Introduction to Deep Generative Modeling

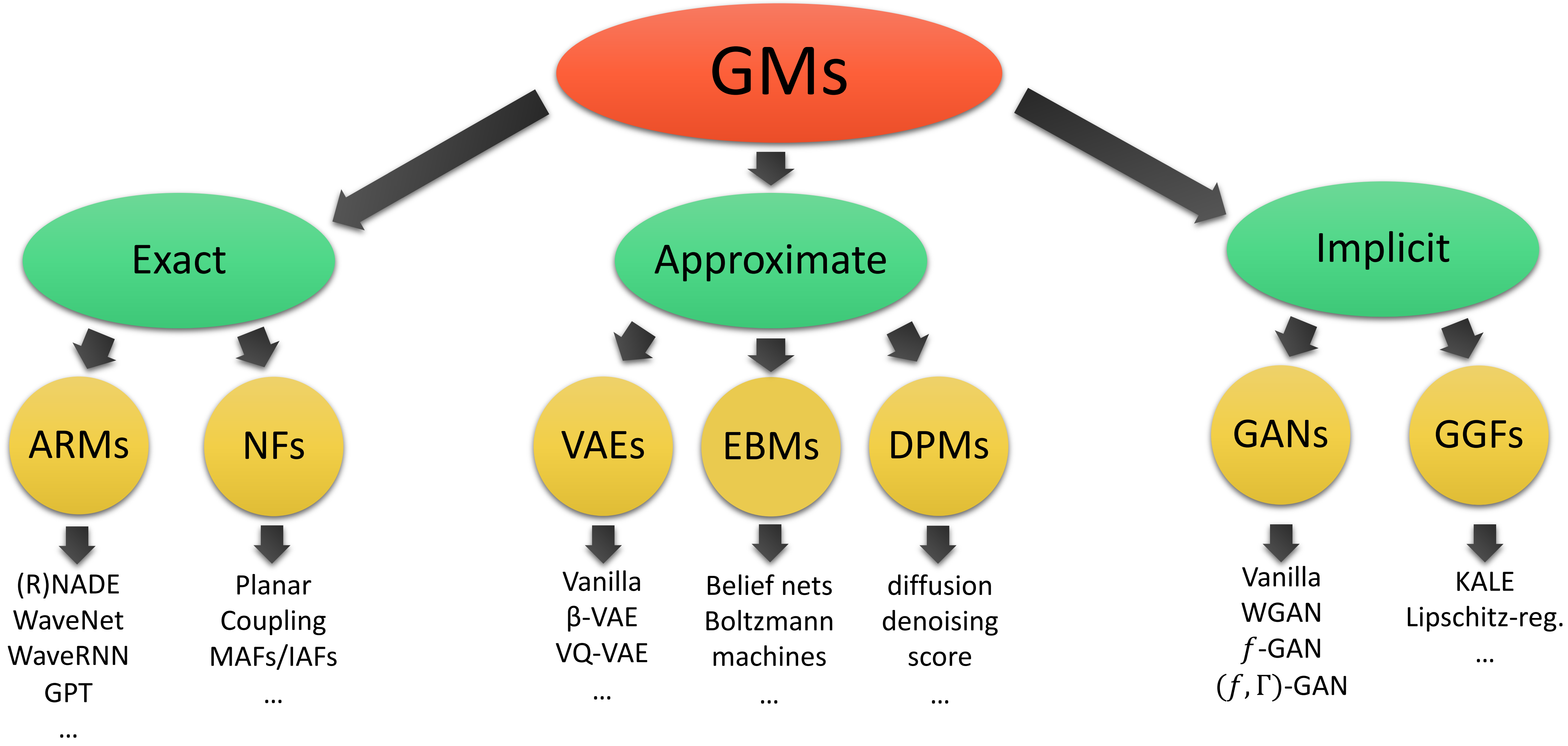
Lecture #3

HY-673 – Computer Science Dep., University of Crete

Professors: Yannis Pantazis & Yannis Stylianou

TAs: Michail Raptakis & Michail Spanakis

Taxonomy of Deep Generative Models According to the Likelihood Function



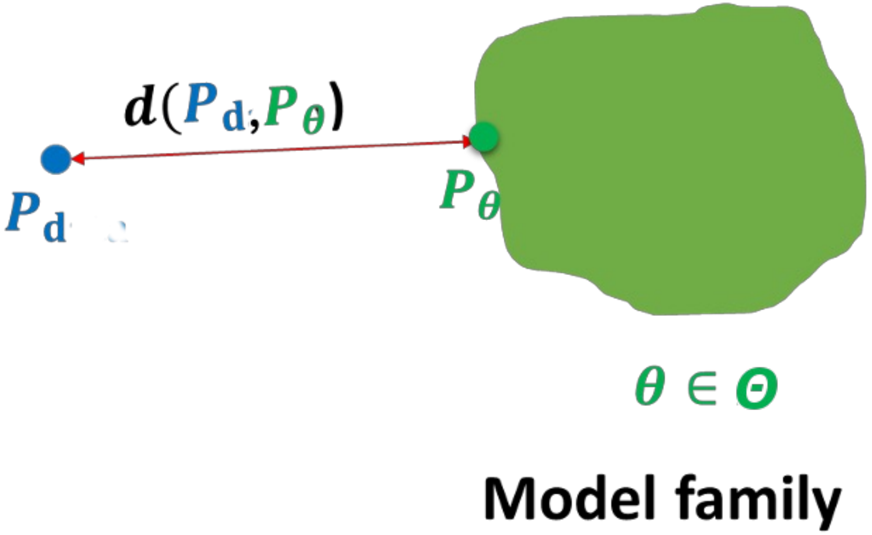
Introduction to Estimator Theory

- What is an estimator?

Let $\mathcal{D} = \{x_1, \ldots, x_n\}$ be a set of data drawn from $p_d(x)$, and $p_\theta(x)$ be a family of models with $\theta \in \Theta$. A point estimator $\hat{\theta} = \hat{\theta}(\mathcal{D})$ is a random variable for which we want:

$$p_{\hat{\theta}}(x) \approx p_d(x)$$

$\mathbf{x}_i \sim P_d$
$i = 1, 2, \ldots, n$



Introduction to Estimator Theory

- How to construct an estimator?
  - Maximum Likelihood Estimation (MLE)
  - Maximum A Posteriory (MAP) Estimation
  - Based on a Probability Distance or a Divergence (implicit)
  - Bayesian Inference (learns a distribution for the estimator's parameters)

Maximum Likelihood Estimator

- Maximum Likelihood Estimator:

$$\hat{\theta}_{\mathrm{MLE}}(\mathcal{D}) = \arg\max_{\theta} p_{\theta}(\mathcal{D}) := p_{\theta}(x_1, \ldots, x_n).$$

- Equivalently, under the i.i.d. assumption:

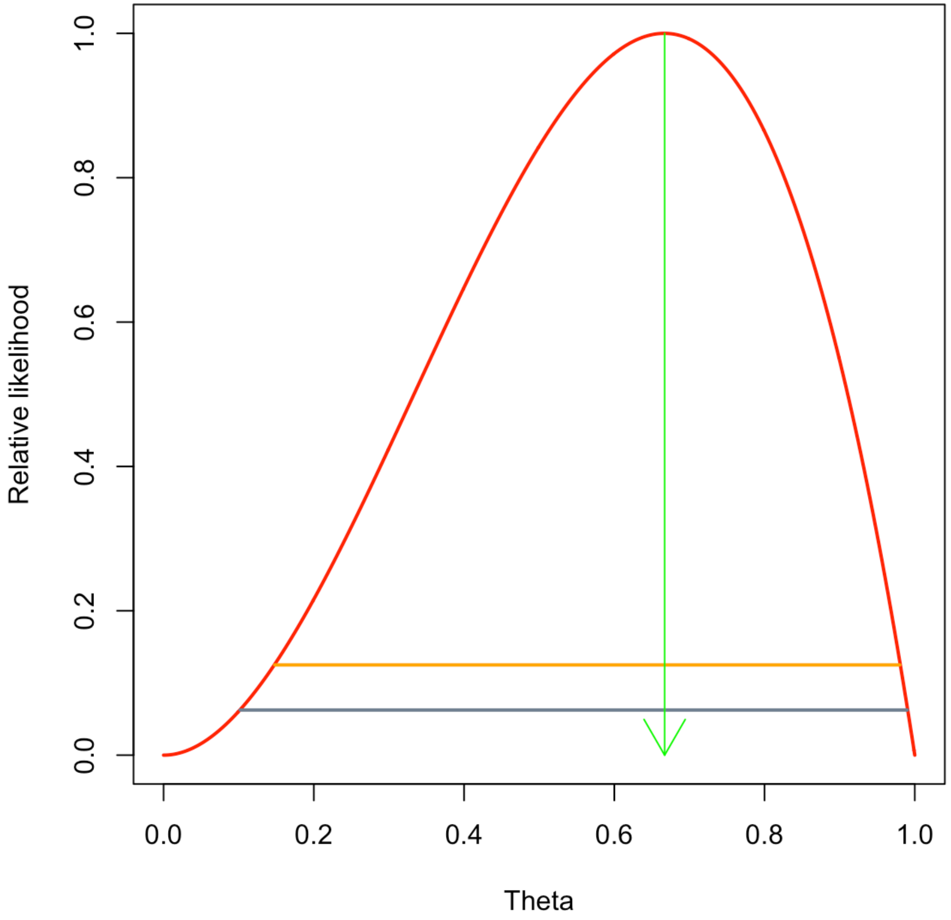
$$\hat{\theta}_{\mathrm{MLE}}(\mathcal{D}) = \arg\max_{\theta} \sum_{i=1}^{n} \log p_{\theta}(x_i) =: L(\theta; \mathcal{D}) \ \big(\equiv L_n(\theta) \big).$$

Maximum Likelihood Estimator

- MLE interpretation:

  - $L_n(\hat{\theta}_1) > L_n(\hat{\theta}_2)$ implies that $\hat{\theta}_1$ is *more likely* to have generated the observed samples $x_1, ..., x_n$.

  - Thus, it provides a ranking of model's fitness/accuracy/matching to the data.



MLE Example #1

- 1d Gaussian, unknown mean $\theta = \mu$, known variance σ^2:

Dataset: $\mathcal{D} = \{x_1, \ldots, x_n\}$,

Model family: $p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}}$.

$$L(\theta, \mathcal{D}) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}}\right) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{1}{2\sigma^2}(x_i - \theta)^2.$$

$$\frac{d}{d\theta} L(\theta; \mathcal{D}) = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \theta)$$

Thus, $\frac{d}{d\theta} L(\hat{\theta}; \mathcal{D}) = 0 \Longrightarrow \boxed{\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i.}$

MLE Example #2

- Exponential distribution:

$$p_\theta(x) = \begin{cases} \theta e^{-\theta x}, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

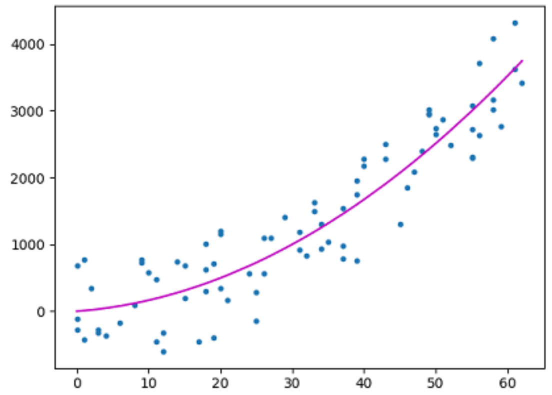
$$L(\theta, \mathcal{D}) = \sum_{i=1}^{n} (\log \theta - \theta x_i)$$

$$\frac{d}{d\theta} L(\theta; \mathcal{D}) = \sum_{i=1}^{n} (\frac{1}{\theta} - x_i)$$

Thus, $\frac{d}{d\theta} L(\hat{\theta}; \mathcal{D}) = 0 \implies \boxed{\hat{\theta} = \frac{n}{\sum_{i=1}^{n} x_i}.}$

MLE Example #3

- Linear model with Gaussian error:



Dataset: $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$,
Model family: $y_i = \theta^T x_i + e_i$ with $e_i \sim \mathcal{N}(0, \sigma^2)$ and $\theta \in \mathbb{R}^d$.

$$L(\theta, \mathcal{D}) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}}\right) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2.$$

Partial derivative or gradient vector: $$\frac{\partial}{\partial\theta} L(\theta; \mathcal{D}) = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \theta^T x_i)x_i^T.$$

Thus, $$\frac{\partial}{\partial\theta} L(\hat{\theta}; \mathcal{D}) = 0 \implies \hat{\theta} = \left(\sum_{i=1}^{n} x_i x_i^T\right)^{-1}\left(\sum_{i=1}^{n} x_i y_i\right).$$

MLE Example #3

- In matrix form:

$$y = X\theta + e \text{ with } y = [y_1, ..., y_n]^T \in \mathbb{R}^n,\ X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times d},$$

$$e = [e_1, ..., e_n]^T \in \mathbb{R}^n \text{ and } e \sim \mathcal{N}(0, \sigma^2 I_n).$$

$$L(\theta, \mathcal{D}) = C - \frac{1}{2\sigma^2}(y - X\theta)^T(y - X\theta).$$

← Maximizing L(θ) is equivalent to minimizing the Sum of Squares (Least Squares)

$$\frac{\partial}{\partial\theta} L(\theta; \mathcal{D}) = -\frac{1}{\sigma^2} X^T(y - X\theta)$$

Exactly the same solution as LS!

$$\text{Thus,}\quad \frac{\partial}{\partial\theta} L(\hat{\theta}; \mathcal{D}) = 0 \implies \boxed{\hat{\theta} = (X^T X)^{-1} X^T y.}$$

MLE Example #4

- Logistic regression with sigmoids a.k.a. binary classification.

Dataset: $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ with $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$,

Model family: $p_\theta(y_i = 1|x_i) = \sigma(\theta^T x_i)$, $p_\theta(y_i = 0|x_i) = 1 - p_\theta(y_i = 1|x_i)$, $\theta \in \mathbb{R}^d$ and $\sigma(z) = \frac{1}{1+e^{-z}}$ be the sigmoid function.

Compact form: $p_\theta(y_i|x_i) = p_\theta(y_i = 1|x_i)^{y_i} p_\theta(y_i = 0|x_i)^{1-y_i}$.

$$L(\theta, \mathcal{D}) = \sum_{i=1}^{n} y_i \log \frac{1}{1 + e^{-\theta^T x_i}} + (1 - y_i) \log \frac{e^{-\theta^T x_i}}{1 + e^{-\theta^T x_i}}$$

Unfortunately, $\frac{\partial}{\partial \theta} L(\hat{\theta}; \mathcal{D}) = 0$ is a non-linear system of equations.

MLE Example #4

- Solution: Iteratively solve for the root of the system of equations.

- Gradient ascent.

  1. Randomly initialize $\theta^0 = \left(\theta_1^0, \ldots, \theta_d^0\right)$.
  2. Compute $\nabla_\theta L(\theta; \mathcal{D})$.
  3. Update $\theta^{t+1} = \theta^t + \alpha_t \nabla_\theta L(\theta; \mathcal{D})$. ← Learning rate
  4. Repeat 2 & 3 until convergence.

- Caution: MLE results in a non-convex optimization problem
  $\Rightarrow$ stack to a local maximum.

Maximum Likelihood Estimator

- We could obtain a histogram (i.e., empirical or sampling distribution) for the estimator as

$$\{\hat{\theta}(\mathcal{D}') : \mathcal{D}' \sim p_d(x)\}$$

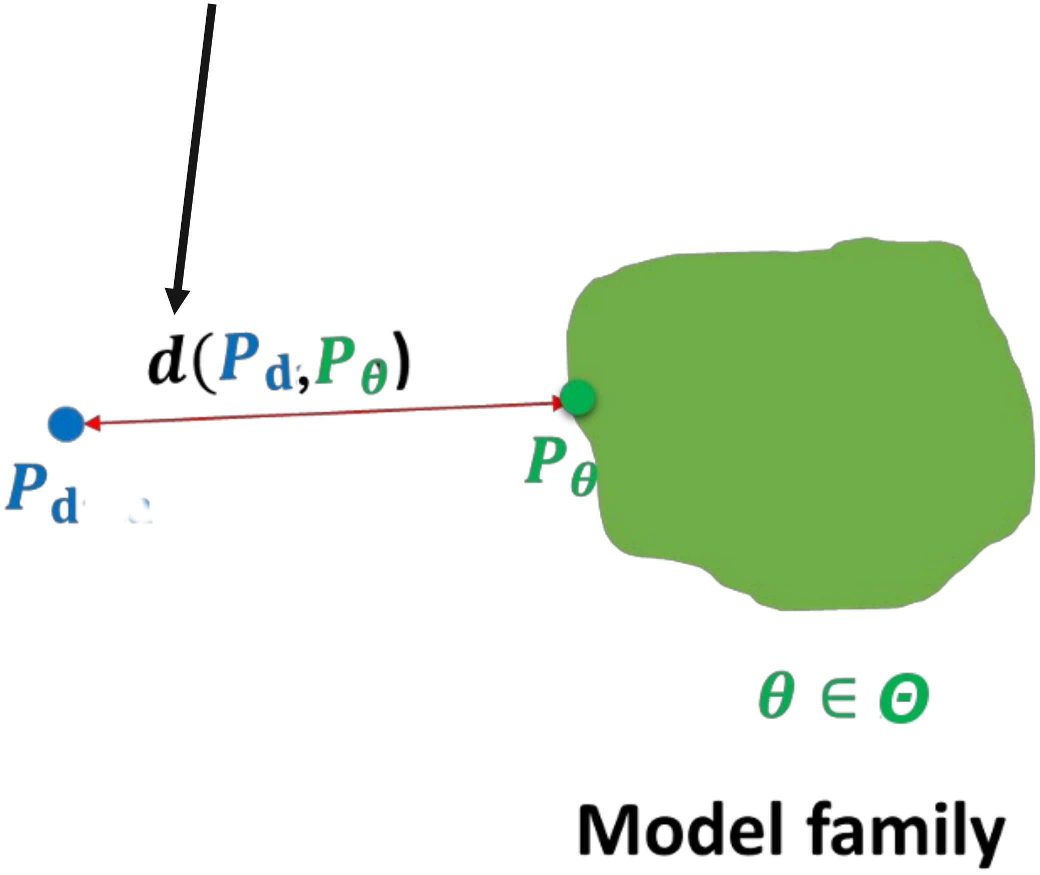
$\hookrightarrow$ The variance of the sampling distribution is a measure of uncertianty about $\hat{\theta}(\mathcal{D})$.

$\hookrightarrow$ One standard approach to approximate the sampling distribution is the bootstrap algorithm.

Kullback-Leibler Divergence (KLD)

- Geometric interpretation:
  MLE is equivalent to minimizing the KLD of $p_d(x)$ w.r.t. $p_\theta(x)$.

$x_i \sim P_d$
$i = 1, 2, \ldots, n$



Maximum Likelihood Estimator

- MLE asymptotics:

$$L_n(\theta) \overset{n\to\infty}{\longrightarrow} \int \log p_\theta(x) p_d(x) dx = \mathbb{E}_{p_d}[\log p_\theta(x)] =: L(\theta; p_d) \ \big(\equiv L(\theta) \big).$$

- MLE is equivalent to minimizing the cross entropy!

$$\arg\max_\theta L(\theta; p_d) = \arg\min_\theta H^\times(p_d || p_\theta).$$

where the cross entropy of probability P with PDF $p(x)$ with respect to probability Q with PDF $q(x)$ is defined as

$$H^\times(P||Q) := \int \log \frac{1}{q(x)} p(x) dx = -\int \log q(x) p(x) dx.$$

Kullback-Leibler Divergence

- MLE is also equivalent to minimizing the KLD of $p_d(x)$ w.r.t. $p_\theta(x)$.

$$\arg\max_\theta L(\theta; p_d) = \arg\min_\theta D_{KL}(p_d || p_\theta)$$

- The Kullback-Leibler divergence (KLD) of P w.r.t. Q is defined as:

$$D_{\mathrm{KL}}(P||Q) := \int \log \frac{p(x)}{q(x)} p(x)dx = \underbrace{\int \log p(x)p(x)dx}_{-H(P)} - \underbrace{\int \log q(x)p(x)dx}_{H^\times(P||Q)}$$

Entropy

Cross Entropy

- Thus,

$$D_{\mathrm{KL}}(P||Q) = -H(P) + H^\times(P||Q).$$

Kullback-Leibler Divergence

- KLD satisfies the divergence property:

$$D_{\mathrm{KL}}(P||Q) \geq 0 \text{ and } D_{\mathrm{KL}}(P||Q) = 0 \iff P = Q.$$

Proof

Jensen's inequality

$$\mathbb{E}_P\left[-\log\frac{q(x)}{p(x)}\right] \geq -\log\left(\mathbb{E}_P\left[\frac{q(x)}{p(x)}\right]\right) = -\log\left(\int\frac{q(x)}{p(x)}p(x)dx\right) = 0$$

- KLD is asymmetric, i.e., $D(P||Q) \neq D(Q||P)$.

- Nevertheless, it offers a notion of a (pseudo-)distance.

Maximum A Posteriori Estimator

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}.$$

1. $p(\mathcal{D}|\theta) = p_\theta(\mathcal{D})$: likelihood.
2. $p(\theta)$: prior probability (prior knowledge).
3. $p(\mathcal{D}) = \int p(\theta')p(\mathcal{D}|\theta')d\theta'$: evidence (usually intractable but with tractable approximations).

Maximum A Posteriori Estimator

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \{\log p_\theta(\mathcal{D}) + \log p(\theta)\}.$$

- 1d Gaussian: $\mathcal{D} = \{x_1, \ldots, x_n\}$, $p_\theta(x) = \mathcal{N}(\theta, \sigma^2)$, $p(\theta) = \mathcal{N}(\theta_0, \sigma_0^2)$.

$$\frac{d}{d\theta} L_{\text{MAP}}(\theta) = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \theta) - \frac{1}{\sigma_0^2}(\theta - \theta_0) = 0$$

$$\implies \hat{\theta}_{\text{MAP}} = \frac{\sum_{i=1}^{n} x_i - \rho\theta_0}{n - \rho}, \quad \rho = \frac{\sigma^2}{\sigma_0^2}.$$

$\hookrightarrow$ What happens as n increases?

Maximum A Posteriori Estimator

- Often, the prior probability acts as regularization.

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \{\log p_\theta(\mathcal{D}) + \log p(\theta)\}.$$

- Linear model: $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, model: $y_i = \theta^T x_i + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, 1)$

  – $p(\theta) = \mathcal{N}(0, \lambda^{-1} I_d) \Rightarrow$ rigde regression a.k.a. (Tikhonov) regularized Least Squares.

  – $p(\theta) = \text{Laplace}(0, \lambda^{-1}) \Rightarrow$ lasso regression (least absolute shrinkage and selection operator).

Estimator Assessment

- Basic toolkit to assess an estimator:
  1. Unbiasedness.
  2. Consistency.
  3. Bias-Variance Trade-Off.
  4. Efficiency.
  5. Fisher Information.
  6. Cramér-Rao Lower Bound (CRLB).

Estimator Assessment

1. Unbiasedness:

Key assumption: $\exists \theta^* \quad \text{s.t.} \quad p_d(x) = p_{\theta^*}(x)$.

- An unbiased estimator is an estimator whose expected values (w.r.t. the data generation distribution) is equal to the parameter:

$$\text{Bias}(\hat{\theta}) = \theta^* - \mathbb{E}_{p_d}[\hat{\theta}].$$

- The *sample mean* $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n x_i$ with $x_i \sim p_{\theta^*}(x) \equiv \mathcal{N}(\theta^*, \sigma^2)$, i.i.d., is an unbiased estimator.

  Proof:

$$\mathbb{E}_{p_d}[\hat{\theta}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{p_{\theta^*}}[x_i] = \frac{1}{n}\sum_{i=1}^n \theta^* = \frac{1}{n} n\theta^* = \theta^*.$$

Estimator Assessment

1. Unbiasedness:

- An asymptotically unbiased estimator is the least requirement for an estimator:

$$\lim_{n\to\infty} \text{Bias}(\hat{\theta}_n) = 0.$$

– Example: Let $\theta = \mathbb{E}_{p_d}[g(x)]$ and $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$ its estimator.

$\hookrightarrow$ $\hat{\theta}_n$ is unbiased and the basic idea of **Monte Carlo** methods.

– Example: Let $\theta = g(\mathbb{E}_{p_d}[x])$ and $\hat{\theta}_n = g\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$ its estimator.

$\hookrightarrow$ $\hat{\theta}_n$ is biased, but asymptotically it is an unbiased estimator.

Estimator Assessment

2. Consistency:

- An unbiased estimator is said to be consistent if the difference between the estimator and the true value becomes smaller as we increase the sample size. Formally:

$$\lim_{n\to\infty} P_{p_d}(|\hat{\theta}_n - \theta^*| > \epsilon) = 0 \ , \quad \forall \epsilon > 0.$$

–Example (consistent): Sample mean $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ with $x_i \sim \mathcal{N}(\theta^*, \sigma^2)$, i.i.d.

Chebyshev's inequality

$$\text{Var}_{p_d}(\hat{\theta}_n) = \frac{\sigma^2}{n} \Rightarrow P_{p_d}(|\hat{\theta}_n - \theta^*| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

–Example (not consistent): $\hat{\theta}_{10} = \frac{1}{10}\sum_{i=1}^{10} x_i$.

Estimator Assessment

2. Consistency and (asymptotic) unbiasedness:

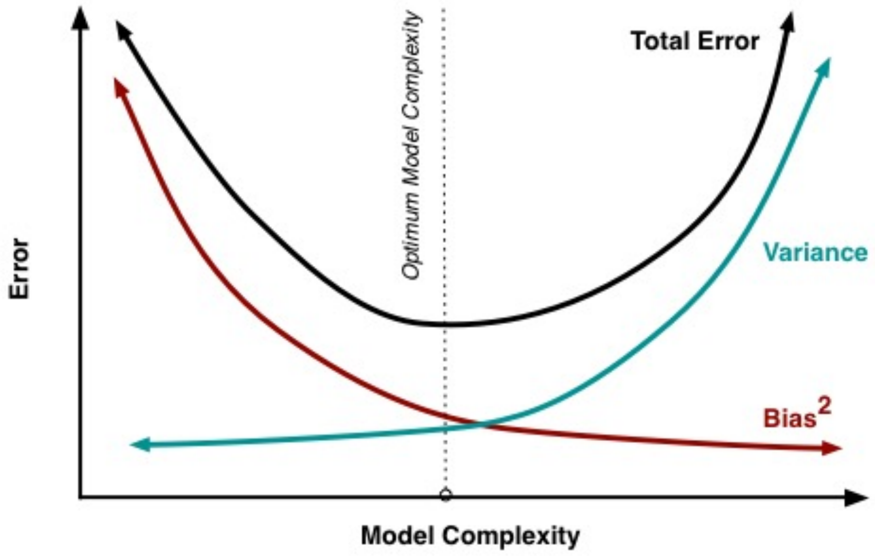
- *Proposition:* If $\mathrm{Var}_{p_d}(\hat{\theta}_n)$ is finite, then, consistency implies asymptotic unbiasness.
- *Proposition:* If $\mathrm{Var}_{p_d}(\hat{\theta}_n)$ tends to 0 as $n \to \infty$, then, asymptotic unbiasness implies consistency.
- *Proposition:* If the Mean Squared Error $\mathrm{MSE}(\hat{\theta}_n) := \mathbb{E}_{p_d}\left[(\hat{\theta}_n - \theta^*)^2\right]$ tends to 0 as $n \to \infty$, then, the estimator $\hat{\theta}_n$ is consistent.

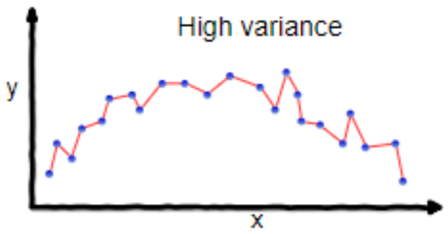
Estimator Assessment

3. Bias-variance trade-off in estimation theory:

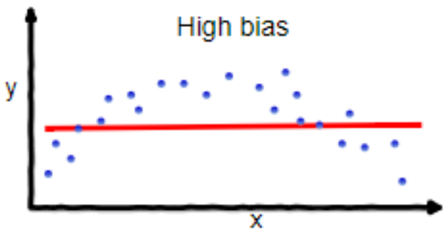
$$\mathrm{MSE}(\hat{\theta}) := \mathbb{E}_{p_d}\left[(\hat{\theta}-\theta^*)^2\right] = \mathrm{Bias}^2_{p_d}(\hat{\theta}) + \mathrm{Var}_{p_d}(\hat{\theta}).$$

Bias-variance trade-off in machine learning:

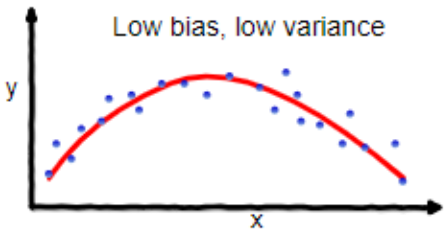




overfitting



underfitting



Good balance

Estimator Assessment

4. Efficiency:

- Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators of θ^*. $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if and only if $\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$.

- An estimator $\hat{\theta}$ is efficient if the variance of the estimator, $\mathrm{Var}(\hat{\theta})$, equals the Cramér-Rao lower bound.

Estimator Assessment

5. Fisher Information:

$$I(\theta) = \mathbb{E}_{p_\theta}\left[\left(\frac{d}{d\theta}\log p_\theta(x)\right)^2\right].$$

$\hookrightarrow\ I(\theta) = \mathrm{Var}_{p_\theta}\left(\frac{d}{d\theta}\log p_\theta(x)\right) = -\mathbb{E}_{p_\theta}\left[\frac{d^2}{d\theta^2}\log p_\theta(x)\right],$

since $-\mathbb{E}_{p_\theta}\left[\frac{d}{d\theta}\log p_\theta(x)\right] = 0.$

–Example: $p_\theta(x) = \mathrm{Bernoulli}(\theta),\ \theta \in [0,1]$: success probability.

$I(\theta) = \cdots = \frac{1}{\theta(1-\theta)}.$

Estimator Assessment

6. Cramér-Rao Lower Bound (CRLB):

- The variance of any unbiased estimator $\hat{\theta}_n$ of θ^* is bounded by the reciprocal of the Fisher information:

$$\mathrm{Var}_{p_{\theta^*}}(\hat{\theta}_n) \geq \frac{1}{I_n(\theta^*)} = \frac{1}{nI(\theta^*)},$$

n: # i.i.d. samples drawn from $p_{\theta^*}(x)$.

- MLE is asymptotically efficient!

References

1. All of statistics: A Concise Course in Statistical Inference (*Chapters 6 & 9*)
   Larry Wasserman, Springer (2004)

2. Probabilistic Machine Learning: An Introduction (*Chapter 4*)
   Kevin P Murphy, The MIT Press (2022)

3. Matrix Calculus:
   `https://en.wikipedia.org/wiki/Matrix_calculus`
   `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`

Introduction to Deep Generative Modeling

Lecture #3

HY-673 – Computer Science Dep., University of Crete
Professors: Yannis Pantazis & Yannis Stylianou
TAs: Michail Raptakis & Michail Spanakis