UiO : Matematisk institutt

Det matematisk-naturvitenskapelige fakultet

STK-4051/9051 Computational Statistics Spring 2021
SGD

Instructor: Odd Kolbjørnsen, oddkol@math.uio.no

Stochastic gradient decent

- Existed for many years (Robbins and Monro, 1951, reprinted 1985)
- Received renewed attention due to its importance in fitting deep neural networks.
- A thourough discussion of the algorithm is given in Bottou et al. (2018) while a broader discussion on stochastic optimization methods in general is given in Spall (2005).
- Aim: minimize some $F(\theta)$ with respect to θ.
- Empirical risk:

$$F(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta) + J(\theta).$$

with many possible options for $f_i(\theta)$, e.g.

$$f_i(\theta) = \begin{cases} (\hat{y}_i - y_i)^2 & \text{Least squares;} \\ I(\hat{y}_i \neq y_i) & \text{Classification error;} \\ -\log f(y_i;\theta) & \text{log-likelihood.} \end{cases}$$

- Alternative: Expected risk

$$F(\theta) = E[f(\theta;\boldsymbol{\varepsilon})], \quad \boldsymbol{\varepsilon} \text{ is some random vector}$$

Main Idea

- $F(\cdot)$ is nice and smooth, a necessary requirement is

$$\boldsymbol{g}(\theta^*) = \frac{\partial}{\partial \boldsymbol{\theta}} F(\theta) \,|_{\theta=\theta^*} = \mathbf{0} \tag{1}$$

- Ordinary gradient descent methods:

$$\theta^{t+1} = \theta^t - \boldsymbol{M}_t^{-1} \boldsymbol{g}(\theta^t), \quad \boldsymbol{M}_t \text{ is some positive definite matrix}$$

- Main problem: gradient might be difficult to compute.
- The stochastic gradient algorithm replaces the gradient by an estimate instead:

$$\theta^{t+1} = \theta^t - \alpha_t \boldsymbol{M}_t^{-1} \boldsymbol{Z}(\theta^t; \boldsymbol{\phi}^t), \quad \boldsymbol{Z}(\theta^t; \boldsymbol{\phi}^t) \approx \boldsymbol{g}(\theta^t) \tag{2}$$

«some stochastic element»

- A class of possibilities are given by

$$\boldsymbol{Z}(\theta^t; \boldsymbol{\phi}^t) = \frac{1}{n_t} \sum_{i \in \mathcal{S}_t} \nabla f_i(\theta^t), \quad \mathcal{S}_t \subset \{1, \dots, n\},\ n_t = |\mathcal{S}_t|$$

"$\phi^t = \mathcal{S}_t$"

- Algorithm:

1: **for** $t = 1, 2, \dots$ **do**
2: Simulate the stochastic gradient $Z(\theta^t; \boldsymbol{\phi}^t)$;
3: Choose a stepsize α^t;
4: Update the new value by $\theta^{t+1} \leftarrow \theta^t - \alpha_t \boldsymbol{M}_t^{-1} Z(\theta^t; \boldsymbol{\phi}^t)$.
5: **end for**

Example

- Logistic regression with n large:

$$Y_i \sim \text{Binomial}(1, p(x_i)), \quad i = 1, ..., n$$

$$p(x) = \frac{\exp(\theta_0 + \theta_1 x)}{1 + \exp(\theta_0 + \theta_1 x)}$$

- Want to minimize

$$F(\boldsymbol{\theta}) = -\sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

$$= -\sum_{i=1}^{n} [y_i(\theta_0 + \theta_1 x_i) - \log(1 + \exp(\theta_0 + \theta_1 x_i))].$$

- Defining

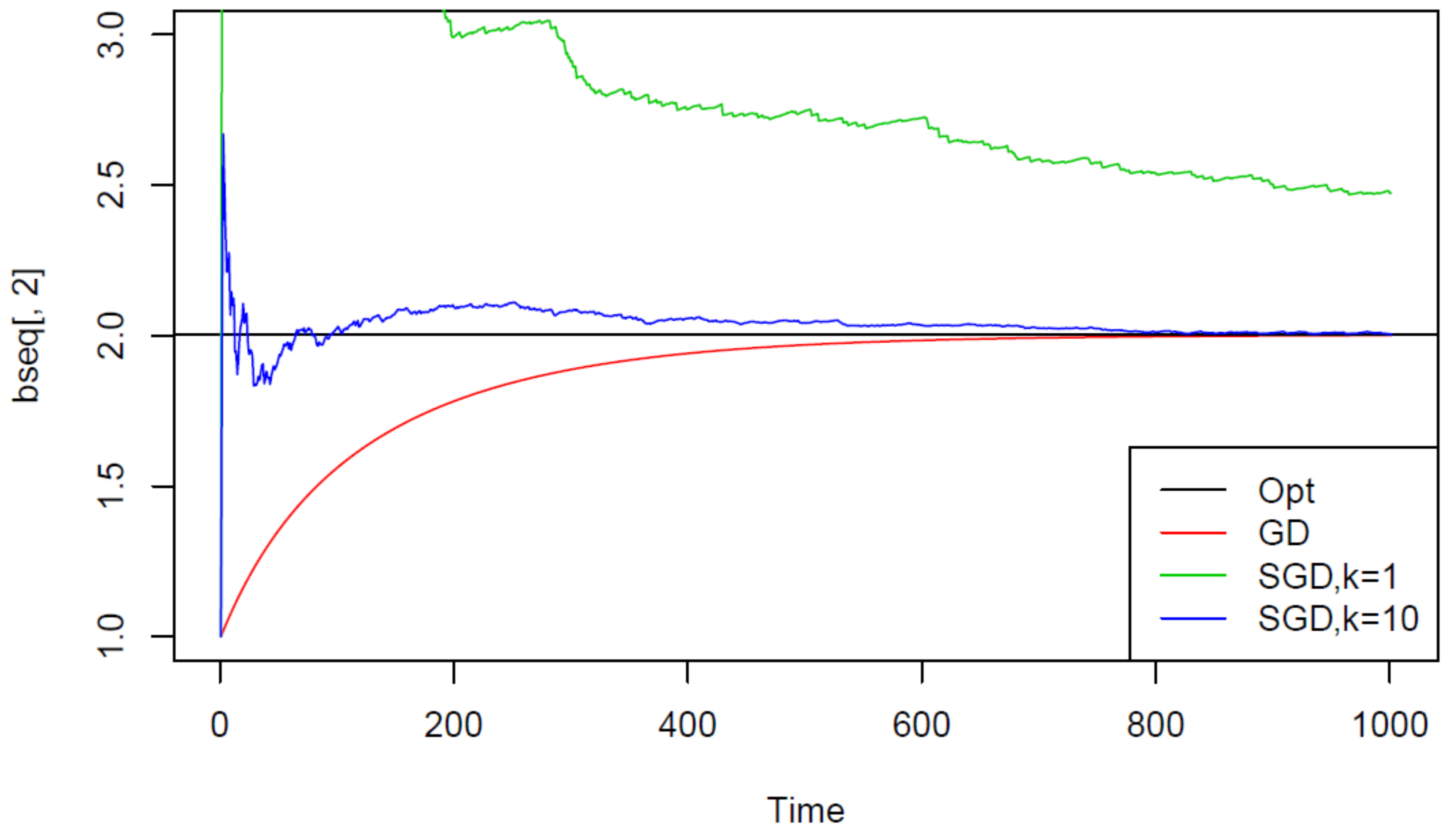
$$f_i(\boldsymbol{\theta}) = -y_i(\theta_0 + \theta_1 x_i) + \log(1 + \exp(\theta_0 + \theta_1 x_i))$$

we have

$$\nabla f_i(\boldsymbol{\theta}) = -\begin{pmatrix} y_i - \frac{\exp(\theta_0+\theta_1 x_i)}{1+\exp(\theta_0+\theta_1 x_i)} \\ [y_i - \frac{\exp(\theta_0+\theta_1 x_i)}{1+\exp(\theta_0+\theta_1 x_i)}] x_i \end{pmatrix}$$

```
#Initialization
b = c(0,1)      #Initial value
N.it = 1000     #Number of iterations
k = 10          #Number of samples for estimating gradient
#SG-loop
for(it in 1:N.it)
{
  i = sample(1:n,k)
  alpha = 10/it
  p.i = exp(b[1]+b[2]*x[i])/(1+exp(b[1]+b[2]*x[i]))
  g = colMeans(cbind(y[i]-p.i,(y[i]-p.i)*x[i]))
  b = b + alpha*g
}
```

Convergence in example



Convergence of SGD

- Want to show that the SGD procedure is consistent

Definition 1.

If $\lim_{t\to\infty} \theta^t = \theta^*$ in probability, irrespective of any arbitrary initial value θ^0, we call the procedure consistent. Here, convergence in probability means that for any $\varepsilon > 0$

$$\lim_{t\to\infty} \Pr(|\theta^t - \theta^*| > \varepsilon) = 0.$$

- Do this in three steps (with some sub-steps on the way)
  1. Prove that L2 convergence gives consistency
  2. Prove that the sequence converge
  3. Prove that we converge to the true parameter

Step 1 L2 convergence gives consistency

Lemma 1.

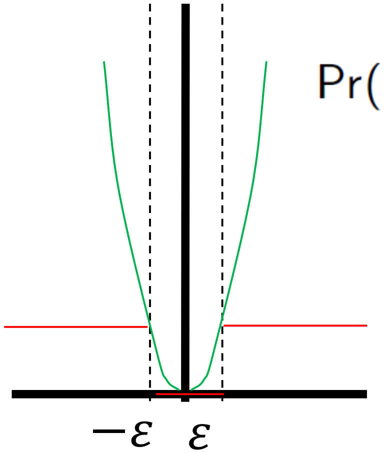
Define

$$b_t = E[||\theta^t - \theta^*||^2].$$

If $\lim_{t\to\infty} b_t = 0$, *then* $\{\theta^t\}$ *is consistent.*

- $\{\theta^t\}$ is stochastic and multidimensional
- $\{b_t\}$ is deterministic and one-dimensional
- Easier to prove convergence with respect to $\{b_t\}$

Defining $p_t(\cdot)$ to be the density of θ^t, we have that



$$\begin{aligned}\Pr(|\theta^t - \theta^*| > \varepsilon) =& \int_z I[(z-\theta^*)^2 > \varepsilon^2] p_t(z)dz \\ \le& \int_z \frac{(z-\theta^*)^2}{\varepsilon^2} p_t(z)dz \\ =& \frac{1}{\varepsilon^2}\int_z (z-\theta^*)^2 p_t(z)dz = \frac{1}{\varepsilon^2} b_t \to 0\end{aligned}$$

Assumptions

- Requirements on the sequence $\{\alpha_t\}$:

$$\alpha_t > 0 \qquad \text{(A-1)}$$

$$\sum_{t=2}^{\infty} \frac{\alpha_t}{\alpha_1 + \cdots + \alpha_{t-1}} = \infty \qquad \text{(A-2)}$$

$$\sum_{t=1}^{\infty} \alpha_t^2 < \infty \qquad \text{(A-3)}$$

Note that (A-2) implies $\sum_{t=1}^{\infty} \alpha_t = \infty$

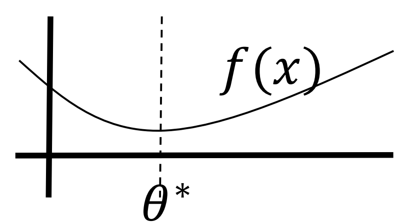
- Requirements on the function $g(x)$ combined with its estimate:

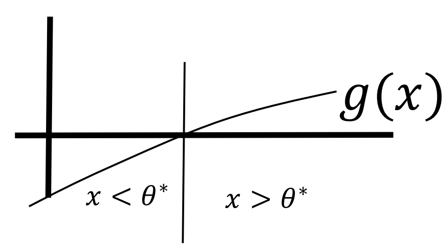
$g(x)$ has same sign as $(x - \theta^*)$

$$\exists \delta \geq 0 \text{ such that } g(x) \leq -\delta \text{ for } x < \theta^* \text{ and } g(x) \geq \delta \text{ for } x > \theta^*. \qquad \text{(A-4)}$$

$$E[Z(\theta;\phi)] = g(\theta) \text{ and } \Pr(|Z(\theta;\phi)| < C) = 1 \qquad \text{(A-5)}$$

The constraint $|Z(\theta;\phi)| < C$ is included to simplify the proof. More general results are available.





Step 2 Prove that the sequence converge

Theorem 1.

Assume (A-1), (A-3), (A-4) *and* (A-5). *Then the sequence*

$$\theta^{t+1} = \theta^t - \alpha_t Z(\theta^t; \phi^t) \tag{3}$$

will converge in probability.

- This result only gives convergence to some value, not necessarily to the optimal value.
- Convergence to the optimal value will be proved later were also (A-2) will be assumed.
- Simplify the notation: Denoting $Z(\theta^t; \phi^t)$ by Z_t.

Recall: Z is the stochastic version of the gradient

$$\boldsymbol{Z}(\boldsymbol{\theta}^t; \boldsymbol{\phi}^t) \approx \boldsymbol{g}(\boldsymbol{\theta}^t)$$

Proof of Theorem 1

$$
\begin{aligned}
b_{t+1} &= E[(\theta^{t+1} - \theta^*)^2] = E[E[(\theta^{t+1} - \theta^*)^2|\theta^t]] = E[E[(\theta^t - \alpha_t Z_t - \theta^*)^2|\theta^t]] \\
&= E[(\theta^t - \theta^*)^2 + \alpha_t^2 E[Z_t^2|\theta^t] - 2\alpha_t(\theta^t - \theta^*)E[Z_t|\theta^t]] \\
&= b_t + \alpha_t^2 E[Z_t^2] - 2\alpha_t E[(\theta^t - \theta^*)g(\theta^t)]
\end{aligned}
$$

$$
e_t = E[Z_t^2] \quad d_t = E[(\theta^t - \theta^*)g(\theta^t)],
$$

we get

$$
b_{t+1} - b_t = \alpha_t^2 e_t - 2\alpha_t d_t.
$$

- By summing the equation above over t, we get

$$
b_{t+1} = b_1 + \sum_{s=1}^{t} \alpha_s^2 e_s - 2\sum_{s=1}^{t} \alpha_s d_s. \qquad (4)
$$

First series has only positive terms:
Since $e_t = E\{Z_t^2\} > 0$,

Second series has only positive terms:
Since by (A-4) : $g(x)$ has same sign as $(x - \theta^*)$, $d_t \geq 0$
Since by (A-1): $\alpha_t > 0$, we then have also $\alpha_t d_t \geq 0$

If we can show that both $\sum_{s=1}^{t} \alpha_s^2 e_s$ and $\sum_{s=1}^{t} \alpha_s d_s$ are bounded, then both series converge by monotone convergence. And thereby also b_t converge

Bounding the two series

$$b_{t+1} = b_1 + \sum_{s=1}^{t} \alpha_s^2 e_s - 2 \sum_{s=1}^{t} \alpha_s d_s.$$

From $|Z(\theta;\phi)| \leq C$ we have

$$\sum_{t=1}^{\infty} \alpha_t^2 e_t \leq C^2 \sum_{t=1}^{\infty} \alpha_t^2 < \infty$$

(A-5): Since $|Z_t| < C$, $\mathrm{e}_t = \mathrm{E}\{|Z_t|^2\} < C^2$

(A-3): $\sum {\alpha_t}^2 < \infty$

$$\sum_{s=t+1}^{\infty} \alpha_s^2 \, \mathrm{e}_s \geq 0$$

$$\sum_{s=1}^{t} \alpha_s d_s = \tfrac{1}{2}\left[b_1 + \sum_{s=1}^{t} \alpha_s^2 e_s - b_{t+1}\right] \leq \tfrac{1}{2}\left[b_1 + \sum_{s=1}^{\infty} \alpha_s^2 e_s\right]$$

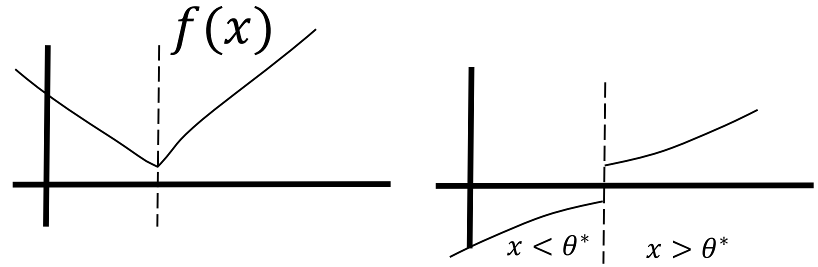
Add two Non-negative finite numbers

$$b_{t+1} = E[(\theta^{t+1} - \theta^*)^2] \geq 0$$

Thus if we remove it we reduce the sum

Both series are bounded and therefore converge

Two main results



Theorem 2.

Assume (A-1), (A-2), (A-3), (A-4) *and* (A-5). *Assume further* $\delta > 0$ *in* (A-4). *Then* $\lim_{t\to\infty} b_t = 0$.

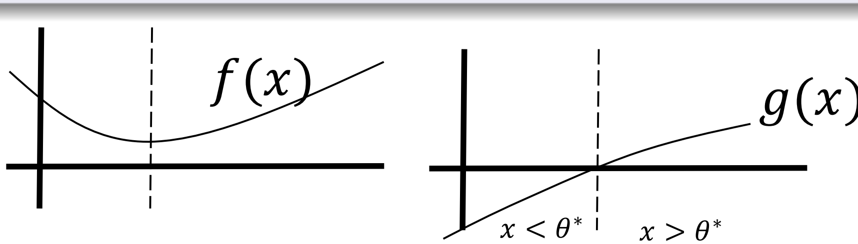
$$\exists\delta \geq 0 \text{ such that } g(x) \leq -\delta \text{ for } x < \theta \text{ and } g(x) \geq \delta \text{ for } x > \theta. \qquad \text{(A-4)}$$

Theorem 3.

Assume (A-1), (A-2), (A-3) *and* (A-5). *Assume further*

$$g(z) \text{ is nondecreasing;} \tag{9}$$
$$g(\theta^*) = 0; \tag{10}$$
$$g'(\theta^*) > 0. \tag{11}$$

Then $\lim_{t\to\infty} b_t = 0$.



Warm up to Theorems

Lemma 2.

Assume (A-1), (A-3), (A-4) *and* (A-5). *Assume* $\{k_t\}$ *is a sequence of nonnegative constants satisfying*

$$k_t b_t \leq d_t, \quad \sum_{t=1}^{\infty} \alpha_t k_t = \infty \tag{5}$$

Then $\lim_{t\to\infty} b_t = 0$.

So if we can find such a k_t-sequence we are done

Proof:

- We have that

$$\sum_{s=1}^{t} \alpha_s d_s = \tfrac{1}{2}\left[b_1 + \sum_{s=1}^{t} \alpha_s^2 e_s - b_{t+1}\right] \leq \tfrac{1}{2}\left[b_1 + \sum_{s=1}^{\infty} \alpha_s^2 e_s\right]$$

$$\sum_{t=1}^{\infty} \alpha_t k_t b_t \leq \sum_{t=1}^{\infty} \alpha_t d_t < \infty \tag{6}$$

  from the proof of the previous Theorem.
- From the second part of (5) there must be an infinite number of b_t's for which $b_t < \epsilon$ for any value of ϵ.
- Since we have already shown that $\lim_{t\to\infty} b_t$ exists, this shows that the limit has to be zero.

Warm up to Theorems cont…

Lemma 3.

Assume (A-1), (A-2), (A-3), (A-4) *and* (A-5). *Assume for some constant* $\delta > 0$ *that*

$$\inf_{z \in [\theta^* - A_t, \theta^* + A_t]} \left[\frac{g(z)}{z - \theta^*}\right] \geq \frac{\delta}{A_t} \text{ for } t > N \tag{7}$$

where

$$A_t = |\theta^1 - \theta^*| + C(\alpha_1 + \cdots + \alpha_{t-1}). \tag{8}$$

Then $\lim_{t\to\infty} b_t = 0$.

This δ need not be the one in (A-4)

- We have that $\theta^t = \theta^1 - \sum_{s=1}^{t-1} \alpha_s Z_s$ so that

$$\begin{aligned} |\theta^t - \theta^*| &= |\theta^1 - \theta^* - \sum_{s=1}^{t-1} \alpha_s Z_s| \\ &\leq |\theta^1 - \theta^*| + \sum_{s=1}^{t-1} \alpha_s |Z_s| \leq |\theta^1 - \theta^*| + \sum_{s=1}^{t-1} \alpha_s C = A_t \end{aligned}$$

  where the second inequality is with probability 1.
- Define

$$k_t = \inf_{x \in [\theta^* - A_n, \theta^* + A_n]} \left[\frac{g(x)}{x - \theta^*}\right] \geq 0 \quad \text{from (A-4)}$$

If we can show:
1. $k_t b_t \leq d_t$
2. $\sum_{t=1}^{\infty} \alpha_t k_t = \infty$

We can use lemma 2

Proof $k_t b_t \le d_t$

$$k_t = \inf_{x \in [\theta^* - A_n, \theta^* + A_n]} \left[\frac{g(x)}{x - \theta^*} \right]$$

$$A_t = |\theta^1 - \theta^*| + C(\alpha_1 + \cdots + \alpha_{t-1})$$

- Define $p_t(\cdot)$ to be the density for θ^t:

$$k_t b_t = k_t E[(\theta^t - \theta^*)^2] = \int_z k_t (z - \theta^*)^2 p_t(z) dz$$

$$= \int_{|z-\theta^*| \le A_t} k_t (z - \theta)^2 p_t(z) dz \le \int_{|z-\theta^*| \le A_t} \frac{g(z)}{z-\theta^*} (z - \theta^*)^2 p_t(z) dz$$

$$= \int_{|z-\theta^*| \le A_t} g(z)(z - \theta^*) p_t(z) dz = E[g(\theta^t)(\theta^t - \theta^*)] = d_t$$

By the construction of A_t The density p_t is supported on this interval

Proof $\sum_{t=1}^{\infty} \alpha_t k_t = \infty$

$$A_t = |\theta^1 - \theta^*| + C(\alpha_1 + \cdots + \alpha_{t-1})$$

- By (A-2), $\sum_{t=1}^{\infty} \alpha_t = \infty$ which implies that for t larger than some T

$$2C(\alpha_1 + \cdots + \alpha_{t-1}) = A_t + C(\alpha_1 + \cdots + \alpha_{t-1}) - |\theta^1 - \theta^*| \geq A_t.$$

This results in that

$$\sum_{t=1}^{\infty} \alpha_t k_t \geq \sum_{t=\min\{N,T\}}^{\infty} \alpha_t k_t \geq \sum_{t=\min\{N,T\}}^{\infty} \frac{\alpha_t \delta}{A_t}$$

$$\geq \sum_{t=\min\{N,T\}}^{\infty} \frac{\alpha_t \delta}{2C(\alpha_1 + \cdots + \alpha_{t-1})} = \infty$$

$$k_t = \inf_{x \in [\theta^* - A_n, \theta^* + A_n]} \left[\frac{g(x)}{x - \theta^*} \right]$$

showing the second requirement in (5).

Theorem 2

Theorem 2.

Assume (A-1), (A-2), (A-3), (A-4) *and* (A-5). *Assume further* $\delta > 0$ *in* (A-4). *Then* $\lim_{t\to\infty} b_t = 0$.

Proof:
We have for any $z \in [\theta - A_t, \theta + A_t]$

$$\frac{g(z)}{z-\theta} \geq \frac{\delta}{|z-\theta|} \geq \frac{\delta}{A_t}$$

implying that (7) is fulfilled which by Lemma 3 imply the result.

Here
«δ» in (A-4)
can be used directly as
«δ» in Lemma 3

$$\alpha_t > 0 \tag{A-1}$$

$$\sum_{t=1}^{\infty} \frac{\alpha_t}{\alpha_1 + \cdots + \alpha_{t-1}} = \infty \tag{A-2}$$

$$\sum_{t=1}^{\infty} \alpha_t^2 < \infty \tag{A-3}$$

$$\exists \delta \geq 0 \text{ such that } g(x) \leq -\delta \text{ for } x < \theta \text{ and } g(x) \geq \delta \text{ for } x > \theta. \tag{A-4}$$

$$E[Z(\theta;\phi)] = g(\theta) \text{ and } \Pr(|Z(\theta;\phi)| < C) = 1 \tag{A-5}$$

Theorem 3

Theorem 3.

Assume (A-1), (A-2), (A-3) *and* (A-5). *Assume further*

$$g(z) \text{ is nondecreasing;} \tag{9}$$

$$g(\theta^*) = 0; \tag{10}$$

$$g'(\theta^*) > 0. \tag{11}$$

Then $\lim_{t\to\infty} b_t = 0$.

- Need to be clever when finding «δ» of Lemma 3

$$\alpha_t > 0 \tag{A-1}$$

$$\sum_{t=1}^{\infty} \frac{\alpha_t}{\alpha_1 + \cdots + \alpha_{t-1}} = \infty \tag{A-2}$$

$$\sum_{t=1}^{\infty} \alpha_t^2 < \infty \tag{A-3}$$

$$\exists\delta \geq 0 \text{ such that } g(x) \leq -\delta \text{ for } x < \theta \text{ and } g(x) \geq \delta \text{ for } x > \theta. \tag{A-4}$$

$$E[Z(\theta;\phi)] = g(\theta) \text{ and } \Pr(|Z(\theta;\phi)| < C) = 1 \tag{A-5}$$

Proof of Theorem 3

- $g'(\theta^*) = \lim_{x\to\theta^*} \frac{g(x)-g(\theta^*)}{x-\theta^*}$ imply

$$\frac{g(x)}{x-\theta^*} = g'(\theta^*) + \varepsilon(x-\theta^*), \quad \text{with } \lim_{t\to 0}\varepsilon(t) = 0$$

giving

$$\varepsilon(x-\theta^*) = \frac{g(x)}{(x-\theta^*)} - g'(\theta^*) \geq -\frac{1}{2}g'(\theta^*)$$

for $|x-\theta^*| < \delta$ and δ small enough. Thereby

$$\frac{g(x)}{x-\theta^*} \geq \frac{1}{2}g'(\theta^*), \quad \text{for } |x-\theta^*| \leq \delta$$

Since $g'(\theta^*) > 0$, we can choose a δ so small that the inequality is fulfilled for all values closer to θ^*

- For $\theta^* + \delta \leq x \leq \theta^* + A_t$, since $g(z)$ is nondecreasing

$$\frac{g(x)}{x-\theta^*} \geq \frac{g(x+\delta)}{A_t} \geq \frac{\delta g'(\theta^*)}{2A_t}$$

while for $\theta^* - A_t \leq x \leq \theta^* - \delta$

$$\frac{g(x)}{x-\theta^*} = \frac{-g(x)}{\theta^*-x} \geq \frac{-g(x-\delta)}{A_t} \geq \frac{\delta g'(\theta^*)}{2A_t}$$

$$A_t = |\theta^1 - \theta^*| + C(\alpha_1 + \cdots + \alpha_{t-1})$$

- Assuming (without loss of generality) $\delta/A_t \leq 1$ gives

$$\frac{g(x)}{x-\theta^*} \geq \frac{\delta g'(\theta^*)}{2A_t} \quad \text{for } 0 < |x-\theta^*| \leq A_t \Rightarrow \quad (7)$$

Here «δ» in Lemma 3 Is: $\frac{\delta g'(\theta^*)}{2}$ where «δ» is selected above