The Law of Large Numbers (LLN)

- Let X_i for $i = 1, 2, \ldots, n$ be independent and identically distributed random variables (i.i.d.) with finite mean $\mathbb{E}(X_i) = \mu$ & variance $\text{Var}(X_i) = \sigma^2$.

- Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

- Note that

$$\mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

$$Var[\bar{X}_n] = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

- Weak LLN: $\lim_{n\to\infty} \Pr[|\bar{X}_n - \mu| \geq \varepsilon] = 0 \; \forall \varepsilon > 0$

- Strong LLN: $\lim_{n\to\infty} \bar{X}_n = \mu \; almost\; surely$

This means that with probability one, the average of any realizations of $x_1, x_2, \ldots$ of the random variables $X_1, X_2, \ldots$ converges to

Example of the Law of Large Numbers

- Assume that we sample $S = \{x_1, x_2, \ldots x_N\}, \quad x_j \in \mathbb{R}^2$
- We consider a parametric model with x_j realizations of $X \sim \mathcal{N}(x_0, \Sigma)$

  where we take both the mean x_0 and the variance $\Sigma \in \mathbb{R}^{2\times 2}$ as unknowns.

- The probability density of X is:

$$\pi(x \mid x_0, \Sigma) = \frac{1}{2\pi\left(\det \Sigma\right)^{1/2}} \exp\left(-\frac{1}{2}(x - x_0)^T \Sigma^{-1}(x - x_0)\right)$$

- Our problem is to estimate x_0 and $\Sigma \in \mathbb{R}^{2\times 2}$

Empirical Mean and Empirical Covariance

- From the law of large numbers, we calculate:

$$x_0 = \mathbb{E}[X] \approx \frac{1}{N}\sum_{j=1}^{N} x_j = \bar{x}$$

- To compute the covariance matrix, note that if $X_1, X_2, \ldots$ are i.i.d. so are $f(X_1), f(X_2), \ldots$ for any function $f : \mathbb{R}^2 \to \mathbb{R}^k$

- Then we can compute:

$$\boldsymbol{\Sigma} = \mathrm{cov}(X) = \mathbb{E}\left[(x - \mathbb{E}[X])(x - \mathbb{E}[X])^T\right] \approx \mathbb{E}\left[(x - \bar{x})(x - \bar{x})^T\right]$$

$$\Rightarrow$$

$$\boldsymbol{\Sigma} \approx \frac{1}{N}\sum_{j=1}^{N}(x_j - \bar{x})(x_j - \bar{x})^T = \bar{\boldsymbol{\Sigma}}$$

- The above formulas define **the empirical mean and empirical covariance.**

The Central Limit Theorem

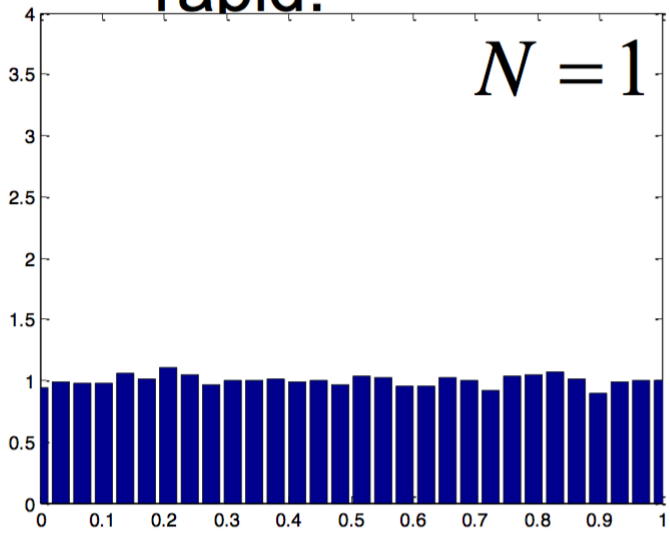
- Let $(X_1, X_2, \ldots X_N)$ be independent and identically distributed (i.i.d.) continuous random variables each with expectation μ and variance σ^2.
- Define: $Z_N = \frac{1}{\sigma\sqrt{N}}(X_1 + X_2 + \ldots + X_N - N\mu) = \frac{\bar{X}_N - \mu}{\frac{\sigma}{\sqrt{N}}}, \bar{X}_N = \frac{1}{N}\sum_{j=1}^{N} X_j$
- As $N -> \infty$, the distribution of Z_N converges to the distribution of a standard normal random variable

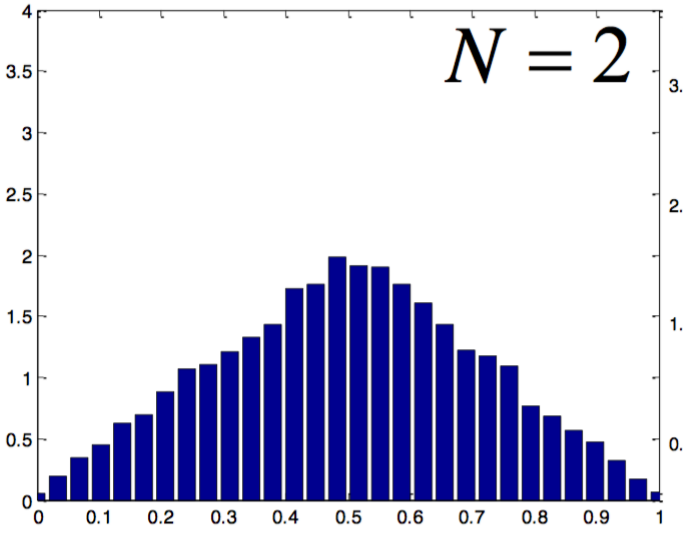
$$\lim_{N\to\infty} P\{Z_N \le x\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

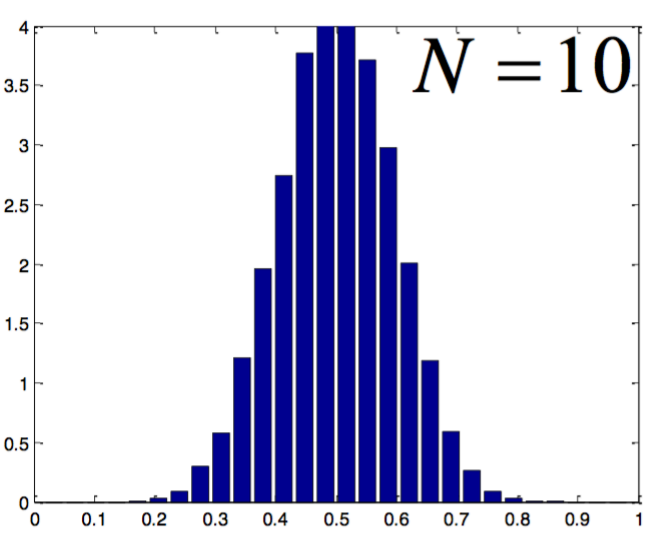
- If $\bar{X}_N = \frac{1}{N}\sum_{j=1}^{N} X_j$, for N large, $\bar{X}_N \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$ $as\ N \to \infty$
- Somewhat of *a justification for assuming Gaussian noise is common*

The CLT and the Gaussian Distribution

- As an example, assume N variables $(X_1, X_2, \ldots X_N)$ each of which has a uniform distribution over $[0, 1]$ and then consider the distribution of the sample mean $(X_1 + X_2 + \ldots + X_N)/N$. For large N, this distribution tends to a Gaussian. The convergence as N increases can be rapid.







Monte Carlo Approximation

- Consider evaluating the integral

$$E_f[h(X)] = \int_{\mathcal{X}} h(x) f(x) dx$$

using the Monte Carlo estimate

$$\hat{h}_J = \frac{1}{J} \sum_{j=1}^{J} h\left(x^{(j)}\right)$$

where $x^{(j)} \overset{iid}{\sim} f(x)$. We know

- SLLN: $\hat{h}_J$ converges almost surely to $E_f[h(X)]$.
- CLT: if h^2 has finite expectation under f, then

$$\hat{h}_J \xrightarrow{d} N(E_f[h(X)], v_J)$$

where

$$v_J = \frac{1}{J} \widehat{V_f[h(X)]} = \frac{1}{J^2} \sum_{i=1}^{J} \left[h\left(x^{(j)}\right) - \hat{h}_J\right]^2.$$

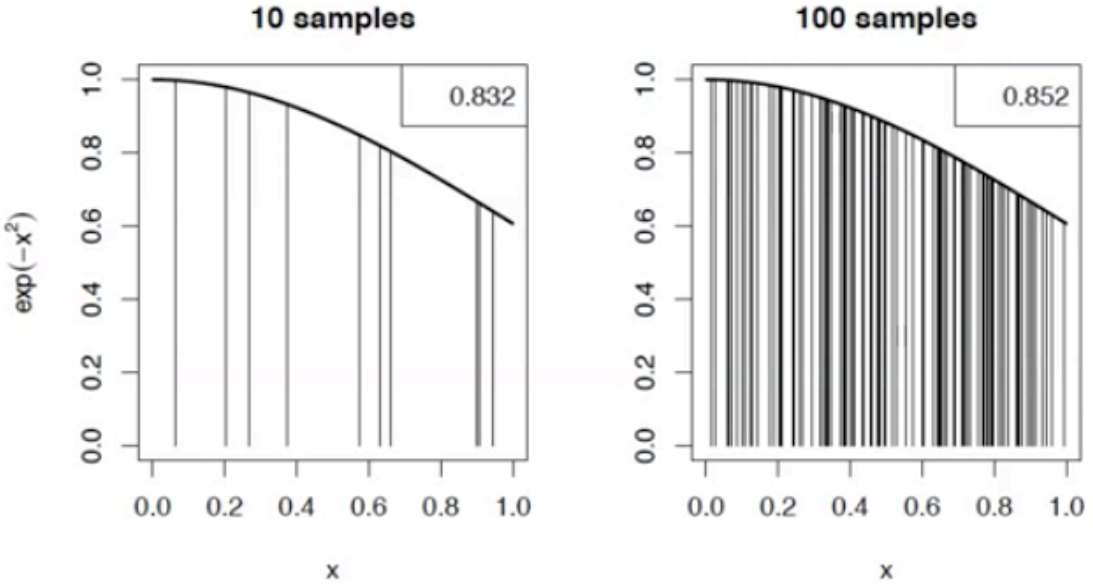
Monte Carlo Approximation

- Suppose you are interested in evaluating *

$$\int_0^1 e^{-x^2/2} dx.$$

Then set

- $h(x) = e^{-x^2/2}$ and
- $f(x) = 1$, i.e. $x \sim \text{Unif}(0, 1)$.



* Jarad Niemi, https://www.youtube.com/watch?v=MKnjsqYVG4Y

MLE Estimate: Multinoulli Distribution

- To compute the maximum likelihood (MLE) estimate of $\boldsymbol{\mu}$, we maximize an augmented log-likelihood

$$\ln p(\mathcal{D} \mid \boldsymbol{\mu}) + \lambda\left(\sum_{k=1}^{K}\mu_k - 1\right) = \sum_{k=1}^{K} m_k \ln \mu_k + \lambda\left(\sum_{k=1}^{K}\mu_k - 1\right)$$

- Setting the derivative wrt μ_k equal to zero: $\mu_K = -\dfrac{m_k}{\lambda}$

- Substitution into the constraint

$$\sum_{k=1}^{K}\mu_k = 1 \Rightarrow -\frac{\sum_{k=1}^{K} m_k}{\lambda} = 1 \Rightarrow \lambda = -\sum_{k=1}^{K} m_k \Rightarrow$$

$$\mu_K = \frac{m_k}{\sum_{k=1}^{K} m_k} = \frac{m_k}{N}$$

As expected, this is the fraction in the N observations of $x_k = 1$

Posterior Inference: Point Estimates

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{m(x)}$$

Maximum A Posteriori estimate (MAP)

$$\theta^* = \arg\max_{\theta} \log\left(\pi(\theta \mid x)\right) = \arg\max_{\theta} \left(\log f(x \mid \theta) + \log \pi(\theta)\right)$$

Posterior Mean

$$\hat{\theta} = \mathbb{E}_{p(\theta|x)}[\theta] = \int \theta \pi(\theta|x) d\theta$$

Posterior Quantiles

$$\Pr\left[\theta > a\right] = \int_a^{\infty} \pi(\theta \mid x) d\theta$$

Appendix: Laplace Approximation

- The Laplace approximation allows a Gaussian approximation of the parameter posterior about the maximum a posteriori (MAP) parameter estimate.
- Consider a data set $\mathcal{D}$ and M models $\mathcal{M}_i$, $i=1,..,M$ with corresponding parameters θ_i, i=1,…M. We compare models using the posteriors:

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

- For large sets of data $\mathcal{D}$ (relative to the model parameters), the parameter posterior is approximately Gaussian around the MAP estimate θ_m^{MAP} (can also use 2nd order Taylor expansion of the log-posterior):

$$p(\theta_m \mid \mathcal{D}, \mathcal{M}_m) \approx (2\pi)^{-d/2} |A|^{1/2} \exp\left(-\frac{1}{2} \left(\theta_m - \theta_m^{MAP}\right)^T A \left(\theta_m - \theta_m^{MAP}\right) \right),$$

$$A_{ij} = -\left. \frac{\partial^2 \log P(\theta_m \mid \mathcal{D}, \mathcal{M}_m)}{\partial \theta_{mi} \partial \theta_{mj}} \right|_{\theta_m^{MAP}}$$

Bayesian Information Criterion

- Start with the Laplace approximation for large data sets N→∞,

$$\log p(\mathcal{D}\mid \mathcal{M}_m) \approx \log p(\mathcal{D}\mid \theta_m^{MAP}, \mathcal{M}_m) + \log p(\theta_m^{MAP}\mid \mathcal{M}_m) + \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|A|$$

- A *N* grows, $\boldsymbol{A}$ grows as $N\boldsymbol{A}_0$ for some fixed matrix $\boldsymbol{A}_0$, thus

$$\log|A| \to \log|NA_0| = \log\left(N^d\,|A_0|\right) = d\log N + \log\left(|A_0|\right) \xrightarrow{N\to\infty} d\log N$$

- Then the Laplace approximation is simplified as:

$$\log p(\mathcal{D}\mid \mathcal{M}_m) \approx \log p(\mathcal{D}\mid \theta_m^{MAP}, \mathcal{M}_m) - \frac{d}{2}\log N \quad (\text{limit } N\to\infty)$$

- Note interesting properties of (the easy to compute) BIC:
  - No dependence on the prior
  - One can use the MLE rather than the MAP estimate of (but use MAP when working with mixtures of Gaussians)
  - If not all parameters are well determined from the data, θ_m *d*=number of effective parameters.