بسم الله الرحمن الرحيم

لما بنيجي نتكلم علي ال linear regression .. و in terms of statistical learning .. القصه كلها ان لماا بيكون عندنا شوية داتا .. هنقول ان الدااتا جايه من شوية قياسات ممكن يبقا فيها نويز .. فانت عندك إنبوت x .. و أوتبوت y .. فانت مش هتلاحظ الأوتبوت بتاع ال function f .. و ديه ال function اللي انت عاوز تتعلمها . انت هت observe ... a corrupted version of y .. اللي هي f(x) + noise .. زي مثلاً ل weather .. prediction .. وانت عندك historical data زي مثلا wind pressure temperature .. و كل يوم عندك ال measured temperature .. بس ال thermometer مش accurate .. بيحطلك noise ... فبشكل عام لأي نوع من انواع الداتا لازم تفترض ان هيبقا في corruption نتيجة وجود الnoise .. فانت محتاج تاخدها في الاعتبار .. لما بتيجي ت model الكلام ده .. واحنا بنتكلم علي ال linear regression .. هنفترض ان ال function f هي linear function .. و النويز هنفترض انها Gaussian noise. . انت ممكن تحط noise تحت أي distribution .. بس Gaussian هو اللي منتشر يعني .. ده كدا ال setting اللي هنتكلم عليه ...

فكدا انت تقدر تعبر عن likelihood distribution يقولنا .. لكل إنبوت في الداتا سيت بتاعتنا .. ايه هو ال likelihood اللي هيقيس أوتبوت معين اللي هو y .. فالدكتور بيقول ان ال likelihood distribution . نقدر ن infere من الموديل اللي عندنا

Noisy Linear Regression

- Assume y is obtained from $\boldsymbol{x}$ by a deterministic function f that has been perturbed (i.e., noisy measurement)

model → $y = f(\bar{\boldsymbol{x}}) + \epsilon$ ($f(\bar{\boldsymbol{x}}) = \boldsymbol{w}^T\bar{\boldsymbol{x}}$, $\epsilon \sim N(0, \sigma^2)$)

- Gaussian noise:

$$\Pr(\boldsymbol{y}|\bar{\boldsymbol{X}}, \boldsymbol{w}, \sigma) = N\left(\boldsymbol{y}\middle|\boldsymbol{w}^T\bar{\boldsymbol{X}}, \sigma^2\right)$$

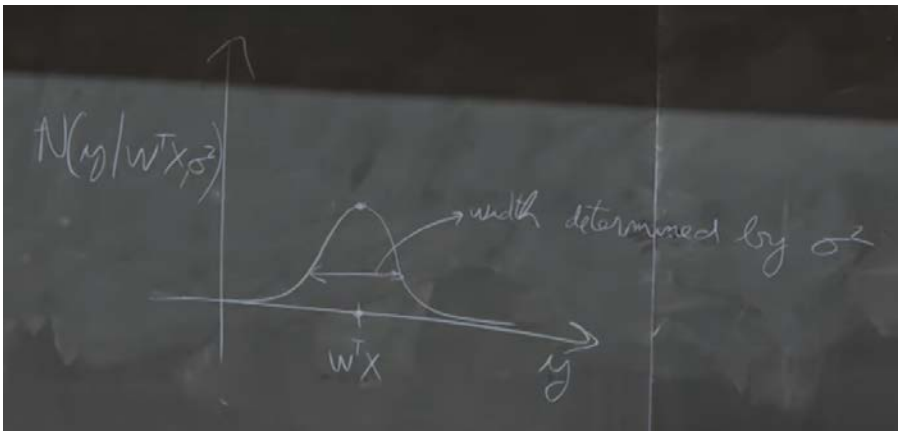
output ↑ input

$$= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\left(y_n - \boldsymbol{w}^T\bar{\boldsymbol{x}}_n\right)^2}{2\sigma^2}}$$

University of Waterloo CS480/680 Spring 2019 Pascal Poupart

الدكتور شرح هو ليه ال Probability هتطلع Gaussian.. عشان انت فارض ان ال function هتبقا linear .. و كمان it's deterministic .. فبكدا انت هتضيف نويز ليها جواسين ديستربيوشن .. فالواي هيبقا ليها فاريانس زي ال noise .. و القيمه اللي هنشوفها هتبقا هي ال wTx .. لو جيت تكتب ال pdf .. هتبقا هي ال Pr(y|X,w, sigma) ...تعال نرسم اللي بيحصل ده عشان يبقا عندنا mental image .. مش mental sickness هه ...

هنرسم curve .. الإكس أكسيز هو ال output y .. عشان هنرسم ال pdf لل Gaussian distribution بالنسبة ل y ... ال max هيبقا عند ال mean و ال variance هيحددلك ال width بتاع الكيرف ...



دلوقت احنا هنعوز ن apply the maximum likelihood .. اللي شفناه المحاضره اللي فاتت .. هناخد ال expression و نقول عاوزين W اللي هتجيب ال max probabiolity لكل ال y في الداتا سيت .. انت عندك داتا .. فيها الإنبوت و الواي .. دلوقت عندك موديل بيتحدد بال W . وال W بيقلك ازاي تقدر تاخد linear combinations من الإنبوتس اللي عندك عشان تعرف ت predict the output .. ودلوقت انت عاوز تختار أحسن W .. عشان تعمل كدا هتقول مين هي أحسن W اللي هتديك اعلي احتمالية لل y اللي احنا عاوزنها ..

Maximum Likelihood

- Possible objective: find best $\boldsymbol{w}^*$ by maximizing the likelihood of the data

$$\boldsymbol{w}^* = argmax_{\boldsymbol{w}} \Pr(\boldsymbol{y}|\bar{\boldsymbol{X}}, \boldsymbol{w}, \sigma)$$

$$= argmax_{\boldsymbol{w}} \prod_n e^{-\frac{(y_n - \boldsymbol{w}^T \bar{\boldsymbol{x}}_n)^2}{2\sigma^2}}$$

$$= argmax_{\boldsymbol{w}} \sum_n -\frac{(y_n - \boldsymbol{w}^T \bar{\boldsymbol{x}}_n)^2}{2\sigma^2}$$

$$= argmin_{\boldsymbol{w}} \sum_n (y_n - \boldsymbol{w}^T \bar{\boldsymbol{x}}_n)^2$$

log

مش هيغير الـ output

- We arrive at the original least square problem!

انما هيعمل Rescaling

هو ليه اصلا بناخد ال log .. عشان احنا عندنا ضرب و هنا ال expression is complicated و ده expression non convex ... لما بيبقي convex بيبقا سهل ان الواحد يتعامل معاه انما لما بيبقي non convex مش بيبقا سهل تتعامل معاه .. فبالتالي بنحتاج ال log عشان كدا .. وديه هي ال leat square problem الي اتكلمنا عنها في ال linear regression ..

طيب ايه الي هيحصل لو قلنا بدل ما نجيب ال W اللي بت maximize ال likelihood .. نجيب ال W اللي بت maximize the posterior .. عن طريق اننا نحسب ال posterior ...

Maximum A Posteriori

- Alternative objective: find $\boldsymbol{w}^*$ with highest posterior probability
- Consider Gaussian prior: $\Pr(\boldsymbol{w}) = N(\boldsymbol{0}, \boldsymbol{\Sigma})$
- Posterior:

prior Likelihood

$$\Pr(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) \propto \Pr(\boldsymbol{w}) \Pr(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$$

$$= k e^{-\frac{\boldsymbol{w}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{w}}{2}} e^{-\frac{\sum_n (y_n - \boldsymbol{w}^T \boldsymbol{x}_n)^2}{2\sigma^2}}$$

الدكتور بعد كدا اتكلم علي وجود ال prior distribution .. ال prior انت محتاج ت express ال distribution بتاعه .. لو فكرت في ال Machine learning كإنك مش عارف ايه هي ال hypothesis .. انا عندي شوية Uncertainty في البدايه .. و ال process بتاعت ال machine learning .. هي process of reducing the uncertainty وبعدين نقدر ن quantify ال uncertainty بتاعتنا عن طريق distribution اللي هو ال prior .. وعملية ال learning هي انك تحسب ال posterior distribution ..و هي اسمها posterior عشان ال distribution ده بيطلع بعد ما نكون بصينا علي الداتا الإكس و الواي .. وده هيدينا الديستربيوشن اللي ليه less uncertainty وده هو ال bayesian learning بيتكلم عليه .. فبالتالي انا محتاج prior distribution اللي بيحدد ال initial uncertainty لل W ... هنا احنا هنستخدم ال domain knowledge .... فانا هخمّن فين ال W و اقوم حاطط Gaussian هناك .. فهنستخدم جاوسين mean zero and covariance sigma ..

Maximum A Posteriori

- Alternative objective: find $\boldsymbol{w}^*$ with highest posterior probability
- Consider Gaussian prior: $\Pr(\boldsymbol{w}) = N(\mathbf{0}, \boldsymbol{\Sigma})$
- Posterior:

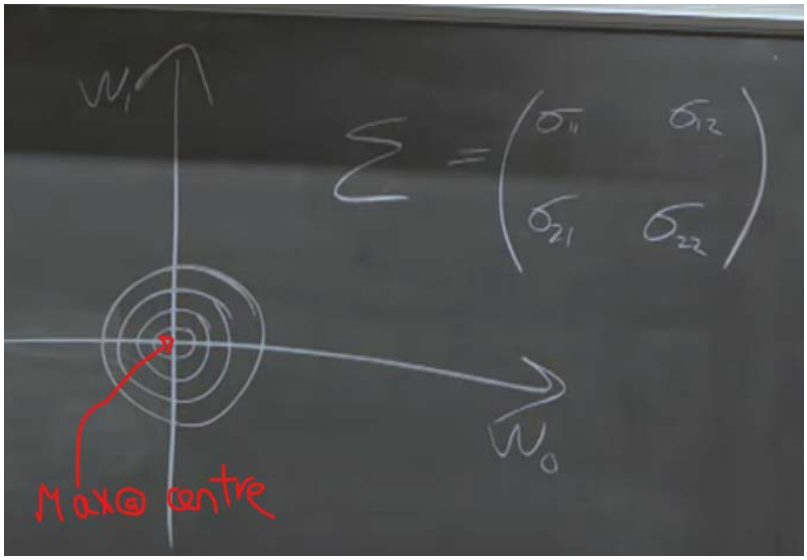
$$\Pr(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{y}) \propto \Pr(\boldsymbol{w}) \Pr(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$$

$$= k e^{-\frac{\boldsymbol{w}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{w}}{2}} e^{-\frac{\Sigma_n (y_n - \boldsymbol{w}^T \boldsymbol{x}_n)^2}{2\sigma^2}}$$

prior — Likelihood — covariance matrix

W → a vector of weights

دلوقت هنرسم مثال لل prior distribution لل W .. وهنقول يعني ان ال W ليها W0 و W1 .. ممكن يكون عندك بعد كدا أكتر من W .. بس يعني خلينا نتكلم علي الاتنين دول دلوقت . . هنا هنرسم ال Gaussian distribution .. فهيبقا surface وكإنو طالع من الصفحه .. فهيبقا عندك contour lines .. ال peak عند النص و كل contor بيعبر عن spherical Gaussian .. وال Width بيتحدد عن طريق ال covariance matrix ... احنا عندنا covariance matrix Segma .. ال Segma هيبقا ليه 4 ارقام ... sigma11, sigma12, sigma21, sigma22 .. علي ال diagonal هيبقا عندك ال variance .. ده اللي هو زي لما اتكلمنا علي 1D .. و ال off diagonal .. هيبقا ال covariance .. وده اللي بيقولنا ازاي ال different weights بت relate لبعض .. فهي ديه الفكره ..



دلوقت معانا خلاص ال expression بتاع ال posterior .. هنعوز نجيبلو ال maximum ..

Maximum A Posteriori

- Optimization:

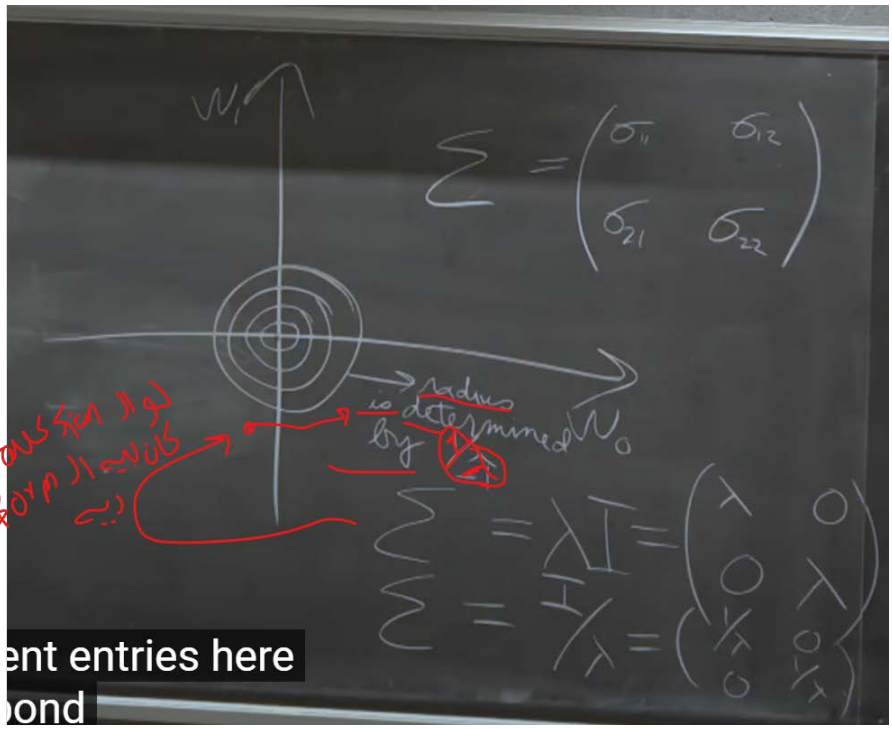
$$\boldsymbol{w}^* = argmax_{\boldsymbol{w}} \Pr(\boldsymbol{w}|\bar{\boldsymbol{X}}, \boldsymbol{y})$$
$$= argmax_{\boldsymbol{w}} \; -\sum_n \left(y_n - \boldsymbol{w}^T \bar{\boldsymbol{x}}_n\right)^2 - \boldsymbol{w}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{w}$$
$$= argmin_{\boldsymbol{w}} \sum_n \left(y_n - \boldsymbol{w}^T \bar{\boldsymbol{x}}_n\right)^2 + \boldsymbol{w}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{w}$$

- Let $\boldsymbol{\Sigma}^{-1} = \lambda \boldsymbol{I}$ then

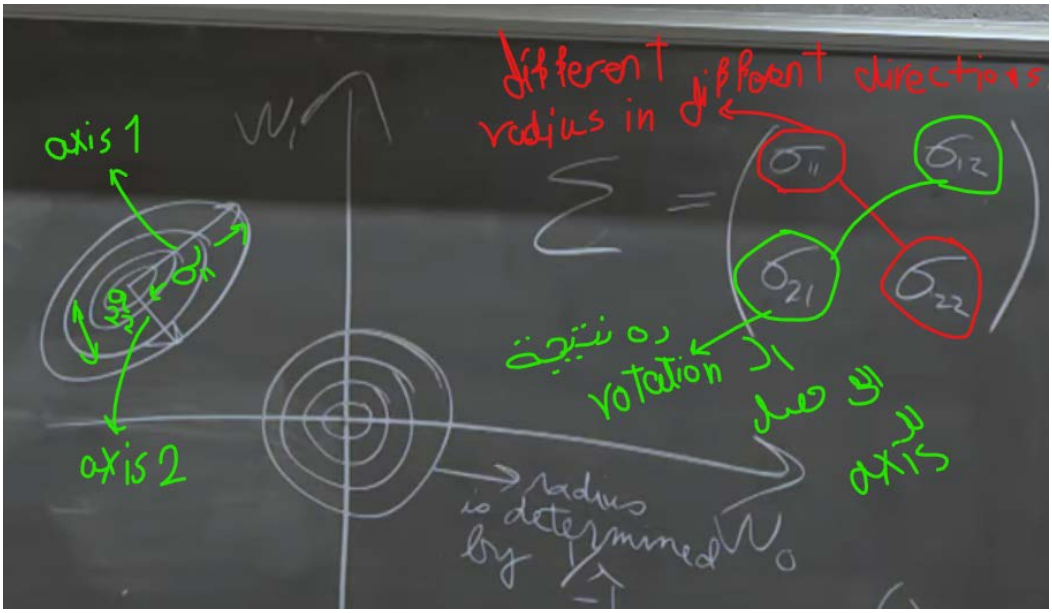
$$= argmin_{\boldsymbol{w}} \sum_n \left(y_n - \boldsymbol{w}^T \bar{\boldsymbol{x}}_n\right)^2 + \lambda ||\boldsymbol{w}||_2^2$$

- We arrive at the original **regularized** least square problem!

دلوقت انت عندك طريقه مختلفه لترجمة ال regularization term .. المحاضره اللي فاتت قلنا ان ال solution is stable عن طريق انك تحاول ت force ال entries of W انها تبقي as small as possible فهي هي لو قلت اني عاوز اقلل ال Norm of W .. اللي هو انا برضو عاوز أخلي ال entries يبقو قريبين من بعض ومن الصفر .. بس مكنش واضح ليه عاوزنها قريبه للصفر . ليه مثلاً مش 10 او 100 .. دده عشان ان ده بيجي نتيجة ان ال prior distribution بيديك ال knowledge اللي هو بنبقا عارفين فيه ان ال solution اللي هو أحسن W .. لازم تبقا قريبه للصفر .. وبعدين تبقا في منطقه فيها bell-shaped distribution .. فلو عندك ال prior distribution بتقولك ان الحلول القريبه من الصفر هي الأحسن ..



لو كان ال Gaussian ليه ال form التانيه اللي هي السجما الكامله



تعال نشوف ازاي ال 2 approaches بيدولنا losses مختلفه .. فدلوقت هن analysis ال Loss function .. وازاي الترمات بتختلف ... يبقا انت عندك different techniques بيدولك نفس ال W .. بس ال solution بيقا impacted by our choice of labda ... لكل لمضه مختلفه هيطلعلنا حلول مختلفه . لو انت given W . هنشوف ايه هو ال expected loss واحنا بنغير اللمضه ازاي ال expected loss هتتغير ... تعال نشوف ال derivations الي بتحصل ..

Expected Squared Loss

- Even though we use a statistical framework, it is interesting to evaluate the expected squared loss

data x,y from a distribution — Euclidean distance

$$E[L] = \int_{x,y} \Pr(x,y)\left(y - \boldsymbol{w}^T\bar{x}\right)^2 dxdy$$

$$= \int_{x,y} \Pr(x,y)\left(y - f(x) + f(x) - \boldsymbol{w}^T\bar{x}\right)^2 dxdy$$

$$= \int_{x,y} \Pr(x,y)\left[\left(y - f(x)\right)^2 + 2\left(y - f(x)\right)\left(f(x) - \boldsymbol{w}^T\bar{x}\right) + \left(f(x) - \boldsymbol{w}^T\bar{x}\right)^2\right] dxdy$$

Expectation with respect to y is 0

$$E[L] = \underbrace{\int_{x,y} \Pr(x,y)\left(y - f(x)\right)^2 dxdy}_{\text{noise (constant)}} + \underbrace{\int_x \Pr(x)\left(f(x) - \boldsymbol{w}^T\bar{x}\right)^2 dx}_{\text{error (depends on } \boldsymbol{w})}$$

exp and this term

Bias-Variance Decomposition

- Decompose squared loss

$$E_S\left[\left(f(\boldsymbol{x}) - \boldsymbol{w}_S^T\bar{\boldsymbol{x}}\right)^2\right]$$
$$= E_S\left[f(\boldsymbol{x}) - E_S\left[\boldsymbol{w}_S^T\bar{\boldsymbol{x}}\right] + E_S\left[\boldsymbol{w}_S^T\bar{\boldsymbol{x}}\right] - \boldsymbol{w}_S^T\bar{\boldsymbol{x}}\right]^2$$
$$= E_S\left[\left(f(\boldsymbol{x}) - E_S\left[\boldsymbol{w}_S^T\bar{\boldsymbol{x}}\right]\right)^2 + 2\left(f(\boldsymbol{x}) - E_S\left[\boldsymbol{w}_S^T\bar{\boldsymbol{x}}\right]\right)\underbrace{\left(E_S\left[\boldsymbol{w}_S^T\bar{\boldsymbol{x}}\right] - \boldsymbol{w}_S^T\bar{\boldsymbol{x}}\right)}_{\text{Expectation is 0}} + \left(E_S\left[\boldsymbol{w}_S^T\bar{\boldsymbol{x}}\right] - \boldsymbol{w}_S^T\bar{\boldsymbol{x}}\right)^2\right]$$
$$= \underbrace{\left(f(\boldsymbol{x}) - E_S\left[\boldsymbol{w}_S^T\bar{\boldsymbol{x}}\right]\right)^2}_{\text{bias}^2} + \underbrace{E_S\left[\left(E_S\left[\boldsymbol{w}_S^T\bar{\boldsymbol{x}}\right] - \boldsymbol{w}_S^T\bar{\boldsymbol{x}}\right)^2\right]}_{\text{variance}}$$

Bias-Variance Decomposition

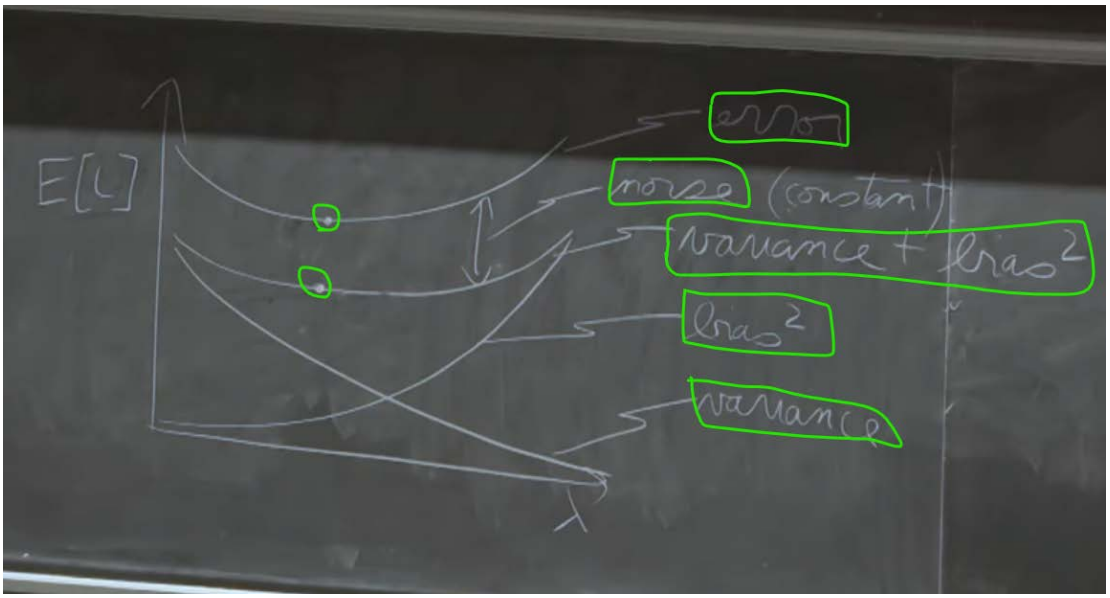
- Hence:

$$E[loss] = (bias)^2 + variance + noise$$

- Picture:

تعال نرسم صوره للرزع ده عشان لما نيجي نحل الاسيمنت نفهم ايه اللي بيحصل ..

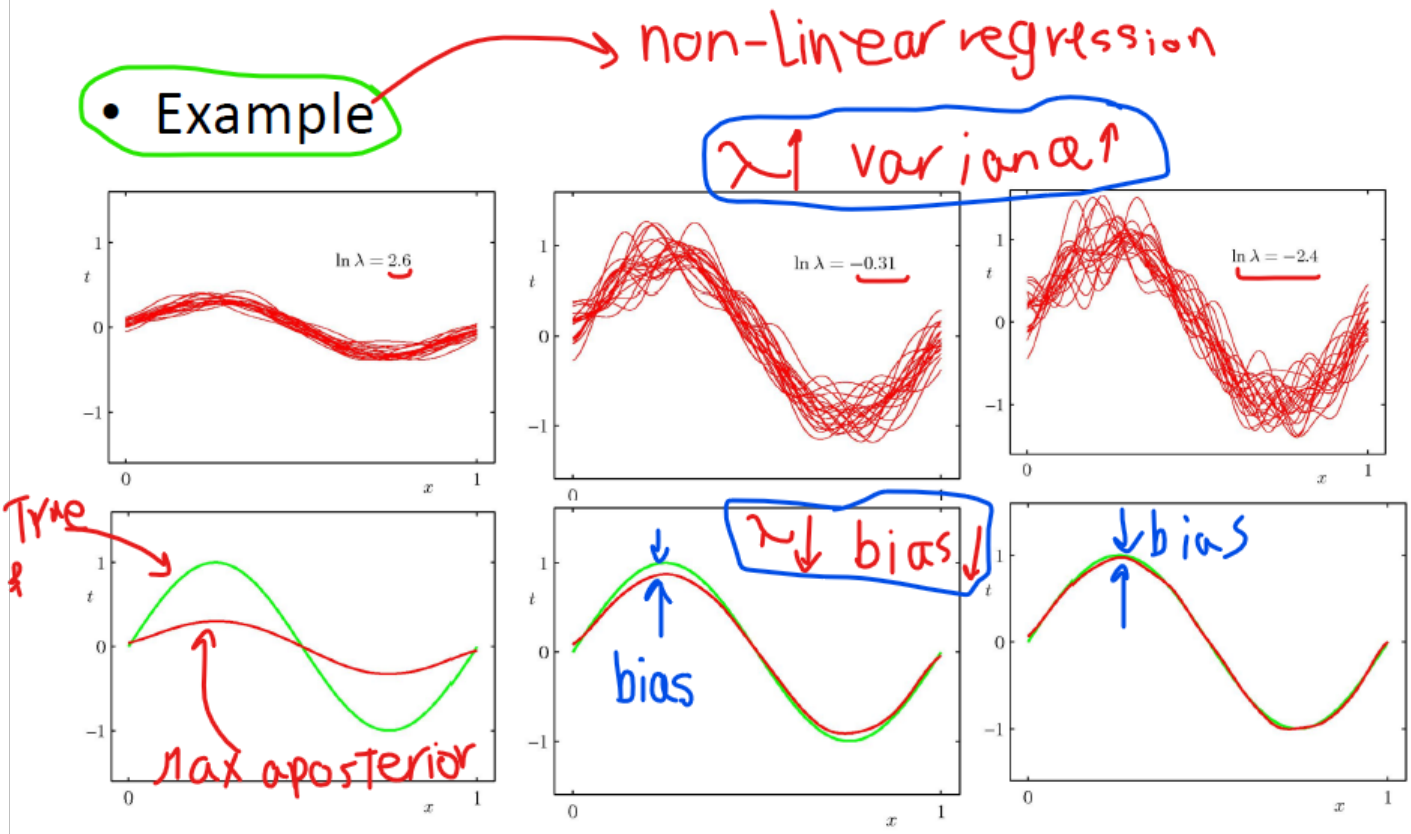
الواي أكسيس هو ال E[loss] .. و الإكس أكسيس هو ال Lambda .. في الاسيمنت انت هتغير اللمضه و تحدد لمضه كويسه باستخدام ال corss validation .. عشان اللمضه هايبر بارمتر ... انت هتلاحظ ال E[Loss] زي ما الدكتور راسمها كدا .. الكيرف ده عليه 3 حاجات ... نويز و بايس تربيع و فارينس .. لو شلت النويز هتلاقي الكيرف نزل لتحت شويه لإنو بقي عباره عن variance + bias ^2 .. في ال practice .. انت هتعوز تشوف لمضه بتجيبلك احسن قيمه لل error فين .. و بعدين تشيل منو النويز .. وهو ده ال sum of bias and variance



الدكتور بعد كدا دخل علي المثال اللي ف السلايد :

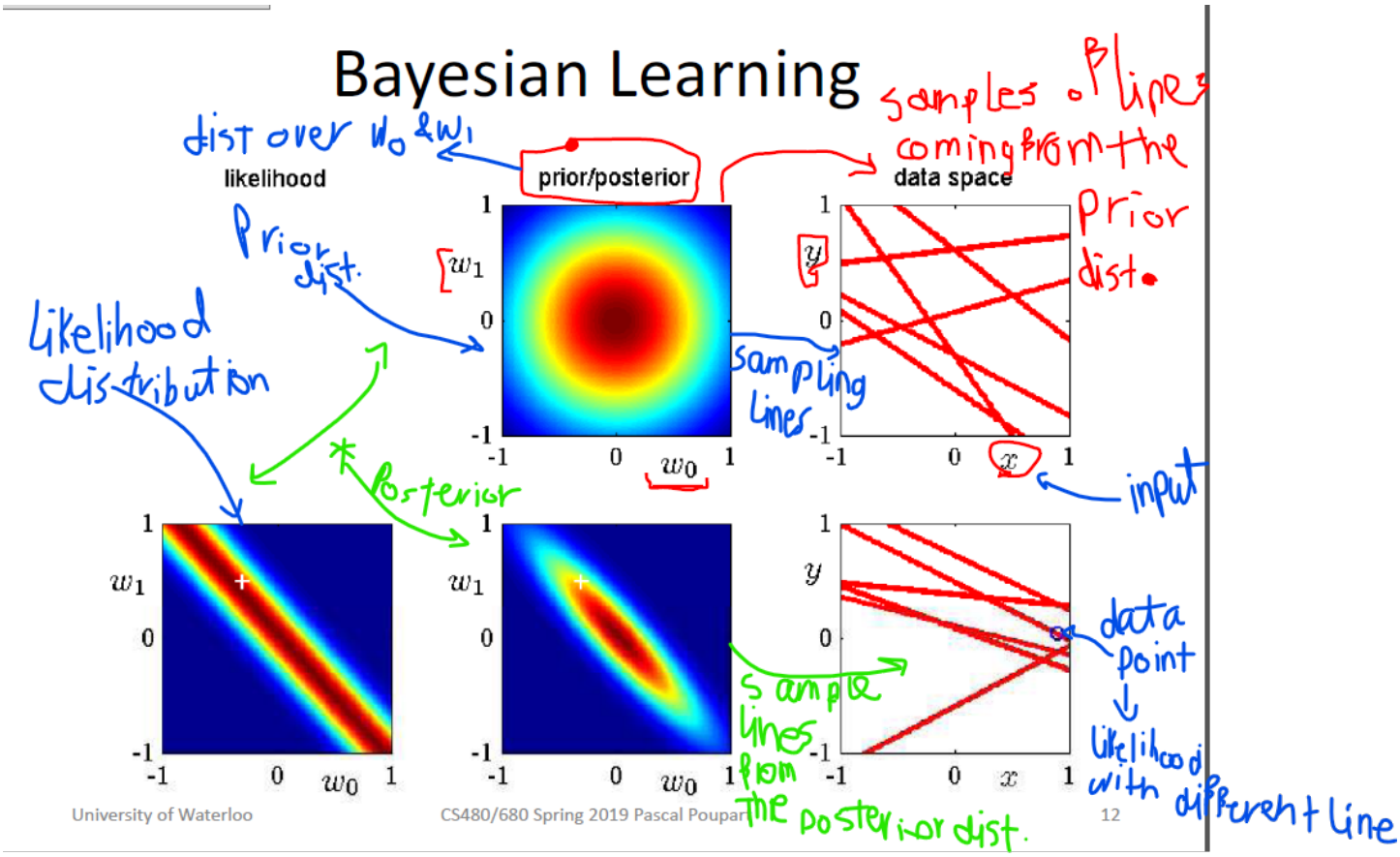
Bias-Variance Decomposition

- Example

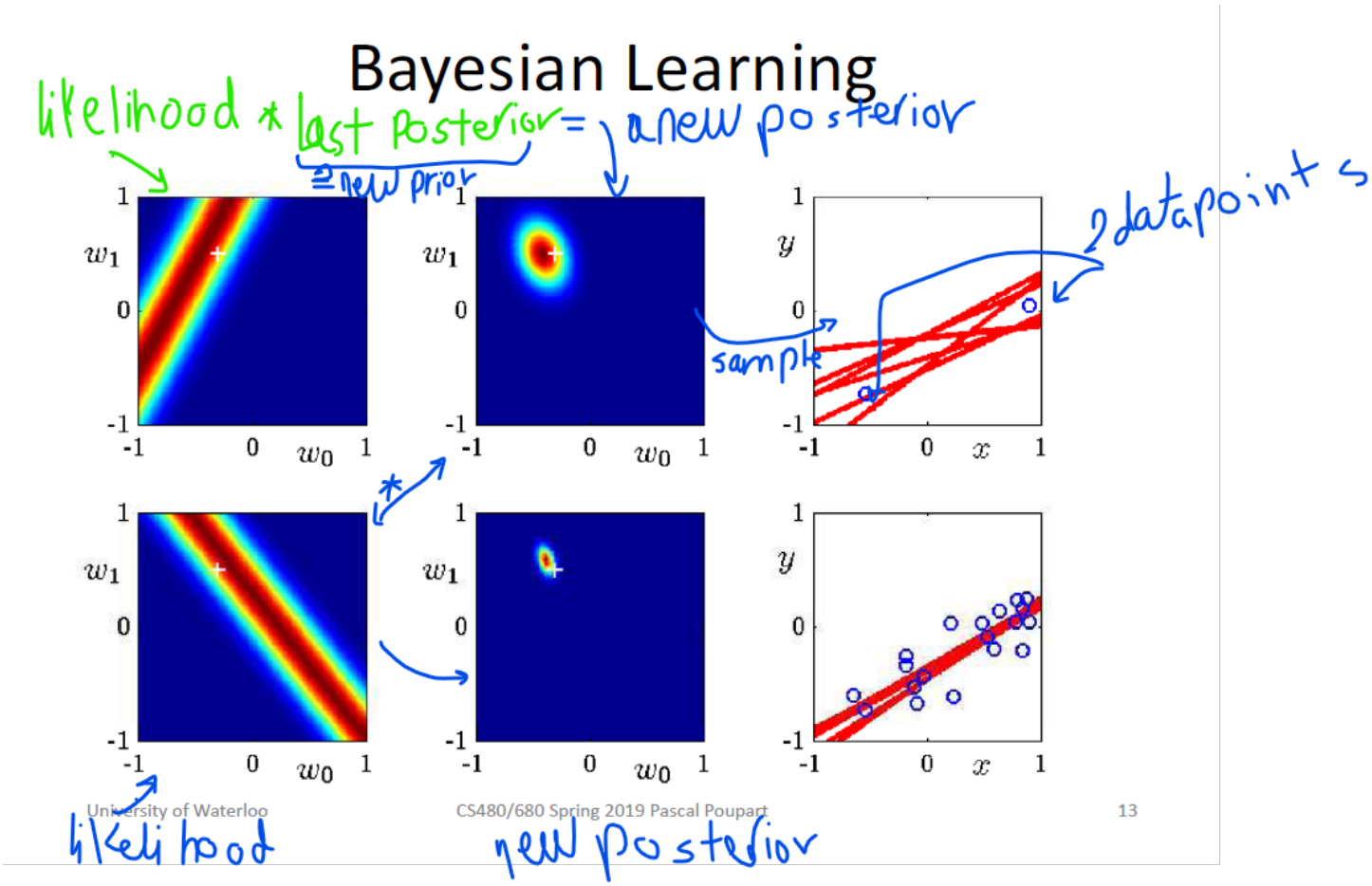


تعال نتكلم علي Bayesian linear regression .. في الحاله ديه بدل ما تاخد ال Hypothesis w اللي بت maximize the posterior .. انت حرفياً ممكن بس تحسب ال posterior و تشتغل بيها علي طول ... انت هنا هتشتغل بال posterior .. هن rewrite ال expression بس مش أكتر ... لما بنعمل Bayesian lerning .. احنا بنبدأ ب prior و نحسبال Posterior .. فهنا انت بتحسب ال posterior distribution ..

تعال نبص علي صوره



الصوره اللي فاتت ديه بعد خطوه واحده من ال observation for data point .. لو كمّلنا علي داتا بوينت تانيه .. هنلاقي الصوره الجايه ..



خد بالك من نقطه مهمه .. كل pair من W0 and W1 .. بيديلك خط .. ده عشان الخط ده مرسوم مع ال x and y values لما بستخدم ال w0 and w1 .

الدكتور بيقول اننا بنعمل predictions in Bayesian learning عن طريق انك تاخد weighted combinations of the predictions made by each hypothesis .. انت عندك داتا سيت X bar .. انت هتعمل prediction for a new input X star .. و هنتوقع ال y star .. ده نوع ال prediction اللي هنعوز نعملو ..

المشكله اني معنديش hypothesis واحده .. خد weighted combinations of the prediction of each mode where the weights are determined by posterior distribution .