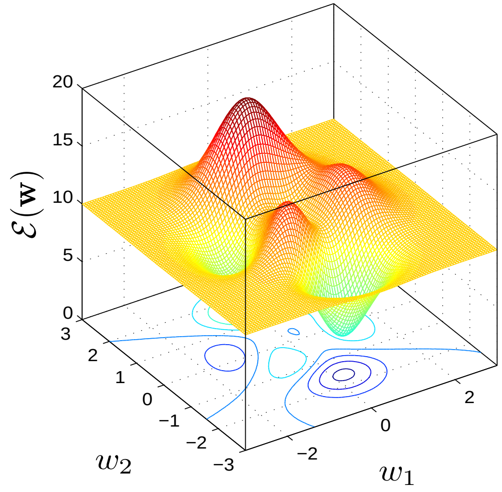
Máxima Descida de Gradiente

Tópicos

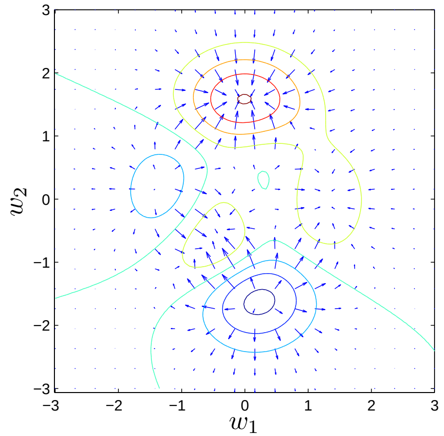
- Máxima descida de gradiente
- Passo de adaptação
- Termo de momento
- Modo *batch* ou *on-line*

Máxima descida de gradiente

- Utilizado para encontrar mínimos de funções (ex. mínimo de $\mathcal{E}(\mathbf{w})$)
- Adaptar os parâmetros $\mathbf{w}$ iterativamente, de modo a que o valor da função a minimizar seja inferior (ou igual) ao seu valor na iteração anterior.
- Gradiente: $\frac{\partial \mathcal{E}(\mathbf{w})}{\partial \mathbf{w}} = \left[\frac{\partial \mathcal{E}(\mathbf{w})}{\partial w_1}, \frac{\partial \mathcal{E}(\mathbf{w})}{\partial w_2}\right]^\top$



Função $\mathcal{E}(\mathbf{w})$ para vários valores de $\mathbf{w} = [w_1, w_2]^\top$



Vectores de gradiente apontam para os máximos de $\mathcal{E}(\mathbf{w})$

Máxima descida de gradiente

- Utilizado para encontrar mínimos de funções (ex. mínimo de $\mathcal{E}(\mathbf{w})$)
- Adaptar os parâmetros $\mathbf{w}$ iterativamente, de modo a que o valor da função a minimizar seja inferior (ou igual) ao seu valor na iteração anterior.
- Gradiente: $\frac{\partial \mathcal{E}(\mathbf{w})}{\partial \mathbf{w}} = \left[\frac{\partial \mathcal{E}(\mathbf{w})}{\partial w_1}, \frac{\partial \mathcal{E}(\mathbf{w})}{\partial w_2}\right]^\top$
- O vector de gradiente indica a direcção de maior crescimento da função $\mathcal{E}(\mathbf{w})$
- Adaptar $\mathbf{w}$ na direcção contrária ao do vector de gradiente:

$$\mathbf{w}(i+1) = \mathbf{w}(i) - \eta \frac{\partial \mathcal{E}(\mathbf{w}(i))}{\partial \mathbf{w}}$$

  onde i é a iteração actual, e η (com $0 < \eta \ll 1$) é uma constante que pondera a actualização de $\mathbf{w}$ (denominado *passo* de actualização)

Máxima descida de gradiente

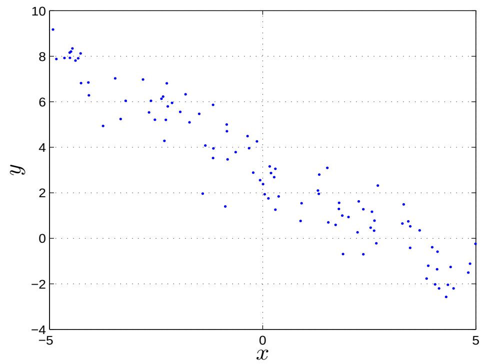
Pseudo-código

1. Inicializar matriz de pesos $\mathbf{w}(0)$
2. Inicializar passo η
3. Calcular erro $\mathcal{E}(\mathbf{w})$
4. Calcular gradiente do erro $\frac{\partial \mathcal{E}(\mathbf{w}(i))}{\partial \mathbf{w}}$
5. Actualizar pesos: $\mathbf{w}(i+1) = \mathbf{w}(i) - \eta \dfrac{\partial \mathcal{E}(\mathbf{w}(i))}{\partial \mathbf{w}}$
6. Voltar ao ponto 3 e repetir um número suficiente de vezes até o valor da função do erro não se alterar significativamente: $\mathcal{E}(\mathbf{w}(i+1)) \approx \mathcal{E}(\mathbf{w}(i))$

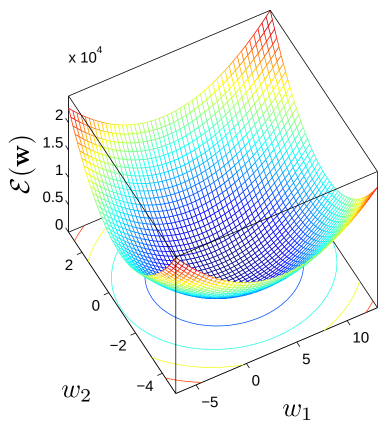
Máxima descida de gradiente

Exemplo: Regressão Linear

- Dados:



- Função do erro: $\mathcal{E} = \frac{1}{N}\sum_{n=1}^{N}(y - \mathbf{w}^\top \mathbf{x})^2$



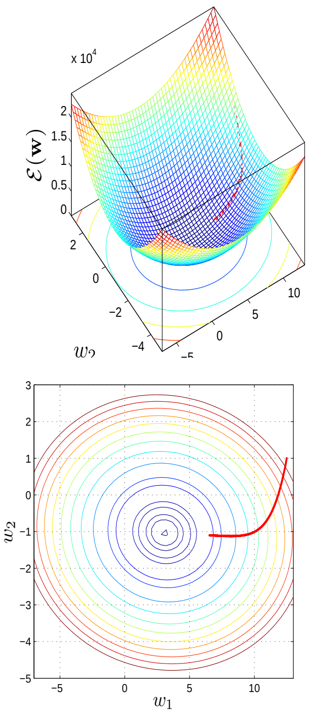
- Modelo: $\hat{y} = w_1 + w_2 x = \mathbf{w}^\top \mathbf{x} = [w_1 \;\; w_2] \begin{bmatrix} 1 \\ x \end{bmatrix}$

- Objectivo: adaptar $\mathbf{w}$ por máxima descida de gradiente
(neste exemplo existe uma solução analítica: $\mathbf{w}_{\text{opt}} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{Y}^\top$)

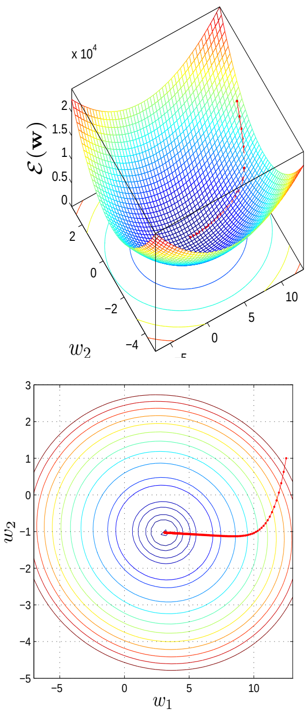
Máxima descida de gradiente

Escolha do passo η

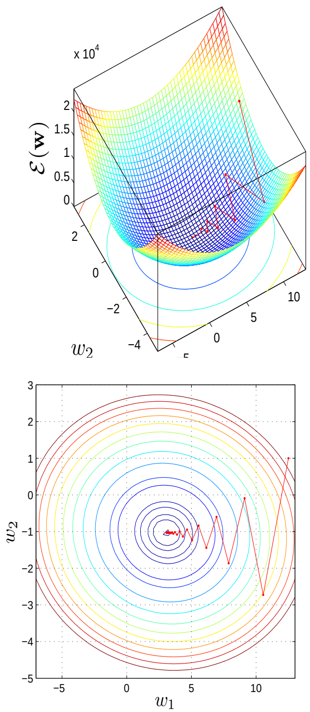
- A escolha do valor do passo η, influencia a convergência do algoritmo
- Valores do passo muito pequenos, o mínimo não é atingido
- Valores do passo muito elevados pode resultar no efeito contrário



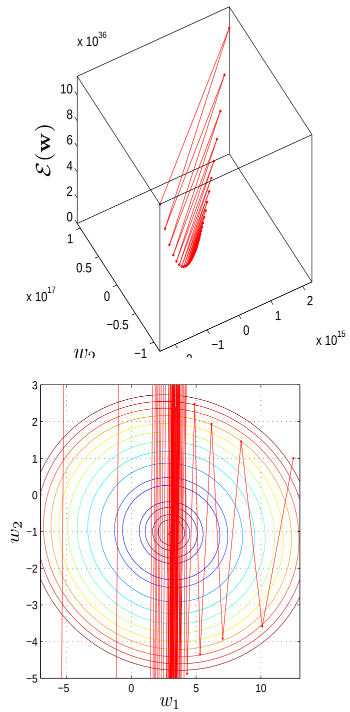
$\eta = 10^{-5}$



$\eta = 10^{-4}$



$\eta = 10^{-3}$



$\eta = 1.2 \times 10^{-3}$

Máxima descida de gradiente

Escolha do passo η

- A escolha do valor do passo η, influencia a convergência do algoritmo
- Valores do passo muito pequenos, o mínimo não é atingido
- Valores do passo muito elevados pode resultar no efeito contrário
- O desejado seria ter valores altos para η quando $\mathbf{w}$ está longe do mínimo, e valores baixos para η quando $\mathbf{w}$ se encontra perto do mínimo
- Existem métodos que se baseiam no vector de gradiente para alterar o valor de η, mas é necessário garantir que a adaptação não fica instável

Máxima descida de gradiente

Termo de momento

- Método para acelerar a convergência da adaptação
- Actualizar $\mathbf{w}$ com uma versão filtrada do gradiente, o termo de momento:

$$\mathbf{z}(i) = \frac{\partial \mathcal{E}\left(\mathbf{w}(i)\right)}{\partial \mathbf{w}} + \alpha \mathbf{z}(i-1) = \frac{\partial \mathcal{E}\left(\mathbf{w}(i)\right)}{\partial \mathbf{w}} + \sum_{k=1}^{+\infty} \alpha^k \frac{\partial \mathcal{E}\left(\mathbf{w}(i-k)\right)}{\partial \mathbf{w}}$$

$$\mathbf{w}(i+1) = \mathbf{w}(i) - \eta \mathbf{z}(i)$$

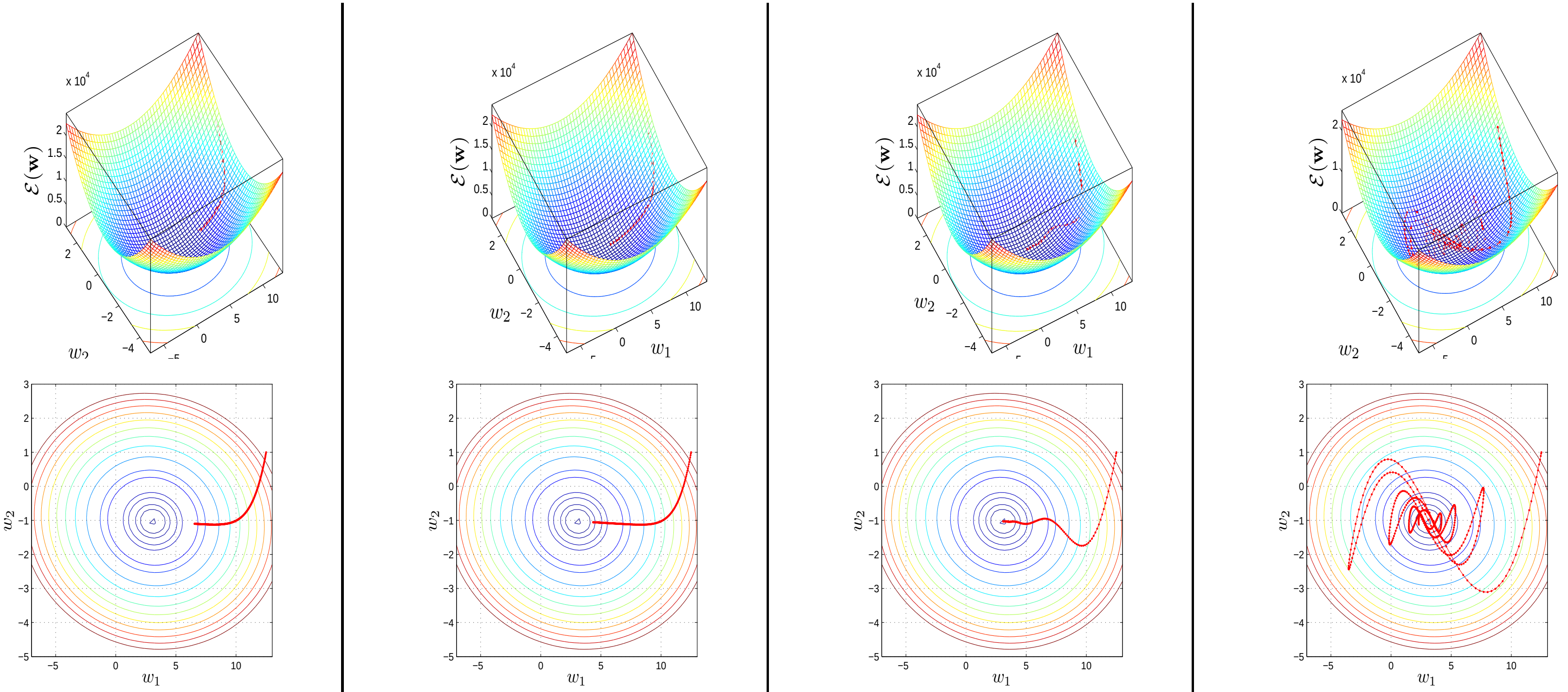
onde $0 \leq \alpha < 1$

- Do ponto de vista de processamento de sinal, o termo de momento é uma filtragem IIR passa-baixo do gradiente
- Se o vector (ou matriz) de gradiente $\mathbf{w}$ "apontar" na mesma direcção (alterar-se pouco) em iterações consecutivas, o termo $\mathbf{z}$ ganha momento e o seu valor aumenta.
- Se gradientes consecutivos tiverem sinais diferentes (apontarem para direcções opostas), o valor de $\mathbf{z}$ diminui, estabilizando assim a convergência

Máxima descida de gradiente

Termo de momento

- Exemplo anterior com $\eta = 10^{-5}$ e diferentes valores de α

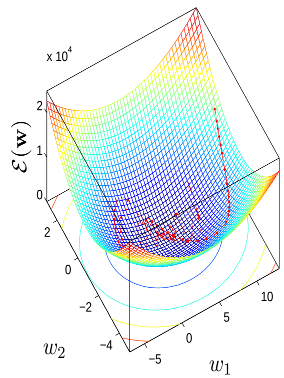


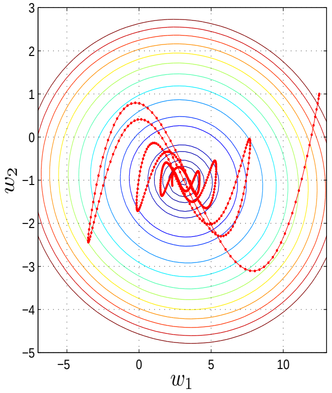
$\alpha = 0$ | $\alpha = 0.5$ | $\alpha = 0.9$ | $\alpha = 0.99$

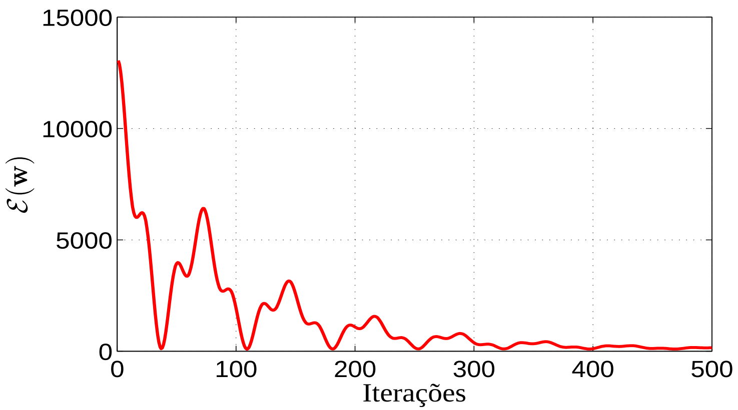
Máxima descida de gradiente

Termo de momento

- Exemplo anterior com $\eta = 10^{-5}$ e $\alpha = 0.99$
- Valores de α perto de 1 podem dificultar a paragem ($\alpha > 1$: instável)







Uma maneira evitar este comportamento é, em cada iteração,

- guardar $\mathbf{w}$ e gradiente se estes corresponderem ao menor valor do erro.
- se o erro na iteração actual for superior ao menor valor do erro: voltar atrás
  - Repor melhores pesos
  - Re-inicializar termo de momento $\mathbf{z}(i-1) = 0$

Máxima descida de gradiente

Métodos de treino:

batch No método *batch* (ou determinístico), todos os $\mathbf{x}$ no conjunto de treino são utilizados no cálculo do erro e do respectivo gradiente. Adaptar os pesos $\mathbf{w}$ baseado no erro quadrático médio e no gradiente:

$$\mathcal{E}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}\left(y_n - \mathbf{w}^\top \mathbf{x}_n\right)^2$$

$$\frac{\partial \mathcal{E}(\mathbf{w})}{\partial \mathbf{w}} = \frac{-2}{N}\sum_{n=1}^{N}\left(y_n - \mathbf{w}^\top \mathbf{x}_n\right)\mathbf{x}_n$$

é um treino em modo *batch*. Nos exemplos anteriores, nos gráficos relativos à escolha de η e de α, as adaptações foram feitas em modo *batch*.

Máxima descida de gradiente

Métodos de treino:

online No método *online* (ou estocástico), o erro $\mathcal{E}(\mathbf{w})$ e o gradiente $\partial\mathcal{E}(\mathbf{w})/\partial\mathbf{w}$ são estimados com um único padrão $\mathbf{x}$

$$\hat{\mathcal{E}}_n(\mathbf{w}) = \left(y_n - \mathbf{w}^\top \mathbf{x}_n\right)^2$$

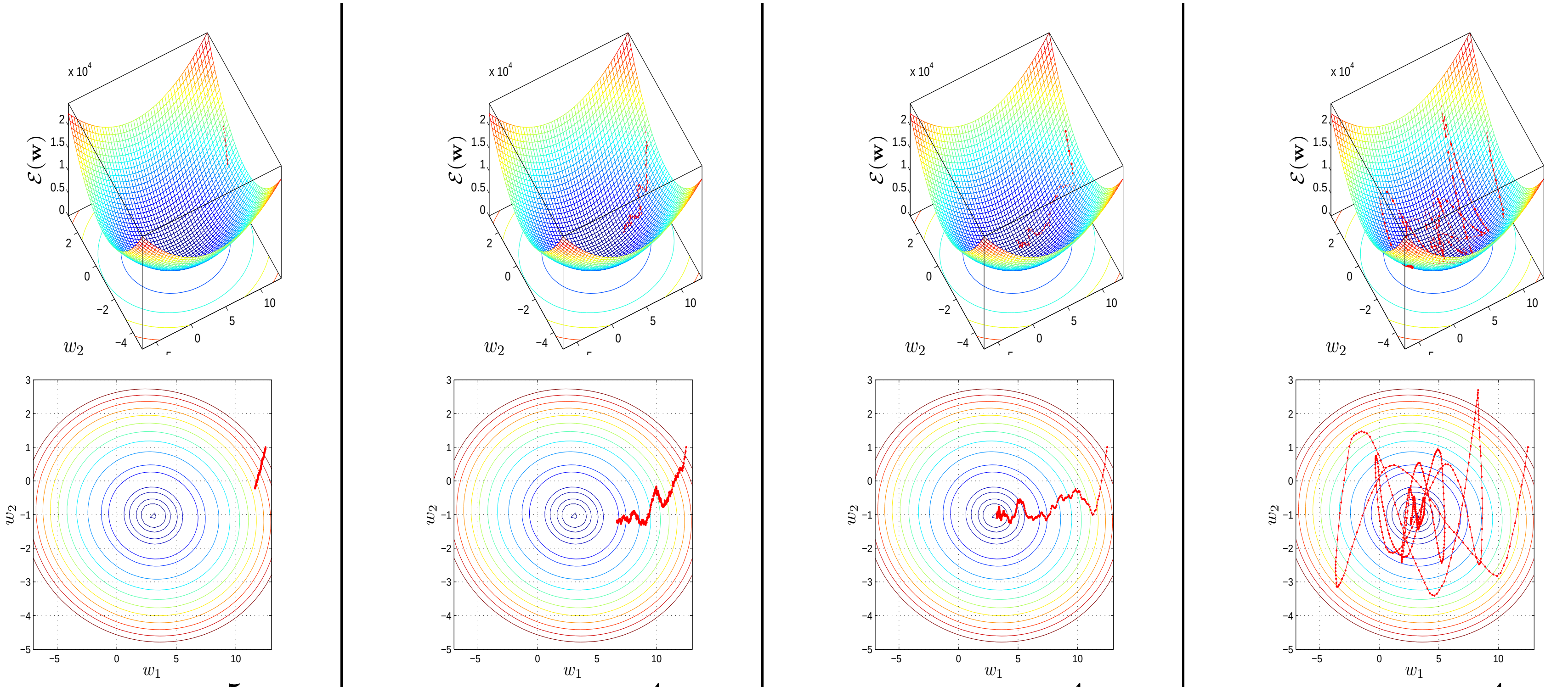
$$\frac{\partial\hat{\mathcal{E}}_n(\mathbf{w})}{\partial\mathbf{w}} = -2\left(y_n - \mathbf{w}^\top \mathbf{x}_n\right)\mathbf{x}_n$$

- O gradiente em modo *online* é uma aproximação (versão ruidosa) do gradiente em modo *batch* (é útil usar o termo de momento para evitar oscilações)
- A adaptação dos pesos é ruidosa (mais errática), o que pode ser útil para sair de mínimos locais
- Para garantir a convergência, é necessário ir diminuindo o passo ao longo do processo iterativo
  $\eta(i+1) = \eta(i)i^{-1}$ ou $\eta(i+1) = \eta(i)\beta^i$ com $0 < \beta < 1$
- Geralmente utilizado quando existem quantidades de dados muito elevadas (convergência mais rápida)

Máxima descida de gradiente

Métodos de treino:

online No método *online* (ou estocástico), o erro $\mathcal{E}(\mathbf{w})$ e o gradiente $\partial\mathcal{E}(\mathbf{w})/\partial\mathbf{w}$ são estimados com um único padrão $\mathbf{x}$



$\eta = 10^{-5}$, $\alpha = 0$ | $\eta = 10^{-4}$, $\alpha = 0$ | $\eta = 10^{-4}$, $\alpha = 0.8$ | $\eta = 10^{-4}$, $\alpha = 0.99$