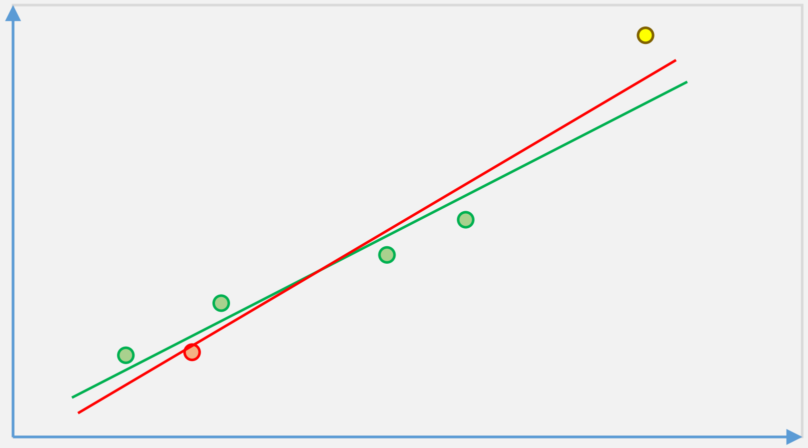
1. What is the difference in terms of the performance between the regression hypotheses based on the objective $\overset{(1)}{\arg_\theta \min \sum_{t=1}^{N} \left[r^{(t)} - h(\boldsymbol{x}^{(t)};\theta)\right]^2}$ and $\arg_\theta \overset{(2)}{\min \sum_{t=1}^{N} \left|r^{(t)} - h(\boldsymbol{x}^{(t)};\theta)\right|}$ respectively?

- 高斯在1829年就證明了Least Square效果優於其他Objective Function，在「Data不存在outlier」的情況，在預測i.i.d(相同趨勢的獨立)未知數據點時，有最小的誤差期望值。
- 在Regression中，(2)式懲罰在某個維度上產生過大的誤差，而(1)式只要能「在某個維度增加誤差e_1，而讓其他維度的誤差總和下降>e_1的量」，就會使整體誤差下降。

(Empirical error)

- 現在考慮如右圖的情況，在一個平面上分布許多點，利用(1)式和(2)式分別擬合後得到兩條趨勢線，紅線為(1)式的結果，綠線為(2)式的。



- 那麼你可以看見由紅線轉為綠線後，綠色點的誤差減少了，而紅色點和黃色點誤差增加了。
- 已知黃色點是outlier，那麼(2)式可以降低outlier的影響，在已知outlier濃度的情況下，甚至有助於找出outlier。
- 因此，在理想的情況下，(1)式較好。反之若對資料有一定了解，例如存在少量outlier，或其他特殊應用(需要根據當時情況判斷)下，(2)式則可能有較好的表現。

2. In logistic regression, show that $l(\boldsymbol{\beta}) = \sum_{t=1}^{N} \left\{ y^{(t)} \boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}^{(t)} - \log\left(1 + e^{\boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}^{(t)}}\right) \right\}$.

$$l(\beta) = \sum_{t=1}^{N} \{ \underbrace{y^{(t)} \log \pi(x^{(t)};\beta)}_{A} + \underbrace{(1-y^{(t)}) \log(1-\pi(x^{(t)};\beta))}_{B} \}$$

$$A = y^{(t)} \log \frac{1}{1+e^{-\beta^T \tilde{x}^{(t)}}} = \boxed{-y^{(t)} \log(1+e^{-\beta^T \tilde{x}^{(t)}})} \rightarrow C$$

$$B = (1-y^{(t)}) \log \frac{1}{e^{\beta^T \tilde{x}^{(t)}}+1} = (1-y^{(t)}) \log \frac{e^{-\beta^T \tilde{x}^{(t)}}}{1+e^{-\beta^T \tilde{x}^{(t)}}}$$

$$= -\log(e^{\beta^T \tilde{x}^{(t)}}+1) - y^{(t)} \big(\log e^{-\beta^T \tilde{x}^{(t)}} \underbrace{\boxed{- \log(1+e^{-\beta^T \tilde{x}^{(t)}})}}_{\rightarrow -C} \big)$$

$$\Rightarrow A+B = y^{(t)} \beta^T \tilde{x}^{(t)} \log e - \log(e^{\beta^T \tilde{x}^{(t)}}+1)$$

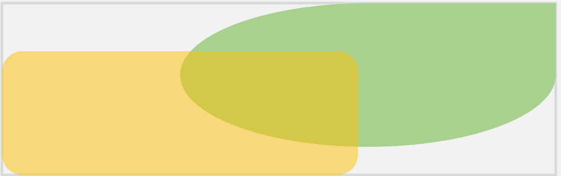
$$\Rightarrow l(\beta) = \sum_{t=1}^{N} \{ y^{(t)} \beta^T \tilde{x}^{(t)} - \log(e^{\beta^T \tilde{x}^{(t)}}+1) \}$$

3. Read Appendix C on the definitions of convex set and functions.

(a) Show that the intersection of convex sets, $\bigcap_{i \in \mathbb{N}} C_i$ where $C_i \subseteq \mathbb{R}^n$, is convex.

(b) Show that the log-likelihood function for logistic regression, $l(\beta)$, is concave.

- 凸集合的充要條件是：一個集合，任兩個集合內的點連線上的所有點都在該集合內。
- 欲證明任意多個凸集合的交集是凸集合，只要證明兩個凸集合的交集是凸集合即可。
- 兩個凸集合的交集可能是空集合，而空集合是凸集合。
- 若兩個凸集合的交集不是空集合：



1. Given two points x_1, x_2, and two sets S_1, S_2
2. Given that $x_1 \in S_1$, $x_2 \in S_1$, $x_1 \in S_2$, $x_2 \in S_2$, $x_3 = \theta x_1 + (1-\theta) x_2$, *where* $\theta \in [0,1]$
3. Then we have $x_3 \in S_1$, $x_3 \in S_2$
4. That is, $x_3 \in S_1 \cap S_2$
5. Thus, the set $S_1 \cap S_2$ is concave, for x_3 is any linear interpolation of two points x_1 and x_2 in $S_1 \cap S_2$, and $x_3 \in S_1 \cap S_2$

3. Read Appendix C on the definitions of convex set and functions.

(a) Show that the intersection of convex sets, $\bigcap_{i \in \mathbb{N}} C_i$ where $C_i \subseteq \mathbb{R}^n$, is convex.

(b) Show that the log-likelihood function for logistic regression, $l(\boldsymbol{\beta})$, is concave.

A. Result of adding 2 concave functions is a concave function.

B. A linear function is both a convex function and a convex function.

- The goal is to proof that $\sum_{t=1}^{N} \left\{ y^{(t)} \boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}^{(t)} - \log\left(1 + e^{\boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}^{(t)}}\right) \right\}$ is concave.
- Linear function $y^{(t)} \boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}^{(t)}$ is concave, by property A.
- $\log\left(1 + e^{\boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}^{(t)}}\right)$ is convex, for its first derivative and second derivative are both greater than 0. Thus, $-\log\left(1 + e^{\boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}^{(t)}}\right)$ is concave.
- Hence, by property B, $y^{(t)} \boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}^{(t)} - \log\left(1 + e^{\boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}^{(t)}}\right)$ is concave.
- As a consequence, $\sum_{t=1}^{N} \left\{ y^{(t)} \boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}^{(t)} - \log\left(1 + e^{\boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}^{(t)}}\right) \right\}$ is concave, by property B.

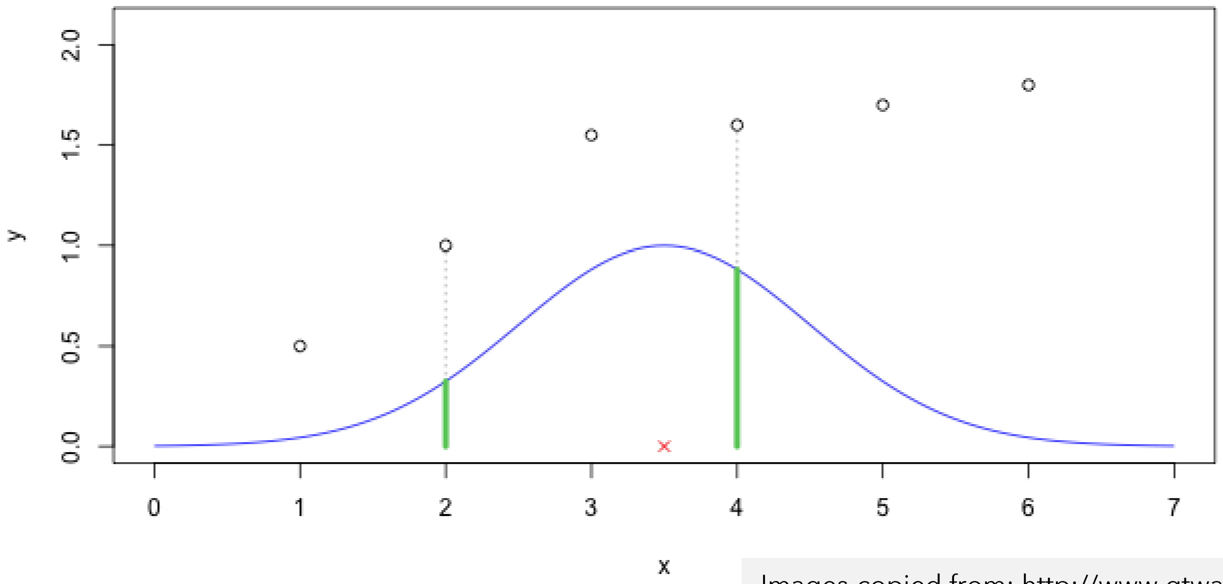
4. Consider the locally weighted linear regression problem with the following objective:

$$\arg\min_{\boldsymbol{w}\in\mathbb{R}^{d+1}} \boxed{\frac{1}{2}} \sum_{i=1}^{N} l^{(i)}\left(\boldsymbol{w}^\top \begin{bmatrix} 1 \\ \boldsymbol{x}^{(i)} \end{bmatrix} - r^{(i)}\right)^2$$

local to a given instance $\boldsymbol{x}'$ whose label will be predicted, where $l^{(i)} = \exp(-\frac{(\boldsymbol{x}'-\boldsymbol{x}^{(i)})^2}{2\tau^2})$ for some constant τ.

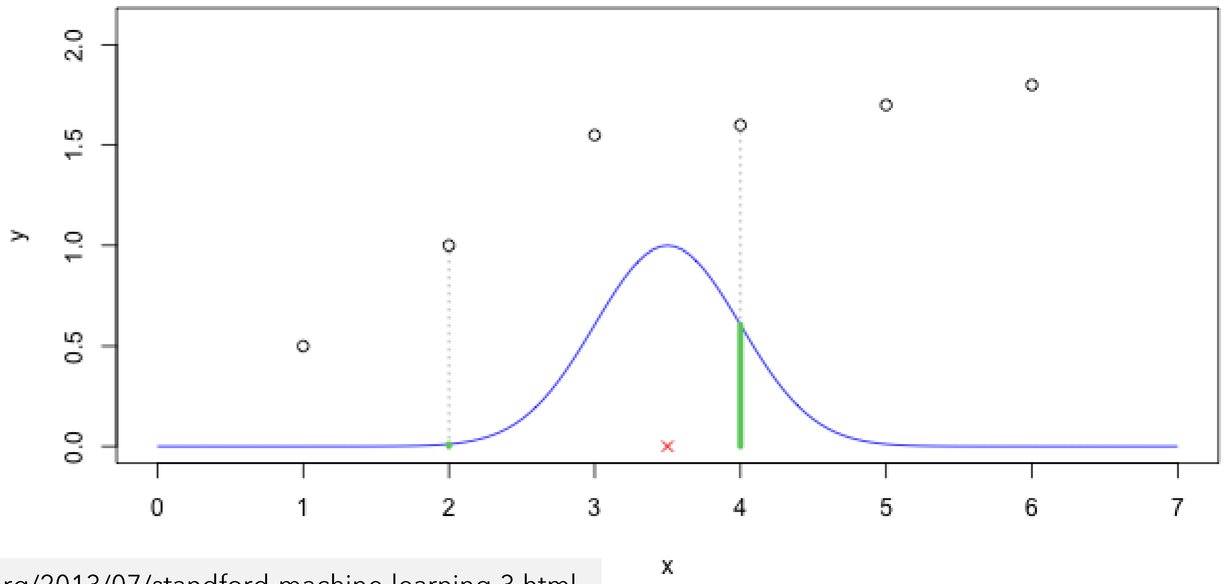
1-D example : x = 3.5 with τ = 1

$y = l^{(i)}$



1-D example : x = 3.5 with τ = 0.5

$y = l^{(i)}$



Images copied from: http://www.gtwang.org/2013/07/standford-machine-learning-3.html

(a) Show that the above objective can be written as the form

$$(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{r})^\top \boldsymbol{L}(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{r}). \qquad \text{Specify clearly what } \boldsymbol{X}, \boldsymbol{r}, \text{ and } \boldsymbol{L} \text{ are.}$$

- ***X*** is all the data, each row is a single data point, which contains 1 and $x^{(i)}$.
- ***w*** is the coefficient array, in which row 1 is w_0, row 2 is w_1, and so on.
- ***r*** is the label array, dimension of which agrees with ***Xw***.
- ***Xw*** – ***r*** is the error term, which means the error (distance) of ***Xw*** and ***r***.
- L is a N × N diagonal matrix, with elements from $\frac{1}{2}l^{(1)}$ to $\frac{1}{2}l^{(n)}$.
  - In fact, we can drop the ½.
- Details as below. You can see that **(*Xw* – *r*)^T L (*Xw* – *r*)** is the objective.

$$\underset{\boldsymbol{X}}{\begin{bmatrix} 1 & \boldsymbol{x}^{(1)} \\ \multicolumn{2}{c}{\boldsymbol{x}^{(2)}} \\ \multicolumn{2}{c}{\boldsymbol{x}^{(3)}} \\ \multicolumn{2}{c}{\dots} \\ \multicolumn{2}{c}{\boldsymbol{x}^{(n)}} \end{bmatrix}} \quad \underset{\boldsymbol{w}}{\begin{bmatrix} w_0 \\ w_1 \\ \dots \end{bmatrix}} \quad \underset{\boldsymbol{Xw}}{\begin{bmatrix} \boldsymbol{x}^{(1)}\boldsymbol{w} \\ \boldsymbol{x}^{(2)}\boldsymbol{w} \\ \boldsymbol{x}^{(3)}\boldsymbol{w} \\ \dots \\ \boldsymbol{x}^{(n)}\boldsymbol{w} \end{bmatrix}} \quad \underset{\boldsymbol{r}}{\begin{bmatrix} r^{(1)} \\ r^{(2)} \\ r^{(3)} \\ \dots \\ r^{(n)} \end{bmatrix}} \quad \underset{\boldsymbol{Xw}\text{-}\boldsymbol{r}}{\begin{bmatrix} \boldsymbol{x}^{(1)}\boldsymbol{w} - r_1 \\ \boldsymbol{x}^{(2)}\boldsymbol{w} - r_2 \\ \boldsymbol{x}^{(3)}\boldsymbol{w} - r_3 \\ \dots \\ \boldsymbol{x}^{(n)}\boldsymbol{w} - r_n \end{bmatrix}} \quad \underset{\boldsymbol{2L}}{\begin{bmatrix} l^{(1)} & & & & \\ & l^{(2)} & & & \\ & & l^{(3)} & & \\ & & & \dots & \\ & & & & l^{(n)} \end{bmatrix}}$$

$$\underset{(\boldsymbol{Xw}-\boldsymbol{r})^T \mathbf{L} (\boldsymbol{Xw}-\boldsymbol{r})}{\begin{bmatrix} \frac{1}{2} l^{(1)}(\boldsymbol{x}^{(1)}\boldsymbol{w} - r^{(1)})^2 + \\ \frac{1}{2} l^{(2)}(\boldsymbol{x}^{(2)}\boldsymbol{w} - r^{(2)})^2 + \\ \frac{1}{2} l^{(3)}(\boldsymbol{x}^{(3)}\boldsymbol{w} - r^{(3)})^2 + \\ \dots + \\ \frac{1}{2} l^{(n)}(\boldsymbol{x}^{(n)}\boldsymbol{w} - r^{(n)})^2 \end{bmatrix}}$$

(b) Give a close form solution to $\boldsymbol{w}$. (Hint: recall that we have $\boldsymbol{w} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{r}$ in linear regression when $l^{(i)} = 1$ for all i)

- Objective of linear regression is $\Sigma_{i=1}^{N}$ ($w_0 + w_1x^{(i)} + \dots - r^{(i)}$)^2 , and its close form solution is $\boldsymbol{w} = (X^T X)^{-1} X^T r$.
- Objective of linear regression is $\frac{1}{2}\,\Sigma_{i=1}^{N}\; l^{(i)}(w_0 + w_1x^{(i)} + \dots - r^{(i)})^2$, which is equal to $\frac{1}{2}\,\Sigma_{i=1}^{N}(w_0\sqrt{l^{(i)}} + w_1 \sqrt{l^{(i)}}x^{(i)} + \dots - \sqrt{l^{(i)}}r^{(i)})^2$.

$\sqrt{(\frac{1}{2}l^{(1)})}$	$\sqrt{(\frac{1}{2}l^{(1)})}\, x^{(1)}$
$\sqrt{(\frac{1}{2}l^{(2)})}\, \boldsymbol{x}^{(2)}$	
$\sqrt{(\frac{1}{2}l^{(3)})}\, \boldsymbol{x}^{(3)}$	
…	
$\sqrt{(\frac{1}{2}l^{(n)})}\, \boldsymbol{x}^{(n)}$	

$\boldsymbol{X}_{weighted}$

- So we can turn $\boldsymbol{X}$ into $\boldsymbol{X}_{weighted}$, just like the right table.
- Thus, $\boldsymbol{X}_{weighted}{}^{T}\, \boldsymbol{X}_{weighted} = \boldsymbol{X}^T\boldsymbol{L}\boldsymbol{X}$.
- And turn $\boldsymbol{r}$ into $\boldsymbol{r}_{weighted}$, just like the right table.
- Thus, $(\boldsymbol{X}^T \boldsymbol{L}\, \boldsymbol{X})^{-1}\, \boldsymbol{X}_{weighted}{}^{T}\, \boldsymbol{r}_{weighted} = \boldsymbol{X}^T\boldsymbol{L}\boldsymbol{r}$.
- As a conclusion, the closed form solution of $\boldsymbol{w}$ is $(\boldsymbol{X}^T \boldsymbol{L}\, \boldsymbol{X})^{-1}\, \boldsymbol{X}^T \boldsymbol{L}\, \boldsymbol{r}$.

$\sqrt{(\frac{1}{2}l^{(1)})}\; r^{(1)}$
$\sqrt{(\frac{1}{2}l^{(2)})}\; r^{(2)}$
$\sqrt{(\frac{1}{2}l^{(3)})}\; r^{(3)}$
…
$\sqrt{(\frac{1}{2}l^{(n)})}\; r^{(n)}$

$\boldsymbol{r}_{weighted}$

(c) Suppose that the training examples $(\boldsymbol{x}^{(i)}, r^{(i)})$ are i.i.d. samples drawn from some joint distribution with the marginal:

$$p(r^{(i)}|\boldsymbol{x}^{(i)}; \boldsymbol{w}) = \frac{1}{\sqrt{2\pi\sigma^{(i)}}} \exp\left(-\frac{(r^{(i)} - \boldsymbol{w}^\top \begin{bmatrix} 1 \\ \boldsymbol{x}^{(i)} \end{bmatrix})^2}{2\sigma^{(i)2}} \right)$$

where $\sigma^{(i)}$'s are constants. Show that finding the maximum likelihood of $\boldsymbol{w}$ reduces to solving the locally weighted linear regression problem above. Specify clearly what the $l^{(i)}$ is in terms of the $\sigma^{(i)}$'s.

- $\arg_{\boldsymbol{w}} \max \ \Pi_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^{(i)}}} \exp\left(-\frac{(r^{(i)} - \boldsymbol{w}^\top \begin{bmatrix} 1 \\ \boldsymbol{x}^{(i)} \end{bmatrix})^2}{2\sigma^{(i)2}} \right)$

$\equiv \arg_{\boldsymbol{w}} \min \ \Pi_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^{(i)}}} \exp\left(\frac{(r^{(i)} - \boldsymbol{w}^\top \begin{bmatrix} 1 \\ \boldsymbol{x}^{(i)} \end{bmatrix})^2}{2\sigma^{(i)2}} \right)$

$\equiv \arg_{\boldsymbol{w}} \min \ \Pi_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^{(i)}}} \ \Pi_{i=1}^{N} \exp\left(\frac{(r^{(i)} - \boldsymbol{w}^\top \begin{bmatrix} 1 \\ \boldsymbol{x}^{(i)} \end{bmatrix})^2}{2\sigma^{(i)2}} \right)$ ...(1)

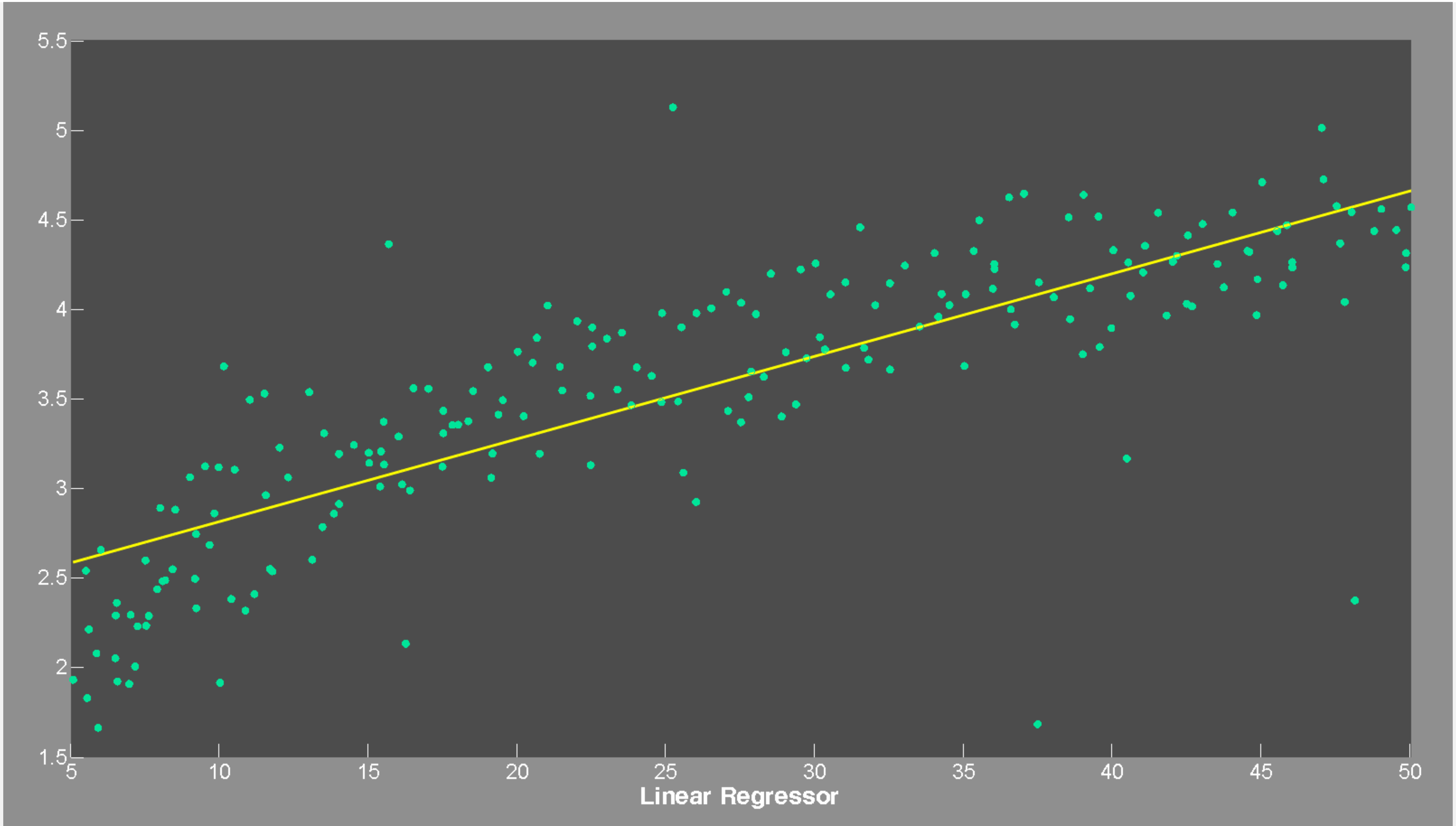
- Since $\Pi_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^{(i)}}}$ is a constant value for a certain sample set,

(1) can be done by $\arg_{\boldsymbol{w}} \min \Pi_{i=1}^{N} \exp\left(\frac{(r^{(i)} - \boldsymbol{w}^\top \begin{bmatrix} 1 \\ \boldsymbol{x}^{(i)} \end{bmatrix})^2}{2\sigma^{(i)2}} \right)$

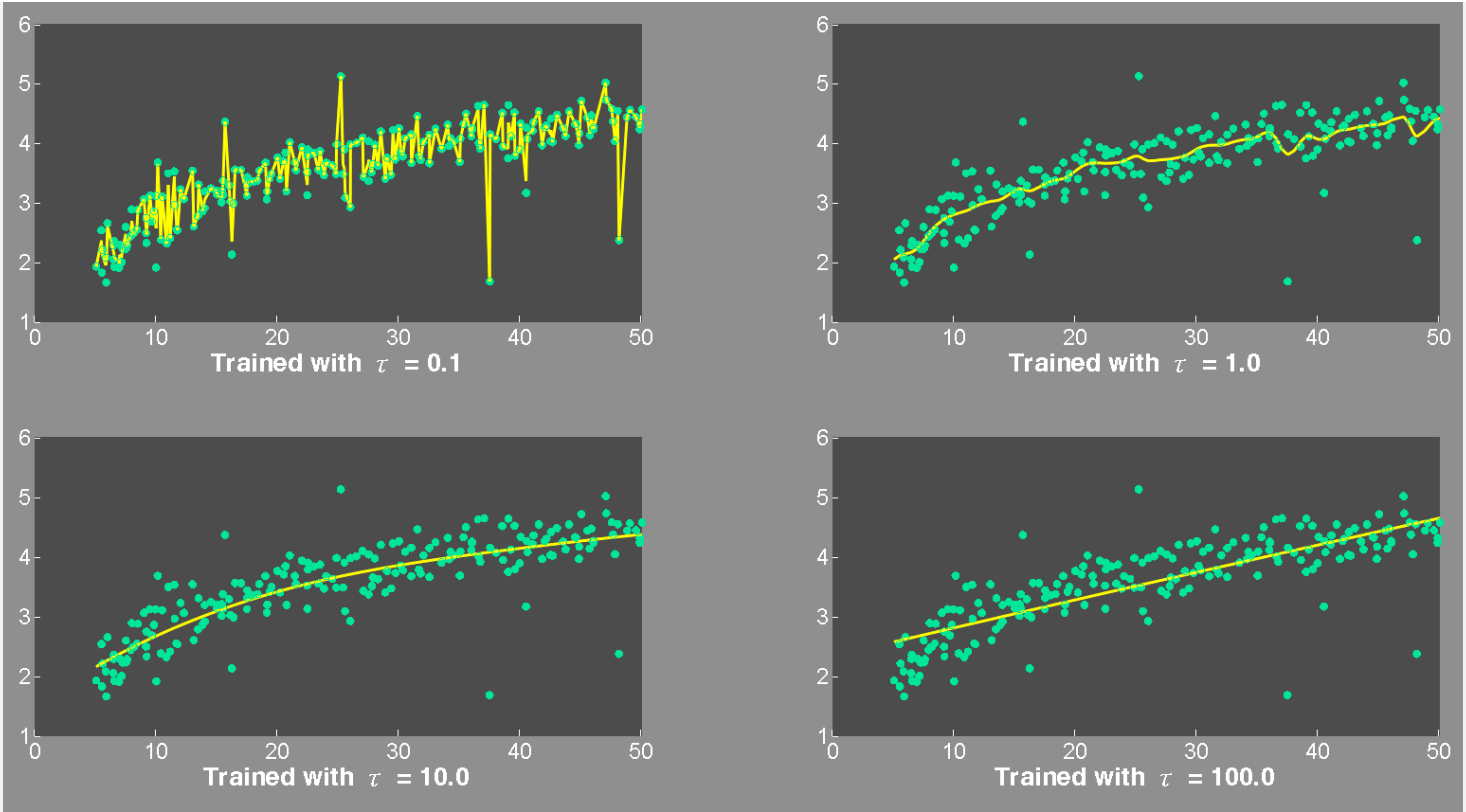
$\equiv \arg_{\boldsymbol{w}} \min \Sigma_{i=1}^{N} \left(\frac{(r^{(i)} - \boldsymbol{w}^\top \begin{bmatrix} 1 \\ \boldsymbol{x}^{(i)} \end{bmatrix})^2}{2\sigma^{(i)2}} \right)$, which is a locally weighted regression problem,

with $l^{(i)} = 1 / (\sigma^{(i)})^2$.

(d) Implement a linear regressor (see the spec for more details) on the provided 1D dataset. Plot the data and your fitted line. (Hint: don't forget the intercept term)

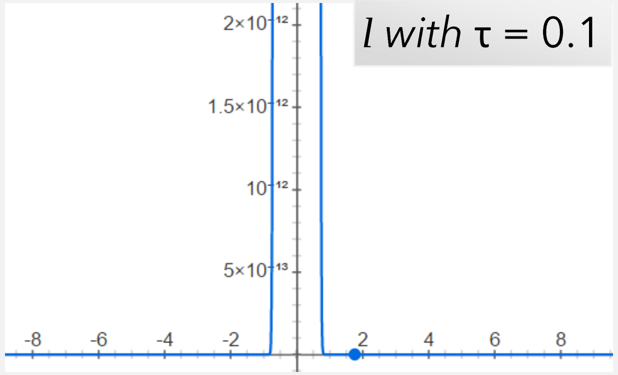


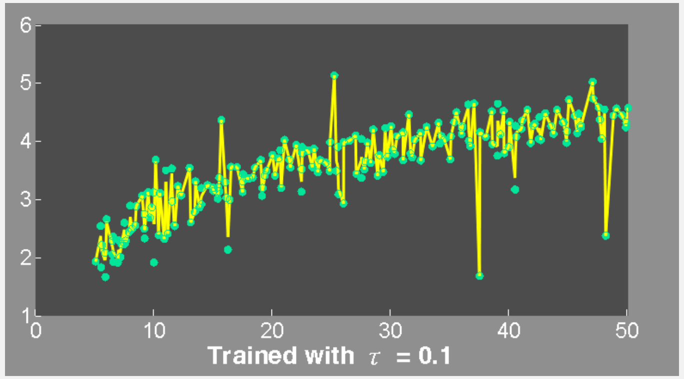
(e) Implement 4 locally weighted linear regressors (see the spec for more details) on the same dataset with $\tau = 0.1$, 1, 10, and 100 respectively. Plot the data and your 4 fitted curves (for different $\boldsymbol{x}'$s within the dataset range).



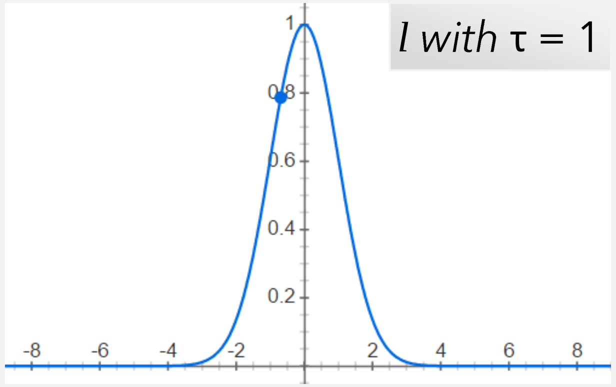
(f) Discuss what happens when τ is too small or large.

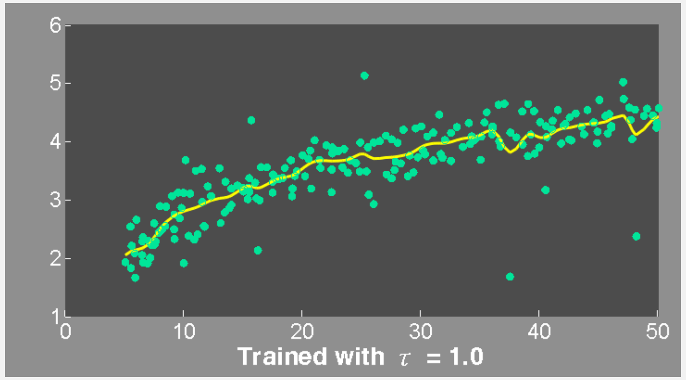
- 這裡的 τ (bandwidth) 與資料在X軸上分布的範圍有關，對於不同的dataset，需要不同的τ



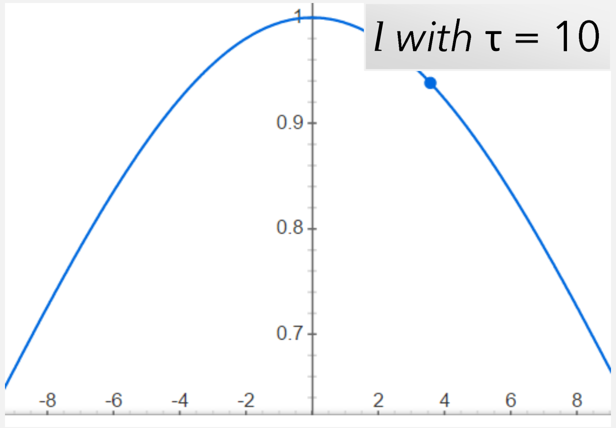


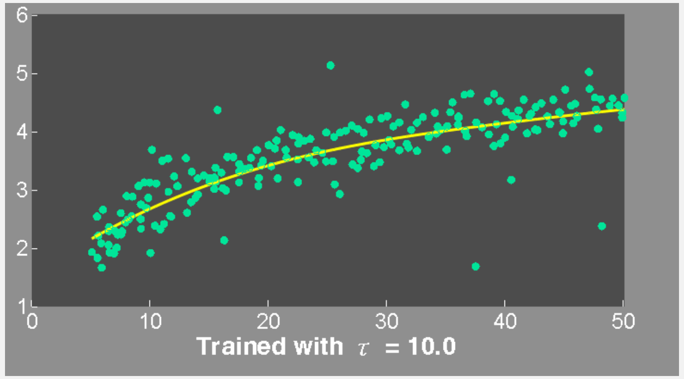
- τ = 0.1時，效果非常差，因為參與考慮的只有非常鄰近的資料，因此發生了過適(over-fitting)



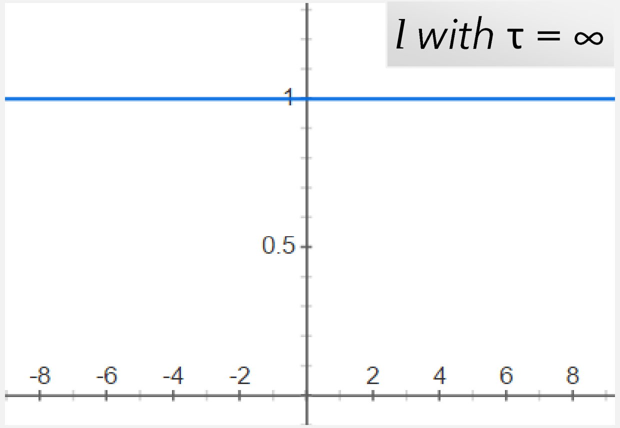


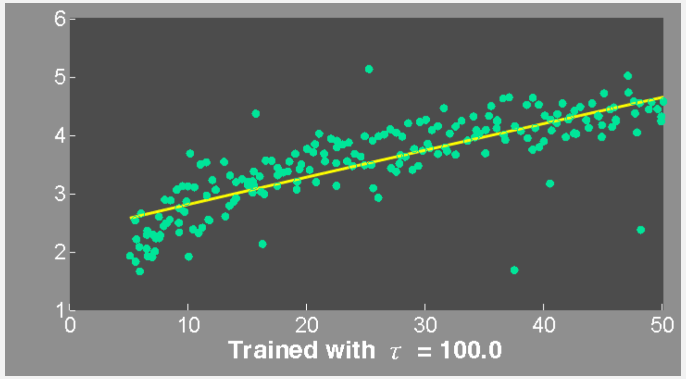
- τ = 1時，效果顯著改善，雖然可以明顯看出受到雜訊影響，不過整體迴歸合理性已經大幅上升





- τ = 10時，曲線更加平滑





- τ = 100，或更大時，
  即喪失local weight的效果，
  退化為普通的線性回歸

- 對於這個dataset而言，0.1是一個過小的 τ ，而100則過大。