# CS136 Lecture 2

Outline

1. Random Variable & Probability
2. Joint . Conditional . Marginal
3. Sum Rule . Product Rule . Bayes Rule
4. Independence
5. Expectation . Mean and Variance

## Independence / Conditional independence

Intuitively, X is independent of Y if the probability of its value never depends on Y's value.

$$P(X=x, Y=y) = P(X=x \mid Y=y) \cdot P(Y=y)$$
$$= P(X=x) \cdot P(Y=y)$$

## Conditional Independence

We says X and Y are conditionally independent given Z if and only if

$$P(Y=y \mid X=x, Z=z) = P(Y=y \mid Z=z)$$

Again if we know X and Y are conditionally independent

$$P(X, Y, Z) = P(X \mid Z) \cdot P(Y \mid Z) \cdot P(Z)$$

Proof:

$$P(X, Y, Z) = P(X \mid Y, Z) \cdot P(Y \mid Z) P(Z) \quad \text{(product rule)}$$
$$= P(X \mid Z) \cdot P(Y \mid Z) P(Z)$$

X, Y is conditional Independent

Sum rule / product rule / Bayes rule

离散型

Sum rule: $p(X) = \sum_y p(X, y) = \sum_y p(X|y) \cdot p(y)$

product rule: $p(X, y) = p(X|y) \cdot p(y)$

连续型:

Sum rule: $p(X) = \int_y p(X, y)\, dy = \int_y p(X|y) \cdot p(y)\, dy$

product rule: $p(X, y) = p(X|y) \cdot p(y)$

Bayes:

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)} = \frac{p(B|A) \cdot p(A)}{\int_A p(B|A) \cdot p(A)\, dA}$$

推导:

$$p(x, y) = p(x|y) \cdot p(y)$$
$$= p(y|x) \cdot p(x)$$
$$\Rightarrow p(y|x) \cdot p(x) = p(x|y) \cdot p(y)$$
$$\Rightarrow p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)}.$$

# Expectation. Mean and Variance

Mean: $E(X) = \sum_{x \in \Omega} p(x) \cdot x$

Variance: $\boxed{Var(X) = \sum_{x \in \Omega} E[(x - E[x])^2]}$

$$= \sum_{x \in \Omega} p(x) \cdot (x - \mu)^2$$

$$= \sum_{x \in \Omega} p(x)(x^2 - 2\mu x + \mu^2)$$

$$= \sum_{x \in \Omega} p(x) x^2 - \sum_{x \in \Omega} p(x) \cdot 2\mu x + \sum_{x \in \Omega} p(x) \cdot \mu^2$$

$$= \sum_{x \in \Omega} p(x) x^2 - 2\mu \underbrace{\sum_{x \in \Omega} p(x) \cdot x}_{\mu} + \mu^2 \underbrace{\sum_{x \in \Omega} p(x)}_{=1}$$

$$= \sum_{x \in \Omega} p(x) x^2 - 2\mu^2 + \mu^2$$

$$Var(X) = E(x^2) - \mu^2$$

$$\boxed{Var(X) = E(x^2) - E(x)^2}$$

# CS 136 Lecture 3 강의

Outline

1. Bernoulli Distribution
2. A spectrum of models for many coin flips
3. ML estimation for iid Bernoulli
4. Pros and Cons of Maximum Likelihood
5. Continous Random Variables. PDFs and CDFs

Bernoulli Distribution.

$$P(X=x) = \begin{cases} \mu & x=1 \\ 1-\mu & x=0 \end{cases}$$

$$= \mu^x (1-\mu)^{1-x}$$

$$P(x|\mu) = \text{BernPMF}(x|\mu) = \mu^x (1-\mu)^{1-x}$$

$x$ is Bernoulli variable $\quad X \sim \text{Bern}(\mu)$

Some Properties:

1. Expectation.

$$E(X) = \sum_{x \in \{0,1\}} x \cdot p(x|\mu)$$

$$= 0 \cdot \mu^0 (1-\mu)^1 + 1 \cdot \mu (1-\mu)^0$$

$$= \mu \qquad E[x] = \mu$$

2. Variance

$$\text{Var} = E(x^2) - E(x)^2$$

$$E(x^2) = \sum_{x \in \{0,1\}} x^2 p(x|\mu)$$

$$= 0 \cdot \mu^0 (1-\mu)^1 + 1 \cdot \mu (1-\mu)^0$$

$$= \mu$$

$$\text{Var} = \mu - \mu^2$$

$$\text{Var}(x) = \mu(1-\mu)$$

A spectrum of models for many coin flips

We perform N coin tosses $\Rightarrow X_1, X_2, \cdots, X_n$

We can model the joint of all tosses in several ways.

| | joint Model | # parameter |
|---|---|---|
| Most general | $P(X_1, X_2, \cdots, X_N)$ | N coins $\Rightarrow 2^N - 1$ free params |
| each toss is independent | $P(X_1)P(X_2)\cdots P(X_N)$ | N one param for each toss. |
| each toss independent and identically distributed | $P(X_1 \mid \mu)P(X_2 \mid \mu)\cdots P(X_N \mid \mu)$ | 1 |

Takeaways: Assumptions can simplify models, but need to carefully decide when appropriate

MLE for iid Bernoulli

Given N observations of coin flip outcomes $x_1, x_2, \cdots, x_N$ where $x_n \in \{0, 1\}$.

Assume an iid (independent & identically distribution model)

$$P(X_1 = x_1, X_2 = x_2 \cdots, X_N) = \prod_{i=1}^{N} \text{BernPMF}(x_n \mid \mu)$$

$$= \prod_{i=1}^{N} \mu^{x_i}(1-\mu)^{1-x_i}$$

maximum likelihood $\Rightarrow$ Find a value of $\mu$ that makes our observations most likely under assume model

$$\hat{\mu} = \arg\max_{\mu} \prod_{i=1}^{N} \mu^{x_i}(1-\mu)^{1-x_i} \quad \mu \in [0, 1]$$

Problem: inaccuracies in real application

the problem is that the product is too small.

solution: use log-likelihood

why: maximizing log-likelihood find the same optimal parameter $\hat{\mu}$ as mL because log is a monotonic increasing function.

solve (g - likelihood (單調遞增)

$$\hat{\mu} = \arg\max_{\mu} \ln \prod_{i=1}^{n} \mu^{x_i} (1-\mu)^{1-x_i}$$

$$= \sum_{i=1}^{n} [x_i \ln \mu + (1-x_i) \ln (1-\mu)]$$

$$\frac{\partial L}{\partial \mu} = \sum_{i=1}^{n} \left[ \frac{1}{\mu} \cdot x_i + \frac{-1}{1-\mu} \cdot (1-x_i) \right]$$

$$= \frac{1}{\mu} \sum_{i=1}^{n} x_i - \frac{n}{1-\mu} + \frac{1}{1-\mu} \sum_{i=1}^{n} x_i \overset{\text{set}}{=} 0$$

$$= \frac{1}{\mu(1-\mu)} \sum_{i=1}^{n} x_i = \frac{n}{1-\mu}$$

$$\sum_{i=1}^{n} x_i = n\mu$$

$$\Rightarrow \boxed{\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}}$$

$\sum_{i=1}^{n} x_i$ ← # of head

$n$ ← # total

Note.

For a constraint optimization problem, two ways to satisfy constraint

(1) solve as if no constraint, see if answer satisfies

(2) method of Lagrange multipliers

# Pros and Cons of Maximum Likelihood

Advantages:

1. MLE are consistent

   This means

   (1) our assumed likelihood distribution matches the true data-generating process

   (2) we have enough data

   Suppose we pick some $\mu^{true}$, and then draw $N$ observations $x_n \sim p(x|\mu^{true})$

   As $N \to \infty$, we can prove that $\hat{\mu}^{ML}(x_1, \cdots, x_N) \to \mu^{true}$

2. MLE are equivalent to different parameterizations

Limitations.

1. Overfitting (bad for small dataset)
2. Difficult of obtaining a solution

   (not all model have closed-form formula)

3. Solution is not always unique

# Continous Random Variables. PDFs and CDFs

## Probability Density Functions

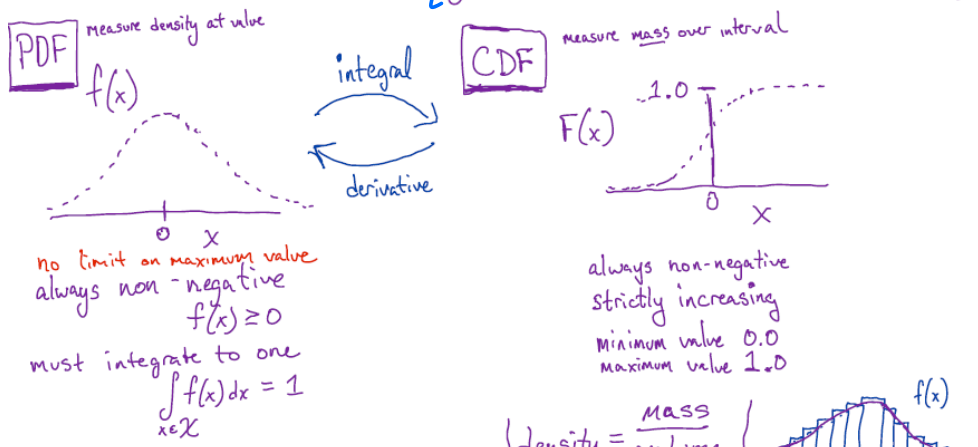
Let random variable $X$ have continous sample space $\mathbb{R}$. Define the cumulative distribution function $F$ for $X$ as

$$F(b) = P(X \leq b)$$

$$P(a \leq X \leq b) = F(b) - F(a)$$

We define the PDF of $X$ as the function satisfying

$$f(b) = \lim_{\varepsilon \to 0} \frac{F(b+\varepsilon) - F(b)}{\varepsilon} = \frac{\partial F(x)}{\partial x}\Big|_{x=b}$$




Remember, PDFs indicate density. $\boxed{\text{density} = \frac{\text{mass}}{\text{volume}}}$




Think of integral as a Riemannian sum as # bins goes to infinity.

$$\int_{x \in \mathcal{X}} f(x)\,dx \approx \lim_{\Delta x \to 0} \sum_{b=1}^{B} f(x_b)\,\Delta x = 1$$

$f(x_b)$: density of bin b; $\Delta x$: volume of bin b; 1: mass

Density function can be larger than one. Implies very small volume. Example: For any $\varepsilon > 0$, Uniform over $(0, +\varepsilon)$ has pdf $\frac{1}{\varepsilon}$

Let $x \in \mathbb{R}$, $y \in \mathbb{R}$ be random variables

We can define joint PDF as $p(x, y)$, which satisfies

- $p(x, y) \geq 0$
- $\int_{y \in \mathbb{R}} \int_{x \in \mathbb{R}} p(x, y)\,dx\,dy = 1$

Sum rule: $p(x) = \int p(x, y)\,dy \quad / \quad p(y) = \int p(x, y)\,dx$

Product rule: $p(x, y) = p(x|y) \cdot p(y)$

$= p(y|x) \cdot p(x)$

# CS136 Lec4

Outline  Reading: 1. 2. 3.  2.1

1. Gamma functions
2. Beta Distributions
3. Beta-Bernoulli model and its posterior
4. MAP Estimation
5. Posterior Predictive Distributions

Prior / likelihood / posterior

$P(\mu)$  $P(x|\mu)$  $P(\mu|x)$

$P(x_1, \cdots, x_N) = \int_0^1 P(\mu, x_1, \cdots, x_N) d\mu$

$P(\mu | x_1, \cdots x_N)$ (posterior)

$= \frac{1}{P(x_1 \cdots x_N)} \cdot P(\mu) \cdot P(x_1, \cdots x_N | \mu)$

prior × likelihood

Beta  Bernoulli

## Gamma Function $\Gamma(x)$

- $(a-1)! = \Gamma(a) = \int_0^{\infty} \mu^{a-1} e^{-\mu} d\mu$
- use $\log \Gamma(x)$ in practice
- $\boxed{\Gamma(x+1) = x \cdot \Gamma(x)}$

## Beta Distributions

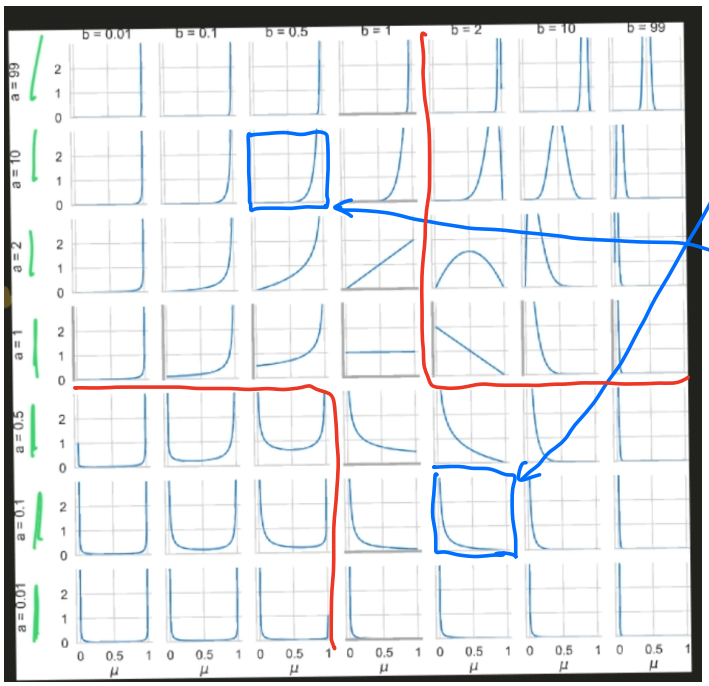
- $\text{BetaPDF}(\mu | a, b) = \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}_{\text{normalizing constant}} \mu^{a-1} (1-\mu)^{b-1}$  $0 \le \mu \le 1$, $a > 0$, $b > 0$

$= C(a, b) \mu^{a-1} (1-\mu)^{b-1}$

- useful facts

$\int_0^1 \text{BetaPDF}(\mu | a, b) d\mu = 1$

$$\boxed{\int_0^1 \mu^{a-1} (1-\mu)^{b-1} d\mu = \frac{1}{C(a,b)} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}}$$




$a<1,\ b<1$ U shape

$a=1,\ b=1$ — uniform

$a<1,\ b>1$

$a>1,\ b<1$

Beta-Bernoulli model and its posterior

Define model over two random variables:

$\mu \in [0,1]$

$x_1, x_2, \ldots, x_n$ with $x_n \in \{0,1\}$

Joint distribution:

$$p(x_1, \ldots, x_N, \mu) = \underbrace{\prod_{n=1}^{N} \text{Bern PMF}(x_n \mid \mu)}_{\text{likelihood}} \cdot \underbrace{\text{Beta PDF}(\mu \mid a, b)}_{\text{"prior"}}$$

$$p(\mu \mid x) = \frac{p(x \mid \mu) \cdot p(\mu)}{p(x)}$$

Given joint model, we can ask about probabilities implied by model:

| | | |
|---|---|---|
| Posterior | $p(\mu \mid x_1, \ldots, x_N)$ | after seeing N flip outcomes, what is likely value of $\mu$ |
| Evidence | $p(x_1, \ldots, x_N)$ | what is the p of seeing N flip outcome? |
| Predictive Posterior | $p(x_N \mid x_1, \ldots, x_{N-1})$ | after seeing N-1 flips, predict next outcome |

Bayes: $p(x) = \int p(x \mid \mu) d\mu$ sum rule

Bayes & sum rule

Posterior of $\mu$ for Beta-Bernoulli

PMF → discrete $p(X=x)$

PDF → continuous $P(a \le x \le b)$

$$p(\mu \mid x_1, \dots, x_N) = \frac{1}{p(x_1, x_2, \dots x_N)} \prod_{n=1}^{N} \underbrace{\text{Bern PMF}(x_n \mid \mu)}_{P(x \mid \mu)} \cdot \underbrace{\text{Beta PDF}(\mu \mid a, b)}_{P(\mu)}$$

$$= \frac{1}{p(x_1, x_2, \dots, x_N)} \prod_{n=1}^{N} \mu^{x_n} (1-\mu)^{1-x_n} \cdot C(a,b)\, \mu^{a-1} (1-\mu)^{b-1}$$

$$= \frac{C(a,b)}{p(x_1, x_2, \dots, x_n)} \prod_{n=1}^{N} \mu^{x_n} (1-\mu)^{1-x_n} \cdot \mu^{a-1} (1-\mu)^{b-1}$$

$$= \frac{C(a,b)}{p(x_1, x_2, \dots, x_n)} \mu^{\sum_n x_n} (1-\mu)^{\sum_n (1-x_n)} \cdot \mu^{a-1} (1-\mu)^{b-1}$$

$$= \text{Const} \cdot \underbrace{\mu^{\sum_n x_n + a - 1}}\ \underbrace{(1-\mu)^{\sum_n (1-x_n) + b - 1}}$$

$$\propto \text{Const} \cdot \mu^{\hat{a}-1} (1-\mu)^{\hat{b}-1}, \quad \text{where } \hat{a} = a + T(x),\ \hat{b} = b + T(x)$$

Posterior over $\mu$ for Beta-Bern is Beta

A general pattern:

If you have an unknown distribution with pdf up to a constant

$$p(\theta \mid \lambda) = \text{const} \cdot f(\theta, \lambda)$$

and you know a distribution family $D$ with pdf

$$p_D(\theta \mid \lambda) = C_D(\lambda) \cdot f(\theta, \lambda)$$

$f$ same as above

$C_D$ is a known function

then your distribution must belong to family $D$

MAP Estimation (Maximum a posteriori estimation)

meaning: most likely value of $\mu$ by posterior density

Suppose we observe $N$ flips outcomes. $X_1, \cdots, X_N$

(1) How to estimate a value for $\mu$?

(2) How to predict the next flip $X_{N+1}$?

Answer:

1) Find $\hat{\mu}$ that is mostly likely under posterior $p(\mu | x_1, x_2 \cdots, x_n)$

2) Predict next flip using likelihood with $\mu = \hat{\mu}$

$$p(X_{N+1}) = \text{Bern PMF}(X_{N+1} | \hat{\mu})$$

Optimization Problem: MAP

$$\hat{\mu} = \arg\max_{\mu \in [0,1]} p(\mu | x_1, x_2, \cdots, x_N)$$

$$= \arg\max_{\mu [0,1]} \frac{1}{p(x_1, x_2, \cdots, x_n)} \cdot p(\mu) \cdot p(x_1, x_2, \cdots, x_N | \mu)$$

$$= \frac{1}{p(x_1, x_2, \cdots, x_n)} \log p(\mu) + \log p(x_1, x_2, \cdots, x_N | \mu)$$

$$= \arg\max_{\mu [0,1]} \log p(\mu) + \sum_{n=1}^{N} \log \text{Beta-Bern PMF}(x_N | \mu)$$

$$= \arg\max_{\mu \in [0,1]} \log p(\mu) + \log \mu^{\sum_n x_n} (1-\mu)^{\sum_n (1-x_n)}$$

$$\vdots$$

$$\Rightarrow \hat{\mu} = \frac{\#(x) + a - 1}{\underbrace{\#(x) + \#(x) + a + b - 2}_{= N}}$$

requires. $\#+a > 1$, $\#+b > 1$

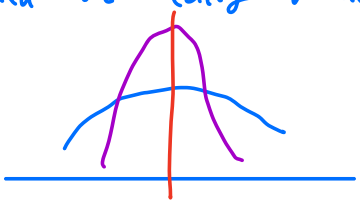
otherwise MAP doesn't exist


beta function

Problems with MAP

(1) does not always exist

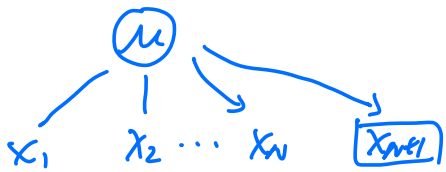
(2) should we really condense whole contribution to one value?



these have same MAP!

But one is less confident

Posterior Predictive Distributions




$p(\mu) = \text{Beta}(\mu \mid a, b)$

$p(x \mid \mu) = \prod_n \text{Bern}(x_n \mid \mu)$

$p(x_{N+1}=1 \mid x_1 \cdots x_N)$

$= \int \underbrace{p(x_{N+1}=1, \mu \mid x_1 \cdots x_N)}_{\Downarrow} \, d\mu$

$p(x_{N+1}, \mu \mid x_1 \cdots x_N)$

$= \underbrace{p(x_{N+1} \mid \mu, x_1, \cdots x_N)}_{\uparrow} \cdot \underbrace{p(\mu \mid x_1 \cdots x_N)}_{\uparrow}$

$=$ by iid, $p(x_{N+1} \mid \mu)$, $\text{Bern}(x_{N+1} \mid \mu)$

posterior given N coins

$= \int p(1 \mid \mu) \cdot \text{Beta}(\mu \mid \hat{a}, \hat{b}) \, d\mu$

$= C(\hat{a}, \hat{b}) \int \mu^1 \mu^{\hat{a}-1} (1-\mu)^{\hat{b}-1} \, d\mu$

$= C(\hat{a}, \hat{b}) \int \mu^{(1+\hat{a})-1} (1-\mu)^{\hat{b}-1} \, d\mu$

$= C(\hat{a}, \hat{b}) \cdot \frac{1}{C(\hat{a}+1, b)}$

$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+1)\cdot\Gamma(b)}{\Gamma(a+1+b)}$

1. Given $p(a,b,c)$

$p(a \mid c) = \int p(a, b \mid c) \, db$

2. product rule

joint: $p(x_{N+1}, \mu \mid x_1 \cdots x_N)$

$= p(x_{N+1} \mid \mu, x_1 \cdots x_N) \cdot p(\mu \mid x_1, \cdots x_N)$

conditional · marginal

$p(xy) = p(x) \cdot p(y)$

$p(x_n, \cdots x_1) = p(x_n \mid x_{n-1} \cdots 1) \cdot p(x_{n-1} \cdots 1)$

joint = marginal × conditional

$\int_0^1 \mu^{a-1} (1-\mu)^{b-1} \, d\mu = \frac{1}{C(a,b)} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$

$\Gamma(x+1) = x \cdot \Gamma(x)$

$$= \frac{\cancel{\Gamma(a+b)}}{\cancel{\Gamma(a)}\,\Gamma(1)} \cdot \frac{\hat{a}\,\cancel{\Gamma(a)}}{(\hat{a}+\hat{b})\cdot\cancel{\Gamma(a+b)}}$$

$$= \frac{\hat{a}}{(\hat{a}+\hat{b})}$$

Thus, by averaging over each possible $\mu$, weighted by its posterior density,

$$p(x_{N+1}=1 \mid x_1, \ldots, x_N) = \frac{\#\text{head} + a}{N + a + b}$$

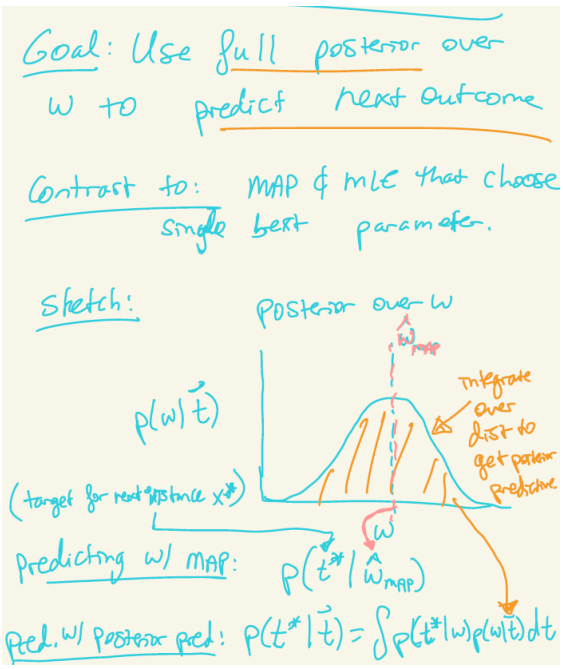
Valid for $a > 0$.
$b > 0$

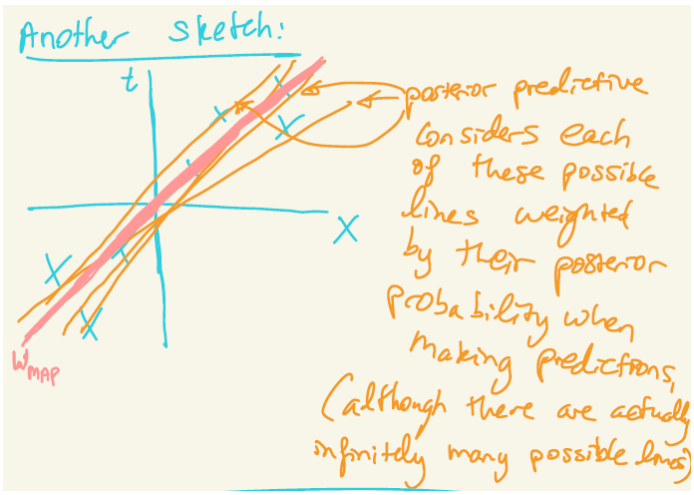
$p(x_{N+1} \mid \hat{\mu})$ MAP

$\int_0^1 p(x_{N+1} \mid \mu)\, p(\mu \mid x_1, \ldots, x_N)\, d\mu$ PPE

| | | |
|---|---|---|
| MLE | likelihood | Find max $\mu$ |
| MAP | Posterior | Find max |
| Posterior prediction | Posterior | Integrate over $\mu$ |

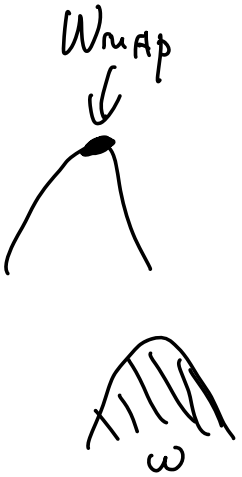
# 补充. MAP 与 Posterior 的区别




$$p(w|t) = \frac{p(t|w)\cdot p(w)}{p(t)} \propto p(t|w)\cdot p(w)$$

MAP: $\arg\max_w \underbrace{p(t|w)}_{\text{Like}} \cdot \underbrace{p(w)}_{\text{prior}} \rightarrow w_{MAP}$




预测下一个 $p(t^*|X, t, w_{MAP})$

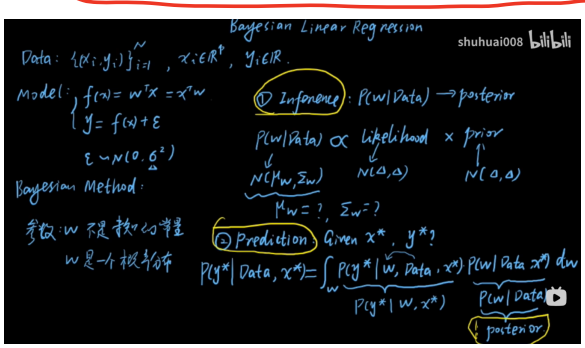
而 full bayesian

$p(w|t) = p(t|w)\cdot p(w)$ (Gaussian)

预测下一个:

$$p(t_*|t) = \int (t_*, w|t)\,dw = \int p(t_*|w,t)\cdot p(w|t)\,dw$$

$$= \int p(t_*|w)\cdot p(w|t)\,dw \quad (t_* \text{ iid})$$




优点: Averaging should be better if we have limited data.

CS 136 Lec 5 A

Outline Reading Sec 2.2

1. From Binary to Categorial Distributions
2. ML estimation for categorical Parameter
3. Dirichlet distribution
4. Dirichlet - Categorical model & its porterior
5. MAP Estimation for Dirichlet - Categorical

From Binary to Categorial Distributions

Binary Distribution → 2 outcomes

Categorical Distribution → V outcomes.

Binary (Bernoulli)

Sample space: $\{0, 1\}$

Parameter: $\mu \in [0, 1]$

PMF: $\mu^x (1-\mu)^{1-x}$

Categorical Distribution

Sample space: $\{1, 2, \cdots, V\}$

Parameters: $\mu = \{\mu_1, \cdots, \mu_V\}$, s.t. $\sum_{v=1}^{V} \mu_v = 1$, $\mu_v \geq 0$

PMF: $p(x = x \mid \mu) = \prod_{v=1}^{V} \mu_v^{[x=v]} = \mu_x$

probability simplex $\mu \in \Delta^V$

notation: $p(x = x \mid \mu) = \mu_x$

$x \in \{1, 2, \cdots, V\}$

$p(x = x \mid \mu) = \prod_{v=1}^{V} \mu_v^{[x=v]} = \mu_x$ ← $\begin{cases} 1 & x = v \\ 0 & x \neq v \end{cases}$

one hot vector

Consider observing $N$ words from vocabulary of size $V$.

Let random variables $x_1, x_2, \cdots, x_N$ indicate the words.

where $x_n = [0\,0\,0\,0 \mid 1\,0\,0\,0]$

$p(x \mid \mu) = \prod_{v=1}^{V} (\mu_v)^{x_n} (1-\mu_v)^{1-x_n}$

可以化简: $p(x \mid \mu) = \prod_{v=1}^{V} \mu_v^{x_v}$

$$x_{nv} = \begin{cases} 1 & \text{if } n \text{ word is type } V \\ 0 & \text{otherwise} \end{cases}$$

Assume $N$ words drawn independent & identically distributed (i.i.d) from Categorical with parameter $\mu \in \Delta^V$

$$p(x_1, \dots, x_N \mid \mu) = \prod_{n=1}^{N} \text{CatPMF}(x_n \mid \mu)$$

$$= \prod_{n=1}^{N} \left( \prod_{v=1}^{V} \mu_v^{x_{nv}} \right) \qquad \prod_{n=1}^{N} \left( \mu_r^{x_{n v_1}} \cdot \mu_r^{x_{n v_2}} \cdot \right) = \prod_{n=1}^{N} \mu_r^{x_{n v_1} + x_{n v_2} \dots}$$

$$= \prod_{v=1}^{V} \mu_r^{\sum_{n=1}^{N} x_{nv}}$$

$\sum_{n=1}^{N} x_{nv}$ ← number of times vocab symbol $v$ appears in our dataset

$$= \prod_{v=1}^{V} \mu_r^{m_v(x)}$$

$$m_v(x) = \sum_n x_{nv}$$

ML estimation for categorical Parameter

$$\hat{\mu} = \arg\max_{\mu \in \Delta^V} \log p(x_1, x_2 \dots x_N \mid \mu)$$

$$= \arg\max_{\mu \in \Delta^V} \log \left( \prod_{v=1}^{V} \mu_v^{m_v(x)} \right)$$

$$= \arg\max_{\mu \in \Delta^V} \sum_{v=1}^{V} m_V(x) \cdot \log \mu_v$$

$$\arg\max_{\mu \in \mathbb{R}^V} \sum_{v=1}^{V} m_v(x) \log \mu_v$$

$$\text{s.t. } 1 - \sum_{i=1}^{V} \mu_v = 0$$

← address with Lagrange multiplier

$$\mu_v \geq 0$$

← address with ignore, then check

Lagrange multiplier method

step 1. $L(\mu,\lambda) = \sum_{i=1}^{V} m_V(x)\log\mu_V + \lambda\left(1-\sum_{i=1}^{V}\mu_V\right)$ $(\lambda\neq 0)$

step 2. $\frac{\partial}{\partial\mu_1}=0 \Rightarrow \frac{m_1}{\mu_1}-\lambda=0$ (1)

$\frac{\partial}{\partial\mu_2}=0 \Rightarrow \frac{m_2}{\mu_2}-\lambda=0$ (2)

$\vdots$

$\frac{\partial}{\partial\mu_V}=0 \Rightarrow \frac{m_V}{\mu_V}-\lambda=0$ (V)

$\frac{\partial}{\partial\lambda}=0 \Rightarrow 1-\sum_{i=1}^{V}\mu_V=0$ (V+1)

Summary (1) ... (V)

$\sum_{V=1}^{V}\frac{m_V}{\mu_V}=\lambda$

$\Rightarrow \frac{1}{\lambda}\sum_{V=1}^{V} m_V = \sum_{V=1}^{V}\mu_V$

plugin in (V+1)

$1-\lambda\sum_{V=1}^{V} m_V = 0 \Rightarrow \frac{1}{\lambda}\sum_{V=1}^{V} m_V = 1$

$\lambda = N = \sum_{V=1}^{V} m_V$

Return back to (1) ~ (V)

$\mu_1^* = \frac{m_1}{\lambda} = \frac{m_1}{N}$

$\mu_2^* = \frac{m_2}{N}$

$\vdots$

$\Rightarrow \mu^* = \left[\frac{m_1}{N}, \frac{m_2}{N}, \cdots \frac{m_V}{N}\right]$

At last. check

$\boxed{\mu^*\geq 0}$

簡化版

$L(\mu,\lambda) = \sum_{i=1}^{V} m_V(x)\log\mu_V + \lambda\left(1-\sum_{i=1}^{V}\mu_V\right)$

$\frac{\partial L(\mu,\lambda)}{\partial\mu_V} = \sum_{V}\frac{m_V}{\mu_V} - \lambda \overset{\text{set}}{=} 0$

$\sum_{V}\frac{m_V}{\mu_V} = \lambda$

$\Rightarrow \mu_V^* = \frac{m_V}{\lambda}$

由限制條件

$1-\sum_{i=1}^{V}\mu_V = 0$

$\Rightarrow 1-\sum_{i=1}^{V}\frac{m_V}{\lambda} = 0$

$\lambda = \sum_{i=1}^{V} m_V = N$ (N為觀測數)

$\mu_V^* = \frac{m_V}{N}$

# Dirichlet distribution

Dirichlet: prior

Random variable $u = [u_1, u_2, \dots u_V]$

Sample space $u \in \Delta^V$

Parameter $\alpha = [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_V] \quad \alpha_{1,2,\dots,V} \geq 0$

PDF:

$$\text{DirPDF}(u|\alpha) = \frac{\frac{\Gamma(\alpha_1+\alpha_2+\cdots\alpha_V)}{\prod_{v=1}^{V}\Gamma(\alpha_v)}\prod_{v=1}^{V} u_v^{(\alpha_v-1)}}{\text{Constant } c(\alpha)}$$

Can recognize that

$$\int_u \frac{\Gamma(\alpha_1+\alpha_2+\cdots\alpha_V)}{\prod_{v=1}^{V}\Gamma(\alpha_v)}\prod_{v=1}^{V} u_v^{(\alpha_v-1)}\, du = 1$$

$$\Rightarrow \int_u u_v^{(\alpha_v-1)} = \frac{1}{c(\alpha)} = \frac{\prod_{v=1}^{V}\Gamma(\alpha_v)}{\Gamma(\alpha_1+\alpha_2\cdots+\alpha_V)}$$

## special cases & visualization

When $V=2$. Dirichlet reduces to Beta.

Beta

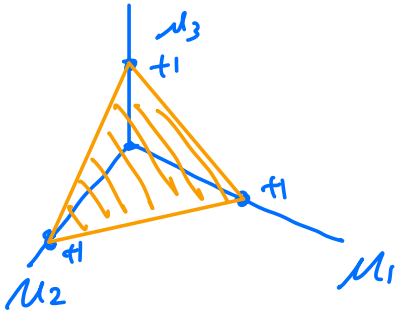
Sample space $u \in [0,1]$

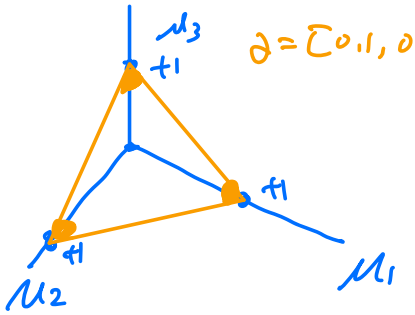
PDF: $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^{a-1}(1-u)^{b-1}$

Dirichlet s.t. $u_1 \geq 0$, $u_2 \geq 0$, $u_1+u_2=1$

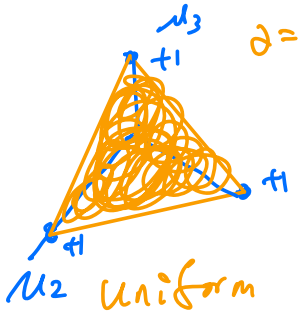
$[u_1, u_2]$

$\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \cdot u_1^{\alpha_1-1} \cdot u_2^{\alpha_2-1}$

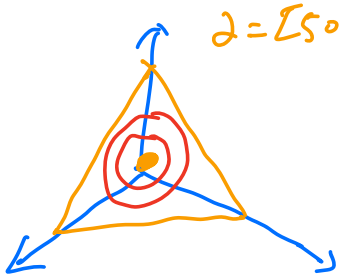



$\alpha = [0.1, 0.1, 0.1]$




$\alpha = [1, 1, 1]$ uniform



$\alpha = [50, 50, 50]$

## Dirichlet - Categorical model & its posterior

Jointly Explain:

$[u_1, u_2, \cdots, u_V]$ unknown probability vector $u \sim Dir(\alpha)$

$x_1, x_2, \cdots, x_N$ $N$ observed words $x_n \overset{iid}{\sim} Cat(u)$

$$p(x_1, x_2, \cdots x_N, u) = \prod_{n=1}^{N} Cat\,PMF(x_n, u) \cdot Dir\,PDF(u \mid \alpha)$$

P.S. Beta - Bernoulli model

$$p(x_1, \cdots, x_N, u) = \underbrace{\prod_{n=1}^{N} Bern\,PMF(x_n \mid u)}_{\text{likelihood}} \cdot \underbrace{Beta\,PDF(u \mid a, b)}_{\text{"prior"}}$$

Posterior of $u$ under Dir - Cat

$$p(u \mid x_1, x_2, \cdots, x_N)$$

$$= \frac{1}{p(x_1, x_2, \cdots x_N)} \cdot p(u) \cdot p(x_1, x_2, \cdots, x_N \mid u)$$

$$= C_1 \cdot Dir\,PDF(u \mid \alpha) \cdot \prod_{n=1}^{N} Cat\,PMF(x_n, u)$$

one-hot

$$= C_1 \cdot \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots \alpha_V)}{\prod_{v=1}^{V} \Gamma(\alpha_v)} \prod_{v=1}^{V} u_v^{(\alpha_v - 1)} \cdot \prod_{n=1}^{N} \prod_{v=1}^{V} u_v^{x_{nv}}$$

$$= C_1 \cdot C_2 \cdot \prod_{v=1}^{V} u_v^{\alpha_v - 1} \cdot \prod_{n=1}^{N} \prod_{v=1}^{V} u_v^{x_{nv}}$$

$$= \prod_{v=1}^{V} \mu_v^{\alpha_v - 1} \cdot \prod_{v=1}^{V} \mu_v^{\sum_{i=1}^{N} x_{nv}}$$

$m_v = \sum_{i=1}^{N} x_{nv}$ ⇒ count of type $v$ in data

$$= \prod_{v=1}^{V} \mu_v^{\alpha_v - 1} \prod_{v=1}^{V} \mu_v^{m_v}$$

$$= \prod_{v=1}^{V} \mu_v^{\alpha_v - 1 + m_v}$$

$$\text{DirPDF}(\mu|\alpha) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots \alpha_V)}{\prod_{v=1}^{V} \Gamma(\alpha_v)} \prod_{v=1}^{V} \mu_v^{(\alpha_v - 1)}$$

Thus, posterior is Dirichlet — Conjugate prior

$$p(\mu | x_1, \ldots, x_N) = \text{DirPDF}(\mu | \hat{\alpha}_1, \ldots, \hat{\alpha}_V)$$

where $\hat{\alpha}_v = \alpha_v + m_v$

= prior pseudo count + count of symbol $v$

## MAP Estimation for Dirichlet - Categorical

$$\mu^{MAP} = \underset{\mu \in \Delta^V}{\arg\max} \log p(\mu | x_1, x_2, \ldots, x_N)$$

$$= \underset{\mu \in \Delta^V}{\arg\max} \log\left( C(\hat{\alpha}) \prod_{v=1}^{V} \mu_v^{\hat{\alpha}_v - 1} \right)$$

$$= \underset{\mu \in \Delta^V}{\arg\max} \not{C} + \sum_{v=1}^{V} (\hat{\alpha}_v - 1) \log \mu_v$$

$$= \underset{\mu \in \Delta^V}{\arg\max} \sum_{v=1}^{V} (\hat{\alpha}_v - 1) \log \mu_v$$

$$\mu_{MAP} = \left[ \frac{\hat{\alpha}_1 - 1}{S}, \frac{\hat{\alpha}_2 - 1}{S}, \ldots, \frac{\hat{\alpha}_V - 1}{S} \right]$$

$$S = \sum_{v} \hat{\alpha}_v - 1$$

# CS 136 Lecture 5 B Lagrange multipliers

Outline Appendix E

1. Recipe & example
2. Why it works?

## Recipe & example

Given an objective function $f$

$$\mathbb{R}^D \to f \to \mathbb{R}$$

$$\theta = [\theta_1, \theta_2 \cdots \theta_D]$$

Goal: Equality constrained optimal value

$$\theta^* = \arg\min_{\theta} f(\theta)$$

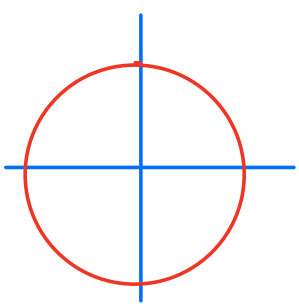
$$\text{s.t. } g(\theta) = 0$$

限制条件式

还有一种为 $g(\theta) \geq 0$. 后面讲

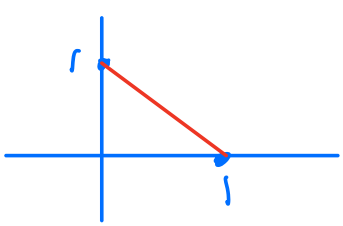
Example 1

Find $\theta$ to maximize $f$, such that $\theta$ lies on unit circle



$$g(\theta) = 1 - \theta_1^2 - \theta_2^2$$

Find $\theta$ to maximize $f$, such that $\theta$ sum to one.




$$g(\theta) = 1 - \sum_{d=1}^{D} \theta_d$$

**Recipe**

**Step 1** Define objective $f$ and constraint $g$

**Step 2** Define expanded objective, with new "multiplier" $\lambda \neq 0$

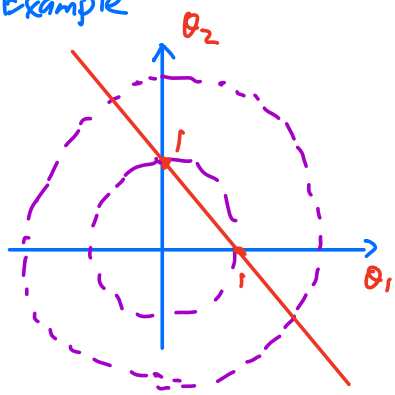
$$\mathcal{L}(\theta, \lambda) = f(\theta) + \lambda g(\theta)$$

**Step 3** (calculus) Setup system of $D+1$ equations, where each partial derivative is zero

$$\frac{\partial}{\partial \theta_1}\mathcal{L} = 0, \quad \frac{\partial}{\partial \theta_2}\mathcal{L} = 0, \quad \cdots \quad \frac{\partial}{\partial \theta_D}\mathcal{L} = 0, \quad \frac{\partial}{\partial \lambda}\mathcal{L} = 0$$

**Step 4** (algebra) Solve the system of equations to find optimal values of $\theta_1^*, \theta_2^*, \cdots \theta_D^*, \lambda^*$

Example




step 1.

$$f(\theta) = 1 - \theta_1^2 - \theta_2^2 \quad \leftarrow \text{objective}$$

$$g(x) = \theta_1 + \theta_2 - 1 \quad \leftarrow \text{constraint}$$

step 2.

$$L(\theta, \lambda) = f(x) + \lambda g(x)$$

$$= 1 - \theta_1^2 - \theta_2^2 + \lambda(\theta_1 + \theta_2 - 1)$$

step 3.

$$\frac{\partial L}{\partial \theta_1} = -2\theta_1 + \lambda \overset{\text{set}}{=} 0$$

$$\frac{\partial L}{\partial \theta_2} = -2\theta_2 + \lambda \overset{\text{set}}{=} 0$$

$$\frac{\partial L}{\partial \lambda} = \theta_1 + \theta_2 - 1 \overset{\text{set}}{=} 0$$

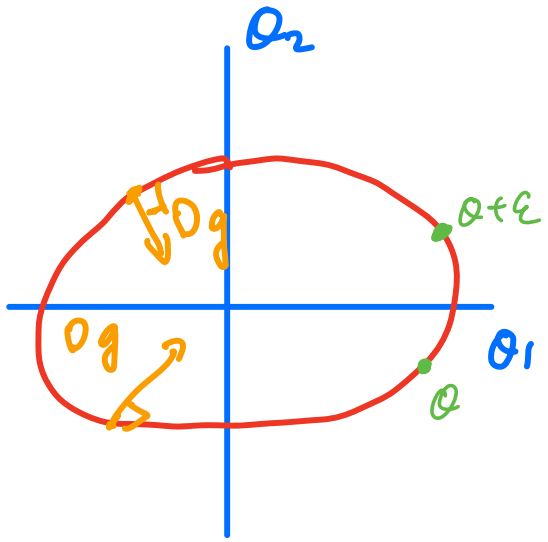
$$\Rightarrow \begin{cases} \theta_1 = \frac{1}{2} \\ \theta_2 = \frac{1}{2} \end{cases}$$

Why it works?

Consider the subset $S$ of possible $\theta$ values that satisfy the constraint.

Lemma 1: Any $\theta$ that satisfy $g(\theta)=0$ will have gradient perpendicular to constraint surface

$$\nabla_\theta g \perp \mathcal{E}$$




Proof:

Examine two "close" points that satisfy

Call these points $\theta$ and $\theta+\varepsilon$

$$g(\theta)=0 \quad (1)$$
$$g(\theta+\varepsilon)=0 \quad (2)$$

Taylor:

$$g(\theta+\varepsilon) = g(\theta) + \varepsilon^T \nabla_\theta g(\theta)$$

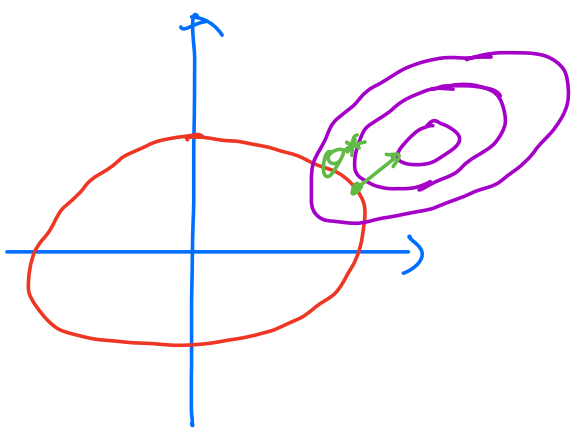
plug in (1):

$$g(\theta+\varepsilon) = \boxed{\varepsilon^T \nabla_\theta g(\theta) = 0}$$

Lemma 2:

At optimal $\theta^*$, the gradient of the objective $f$ will be perpendicular to constraint surface

$$\nabla_\theta f|_{\theta=\theta^T} \perp \varepsilon$$




punchline:

At optima $\theta^*$, $\nabla_\theta f$ and $\nabla_\theta g$ are orthogonal to $\varepsilon$, thus they must be parallelled to each other

If $\nabla f$, $\nabla g$ are parallel, then $\exists$ some $\lambda \neq 0$ s.t.

$$\nabla f + \lambda \nabla g = 0$$

# Why it works (cont'd)?

Thus, any optimal value $\theta^*$ must satisfy these $D+1$ equations

$\theta^*$ must be optimal

$$\vec{\nabla}_f + \lambda \vec{\nabla}_g = \vec{0} \quad \rightarrow \quad \begin{bmatrix} \frac{\partial}{\partial \theta_1} f \\ \vdots \\ \frac{\partial}{\partial \theta_D} f \end{bmatrix} + \lambda \begin{bmatrix} \frac{\partial}{\partial \theta_1} g \\ \vdots \\ \frac{\partial}{\partial \theta_D} g \end{bmatrix} = 0 \qquad \begin{matrix} (1) \\ (2) \\ \\ (D) \end{matrix}$$

$\theta^*$ must satisfy constraints

$$g(\theta^*) = 0 \quad \rightarrow \quad \frac{\partial}{\partial \lambda} \left[ \lambda \, g(\theta) \right] = 0 \qquad (D+1)$$

Can view the whole system as partial derivatives of expanded objective

$$\alpha(\theta, \lambda) = f(\theta) + \lambda g(\theta)$$

where $\theta \in \mathbb{R}^D$, $\lambda \neq 0$ are unknown, to be solved for.