza $R_n(w)$:

$$R_n(w) = \frac{1}{n} \sum_{i=1}^n \ell(h(x_i; w), y_i)$$

Que resulta un estimador insesgado de R para cada $w \in \mathbb{R}^d$ por la desigualdad Hoeffding

No obstante, este subconjunto \mathcal{H} parametrizado además debe minimizar $L(h^*, w) = |R(h^*) - R_n(h_w)|$ donde h^* es la función óptima objetivo y h_w es la minimizante de R_n sobre el subconjunto \mathcal{H} .

Marco teórico del problema

Dada $\ell : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \rightarrow \mathbb{R}$ una métrica en \mathbb{R}^{d_y} , el objetivo se reduce a minimizar $R(w)$ donde:

$$R(w) = \int_{\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}} \ell(h(x; w), y) d\mathbb{P}(x, y) = \mathbb{E}[\ell(h(x; w), y)]$$

Al no conocer \mathbb{P} se optimiza $R_n(w)$:

$$R_n(w) = \frac{1}{n} \sum_{i=1}^n \ell(h(x_i; w), y_i)$$

Que resulta un estimador insesgado de R para cada $w \in \mathbb{R}^d$ por la desigualdad Hoeffding

No obstante, este subconjunto \mathcal{H} parametrizado además debe minimizar $L(h^*, w) = |R(h^*) - R_n(h_w)|$ donde h^* es la función óptima objetivo y h_w es la minimizante de R_n sobre el subconjunto \mathcal{H} .

Marco teórico del problema

En conclusión, la obtención de dicha h óptima se separa en dos problemas no disjuntos:

- ① Encontrar \mathcal{H} parametrizada por $w \in \mathbb{R}^d$ tal que $L(h^*, w^*)$ sea mínima, donde $w^* = \arg \min_{w \in \mathbb{R}^d} R_n(w)$.
- ② Dado \mathcal{H} parametrizado, hallar $w^* = \arg \min_{w \in \mathbb{R}^d} R_n(w)$.

El problema 1 suele tener diferentes enfoques pero ninguno estrictamente teórico, sino que mas bien son basados en el conocimiento a priori del problema

Nos vamos a enfocar en el problema 2 viendo diferentes algoritmos existentes para resolverlo y sus propiedades de convergencia

Marco teórico del problema

En conclusión, la obtención de dicha h óptima se separa en dos problemas no disjuntos:

- ① Encontrar \mathcal{H} parametrizada por $w \in \mathbb{R}^d$ tal que $L(h^*, w^*)$ sea mínima, donde $w^* = \arg \min_{w \in \mathbb{R}^d} R_n(w)$.
- ② Dado \mathcal{H} parametrizado, hallar $w^* = \arg \min_{w \in \mathbb{R}^d} R_n(w)$.

El problema 1 suele tener diferentes enfoques pero ninguno estrictamente teórico, sino que mas bien son basados en el conocimiento a priori del problema

Nos vamos a enfocar en el problema 2 viendo diferentes algoritmos existentes para resolverlo y sus propiedades de convergencia

Algoritmos de primer orden

Comunmente para encontrar $\arg \min_{w \in \mathbb{R}^d} F(w)$ se utilizan algoritmos iterativos de primer orden; es decir, algoritmos que se pueden representar por $g : \mathbb{R}^d \rightarrow \mathbb{R}^d$ tal que $w_n = \underbrace{g \circ \dots \circ g}_{n \text{ veces}}(w_0)$ donde

$g(w) = N(w, F(w), \nabla F(w))$ para alguna $N \in C^\infty(\mathbb{R}^{d_y})$. Se suelen dividir en dos grandes grupos:

De tipo *batch*, donde para cada iteración se utilizan todo el conjunto de datos $\{x_i, y_i\}$. Un ejemplo de esta categoría es el *descenso de gradiente* (GD) dado por $w_n = g(w_{n-1}) = w_{n-1} - \alpha_n \sum_{i=1}^N \nabla F(x_i, y_i)$.

De tipo *estocástico*, donde se elije en cada iteración al azar un subconjunto $S \subset \{x_i, y_i\}$ para calcular g . Un ejemplo de esta categoría es el *descenso estocástico de gradiente* (SG) dado por $w_n = g(w_{n-1}) = w_{n-1} - \alpha_n \nabla F(x_i, y_i)$ para un i elegido al azar.

Algoritmos de primer orden

Comunmente para encontrar $\arg \min_{w \in \mathbb{R}^d} F(w)$ se utilizan algoritmos iterativos de primer orden; es decir, algoritmos que se pueden representar por $g : \mathbb{R}^d \rightarrow \mathbb{R}^d$ tal que $w_n = \underbrace{g \circ \dots \circ g}_{n \text{ veces}}(w_0)$ donde

$g(w) = N(w, F(w), \nabla F(w))$ para alguna $N \in C^\infty(\mathbb{R}^{d_y})$. Se suelen dividir en dos grandes grupos:

De tipo *batch*, donde para cada iteración se utilizan todo el conjunto de datos $\{x_i, y_i\}$. Un ejemplo de esta categoría es el *descenso de gradiente* (GD) dado por $w_n = g(w_{n-1}) = w_{n-1} - \alpha_n \sum_{i=1}^N \nabla F(x_i, y_i)$.

De tipo *estocástico*, donde se elige en cada iteración al azar un subconjunto $S \subset \{x_i, y_i\}$ para calcular g . Un ejemplo de esta categoría es el *descenso estocástico de gradiente* (SG) dado por $w_n = g(w_{n-1}) = w_{n-1} - \alpha_n \nabla F(x_i, y_i)$ para un i elegido al azar.

Algoritmos de primer orden

Comunmente para encontrar $\arg \min_{w \in \mathbb{R}^d} F(w)$ se utilizan algoritmos iterativos de primer orden; es decir, algoritmos que se pueden representar por $g : \mathbb{R}^d \rightarrow \mathbb{R}^d$ tal que $w_n = \underbrace{g \circ \dots \circ g}_{n \text{ veces}}(w_0)$ donde

$g(w) = N(w, F(w), \nabla F(w))$ para alguna $N \in C^\infty(\mathbb{R}^{d_y})$. Se suelen dividir en dos grandes grupos:

De tipo *batch*, donde para cada iteración se utilizan todo el conjunto de datos $\{x_i, y_i\}$. Un ejemplo de esta categoría es el *descenso de gradiente* (GD) dado por $w_n = g(w_{n-1}) = w_{n-1} - \alpha_n \sum_{i=1}^N \nabla F(x_i, y_i)$.

De tipo *estocástico*, donde se elige en cada iteración al azar un subconjunto $S \subset \{x_i, y_i\}$ para calcular g . Un ejemplo de esta categoría es el *descenso estocástico de gradiente* (SG) dado por $w_n = g(w_{n-1}) = w_{n-1} - \alpha_n \nabla F(x_i, y_i)$ para un i elegido al azar.

Hoja de ruta

- 1 Introducción
- 2 Convergencia de algoritmos de tipo batch
 - Convergencia Puntual
 - Convergencia c.t.p.
 - Complejidad exponencial
- 3 Algoritmos estocásticos
 - Convergencia en L1
 - Convergencia en c.t.p.
- 4 Conclusiones y continuación
- 5 Referencias

Definiciones

Definición (Débilmente convexo)

Decimos que $F : \mathbb{R}^d \rightarrow \mathbb{R}$, tal que $F \in C^1$ es *débilmente convexo* si cumple las siguientes dos propiedades:

- Existe un único w^* tal que $F_{inf} := F(w^*) \leq F(w)$ para todo $w \in \mathbb{R}^n$.
- Para todo $\epsilon > 0$ vale que $\inf_{\|w - w^*\|^2 > \epsilon} (w - w^*)^\top \nabla F(w) > 0$

Definición (Condición de Robbins - Monro)

Si consideramos un algoritmo de primer orden, decimos que los incrementos $\{\alpha_k\} \subset \mathbb{R}_+$ (de existir) cumplen la condición de *Robbins - Monro* si $\{\alpha_k\} \in \ell^2 \setminus \ell^1$:

Más definiciones

Definición (Condición de *Polyak-Lojasiewicz*)

Decimos que una función $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^1$ es *PL*-convexa, o cumple la condición de *Polyak-Lojasiewicz* si existe $\mu > 0$ tal que para todo $x \in \mathbb{R}^d$ vale:

$$\frac{1}{2} \|\nabla F(x)\|_2^2 \geq \mu (F(x) - F_{\inf})$$

Resultados de convergencia puntual para GD

Teorema

Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^1$ la función objetivo, asumamos que F es débilmente convexo, w^* su mínimo y que existen $A, B \geq 0$ tal que para todo $w \in \mathbb{R}^d$ vale que:

$$\|\nabla F(w)\|^2 \leq A + B \|w - w^*\|^2$$

Luego si consideramos el algoritmo de descenso de gradiente por batch tal que los incrementos $\{\alpha_k\}$ cumplen la condición Robbins - Monro entonces:

$$w_k \xrightarrow[k \rightarrow \infty]{} w^*$$

Idea de la demostración

Sea $h_k = \|w_k - w^*\|^2$, entonces definiendo:

$$\mu_k = \prod_{j=1}^{k-1} \frac{1}{1 + \alpha_j^2 B}$$

Vale que $\alpha_k (w_k - w^*) \nabla F(w_k) \leq \mu_{k+1} h_{k+1} - \mu_k h_k \leq \alpha_k^2 A \mu_k \leq \alpha_k^2 A$.

Luego vemos que $\{h_k\}$ converge y además $\{\alpha_k (w_k - w^*) \nabla F(w_k)\}$ es sumable.

Finalmente, como los incrementos cumplen la condición de Robbins Monro uno obtiene que $w_k \xrightarrow[k \rightarrow \infty]{} w^*$.

Idea de la demostración

Sea $h_k = \|w_k - w^*\|^2$, entonces definiendo:

$$\mu_k = \prod_{j=1}^{k-1} \frac{1}{1 + \alpha_j^2 B}$$

Vale que $\alpha_k (w_k - w^*) \nabla F(w_k) \leq \mu_{k+1} h_{k+1} - \mu_k h_k \leq \alpha_k^2 A \mu_k \leq \alpha_k^2 A$.

Luego vemos que $\{h_k\}$ converge y ademas $\{\alpha_k (w_k - w^*) \nabla F(w_k)\}$ es sumable.

Finalmente, como los incrementos cumplen la condicion de Robbins Monro uno obtiene que $w_k \xrightarrow[k \rightarrow \infty]{} w^*$.

Resultados de convergencia puntual para GD

Teorema

Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ la función objetivo tal que $F \in C^1$, F es L -Lipschitz y μ -convexa; entonces el algoritmo descenso de gradiente por batch con incremento fijo $\alpha_k = \frac{1}{L}$ cumple:

$$F(w_k) - F_{inf} \leq \left(1 - \frac{\mu}{L}\right)^k (F(w_1) - F_{inf})$$

Idea de la demostración

Por las implicancias de ser L –Lipschitz y PL –convexa tenemos:

$$F(w_{k+1}) - F(w_k) \leq -\frac{1}{2L} \|\nabla F(w_k)\|_2^2 \leq -\frac{\mu}{L} (F(w_k) - F_{inf})$$

Luego inductivamente:

$$F(w_{k+1}) - F_{inf} \leq \left(1 - \frac{\mu}{L}\right)^k (F(w_1) - F_{inf})$$

Caso no convexo

Bajo qué casos el algoritmo GD converge (en alguna forma) con objetivos no convexos?

Para responder esto, sea $g : M \rightarrow M$ la fórmula del algoritmo de primer orden en $M \subset \mathbb{R}^N$ una subvariedad sin borde de dimensión d .

Definición

Sea $f : M \rightarrow \mathbb{R}$ tal que $f \in C^2$ y $x^* \in \mathbb{R}^d$, luego decimos que x^* es un *punto silla estricto* de f si es un punto crítico y $\lambda_{\min}(\nabla^2 f(x^*)) < 0$

Notaremos M^* al conjunto de puntos silla estrictos de f .

Definición

Sea:

$$\mathcal{A}_g^* := \left\{ x : g(x) = x \quad \max_i |\lambda_i(Dg(x))| > 1 \right\}$$

A este conjunto lo llamaremos el conjunto de *puntos fijos inestables*

Caso no convexo

Bajo qué casos el algoritmo GD converge (en alguna forma) con objetivos no convexos?

Para responder esto, sea $g : M \rightarrow M$ la fórmula del algoritmo de primer orden en $M \subset \mathbb{R}^N$ una subvariedad sin borde de dimensión d .

Definición

Sea $f : M \rightarrow \mathbb{R}$ tal que $f \in C^2$ y $x^* \in \mathbb{R}^d$, luego decimos que x^* es un *punto silla estricto* de f si es un punto crítico y $\lambda_{\min}(\nabla^2 f(x^*)) < 0$

Notaremos M^* al conjunto de puntos silla estrictos de f .

Definición

Sea:

$$\mathcal{A}_g^* := \left\{ x : g(x) = x \quad \max_i |\lambda_i(Dg(x))| > 1 \right\}$$

A este conjunto lo llamaremos el conjunto de *puntos fijos inestables*

Caso no convexo

Bajo qué casos el algoritmo GD converge (en alguna forma) con objetivos no convexos?

Para responder esto, sea $g : M \rightarrow M$ la fórmula del algoritmo de primer orden en $M \subset \mathbb{R}^N$ una subvariedad sin borde de dimensión d .

Definición

Sea $f : M \rightarrow \mathbb{R}$ tal que $f \in C^2$ y $x^* \in \mathbb{R}^d$, luego decimos que x^* es un *punto silla estricto* de f si es un punto crítico y $\lambda_{\min}(\nabla^2 f(x^*)) < 0$

Notaremos M^* al conjunto de puntos silla estrictos de f .

Definición

Sea:

$$\mathcal{A}_g^* := \left\{ x : g(x) = x \quad \max_i |\lambda_i(Dg(x))| > 1 \right\}$$

A este conjunto lo llamaremos el conjunto de *puntos fijos inestables*

Caso no convexo

Teorema

Sea $g \in C^1(M)$ tal que $\det(Dg(x)) \neq 0$ para todo $x \in M$, luego el conjunto de puntos iniciales que convergen por g a un punto fijo inestable tiene medida cero:

$$\mu \left(\left\{ x_0 : \lim_k g^k(x_0) \in \mathcal{A}_g^* \right\} \right) = 0$$

Corolario

Bajo las mismas hipótesis si agregamos que $M^ \subseteq \mathcal{A}_g^*$ entonces*

$$\mu \left(\left\{ x_0 : \lim_k g^k(x_0) \in M^* \right\} \right) = 0$$

Caso no convexo

Teorema

Sea $g \in C^1(M)$ tal que $\det(Dg(x)) \neq 0$ para todo $x \in M$, luego el conjunto de puntos iniciales que convergen por g a un punto fijo inestable tiene medida cero:

$$\mu \left(\left\{ x_0 : \lim_k g^k(x_0) \in \mathcal{A}_g^* \right\} \right) = 0$$

Corolario

Bajo las mismas hipótesis si agregamos que $M^ \subseteq \mathcal{A}_g^*$ entonces*

$$\mu \left(\left\{ x_0 : \lim_k g^k(x_0) \in M^* \right\} \right) = 0$$

Algoritmos de tipo batch estándar

Además del algoritmo GD, en la optimización de tipo batch existen dos algoritmos muy usuales:

El algoritmo de punto próximo está dado por la iteración:

$$x_{k+1} = g(x_k) \triangleq \arg \min_{z \in M} f(z) + \frac{1}{2\alpha} \|x_k - z\|_2^2$$

Por otro lado, si definimos $g_i(x) = x - \alpha \sum_{j \in S_i} e_j^T \nabla f(x)$ entonces el algoritmo de descenso de coordenadas por bloques está dado por:

$$x_{k+1} = g(x_k) \triangleq g_b \circ g_{b-1} \circ \cdots \circ g_1(x_k)$$

Algoritmos de tipo batch estándar

Además del algoritmo GD, en la optimización de tipo batch existen dos algoritmos muy usuales:

El algoritmo de punto próximo está dado por la iteración:

$$x_{k+1} = g(x_k) \triangleq \arg \min_{z \in M} f(z) + \frac{1}{2\alpha} \|x_k - z\|_2^2$$

Por otro lado, si definimos $g_i(x) = x - \alpha \sum_{j \in S_i} e_j^T \nabla f(x)$ entonces el algoritmo de descenso de coordenadas por bloques está dado por:

$$x_{k+1} = g(x_k) \triangleq g_b \circ g_{b-1} \circ \cdots \circ g_1(x_k)$$

Algoritmos de tipo batch estándar

Además del algoritmo GD, en la optimización de tipo batch existen dos algoritmos muy usuales:

El algoritmo de punto próximo está dado por la iteración:

$$x_{k+1} = g(x_k) \triangleq \arg \min_{z \in M} f(z) + \frac{1}{2\alpha} \|x_k - z\|_2^2$$

Por otro lado, si definimos $g_i(x) = x - \alpha \sum_{j \in S_i} e_j^T \nabla f(x)$ entonces el algoritmo de descenso de coordenadas por bloques está dado por:

$$x_{k+1} = g(x_k) \triangleq g_b \circ g_{b-1} \circ \cdots \circ g_1(x_k)$$

Un marco de demostración común

Teorema

Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^2$ la función objetivo con Hessiano acotado con constante L , w^ algún mínimo local de F ; entonces el algoritmo descenso de gradiente por batch con incremento fijo $\alpha < \frac{1}{L}$ cumple:*

$$w_k \xrightarrow[k \rightarrow \infty]{\text{c.t.p.}} w^*$$

Teorema

Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^2$ la función objetivo con Hessiano acotado con constante L , w^ algún mínimo local de F ; entonces el algoritmo punto próximo con incremento fijo $\alpha < \frac{1}{L}$ cumple:*

$$w_k \xrightarrow[k \rightarrow \infty]{\text{c.t.p.}} w^*$$

Un marco de demostración común

Teorema

Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^2$ la función objetivo con Hessiano acotado por bloques con constante L_b , w^ algún mínimo local de F ; entonces el algoritmo descenso de gradiente por coordenadas con incremento fijo $\alpha < \frac{1}{L_b}$ cumple:*

$$w_k \xrightarrow[k \rightarrow \infty]{\text{c.t.p.}} w^*$$

Idea de las demostraciones

En los tres casos las hipótesis llevan a demostrar que:

- ① $\det(Dg)(x) \neq 0$
- ② $M^* \subset \mathcal{A}_g^*$

Luego con eso uno concluye que $\mu\left(\left\{x_0 : \lim_k g^k(x_0) \in M^*\right\}\right) = 0$ por lo que el conjunto de puntos iniciales tales que el algoritmo en cuestión converge a un punto silla estricto en 0. Como ya sabemos que el algoritmo no converge a máximos locales se concluye que:

$$w_k \xrightarrow[k \rightarrow \infty]{c.t.p.} w^*$$

Donde w^* es mínimo local.

Convergencia exponencial de GD

¿El descenso de gradiente inicializado aleatoriamente generalmente escapa de los puntos de silla en tiempo polinomial?

Definición

Dado $B \in \mathbb{R}^d$ decimos que $B \in \text{poly}(d)$ si existe $p \in \mathbb{R}[X]$ tal que $p(d) = B$. Asimismo decimos que una iteración de un algoritmo w_k esta a $\Omega(f(k))$ de w^* si existe $K \in \mathbb{N}$ tal que $|w^* - w_k| \geq Kf(k)$

Teorema

Consideremos el algoritmo descenso de gradiente por batch con w_0 elegido uniformemente en $[-1, 1]^d$; luego existe $F : \mathbb{R}^d \mapsto \mathbb{R}$ función objetivo B -acotada, l -Lipschitz, μ -Lipschitz en el Hessiano con $B, l, \mu \in \text{poly}(d)$ tal que si $\alpha_k = \alpha \leq \frac{1}{l}$ entonces w_k va a estar a $\Omega(1)$ de cualquier mínimo para todo $k \leq e^{\Omega(d)}$

Convergencia exponencial de GD

¿El descenso de gradiente inicializado aleatoriamente generalmente escapa de los puntos de silla en tiempo polinomial?

Definición

Dado $B \in \mathbb{R}^d$ decimos que $B \in \text{poly}(d)$ si existe $p \in \mathbb{R}[X]$ tal que $p(d) = B$. Asimismo decimos que una iteración de un algoritmo w_k esta a $\Omega(f(k))$ de w^* si existe $K \in \mathbb{N}$ tal que $|w^* - w_k| \geq Kf(k)$

Teorema

Consideremos el algoritmo descenso de gradiente por batch con w_0 elegido uniformemente en $[-1, 1]^d$; luego existe $F : \mathbb{R}^d \mapsto \mathbb{R}$ función objetivo B -acotada, l -Lipschitz, μ -Lipschitz en el Hessiano con $B, l, \mu \in \text{poly}(d)$ tal que si $\alpha_k = \alpha \leq \frac{1}{l}$ entonces w_k va a estar a $\Omega(1)$ de cualquier mínimo para todo $k \leq e^{\Omega(d)}$

Convergencia exponencial de GD

¿El descenso de gradiente inicializado aleatoriamente generalmente escapa de los puntos de silla en tiempo polinomial?

Definición

Dado $B \in \mathbb{R}^d$ decimos que $B \in \text{poly}(d)$ si existe $p \in \mathbb{R}[X]$ tal que $p(d) = B$. Asimismo decimos que una iteración de un algoritmo w_k esta a $\Omega(f(k))$ de w^* si existe $K \in \mathbb{N}$ tal que $|w^* - w_k| \geq Kf(k)$

Teorema

Consideremos el algoritmo descenso de gradiente por batch con w_0 elegido uniformemente en $[-1, 1]^d$; luego existe $F : \mathbb{R}^d \mapsto \mathbb{R}$ función objetivo B -acotada, l -Lipschitz, μ -Lipschitz en el Hessiano con $B, l, \mu \in \text{poly}(d)$ tal que si $\alpha_k = \alpha \leq \frac{1}{l}$ entonces w_k va a estar a $\Omega(1)$ de cualquier mínimo para todo $k \leq e^{\Omega(d)}$

Intuición acerca de la demostración: Parte 1

Escapar de dos puntos silla consecutivos

Sean $L > \gamma > 0$ y $f \in [0, 3] \times [0, 3]$ dada por:

$$f(x_1, x_2) = \begin{cases} -\gamma x_1^2 + Lx_2^2 & \text{si } (x_1, x_2) \in [0, 1] \times [0, 1] \\ L(x_1 - 2)^2 - \gamma x_2^2 & \text{si } (x_1, x_2) \in [1, 3] \times [0, 1] \\ L(x_1 - 2)^2 + L(x_2 - 2)^2 & \text{si } (x_1, x_2) \in [1, 3] \times [1, 3] \end{cases}$$

Notemos que f tiene dos puntos silla estrictos en $(0, 0)$ y $(2, 0)$, mientras que tiene un óptimo en $(2, 2)$.

Intuición acerca de la demostración: Parte 1

Escapar de dos puntos silla consecutivos

Sean $L > \gamma > 0$ y $f \in [0, 3] \times [0, 3]$ dada por:

$$f(x_1, x_2) = \begin{cases} -\gamma x_1^2 + Lx_2^2 & \text{si } (x_1, x_2) \in [0, 1] \times [0, 1] \\ L(x_1 - 2)^2 - \gamma x_2^2 & \text{si } (x_1, x_2) \in [1, 3] \times [0, 1] \\ L(x_1 - 2)^2 + L(x_2 - 2)^2 & \text{si } (x_1, x_2) \in [1, 3] \times [1, 3] \end{cases}$$

Notemos que f tiene dos puntos silla estrictos en $(0, 0)$ y $(2, 0)$, mientras que tiene un óptimo en $(2, 2)$.

Intuición acerca de la demostración: Parte 1

Escapar de dos puntos silla consecutivos

Sean $L > \gamma > 0$ y $f \in [0, 3] \times [0, 3]$ dada por:

$$f(x_1, x_2) = \begin{cases} -\gamma x_1^2 + Lx_2^2 & \text{si } (x_1, x_2) \in [0, 1] \times [0, 1] \\ L(x_1 - 2)^2 - \gamma x_2^2 & \text{si } (x_1, x_2) \in [1, 3] \times [0, 1] \\ L(x_1 - 2)^2 + L(x_2 - 2)^2 & \text{si } (x_1, x_2) \in [1, 3] \times [1, 3] \end{cases}$$

Notemos que f tiene dos puntos silla estrictos en $(0, 0)$ y $(2, 0)$, mientras que tiene un óptimo en $(2, 2)$.

Intuición acerca de la demostración: Parte 2

Sean $U = [0, 1]^2$, $V = [1, 3] \times [0, 1]$ y $W = [1, 3]^2$ entornos respectivos de los tres puntos críticos, supongamos que $w_0 = (x_1^0, x_2^0) \in U$ y definamos:

$$k_1 = \inf_{x_1^k \geq 1} k = \min_{x_1^k \geq 1} k$$

$$k_2 = \inf_{x_2^k \geq 1} k = \min_{x_2^k \geq 1} k$$

Notemos que como la dirección de escape en $(0, 0)$ es por x_1 y *luego* por x_2 (por el cambio de comportamiento de f) podemos concluir que k_1, k_2 están bien definidos y que $k_2 \geq k_1 \geq 0$.

La idea base es probar que $k_2 = Ck_1$ con $C > 1$.

Intuición acerca de la demostración: Parte 2

Sean $U = [0, 1]^2$, $V = [1, 3] \times [0, 1]$ y $W = [1, 3]^2$ entornos respectivos de los tres puntos críticos, supongamos que $w_0 = (x_1^0, x_2^0) \in U$ y definamos:

$$k_1 = \inf_{x_1^k \geq 1} k = \min_{x_1^k \geq 1} k$$

$$k_2 = \inf_{x_2^k \geq 1} k = \min_{x_2^k \geq 1} k$$

Notemos que como la dirección de escape en $(0, 0)$ es por x_1 y *luego* por x_2 (por el cambio de comportamiento de f) podemos concluir que k_1, k_2 están bien definidos y que $k_2 \geq k_1 \geq 0$.

La idea base es probar que $k_2 = Ck_1$ con $C > 1$.

Intuición acerca de la demostración: Conclusiones

Esta f que presentamos tiene varios problemas:

- ① No es continua ni mucho menos C^2
- ② No podemos asegurar que f sea l -Lipschitz o μ -Lipschitz en el hessiano
- ③ Los puntos críticos están en el borde del dominio, lo que no es ideal
- ④ No está definida en todo \mathbb{R}^d

La clave va a ser usar splines para resolver los primeros puntos, espejar f para hacer los puntos extremales interiores, asignar d puntos críticos similares para generar el tiempo exponencial en d y extender esa función \tilde{f} a \mathbb{R}^d con el Teorema de extensión de Whitney. Aunque la demostración es larga y tediosa, la idea clave es la vista aquí.

Intuición acerca de la demostración: Conclusiones

Esta f que presentamos tiene varios problemas:

- ① No es continua ni mucho menos C^2
- ② No podemos asegurar que f sea l -Lipschitz o μ -Lipschitz en el hessiano
- ③ Los puntos críticos están en el borde del dominio, lo que no es ideal
- ④ No está definida en todo \mathbb{R}^d

La clave va a ser usar splines para resolver los primeros puntos, espejar f para hacer los puntos extremales interiores, asignar d puntos críticos similares para generar el tiempo exponencial en d y extender esa función \tilde{f} a \mathbb{R}^d con el Teorema de extensión de Whitney. Aunque la demostración es larga y tediosa, la idea clave es la vista aquí.

Hoja de ruta

- 1 Introducción
- 2 Convergencia de algoritmos de tipo batch
 - Convergencia Puntual
 - Convergencia c.t.p.
 - Complejidad exponencial
- 3 Algoritmos estocásticos
 - Convergencia en L1
 - Convergencia en c.t.p.
- 4 Conclusiones y continuación
- 5 Referencias

Descenso estocástico

En el contexto estocástico vamos a analizar el algoritmo de descenso estocástico generalizado (DE) dado por:

Algorithm 1: Descenso Estocastico (DE)

```

1 Input:  $w_1 \in \mathbb{R}^d$  el inicio de la iteración,  $\{\xi_k\}$  iid
2 for  $k \in \mathbb{N}$  do
3   Generar una muestra de la variable aleatoria  $\xi_k$ 
4   Calcular el vector estocástico  $g(w_k, \xi_k)$ 
5   Elegir  $\alpha_k > 0$ 
6    $w_{k+1} \leftarrow w_k - \alpha_k g(w_k, \xi_k)$ 
  
```

Donde $g(w_k, \xi_k)$ puede ser varias estimaciones del gradiente como por ejemplo:

$$g(w_k, \xi_k) = \begin{cases} \nabla f(w_k, \xi_k) \\ \frac{1}{n_k} \sum_{i=1}^{n_k} \nabla f(w_k, \xi_{k,i}) \end{cases}$$

Lemas fundamentales

Definamos ahora $\mathbb{E}_{\xi_k} [\cdot] := \mathbb{E}_{P_k} [\cdot | w_k]$ la esperanza condicional bajo la distribución de ξ_k dado w_k .

Hipótesis (Acotaciones al primer y segundo momento de g)

Supongamos que dada F función objetivo y g la estimación del gradiente en DE vale:

- ① *Existe $U \subset \mathbb{R}^d$ tal que $\{w_k\} \subset U$ y que existe F_{inf} tal que $F|_U \geq F_{inf}$*
- ② *Existen $\mu_G \geq \mu \geq 0$ tal que para todo $k \in N$ valen:*

$$\nabla F(w_k)^T \mathbb{E}_{\xi_k} [g(w_k, \xi_k)] \geq \mu \|\nabla F(w_k)\|_2^2$$

Y

$$\|\mathbb{E}_{\xi_k} [g(w_k, \xi_k)]\|_2 \leq \mu_G \|\nabla F(w_k)\|_2$$

- ③ *Existen $M, M_V \geq 0$ tal que para todo $k \in \mathbb{N}$:*

$$\mathbb{V}_{\xi_k} [g(w_k, \xi_k)] \leq M + M_V \|\nabla F(w_k)\|_2^2$$

Lemas fundamentales

Notemos que si vale la hipótesis en los momentos de g entonces:

$$\mathbb{E}_{\xi_k} [\|g(w_k, \xi_k)\|_2^2] \leq M + M_G \|\nabla F(w_k)\|_2^2 \quad M_G := M_V + \mu_G^2 \geq \mu^2 \geq 0$$

Es más, si g es además un estimador insesgado de ∇F entonces

$$\mu = \mu_G = 1$$

Lema

Bajo la hipótesis en los momentos de g y si F es Lipschitz entonces las iteraciones del algoritmo DE satisfacen para todo $k \in \mathbb{N}$:

$$\mathbb{E}_{\xi_k} [F(w_{k+1})] - F(w_k) \leq - \left(\mu - \frac{1}{2} \alpha_k L M_G \right) \alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2} \alpha_k^2 L M$$

Lemas fundamentales

Notemos que si vale la hipótesis en los momentos de g entonces:

$$\mathbb{E}_{\xi_k} \left[\|g(w_k, \xi_k)\|_2^2 \right] \leq M + M_G \|\nabla F(w_k)\|_2^2 \quad M_G := M_V + \mu_G^2 \geq \mu^2 \geq 0$$

Es más, si g es además un estimador insesgado de ∇F entonces

$$\mu = \mu_G = 1$$

Lema

Bajo la hipótesis en los momentos de g y si F es Lipschitz entonces las iteraciones del algoritmo DE satisfacen para todo $k \in \mathbb{N}$:

$$\mathbb{E}_{\xi_k} [F(w_{k+1})] - F(w_k) \leq - \left(\mu - \frac{1}{2} \alpha_k L M_G \right) \alpha_k \|\nabla F(w_k)\|_2^2 + \frac{1}{2} \alpha_k^2 L M$$

Convergencia en L1

Teorema

Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^1$ la función objetivo tal que existe F_{inf} valor mínimo, F es L -Lipschitz, F es fuertemente convexa y supongamos además que g tiene varianza acotada; entonces el algoritmo descenso estocástico de gradiente generalizado con incremento fijo

$0 < \alpha_k = \alpha \leq \frac{\mu}{LM_G}$ cumple:

$$\mathbb{E}[F(w_k) - F_{inf}] \xrightarrow{k \rightarrow \infty} \frac{\alpha LM}{2c\mu}$$

Esquema de la demostración

Usando que F es Lipschitz y fuertemente convexa con la condición sobre α tenemos para todo $k \in \mathbb{N}$ vale:

$$\mathbb{E}_{\xi_k} [F(w_{k+1}) - F(w_k)] \leq -\alpha\mu c (F(w_k) - F_{inf}) + \frac{1}{2}\alpha^2 LM$$

Luego si restamos F_{inf} , tomamos esperanza total uno ve inductivamente que:

$$\mathbb{E} [F(w_{k+1}) - F_{inf}] - \frac{\alpha LM}{2c\mu} \leq (1 - \alpha c\mu)^k \left(F(w_1) - F_{inf} - \frac{\alpha LM}{2c\mu} \right)$$

Convergencia en L1

Teorema

Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^1$ la función objetivo tal que existe F_{\inf} valor mínimo, F es L -Lipschitz, F es fuertemente convexa; supongamos además que g tiene varianza acotada y que los incrementos α_k cumplen:

$$\alpha_k = \frac{\beta}{\gamma + k} \quad \text{para algún } \beta > \frac{1}{c\mu} \text{ y } \gamma > 0 \text{ tal que } \alpha_1 \leq \frac{\mu}{LM_G}$$

Luego el algoritmo descenso estocástico de gradiente generalizado cumple::

$$\mathbb{E}[R(w_k) - R^*] = \mathcal{O}\left(\frac{1}{k}\right)$$

Esquema de la demostración

Usando que F es Lipschitz y fuertemente convexa con la condición sobre α tenemos para todo $k \in \mathbb{N}$ vale:

$$\mathbb{E}_{\xi_k} [F(w_{k+1})] - F(w_k) \leq -\alpha_k c \mu (F(w_k) - F(w_*)) + \frac{1}{2} \alpha_k^2 LM$$

Luego restando F_{inf} , tomando esperanza y reordenando vale inductivamente que:

$$\mathbb{E} [F(w_{k+1}) - F_{inf}] \leq \frac{\eta}{\gamma + k + 1}$$

Convergencia en L1

Teorema

Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^1$ la función objetivo tal que existe F_{\inf} valor mínimo, F es L -Lipschitz, F es fuertemente convexa; supongamos además que existe $M \geq 0$ y $\xi \in (0, 1)$ tal que para todo $k \in \mathbb{N}$:

$$\mathbb{V}_{\xi_k} [g(w_k, \xi_k)] \leq M\xi^{k-1}$$

Luego el algoritmo descenso estocástico de gradiente generalizado con incremento fijo $0 < \alpha_k = \alpha \leq \min \left\{ \frac{\mu}{L\mu_G^2}, \frac{1}{\mu} \right\}$ cumple:

$$\mathbb{E} [R(w_k) - R^*] = \mathcal{O}(\rho^k)$$

Esquema de la demostración

Por el lema, la condición de α y las hipótesis sobre F y g vale que:

$$\mathbb{E}[F(w_{k+1}) - F_{inf}] \leq (1 - c\mu\alpha) \mathbb{E}[F(w_k) - F_{inf}] + \frac{1}{2}\alpha^2 LM\xi^{k-1}$$

E inductivamente se prueba que:

$$\mathbb{E}[F(w_{k+1}) - F_{inf}] \leq \omega\rho^k$$

Convergencia en L1

Teorema

Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^2$ la función objetivo, F es L -Lipschitz y $w \mapsto \|\nabla F(w)\|_2^2$ sea L -Lipschitz; supongamos además que g tiene varianza acotada y que los incrementos cumplen la condición de Robbins Monro. Luego el algoritmo descenso estocástico de gradiente generalizado con incrementos decrecientes cumple:

$$\lim_{k \rightarrow \infty} \mathbb{E} \left[\|\nabla F(w_k)\|^2 \right] = 0$$

Convergencia c.t.p.

Teorema

Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ tal que $F \in C^1$ la función objetivo tal que existe F_{\inf} valor mínimo, F es débilmente convexo; supongamos además que g tiene varianza acotada y que los incrementos cumplen la condición de Robbins Monro. Luego el algoritmo descenso estocástico de gradiente generalizado con incremento decrecientes cumple:

$$w_k \xrightarrow[k \rightarrow \infty]{\text{c.t.p.}} w^*$$

$$(w_k - w^*)^T \nabla F(w_k) \xrightarrow[k \rightarrow \infty]{\text{c.t.p.}} 0$$

Convergencia c.t.p.: Esquema de la Demostración

Sea $h_k := \|w_k - w^*\|^2$ y definamos la filtración

$\mathcal{P}_k = \{\xi_0, \dots, \xi_{k-1}, w_0, \dots, w_k, \alpha_0, \dots, \alpha_k\}$. Incorporando las hipótesis se puede ver que:

$$\mathbb{E} [h_{k+1} - (1 + \alpha_k^2 B) h_k | \mathcal{P}_k] \leq -2\alpha_k (w_k - w^*)^T \nabla F(w_k) + \alpha_k^2 A$$

Luego, definiendo μ_k como antes uno llega a:

$$\alpha_k (w_k - w^*)^T \nabla F(w_k) \leq \mathbb{E} [\delta_k^{\mu_h} (\mu_{k+1} h_{k+1} - \mu_k h_k)] \leq \alpha_k^2 \mu_k A$$

Usando el teorema de convergencia de quasimartingalas uno ve que h_k converge, con lo que concluimos que:

$$\sum_{k=1}^{\infty} \alpha_k (w_k - w^*)^T \nabla F(w_k) < \infty \quad \text{ctp}$$

Y de aquí se concluye el resultado.

Convergencia c.t.p. : Hipótesis

Hipótesis

Sea $F : \mathbb{R}^d \rightarrow \mathbb{R}$ función objetivo y g un estimador insesgado de ∇F tal que ambos cumplen:

- ① $F \in C^3$
- ② Existe $w^* \in \mathbb{R}^d$ tal que $F_{\inf} = F(w^*) \leq F(w)$ para todo $w \in U$ entorno.
- ③ $F(w) \geq 0$ para todo $w \in \mathbb{R}^d$
- ④ Los incrementos cumplen la condición de Robbins Monro
- ⑤ Existe $B_2 \geq 0$ tal que:

$$\mathbb{E} \left[\|g(w_k, \xi_k)\|^2 \right] \leq B_2 \|w\|^2$$

Convergencia c.t.p. : Hipótesis continuación

Hipótesis

⑥ Para $j = 3, 4$ existen $A_j, B_j \geq 0$ tal que:

$$\mathbb{E} \left[\|g(w_k, \xi_k)\|^j \right] \leq A_j + B_j \|w\|^j$$

⑦ Existe $D > 0$ tal que:

$$\inf_{(w)^2 > D} w^T \nabla F(w) > 0$$

Convergencia c.t.p. : Teorema

Teorema

Bajo todas las hipótesis mencionadas anteriormente, el algoritmo descenso estocástico de gradiente generalizado con incrementos decrecientes cumple:

$$F(w_k) \xrightarrow[k \rightarrow \infty]{c.t.p.} F_\infty$$

$$\nabla F(w_k) \xrightarrow[k \rightarrow \infty]{c.t.p.} 0$$

Hoja de ruta

- 1 Introducción
- 2 Convergencia de algoritmos de tipo batch
 - Convergencia Puntual
 - Convergencia c.t.p.
 - Complejidad exponencial
- 3 Algoritmos estocásticos
 - Convergencia en L1
 - Convergencia en c.t.p.
- 4 Conclusiones y continuación
- 5 Referencias

Algunas conclusiones

- ① Los métodos determinísticos sufren la lentitud natural de tener un solo *epoch* por set de datos.
- ② El descenso de gradiente tiene complejidad exponencial para alcanzar el mínimo.
- ③ Los métodos estocásticos bajo condiciones de convexidad muy leves convergen rápidamente en esperanza y *casi todo punto*.
- ④ Esa convergencia es independiente del número de datos, lo que los habilita para el régimen del *big data*.
- ⑤ Los algoritmos determinísticos tradicionales están perdiendo preferencia por sus propiedades teóricas desfavorables.

Posibles líneas de investigación

- Analizar qué tan restrictivas son las condiciones que pedimos en esta Tesis a la función objetivo, así como algoritmos que permitan la validación de las hipótesis con la menor cantidad de derivadas a calcular posibles.
- Profundizar el estudio de la complejidad y convergencia a mínimos en los algoritmos mixtos que surgen de intentar juntar características de los algoritmos más usuales.
- Estudiar si SG, u otros DE, cumplen las propiedades vistas en **Jin et. al. 2017** y **Zhang et. al. 2017** ; es de particular interés un algoritmo que cumpla ambas y además no sea costoso computacionalmente o admita un cálculo distribuido.

Hoja de ruta

- 1 Introducción
- 2 Convergencia de algoritmos de tipo batch
 - Convergencia Puntual
 - Convergencia c.t.p.
 - Complejidad exponencial
- 3 Algoritmos estocásticos
 - Convergencia en L1
 - Convergencia en c.t.p.
- 4 Conclusiones y continuación
- 5 Referencias

Referencias I



Lee et al. *Gradient descent only converges to minimizers*. Conference on learning theory, 2016, págs. 1246-1257.



Leon Bottou. "Online Learning and Stochastic Approximations". En: (mayo de 1999).



Leon Bottou, Frank E. Curtis y Jorge Nocedal. "Optimization Methods for Large-Scale Machine Learning". En: 60 (jun. de 2016).



Wassily Hoeffding. *PROBABILITY INEQUALITIES FOR SUMS OF BOUNDED RANDOM VARIABLES*. 1962.



G. Piliouras M. Simchowitz M.I. Jordan "J. Lee I. Panageas y B. Recht". "First-order methods almost always avoid saddle points". En: (2017).



Chi Jin y col. "How to Escape Saddle Points Efficiently". En: *ICML*. 2017.

Referencias II



S. Łojasiewicz. *A topological property of real analytic subsets*. Coll. du CNRS, Les equations aux derivees partielles, 1963, págs. 87-89.



B. T. Polyak. *Gradient methods for minimizing functionals*. Zh. Vychisl. Mat. Fiz, 1963, págs. 643-653.



H. Robbins y S. Monro. *A Stochastic Approximation Model*. Vol. 22(3). The Annal of Mathematical Statistics, 1951, págs. 400-407.



Yuchen Zhang, Percy Liang y Moses Charikar. "A Hitting Time Analysis of Stochastic Gradient Langevin Dynamics". En: *COLT*. 2017.

Gracias por la atención!