MLE, BIC, AIC, and a little optimization

Rob McCulloch

1. Introduction

When we did Naive Bayes we had to estimate

$$p(X_i = x_i \mid Y = y) \text{ (or } p(x_i \mid y) \text{).}$$

How did we do it?

We simply used *the observed frequency*:

To estimate $p(X_i = x_i \mid Y = y)$:
in the training data, out of the times $Y = y$,
what fraction of observations have $X_i = x_i$.

If $X_i \sim \text{Bern}(p)$, we estimate p with the observed fraction of times $x_i = 1$.

We call p the *parameter* of the *statistical model* $X \sim \text{Bern}(p)$.

We consider a variety of statistical models and need to estimated the associated parameters.

For example, if we assume $Y_i \sim N(\mu, \sigma^2)$ then we have to estimate (μ, σ^2).

While the observed conditional frequency seems very reasonable for esitmating probabilities, we want a general approach to estimating the parameters of a statisical model.

Maximimum likelihood is a very general appproach which we will review.

This will also allow us to learn about BIC and AIC.
BIC and AIC are general approaches that allow us to choose simple models without having to do train/test splits.

In principle this is fantastic, but the approaches are just approximately correct.

Along the way, we will also review some very basic ideas from optimization.

2. Finding a Minimum, one variable

Let f be a function of a single variable, so that $f(x)$ is an number for $x \in C \subset R$.

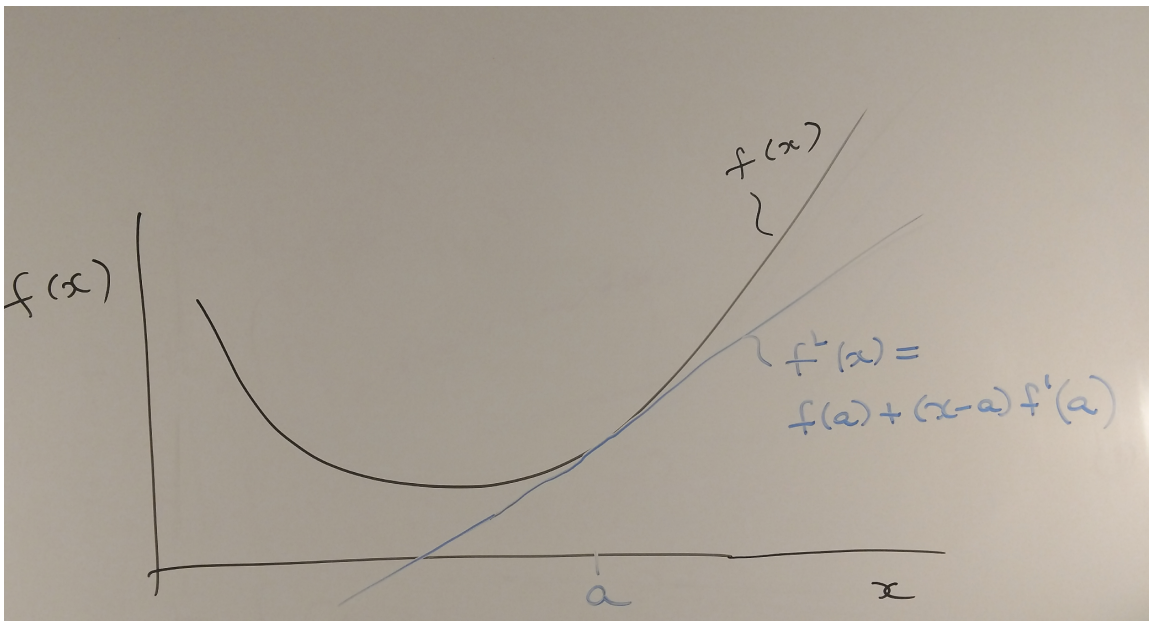
x_0 is a local minimum if $f(x) \geq f(x_0)$ for all x close to x_0.

x_0 is a global minimum if $f(x) \geq f(x_0)$ for all $x \in C$.

Recall:

The derivative gives you a linear approximation to the function:

$$f(x) - f(a) \approx (x - a)f'(a).$$



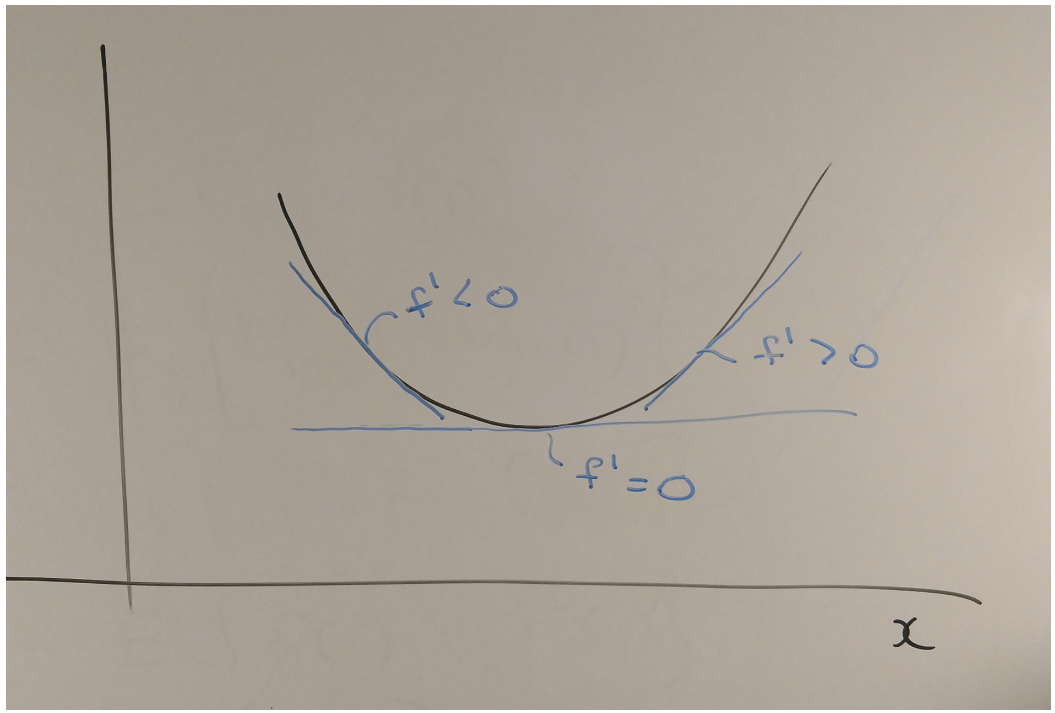
For x close to a, $f(x) \approx f^L(x)$.

Neccessary Condition:

If x_0 is a local min (or max) then $f'(x_0) = 0$.

Sufficient Condition:

If $f'(x_0) = 0$ and $f''(x_0) >$, then x_0 is a local minimum.

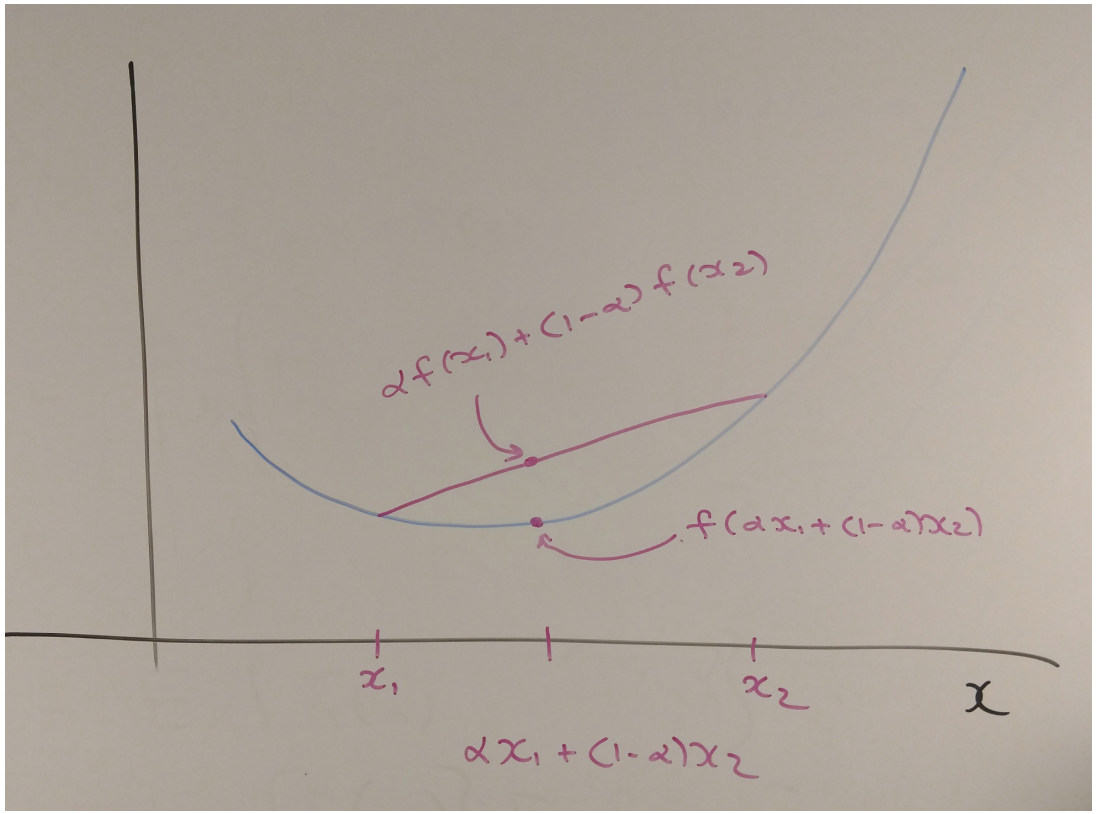


At a local minimum, the derivative is increasing.

Global Sufficient Condition

f is convex if

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha)f(x_2), \alpha \in [0,1].$$



If f is convex and $f'(x_0) = 0$, then x_0 is a global minimum.

We us optimization *a lot* in Machine Learning.

In particular, learning on the training data is often done by some kind of optimization.

For example, in the model $y_i \approx \beta' x_i$ we learn (*estimate*) β by solving

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} (y_i - \beta' x_i)^2$$

We will spend a chunk of time on versions of this problem.

3. Maximum Likelihood, the Bernoulli

Suppose we have a statistical model

$$Y \sim f(y \mid \theta)$$

where θ is the parameter (possibly a vector).

Given data $Y = y$ how can we estimate θ?

Maximum Likelihood:

Choose the θ that makes what you have seen most likely:

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\, f(y \mid \theta)$$

In the iid case, we have $Y = (Y_1, Y_2, \ldots, Y_n)$ with

$$Y_i \sim f(y \mid \theta) \;\; iid,$$

so

$$f(y \mid \theta) = \prod_{i=1}^{n} f(y_i \mid \theta),$$

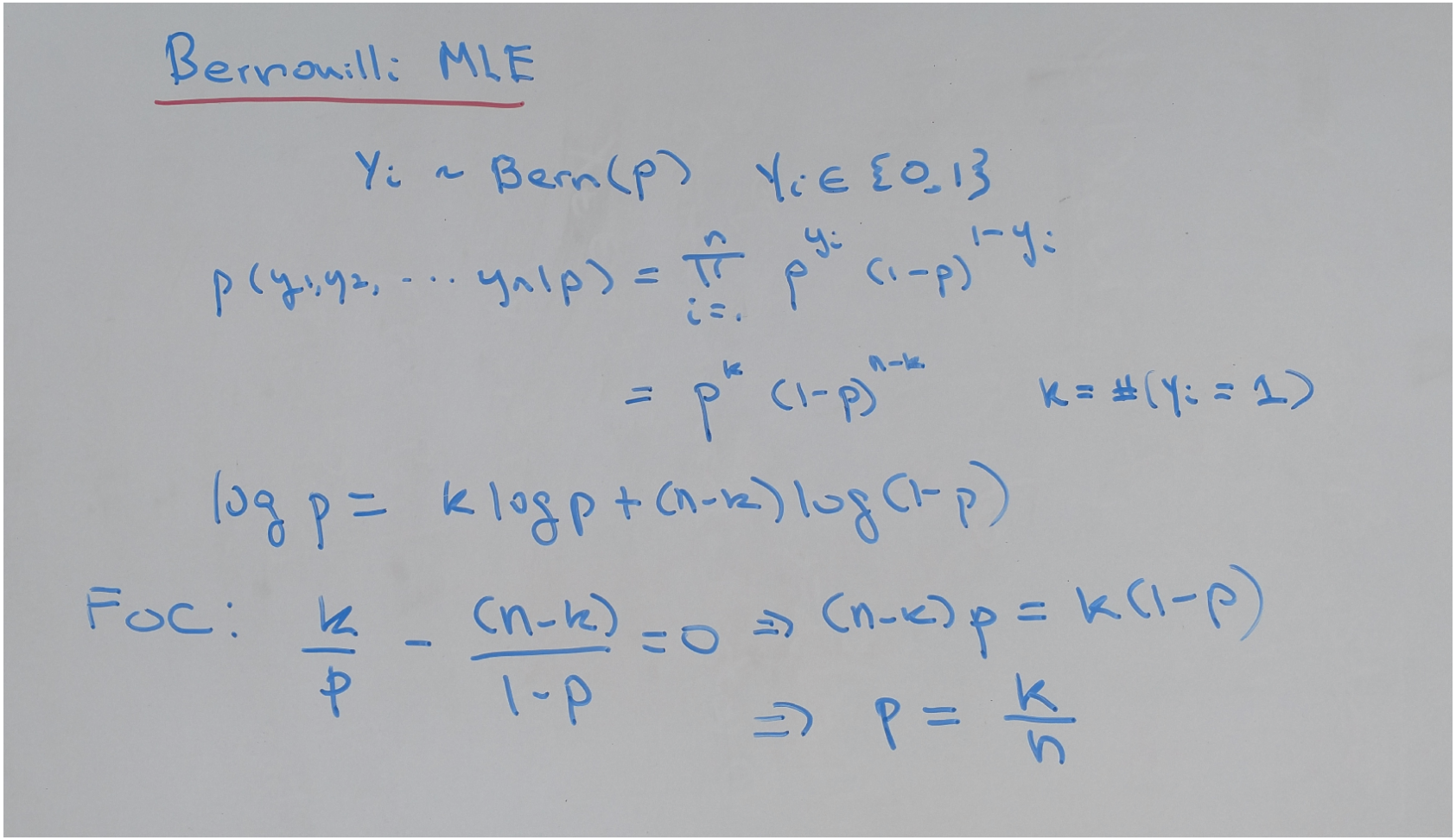
and the MLE is

$$\hat{\theta} = \underset{\theta}{\text{argmax}} \prod_{i=1}^{n} f(y_i \mid \theta).$$

Note:

$f(y \mid \theta)$ viewed as a function of θ for a fixed y is called the likelihood function.

In practice we often maximize the log of the likelihood or minimize the negative of the log likelihood.



FOC: "first order condition", $f' = 0$.

So, *the observed sample frequency is the MLE!*

4. Projecting onto a vector

Let x and $y \in R^n$.

So, for example, $x = (x_1, x_2, \ldots, x_n)'$.

We will find the solution to the following problem very useful:

$$\min_{\beta \in R} ||y - \beta x||^2$$

where $||x||^2 = \sum x_i^2$.

Recall:

$x, y \in R^n$,

The inner product is

$$< x, y >= x'y = y'x = \sum x_i y_i.$$

The L^2 or Euclidean norm (squared) is

$$||x||^2 =< x, x >= x'x = \sum x_i^2$$

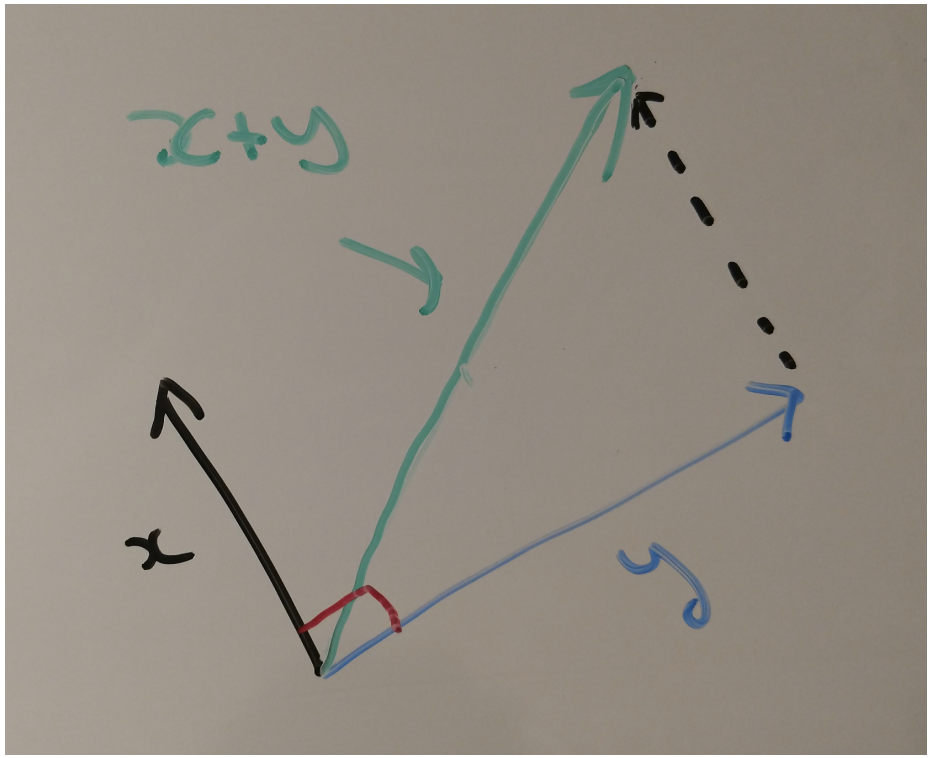
x and y are orthogonal if

$$< x, y >= 0$$

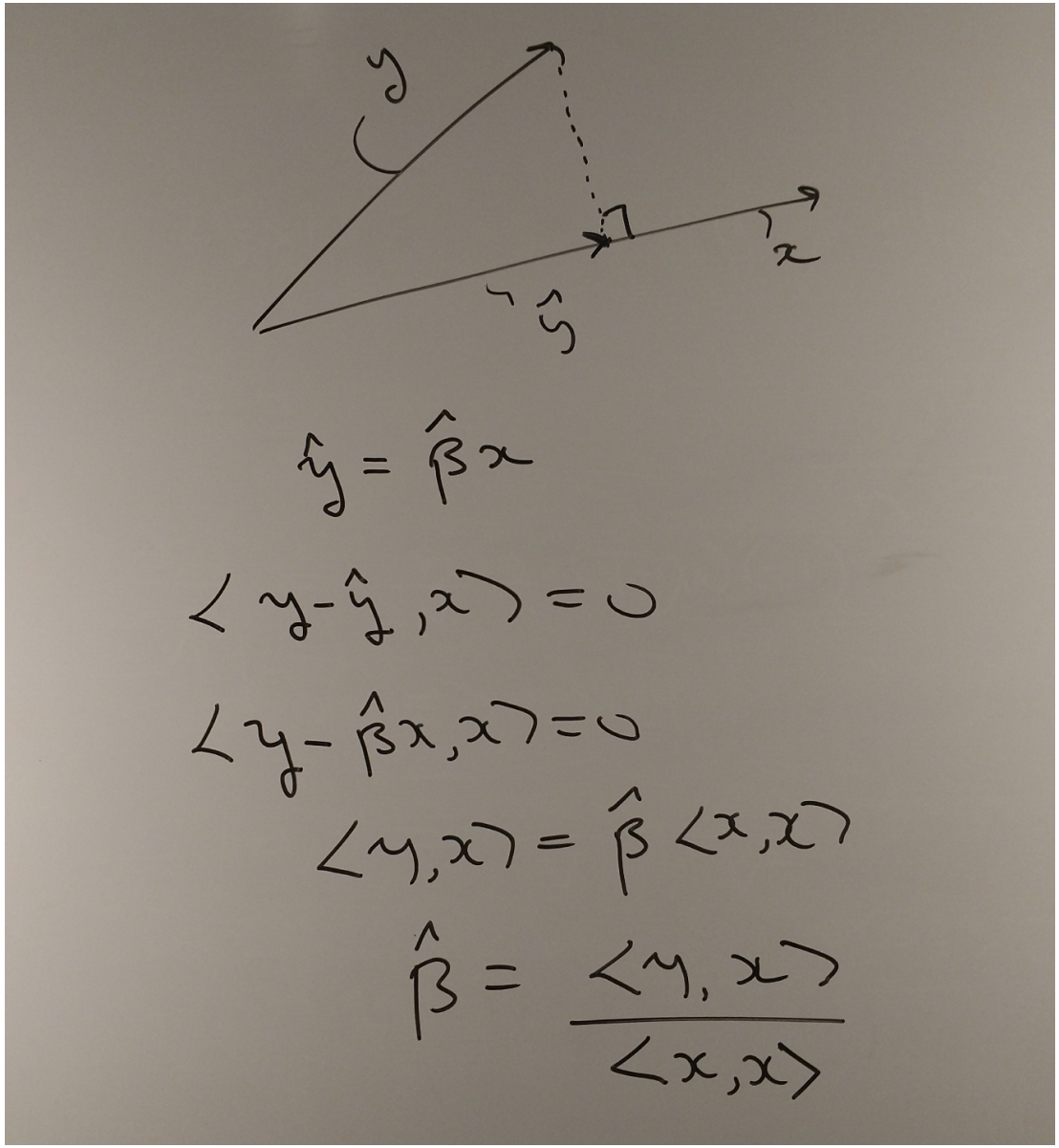
Note:

If x and y are orthogonal:

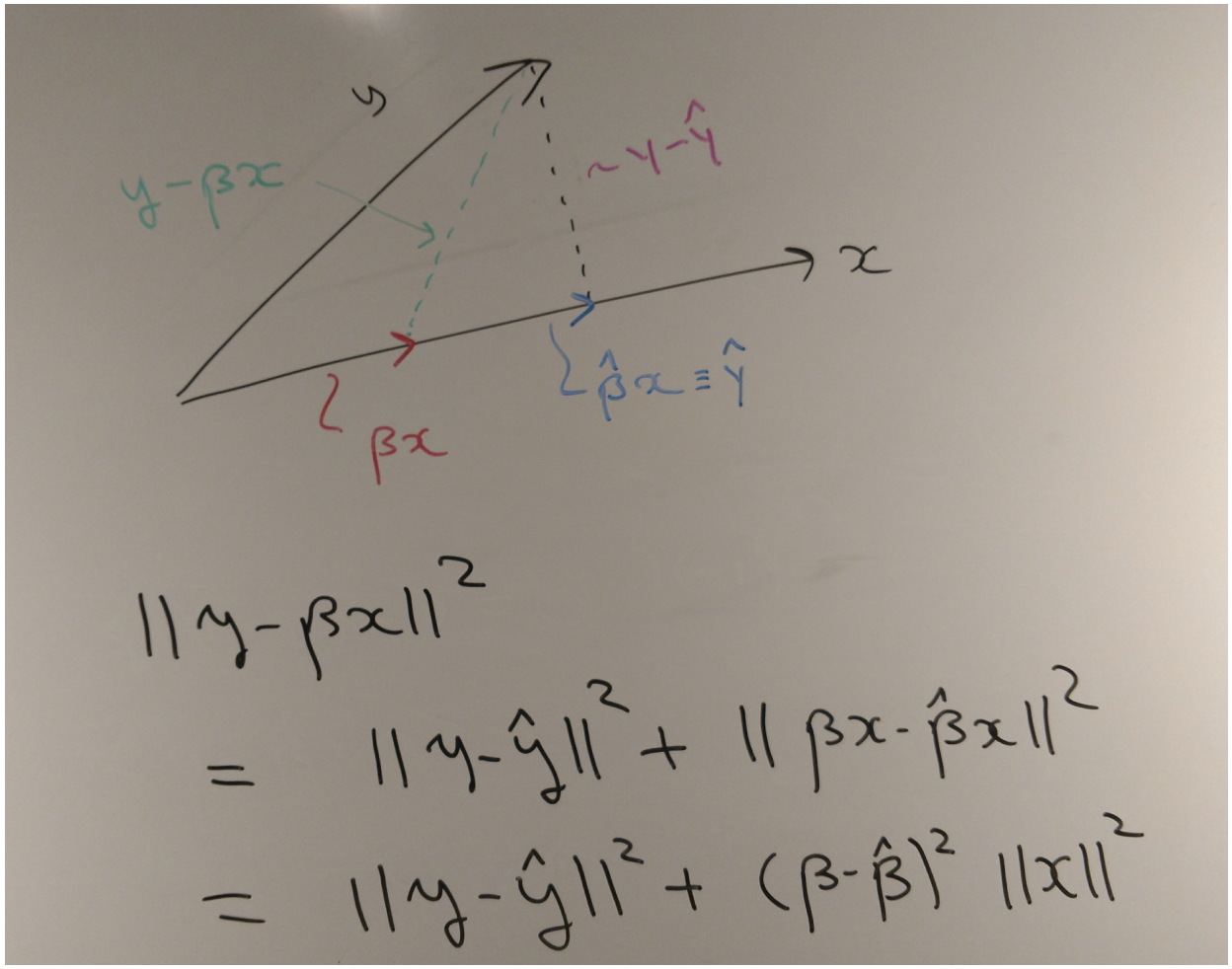
$$\begin{aligned} ||x+y||^2 &= < x+y, x+y > \\ &= < x, x > +2 < x, y > + < y, y > \\ &= ||x||^2 + ||y||^2 \end{aligned}$$



$\hat{y}$ is the orthogonal projection of y onto x.



To solve our problem we have



So that obviously the min is obtained at $\beta^* = \hat{\beta}$.

5. Finding a Minimum, Several Variables

Now suppose $x = (x_1, x_2, \ldots, x_p)'$

and $f(x) = f(x_1, x_2, \ldots, x_p) \in R$.

How do we solve:

$$\min_x f(x)$$

The Gradient:

$$\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_p} \right]$$

where

$$\frac{\partial f(x)}{\partial x_i}$$

is what you get by holding all the x_j, $j \neq i$ fixed, and then differentiating with respect to x_i.

The gradient is a multivariate derivative in that (skipping some technical details):

$$f(x) \approx f(a) + \nabla f(a)(x - a)$$

Note that $\nabla f(x)$ is a row vector so that the product above makes sense with x a column vector.

An alternative notation is:

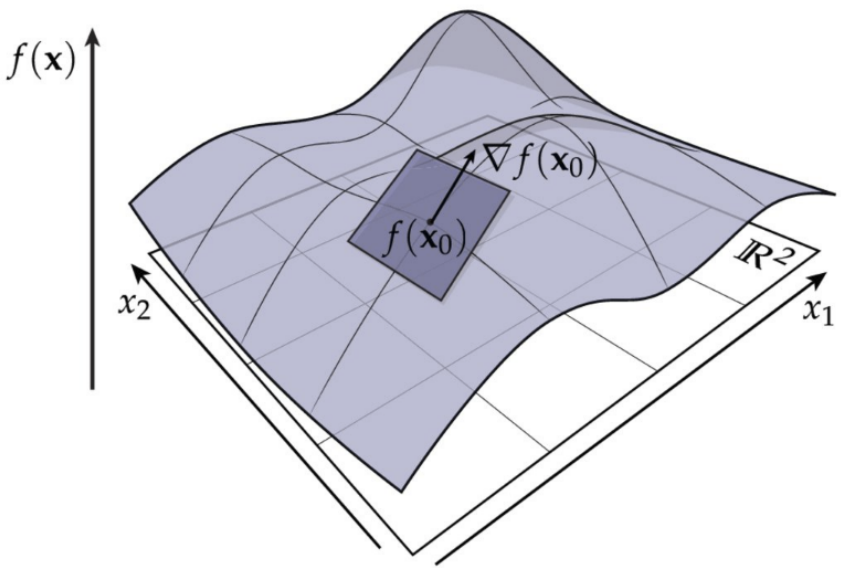
$$f(x) \approx f(a) + < \nabla f(a), (x - a) >$$

Stolen off the web:

Gradient as Best Linear Approximation

Another way to think about it: at each point x0, gradient is the vector $\nabla f(\mathbf{x}_0)$ that leads to the best possible approximation

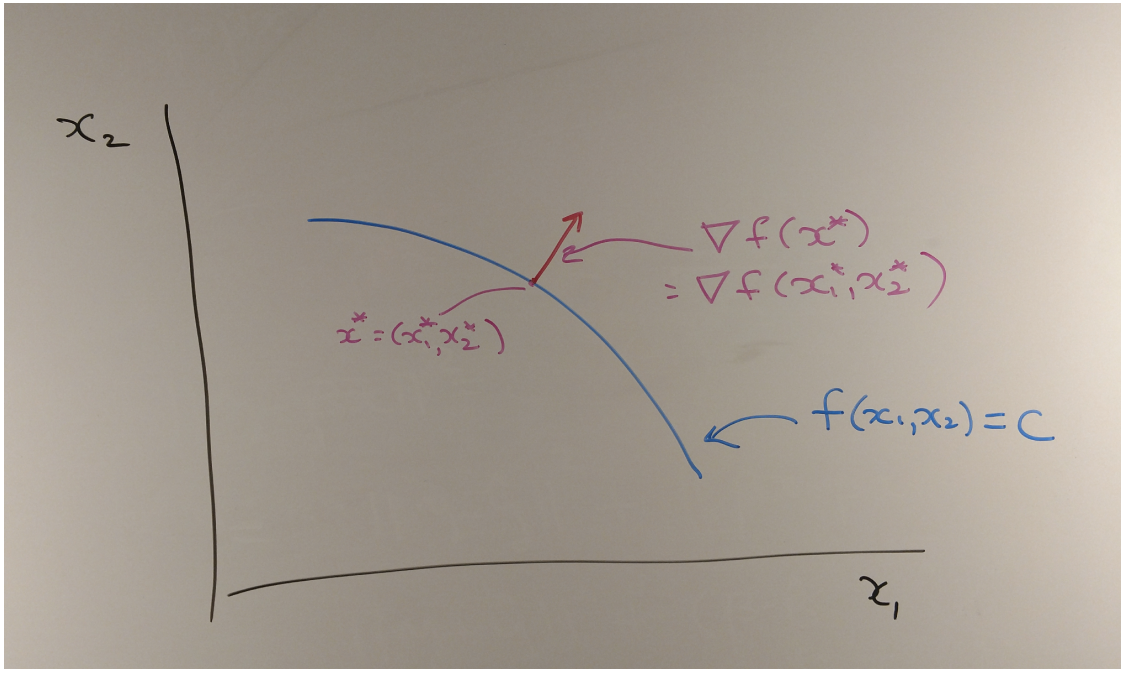
$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle$$



Starting at x_0, this term gets:

- **bigger if we move in the direction of the gradient,**
- **smaller if we move in the opposite direction, and**
- **doesn't change if we move orthogonal to gradient.**

We can visualize the gradient using the *contours* of f.
A *contour* is the set $\{x : f(x) = c\}$.



- If you want to increase f as fast as possible, go in the direction of the gradient ∇f.
- If you want to decrease f as fast as possible, go in the direction of the negative gradient $-\nabla f$.
- If you want to move without changing f to in a direction orthogonal to the gradient.

Neccessary Condition for a local min/max:

If x^* is a local min (or max) then we must have

$$\nabla f(x^*) = 0$$

Again f is convex if,

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha)f(x_2), \alpha \in [0,1].$$

exactly as before except that now x denotes a vector $\in R^p$.

As before, if f is convex, then a local minimum is a global minimum.

Let's defer multivariate conditions for convexity and the "second derivative".

6. Maximum Likelihood, the normal

Suppose

$$Y_i \sim N(\mu, \sigma^2), \quad \text{iid}$$

what is the MLE of $\theta = (\mu, \sigma^2)$?

$$f(y|\mu,\sigma^2) = \prod f(y_i|\mu,\sigma^2)$$

$$= \prod \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$$

$$= (2\pi)^{\frac{-n}{2}} \sigma^{-n} e^{-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2}$$

$$-\log L(\mu,\sigma^2) = \frac{n}{2}\log(2\pi) + n\log\sigma + \frac{1}{2\sigma^2}\sum(y_i-\mu)^2$$

Let $v = \sigma^2$

$$= \frac{n}{2}\log(2\pi) + \frac{n}{2}\log(v) + \frac{1}{2v}\sum(y_i-\mu)^2$$

$$-2\log L(\mu,v) = n\log(2\pi) + n\log(v) + \frac{1}{v}\sum(y_i-\mu)^2$$

$$\underline{1} = (1, 1, 1, \ldots 1)'$$

$$\sum (y_i - \mu)^2 = \| y - \mu \underline{1} \|^2$$

$$\hat{\mu} = \frac{\langle y, \underline{1} \rangle}{\langle \underline{1}, \underline{1} \rangle} = \frac{\sum y_i}{n} = \bar{y}$$

$$\sum (y_i - \mu)^2 = \| y - \bar{y} \underline{1} \|^2 + \| \mu \underline{1} - \bar{y} \underline{1} \|^2$$

$$= \sum (y_i - \bar{y})^2 + n (\bar{y} - \mu)^2$$

$S = \sum(y_i - \bar{y})^2$.

$$-2\log L = c + n\log(v) + \frac{1}{v}\left[S + n(\bar{y} - \mu)^2\right]$$

$$\frac{\partial}{\partial \mu} = \frac{n}{v} 2(\bar{y} - \mu)(-1)$$

$$\Rightarrow \mu^* = \bar{y}$$

$$\frac{\partial}{\partial v}(\text{at } \mu^*) = \frac{n}{v} - \frac{S}{v^2}$$

$$v^* = \frac{S}{n} = \frac{\sum(y_i - \bar{y})^2}{n}$$

7. Simple Linear Regression

The simple linear regression model is:

$$Y_i \mid x_i, \beta_0, \beta_1, \sigma^2 \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

or,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2), \text{ iid.}$$

$$\sum(x_i-\bar{x}) = n\frac{\sum x_i}{n} - n\bar{x} = 0$$

Note: $\sum(x_i-\bar{x})(y_i-\bar{y}) = \sum(x_i-\bar{x})y_i - \bar{y}\sum(x_i-\bar{x})$

$= \sum(x_i-\bar{x})y_i$

$$p(y_1,y_2,\ldots y_n|\beta_0,\beta_1,\sigma^2) = \prod_i \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma} e^{-\frac{1}{2\sigma^2}(y_i-\beta_0-\beta_1 x_i)^2}$$

$$= (2\pi)^{-\frac{n}{2}} V^{-\frac{n}{2}} e^{-\frac{1}{2V}\sum(y_i-\beta_0-\beta_1 x_i)^2}$$

$$-2\log L = n\log(2\pi) + n\log(V) + \frac{1}{V}\sum(y_i-\beta_0-\beta_1 x_i)^2$$

$$\frac{\partial}{\partial\beta_0} = 0: \quad \sum(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i) = 0$$

$$\sum y_i - n\hat{\beta}_0 - \hat{\beta}_1\sum x_i = 0$$

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}}$$

$$\frac{\partial}{\partial\beta_1} = 0: \quad \sum(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)x_i = 0$$

$$\sum((y_i-\bar{y})-\hat{\beta}_1(x_i-\bar{x}))x_i = 0$$

$$\sum(y_i-\bar{y})(x_i-\bar{x}) - \hat{\beta}_1\sum(x_i-\bar{x})^2 = 0$$

$$\boxed{\hat{\beta}_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}}$$

$$\frac{\partial}{\partial V} = 0: \quad \boxed{\hat{\sigma}^2 = \hat{V} = \frac{\sum(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2}{n}}$$

8. The Multinoulli MLE

The fundamental Bernoulli random variable considers the case where something is about to happen or not and we code one possibility up as a 1 and the other as a 0.

The Multinoulli distribution consider the more general case where there is a a set of k possible outcomes.

For example, if we survey a customer and ask them to rate our product on a 1-5 scale then there are 5 possible outcomes.

Let $\{1, 2, \ldots, k\}$ denote the possible outcomes for Y.

Let

$$p = (p_1, p_2, \ldots, p_k)$$

with

$$P(Y = j \mid p) = p_j$$

Then

$$Y \sim \text{Multinoulli}(p)$$

Given $Y_i \sim \text{Multinoulli}(p)$ we want to compute the MLE of p.

$$Y_{ij} = \begin{cases} 1 & \text{if } Y_i = j \\ 0 & \text{else} \end{cases} \qquad \begin{matrix} i = 1,2,\cdots n \\ j = 1,2,\cdots k \end{matrix}$$

$$p(y_1, y_2, \cdots y_n | p) = \prod_i p_1^{y_{i1}} p_2^{y_{i2}} \cdots p_k^{y_{ik}}$$

$$= p_1^{m_1} p_2^{m_2} \cdots p_k^{m_k}$$

$$m_j = \sum_i Y_{ij} = \left[\#\text{ of times } Y_i = j \right]$$

How do we maximize this likelihood?

With just two possible outcomes we had one variable,
$p = P(Y = 1)$.

Now we have $p_j, j = 1, 2, \ldots, k$ with the constraint $\sum p_j = 1$.

We also have $0 \leq p_j \leq 1$, but we won't have to worry about this.

We could let $p_k = 1 - \sum_{j=1}^{k-1}$ and then optimize over
$(p_1, p_2, \ldots, p_{k-1})$.

But, it is easier to use *lagrange multipliers*.

9. Lagrange Multiplier

Let $x \in R^p$.

We want to solve:

$$\min_x f(x), \quad \text{subject to} \quad g(x) = 0$$

Let

$$\mathcal{L}(x, \lambda) = f(x) + \lambda g(x)$$

and then minimize $\mathcal{L}$ unconstrained over (x, λ).

Differentiating $\mathcal{L}$ with respect to λ gives:

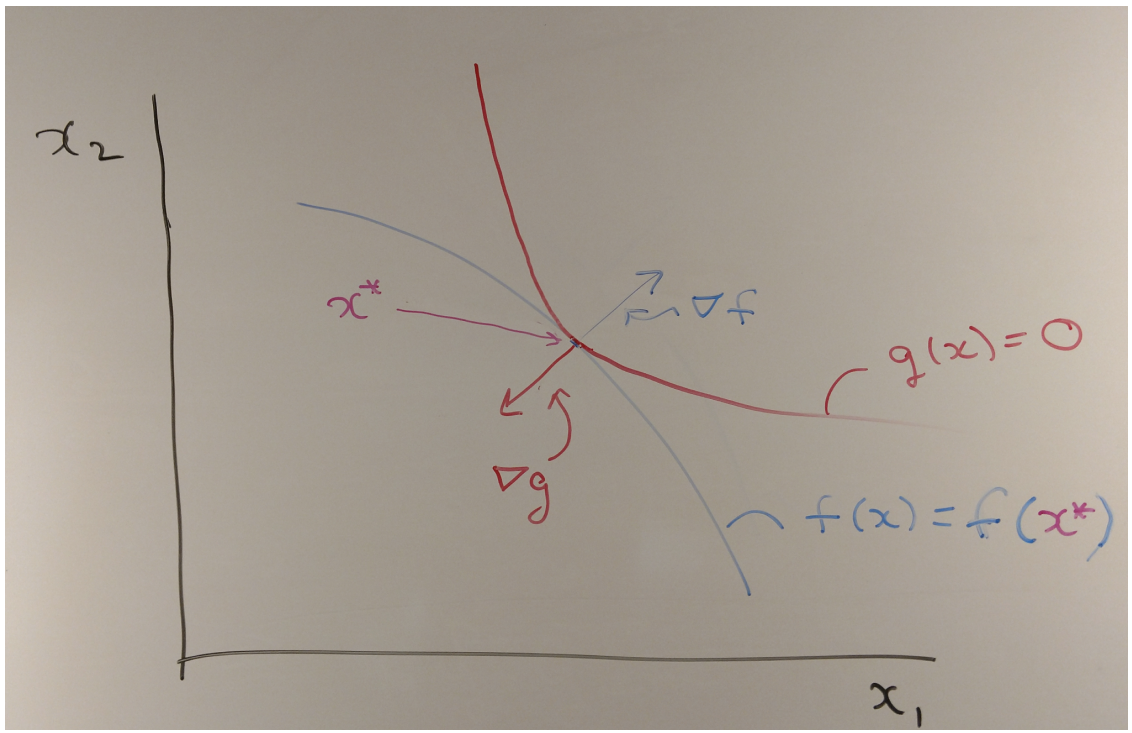
$$g(x) = 0$$

at the min/max.

Differentiating $\mathcal{L}$ with respect to x give:

$$\nabla f(x) + \lambda \nabla g(x) = 0$$

at a local min (or max).
Because of the constraint $g(x) = 0$ you can only move orthogonal to ∇g.



But, $\nabla f \propto \nabla g$, tells you that "small" moves orthogonal to ∇g will not change f so it is a local minimum or maximum.

10. The Multinoulli MLE again

To obtain the Multinoulli MLE we will have

$$L(p) = \prod p_j^{m_j}$$

and we maximize this subject to

$$\sum p_j = 1.$$

We will max the log likelihood:

$$\mathcal{L}(p, \lambda) = \sum_j m_j log(p_j) + \lambda(\sum_j p_j - 1)$$

$$L = \sum m_k \log p_k + \lambda\left(\sum p_k - 1\right)$$

$$\frac{\partial L}{\partial p_k} = \frac{m_k}{p_k} + \lambda$$

$$\Rightarrow \quad p_k \propto m_k$$

$$\Rightarrow \quad \hat{p}_k = \frac{m_k}{\sum m_k} = \frac{m_k}{n}.$$

The MLE is the observed sample frequency.

11. KKT

We will have occasion to consider constraint sets of the form

$$g(x) \leq 0$$

rather than just

$$g(x) = 0$$

The Karush-Kuhn-Tucker conditions cover both inequality and equality constraints.

We'll see how things change with one inequality constraint and then state the general result.

KKT:

To minimize $f(x)$ subject to $g(x) \leq 0$, form

$$L(x, \alpha) = f(x) + \alpha g(x)$$

and then solve

$$min_x max_{\alpha, \alpha \geq 0} \; L(x, \alpha).$$

With $\alpha \geq 0$ we must have $g(x) \leq 0$, since otherwise we could get a max of infinity.

Also note that at the solution:

$$\alpha^* \, g(x^*) = 0.$$

This captures the fact that there are two possibilities:

- If the constraint is *binding* then $g(x^*) = 0$ and we can have $\alpha^* > 0$.
- If the constraint is not binding so that $g(x^*) < 0$ then the max over non-negative α is clearly obtained at $\alpha^* = 0$.

If $g(x) < 0$ ($\alpha = 0$) at the optimal value then the constraint is not binding and we can just use our usual solve $\nabla f = 0$ approach.

If $g(x) = 0$ ($\alpha > 0$) then the KKT result says we can solve the unconstrained problem of minimizing:

$$\min f(x) + \alpha\, g(x).$$

As before, the term

$$\min f(x) + \alpha\, g(x)$$

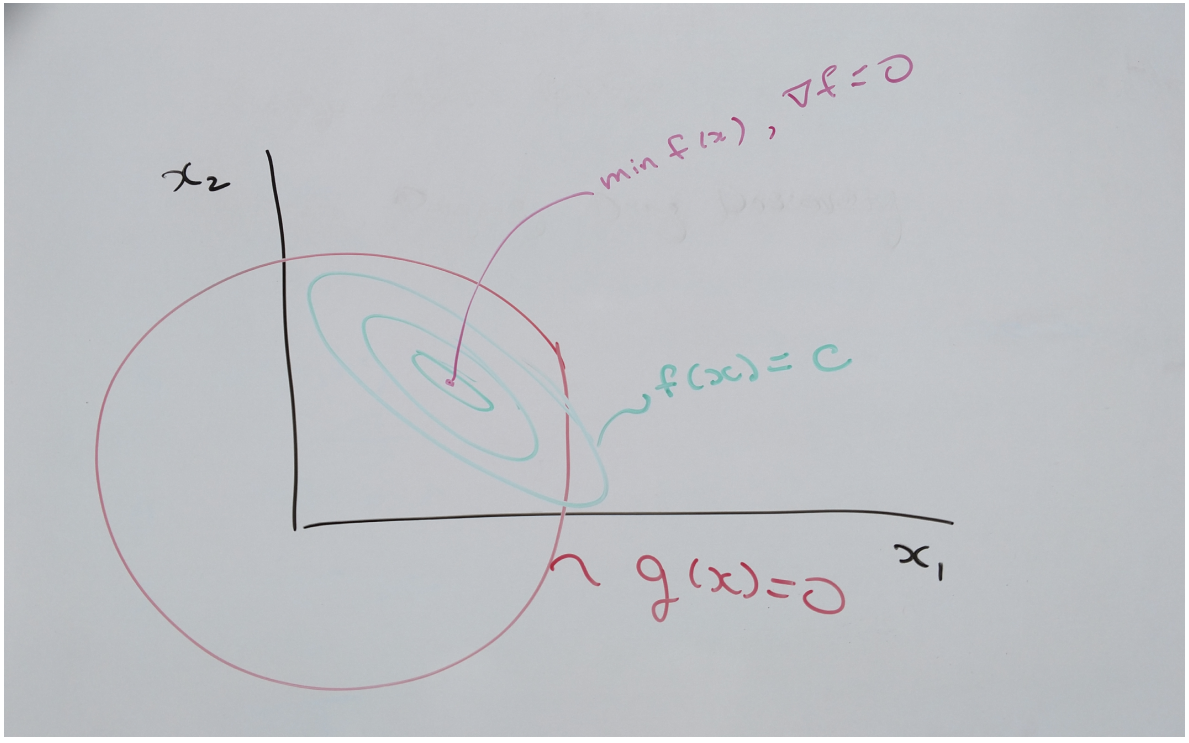
is called the "lagrangian" and α is the lagrange multiplier.

The FOC (first order condition) associate with the lagrangian is:

$$\nabla f(x) + \alpha\, \nabla g(x) = 0.$$

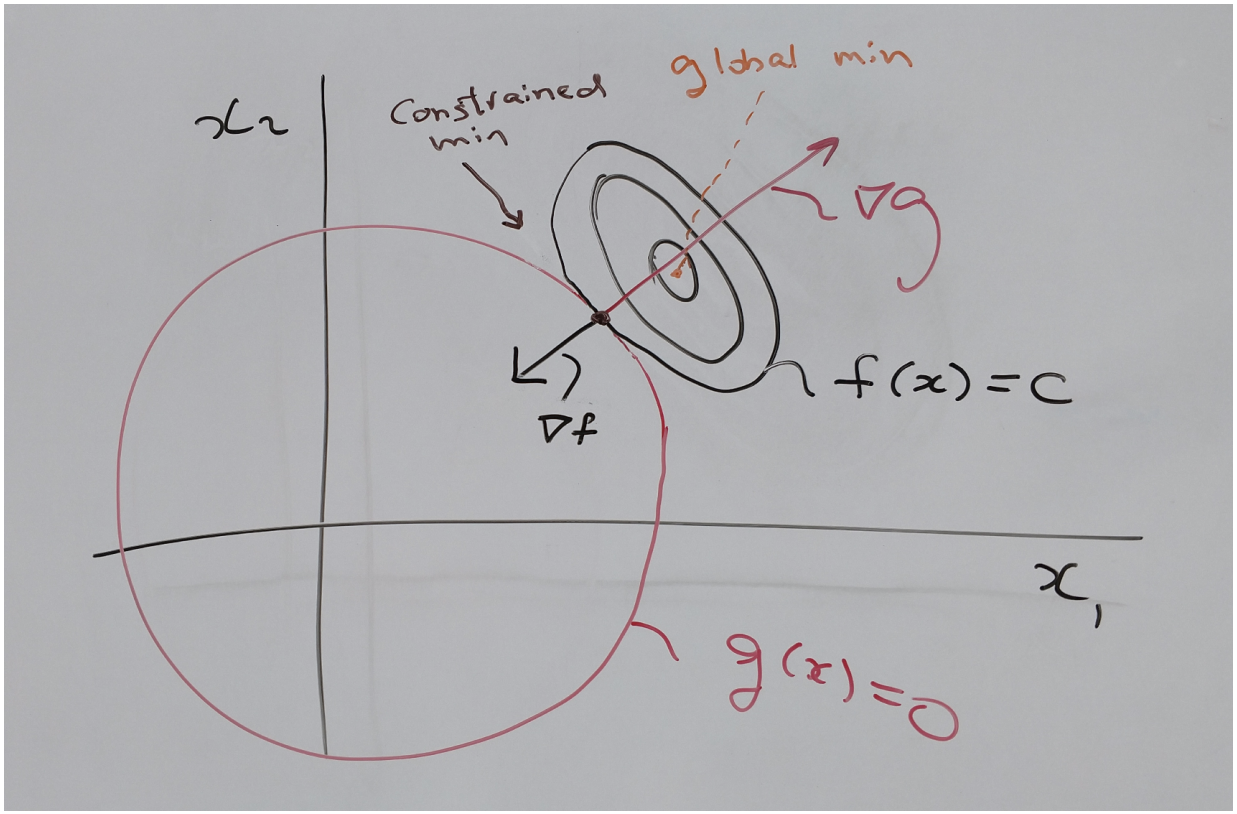
Here is the case where the constraint is not binding.

The global min is in the interior of the set $g(x) \leq 0$.



Here is the key picture for the case where the constraint is binding.

Remember, ∇f is the direction in which f goes up the fastest!!
∇f points perpendicularly to the contour of f.



It is intuitive that $\nabla f + \alpha \nabla g = 0$ with $\alpha > 0$.

The general form of the KKT theorem.

Just notice that with equality constraints you don't know the sign of the constraint coefficient.

$$\min f(x)$$

$$\text{s.t.}: \begin{cases} h_i(x) = 0 \\ g_j(x) \leq 0 \end{cases}$$

$$L(x, \lambda, \alpha) = f(x) + \sum \lambda_i h_i(x) + \sum \alpha_j g_j(x)$$

$$\min_{x} \max_{\lambda} \max_{\alpha, \alpha \geq 0} L(x, \lambda, \alpha)$$

Example:

What happens when we do

$$\min_{x:||x||<c} a'x$$

What happens when we do

$$\max_{x:||x||<c} a'x$$

12. Deviance, BIC and AIC

We have used versions of train/test data splits to deal with the bias-variance tradeoff.

The bias-variance tradeoff teaches us that our in sample fit may not tell us about our out of sample predictive performance, which is what we really care about.

But, any train/test split strategy we employ may be a very crude approximation to our true expected loss!!

The bias-variance tradeoff tells us that the most "complicated" model may not be the best model!

Can we *theoretically* develop procedures that guide us towards simpler models?

AIC and BIC are widely used attempts in this direction.

Something you may have seen that is in the same vein, is adjusted R^2 in linear regression.

You may have been taught that $R^2 = \text{cor}(y, \hat{y})^2$, always goes up when you add x variables to a multiple regression.

You may have been taught that this is a problem!!

The *adjusted* R^2 is supposed to account for this, so that people are often told to look at the adjusted R^2 instead of R^2.

BIC:

The BIC is the *Bayesian information criterion*.

The BIC is an approximation to the Bayesian approach to model selection which gives you an automatic penalty for complex models.

The Bayesian Approach to Model Selection:

(See section 7.7 of ELS)

Suppose we have a set of candidate models

$\mathcal{M}_m$, $m = 1, 2, \ldots, M$.

Model $\mathcal{M}_m$ has parameter vector θ_m associate with it and

$$p(Z \mid \theta_m, \mathcal{M}_m)$$

represents the model of the data Z under model $\mathcal{M}_m$.

Example:

Consider the situation where we observe (x_i, y_i) and we want to see of y is linearly related to x.
We condition on the observed $\{x_i\}$ and supress x in the notation so that $Z = \{y_i\}_{i=1}^n$.
$M = 2$.

$\mathcal{M}_1$:

$$Y_i = \beta_0 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \text{ iid.}$$

$$\theta_1 = (\beta_0, \sigma).$$

$\mathcal{M}_2$:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \text{ iid.}$$

$$\theta_2 = (\beta_0, \beta_1, \sigma).$$

Note: $\mathcal{M}_1$ is equivalent to $Y_i \sim N(\mu, \sigma^2)$, iid with μ in place of β_0.

The Bayesian approach to model selection simply puts prior probabilities on each model $p(\mathcal{M}_m)$, and *prior* distributions on each model parameter $p(\theta_m \mid \mathcal{M}_m)$, and then computes the posterior probability

$$p(\mathcal{M}_m \mid Z).$$

We have:

$$p(\mathcal{M}_m \mid Z) \propto p(\mathcal{M}_m)\, p(Z \mid \mathcal{M}_m),$$

and

$$p(Z \mid \mathcal{M}_m) = \int p(Z \mid \theta_m, \mathcal{M}_m)\, p(\theta_m \mid \mathcal{M}_m)\, d\theta_m$$

The beauty of the Bayesian approach is its conceptual simplicity.

But, choosing all the prior information and doing all the integrals can be very challenging in practice and a lot of research in Bayesian statistics over the past few decades can be described as attempts at these difficult problems.

The BIC gives a simple asymptotic approximation to the Bayesian $p(\mathcal{M}_m \mid Z)$.

Asymptotic in the sense that it is supposed to be approximately correct when the sample size is large enough.

The Deviance:

Let $\hat{\theta_m}$ be the MLE under model $\mathcal{M}_m$.

Let

$$\hat{L}_m = p(Z \mid \hat{\theta}_m, \mathcal{M}_m),$$

the maximized likelihood under model $\mathcal{M}_m$.

Then the deviance is

$$D_m = -2 \log(\hat{L}_m).$$

and the BIC is

$$BIC_m = D_m + log(n)\, d_m$$

where d_m is the dimension of θ_m and n is the sample size.

The approximation to the Bayesian posterior model probability is then,

$$p(\mathcal{M}_m \mid Z) \approx \frac{e^{-\frac{1}{2}BIC_m}}{\sum_{j=1}^{M} e^{-\frac{1}{2}BIC_j}}$$

So,

A small BIC makes a model more likely !!!

$$BIC = D + d\, log(n) = -2log(\hat{L}) + log(n)\, d$$

A complex model:
Will have a big $\hat{L}$ and hence a small $-2log(\hat{L})$.
Will have a big d and hence a big $d\, log(n)$.

A simple model:
Will have a small $\hat{L}$ and hence a large $-2log(\hat{L})$.
Will have a small d and hence a small $log(n)\, d$.

$$BIC = D + d\,log(n) = -2log(\hat{L}) + log(n)\,d$$

The Deviance:

Measures the in sample fit, with a smaller deviance indicating a better fit.

Complexity Penalty:

The term $d\,log(n)$ is a "complexity penalty" in that a higher dimensional parameter θ corresponds to a more complex model. BIC charges you $log(n)$ for a parameter.

As you add parameters, the deviance will go down, but the complexity penalty will go up, giving you a "U".

AIC:

The AIC is "an information criterion" or, "the Akaike information criterion".

$$AIC = D + d\,2 = -2log(\hat{L}) + 2\,d$$

Use: Choose the model with the smallest AIC.

The AIC charges you 2 for a parameter!!

Clearly, for non-tiny n, the BIC charges more for a parameter so it will give you a smaller model.

Like BIC, AIC has an asymptotic justification, but the theoretical starting point is based on information theory rather than Bayesian thinking.

Which is better?

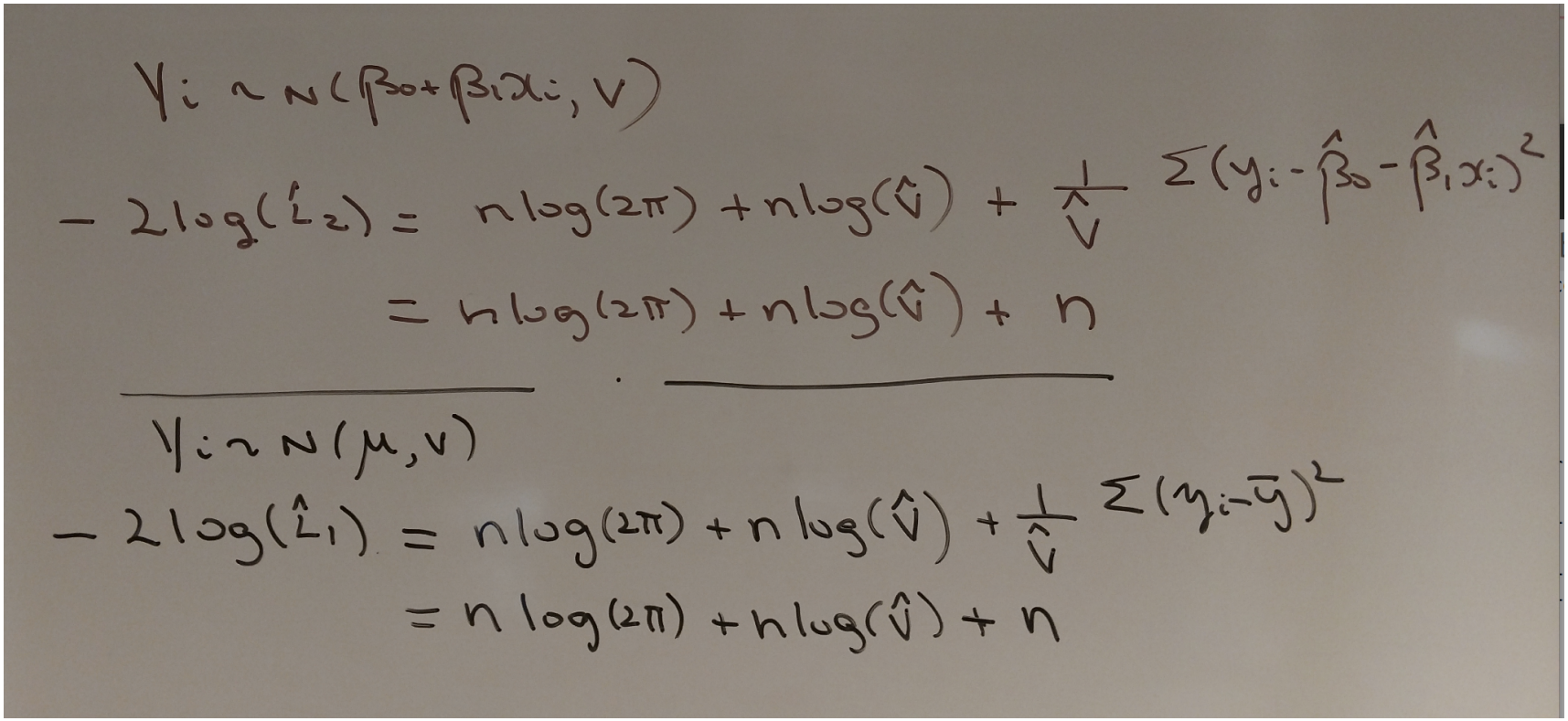
Often in practice people will report the result of both and if they agree, feel better about it.

How could it possibly be that the right way to think about it is a charge per parameter !!??

Seems crazy to me, but, I have to concede it sometimes works great in practice!

Example:

How do AIC and BIC compare our simple models $\mathcal{M}_1$ and $\mathcal{M}_2$?



So it boils down to a comparison of the $\hat{V} = \hat{\sigma}^2$.

Under model 1, the iid normal model:

$$\hat{V}_1 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Under model 2, the simple linear regession model with iid normal errors:

$$\hat{V}_2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Note:

In regression,

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

is the *residual sum of squares*.

Let $e_i = y_i - \hat{y}$ with $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Then

$$\text{RSS} = \sum e_i^2.$$

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Let $e_i = y_i - \hat{y}_i$.

Then:

$$\beta_0 + \beta_1 x_i \approx \hat{y}_i.$$

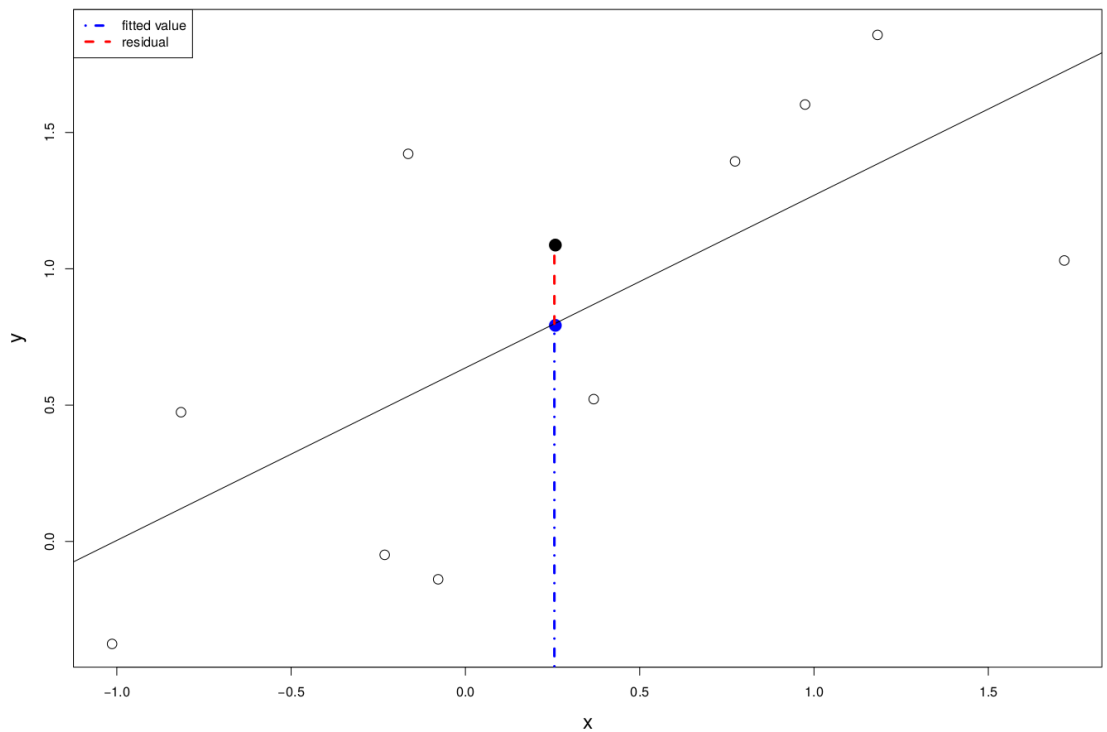
$$\epsilon_i = (y_i - (\beta_0 + \beta_1 x_i)) \approx (y_i - \hat{y}_i) = e_i \ .$$

For each observation:

$$y_i = \hat{y}_i + e_i$$

$\hat{y}_i$:
estimate of part of y x tells you about.

e_i:
estimate of part of y x can't tell you about.



If we let

$$\text{RSS}_1 = \sum (y_i - \bar{y})^2,$$

and,

$$\text{RSS}_1 = \sum e_i^2$$

then,

$$\text{BIC}_i = C + log(\text{RSS}_i) + d_i log(n)$$

where $d_1 = 2$ and $d_2 = 3$ and C is a constant.

So,

In our familiar basic models, the deviance reduces to quantities that depend on our familiar measures of in sample fit.

The point is that the BIC/AIC approach generalizes to any set of parametric models with a built in model complexity penalty.