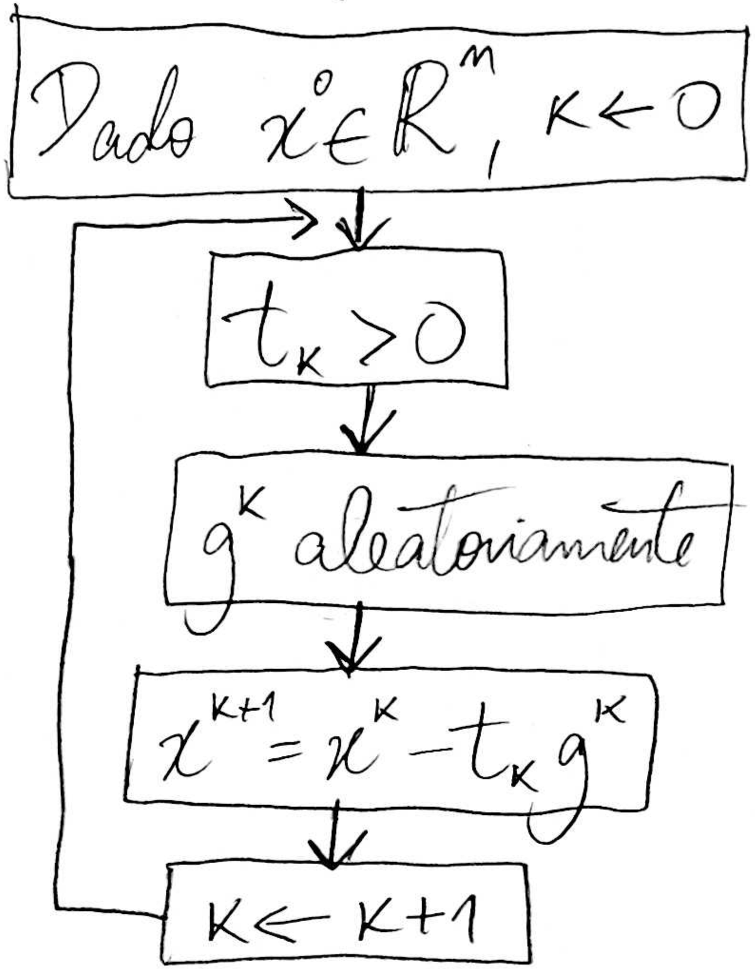
Convergência do método do gradiente estocástico (1



Sequência gerada: $\{x^k\}$.

Problema: $\min_x f(x)$

HIPÓTESES:

H1) $E(g^k | x^k) = \nabla f(x^k)$

(sem viés)

H2) $\exists L > 0$ tal que

$$E(\|g^k\|^2 | x^k) \leq L^2, \ \forall k.$$

<u>Passo decrescente:</u>

$$(1) \quad t_k \to 0^+ , \quad \sum_{k=0}^{\infty} t_k = \infty , \quad \sum_{k=0}^{\infty} t_k^2 < \infty .$$

<u>Teorema:</u> Suponha f convexa, H1, H2 válidas e que f admita minimizador x^*. Seja $\{t_k\}$ como em (1). Então

$$\lim_{k\to\infty} E(f_k) = f^* ,$$

onde $f_k = \min_{0\le j\le k} f(x^j)$ e $f^* = \min_x f(x) = f(x^*)$.

Prova: Para cada $k \geq 0$ temos

$$E(\|x^{k+1}-x^*\|^2 \mid x^k) = E(\|x^k - t_k g^k - x^*\|^2 \mid x^k)$$

$$= E(\|x^k - x^*\|^2 - 2t_k (g^k)^t (x^k - x^*) + t_k^2 \|g^k\|^2 \mid x^k)$$

$$= E(\|x^k - x^*\|^2 \mid x^k) - 2t_k E((g^k)^t (x^k - x^*))$$

$$+ t_k^2 E(\|g^k\|^2 \mid x^k)$$

Note que $E(\|x^k - x^*\|^2 \mid x^k) = \|x^k - x^*\|^2$ visto que, dado x^k, $\|x^k - x^*\|^2$ fica determinado.

4

Também, $E((g^k)^t(x^k - x^*)|x^k) = E(\sum g_i^k(x_i^k - x_i^*)|x^k)$

$$= \sum E(g_i^k \underbrace{(x_i^k - x_i^*)}_{\text{cte em } x^k} | x^k) = \sum E(g_i^k | x^k)(x_i^k - x_i^*)$$

$$= E(g^k | x^k)^t (x^k - x^*).$$

Assim,

$$E(\|x^{k+1} - x^*\|^2 | x^k) = \|x^k - x^*\|^2 - 2t_k E(g^k|x^k)^t (x^k - x^*)$$
$$+ t_k^2 E(\|g^k\|^2 | x^k)$$

Pelas hipóteses H1 e H2, obtemos

$$E(\|x^{k+1}-x^*\|^2 \mid x^k) \leq \|x^k-x^*\|^2 - 2t_k\left[\nabla f(x^k)^t(x^k-x^*)\right] + t_k^2L^2,$$

$$\underset{f \text{ convexa}}{\leq} \|x^k-x^*\|^2 - 2t_k(f(x^k)-f^*) + t_k^2L^2.$$

Tomando a esperança em ambos os lados da desigualdade e lembrando que $E(E(W|Z)) = E(W)$ e que $W \leq Z \Rightarrow E(W) \leq E(Z)$, obtemos

$$E(\|x^{k+1}-x^*\|^2) \leq E(\|x^k-x^*\|^2) - 2t_k(E(f(x^k)) -$$

$$-f^*) + t_k^2 L^2.$$

Aplicando essa desigualdade sucessivas vezes,

$$E(\|x^{k+1}-x^*\|^2) \leq E(\|x^0-x^*\|^2) - 2\sum_{j=0}^{k} t_j (E(f(x^j))$$

$$-f^*) + L^2 \sum_{j=0}^{k} t_j^2$$

$$\Rightarrow \sum_{j=0}^{k} t_j (E(f(x^j)) - f^*) \leq \frac{1}{2}\left[E(\|x^0-x^*\|^2) + L^2 \sum_{j=0}^{k} t_j^2\right].$$

Agora, usando o fato $E(\min\{W, Z\}) \leq \min\{E(W), E(Z)\}$, temos

$$E(f_k) = E\left(\min_{0 \leq j \leq k} f(x^j)\right) \leq \min_{0 \leq j \leq k} E(f(x^j)),$$

e logo

$$0 \leq E(f_k) - f^* \leq \frac{\|x^0 - x^*\|^2 + L^2 \sum_{j=0}^{k} t_j^2}{2 \sum_{j=0}^{k} t_j}$$

(note que $E(\|x^0 - x^*\|^2) = \|x^0 - x^*\|^2$ pois x^0 e

x^* são determinísticos). Fazendo $k \to \infty$ obtemos $\lim_{k\to\infty} E(f_k) = f^*$ das hipóteses sobre $\{t_k\}$. ∎

<u>Obs:</u> compare com a prova de convergência do método do gradiente incremental.

<u>Passo constante $t_k = t$, $\forall k$:</u> assim como no gradiente incremental, é possível provar convergência do método do gradiente

estocástico para funções convexas e passo constante (o SG "básico" usa $t_k = cte$ na prática).

Vamos nos inspirar nas provas do gradiente estocástico para passo decrescente e do subgradiente para passo constante.

Seguindo a prova anterior, note que

$$E(\|x^{k+1} - x^*\|^2) \leq E(\|x^k - x^*\|^2) - 2t_k\left(E(f(x^k)) - f^*\right) + t_k^2 L^2 \qquad (2)$$

permanece válido independentemente da escolha de t_k. Essa desigualdade é uma versão do lema do método do subgradiente com esperanças (compare!).

Seguindo a prova de convergência do subgradiente para o caso $f^* > -\infty$ (estamos supondo aqui que f admite minimizadores), suponha que

$$\liminf_{k\to\infty} E(f(x^k)) > f^* + \frac{tL^2}{2} + 2\varepsilon \qquad (3),$$

para algum $\varepsilon > 0$. Aqui, $t_k = t > 0$, $\forall k$.

Também, para todo $k \gg 1$ temos

$$E(f(x^k)) \geq \liminf_{k\to\infty} E(f(x^k)) - \varepsilon. \quad (4)$$

Somando (3) e (4) obtemos, para um certo k_0,

$$E(f(x^k)) - f^* \geq \frac{tL^2}{2} + \varepsilon, \quad \forall k \geq k_0.$$

Assim, da inequação (2) com $t_k = t$ temos

12/

$$E(\|x^{k+1} - x^*\|^2) \leq E(\|x^k - x^*\|^2) - 2t\left(\frac{tL^2}{2} + \varepsilon\right) + t^2 L^2$$

$$= E(\|x^k - x^*\|^2) - 2t\varepsilon \quad , \forall k \geq k_0 .$$

Aplicando essa desigualdade sucessivas vezes obtemos

$$E(\|x^{k+1} - x^*\|^2) \leq E(\|x^{k_0} - x^*\|^2) - 2(k+1-k_0)t\varepsilon .$$

Tomando $k \gg k_0$ obtemos uma contradição pois $-2(k+1+k_0)t\varepsilon \to -\infty$ e $E(W) \geq 0$ quando $W \geq 0$.

a contradição da suposição (3), isto é, que $\exists \varepsilon > 0$ tal que

$$\liminf_{k \to \infty} E(f(x^k)) > f^* + \frac{tL^2}{2} + 2\varepsilon.$$

Em outras palavras, vale

$$E(f_k) - f^* \leq \frac{tL^2}{2}.$$

Compare com o resultado do método do subgradiente: só há garantia de chegar próximo do valor ótimo !!!

Isso demonstra o seguinte teorema:

<u>Teorema:</u> Suponha f convexa, H1, H2 válidas e que f admita minimizador x^*. Suponha ainda que $t_k = t > 0$, $\forall k$. Então

$$\lim_{k \to \infty} E(f_k) \leq f^* + \frac{tL^2}{2},$$

onde $f_k = \min_{0 \leq j \leq k} f(x^k)$ e $f^* = \min_x f(x) = f(x^*)$.