CS 217: Artificial Intelligence and Machine Learning **Jan-Apr 2024**

Lecture 5: Maximum Aposteriori (MAP) Estimate

Lecturer: Swaprava Nath *Scribe(s): SG9 & SG10*

5.1 Gradient Descent

Gradient descent is a method for unconstrained mathematical optimization. It is a **first-order iterative algorithm** for finding a local minimum of a differentiable multivariate function.

The idea is to take repeated steps in the opposite direction of the gradient (or approximate gradient) of the function at the current point because this is the **direction of steepest descent**. It is particularly useful in machine learning for minimizing the cost or loss function.

- Initialization : $w \leftarrow w_o$
- Iterate until convergence : $\|\nabla_w E\|_2 < \epsilon$
- Minimize the error function i.e. $E(D, w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2$
- Gradient Descent is excellent in terms of accuracy (especially for convex functions) but expensive in terms of computation

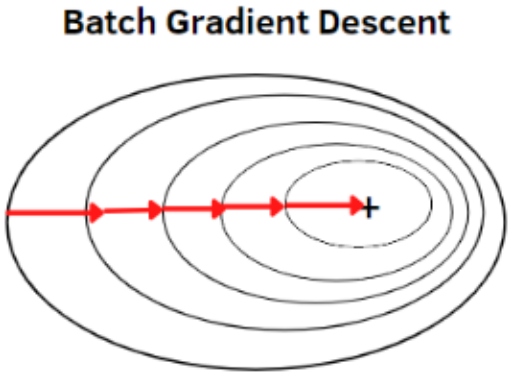


Figure 5.1: Gradient Descent

Algorithm 1 Gradient Descent(X,Y,w,η)

$\epsilon \leftarrow 1e-15$ ▷ Set ϵ as the limit for convergence
$old\_loss \leftarrow 0$
while $abs(old\_loss - f(X,Y,w)) > \epsilon$ **do**
  $old\_loss \leftarrow f(X,Y,w)$
  $dw \leftarrow \nabla f(X,Y,w)$
  $w \leftarrow w - \eta * dw$
end while

5.2 Stochastic Gradient Descent

Stochastic gradient descent is an iterative method for optimizing an objective function with suitable smoothness properties. It can be regarded as a stochastic approximation of gradient descent optimization, since it replaces the actual gradient (calculated from the entire data set) by an estimate thereof (calculated from a randomly selected subset of the data). Especially in high-dimensional optimization problems this **reduces the very high computational burden, achieving faster iterations in exchange for a lower convergence rate**.

- Reduce the computation required as compared to Gradient Descent
- Works faster and very well in practice especially for large datasets
- This randomness can help prevent overfitting by preventing the algorithm from getting stuck in local minima
- In cases where the objective function is smooth and well-behaved, the frequent noise introduced by SGD may not be necessary

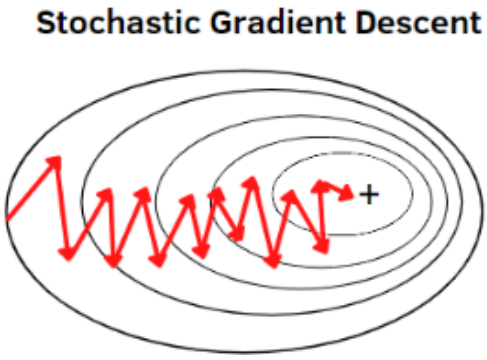


Figure 5.2: Stochastic Gradient Descent

Algorithm 2 Stochastic Gradient Descent

$w \leftarrow w_0$
while $||\nabla_w E|| < \epsilon$ **do** ▷ Setting ϵ for limit for convergence
  $i \leftarrow random \in \{1, 2, ...n\}$
  $w \leftarrow w - \eta \nabla_w E(w, Xi, Yi)$
end while

5.3 Mini Batch Gradient Descent

Mini-batch gradient descent is a variant of gradient descent algorithm. The idea behind this algorithm is to divide the training data into batches, which are then processed sequentially. In each iteration, we update the weights of all the training samples belonging to a particular batch together. This process is repeated with different batches until the whole training data has been processed. Compared to batch gradient descent, the main benefit of this approach is that it can reduce computation time and memory usage significantly as compared to processing all training samples in one shot

- Introduces a certain level of noise, which can have a regularizing effect and help the model generalize better, potentially avoiding overfitting.
- It allows for more flexibility in adjusting the learning rate compared to SGD. Learning rates can be adjusted dynamically based on the characteristics of the optimization.
- The selection of an appropriate batch size is a hyperparameter that needs to be tuned. Different batch sizes may affect the convergence speed and generalization performance of the algorithm.

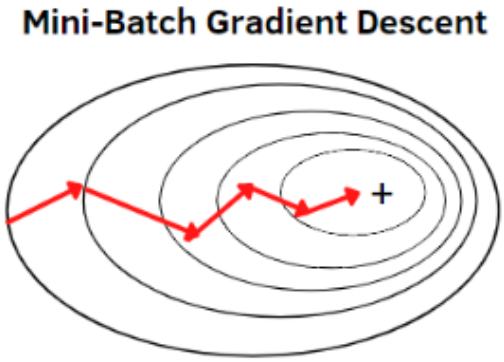


Figure 5.3: Mini Batch Gradient Descent

Algorithm 3 Mini Batch Gradient Descent

$w \leftarrow w_0$
while $||\nabla_w E|| < \epsilon$ **do** ▷ Setting ϵ limit for convergence
  $B \leftarrow random \subset \{1, 2, ...n\}$
  $w \leftarrow w - \eta \sum_{i \in B} \nabla_w E_i$
end while

5.4 Maximum Likelihood Estimate

- Consider D to be a dataset, we can represent it as $D = \{(X_i, y_i)_{i \in \{1,2,...n\}}\}$
- Each X_i may be a vector consisting of a lot of values Consider $y_i = w^T X_i + \epsilon_i$
- Define w as a parameter θ

- We want to find the parameter θ under which, the data is most likely to have occurred That is,

$$\theta_{MLE} = \arg\max_{\theta} P(D|\theta)$$

Coin Toss Example: A coin is tossed n times and y_j is the j^{th} outcome. y_j is Bernoulli random variable

$$y_j = \begin{cases} 1 & \text{with probability } \theta_j; \\ 0 & \text{with probability } 1-\theta_j; \end{cases}$$

$$P(y_j|\theta) = \theta^{y_j}(1-\theta)^{1-y_j}$$

The likelihood is given by,

$$P(y_j|\theta) = \prod_{j=1}^{n} \theta^{y_j}(1-\theta)^{1-y_j}$$

For getting the MLE, maximize the function with respect to θ

5.5 Maximum A Posteriori Estimation (MAP)

In Bayesian statistics, a maximum a posteriori probability (MAP) estimate is an estimate of an unknown quantity, that equals the mode of the posterior distribution. The MAP can be used to obtain a point estimate of an unobserved quantity on the basis of empirical data. It is closely related to the method of maximum likelihood (ML) estimation, but employs an augmented optimization objective which incorporates a prior distribution (that quantifies the additional information available through prior knowledge of a related event) over the quantity one wants to estimate

Prior, Likelihood, and Posterior in MAP Estimation

1. **Prior ($P(\theta)$):**
   - **Definition:** The prior represents our beliefs or knowledge about the parameters before observing any data. It encapsulates our initial assumptions or existing information about the parameters.
   - **Mathematically:** Denoted as $P(\theta)$, where θ is the parameter. It provides a probability distribution for the parameter before incorporating any new data.
   - **Example:** If we are estimating the probability of success in a coin toss, our prior might express our belief that the coin is fair, leading to a prior distribution centered around 0.5.

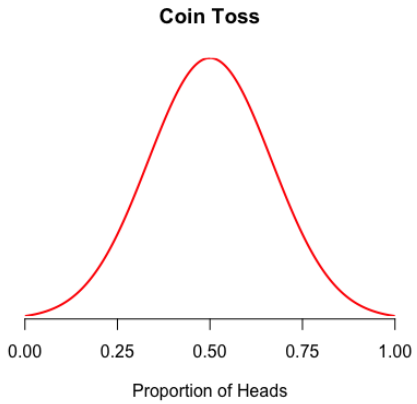


Figure 5.4: Example : Prior Distribution for a Fair Coin

2. **Likelihood ($P(D|\theta)$):**

   - **Definition:** The likelihood represents the probability of observing the given data given a specific set of parameters. It quantifies how well the parameters explain the observed data.
   - **Mathematically:** Denoted as $P(D|\theta)$, where D is the observed data and θ is the parameter. The likelihood describes the data-generation process under the assumed parameter values.
   - **Example:** If we are estimating the probability of success in a coin toss, the likelihood might express how likely it is to observe a certain sequence of heads and tails given a particular probability of success.

3. **Posterior ($P(\theta|D)$):**

   - **Definition:** The posterior is the updated probability distribution for the parameters after incorporating the observed data. It combines the prior information with the new evidence from the likelihood.
   - **Mathematically:** Denoted as $P(\theta|D)$, where θ is the parameter and D is the observed data. The posterior is proportional to the product of the prior and the likelihood.
   - **Example:** Continuing with the coin toss example, the posterior distribution would reflect our updated beliefs about the probability of success after observing a specific sequence of heads and tails.

MAP Estimate: The MAP estimate seeks to find the parameter values that maximize the posterior distribution. Mathematically, it can be expressed as:

$$\text{MAP Estimate: } \theta_{\text{MAP}} = \arg\max_{\theta} P(\theta|D)$$

Now, using Bayes' theorem, we can express the posterior distribution in terms of the likelihood and the prior:

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)}$$

where $P(\theta|D) =$ Posterior, i.e., how probable is our estimate of the parameter given the observed evidence (data)
$P(D|\theta) =$ Likelihood, i.e., how probable is the evidence given our estimate
$P(\theta) =$ Prior, i.e., how probable was our estimate before observing the evidence
$P(D) =$ Marginal, i.e., how probable is the new evidence under all possible estimates

Taking the logarithm of the posterior distribution, we have:

$$\log P(\theta|D) \propto \log P(D|\theta) + \log P(\theta)$$

Now, comparing this with the MLE, you can see that the MAP includes an additional term, $\log P(\theta)$, which represents the contribution of the prior distribution.

Therefore, the equations for θ_{MLE} and θ_{MAP} are as follows:

$$\theta_{\text{MLE}} = \arg\max_{\theta} P(D|\theta)$$

$$\theta_{\text{MAP}} = \arg\max_{\theta} \Big(\log P(D|\theta) + \log P(\theta)\Big)$$

This estimate balances the information from the prior and the likelihood, providing a point estimate for the parameter(s) based on both prior knowledge and observed data.

Note: If $P(\theta)$ is constant (i.e., follows a **uniform distribution**), then $(\theta_{\text{MAP}} = \theta_{\text{MLE}})$

Example : Maximum Likelihood Estimate of Bias of a Coin

Assuming a biased coin with probability of heads θ, the probability mass function for k heads in n tosses is given by the binomial distribution:

$$P(X = k) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \sim \text{Binomial}(n, k) \tag{5.1}$$

The log-likelihood function $LL(\theta)$ is the logarithm of the likelihood function:

$$LL(\theta) = \log\left(\binom{n}{k} \theta^k (1-\theta)^{n-k}\right) \tag{5.2}$$

Now, let's find the maximum likelihood estimate (MLE) by maximizing $LL(\theta)$ with respect to θ.

$$LL(\theta) = \log\left(\binom{n}{k} \theta^k (1-\theta)^{n-k}\right) \tag{5.3}$$

$$= \log\binom{n}{k} + k\log(\theta) + (n-k)\log(1-\theta) \tag{5.4}$$

To find the MLE, take the derivative of $LL(\theta)$ with respect to θ and set it to zero:

$$\frac{dL}{d\theta} = \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0 \tag{5.5}$$

Solving for θ, we get:

$$\frac{k}{\theta} = \frac{n-k}{1-\theta} \tag{5.6}$$
$$k(1-\theta) = \theta(n-k) \tag{5.7}$$
$$k - k\theta = n\theta - k\theta \tag{5.8}$$
$$k = n\theta \tag{5.9}$$
$$\theta = \frac{k}{n} \tag{5.10}$$

Therefore, the Maximum Likelihood Estimate (MLE) for the bias of the coin, given k heads in n tosses, is

$$\boxed{\theta_{\text{MLE}} = \frac{k}{n}} \tag{5.11}$$

5.6 Beta Distribution

The beta distribution, denoted by Beta(α, β), is a continuous probability distribution defined on the interval $(0, 1)$. The probability density function (PDF) of the beta distribution is given by:

$$f(\theta; \alpha, \beta) = C \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where C is the normalization constant, and α and β determine the shape of the distribution. When $\alpha = \beta = 1$, the beta distribution is the uniform distribution on $(0, 1)$.

Note: Setting both parameters α and β to 1 in the beta distribution results in the distribution converging to the uniform distribution on the interval $(0, 1)$.
Use this link to see how the function looks like for arbitrary α, β Beta_Distribution

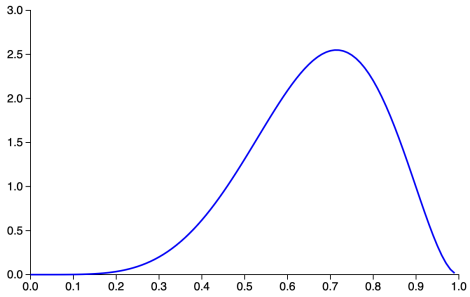
Normalization Constant

The normalization constant C is given by:

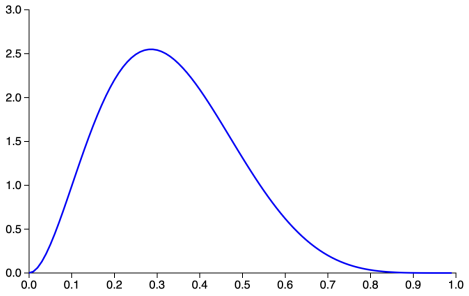
$$C = \frac{1}{\text{B}(\alpha, \beta)} = \frac{1}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\, d\theta}$$

Mean and Variance

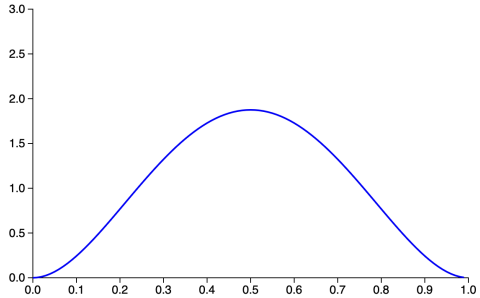
The mean (μ) and variance (σ^2) of the beta distribution are given by:



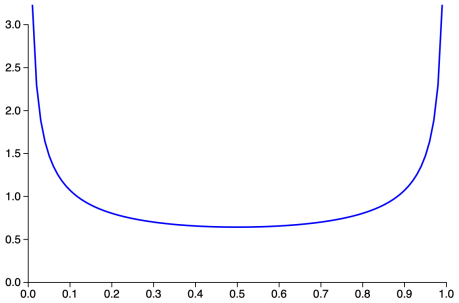
(a) $\alpha = 6, \beta = 3$



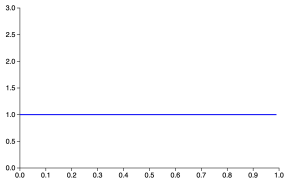
(b) $\alpha = 3, \beta = 6$



(c) $\alpha = \beta = 3$



(d) $\alpha = \beta = 0.5$



(e) $\alpha = \beta = 1$

Figure 5.5: Beta Distribution for different α, β

$$\mu = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

5.7 Conjugate Prior

In Bayesian statistics, a prior distribution $P(\theta)$ is said to be a *conjugate prior* for a likelihood function $P(D|\theta)$ if the resulting posterior distribution $P(\theta|D)$ belongs to the same family of distributions as the prior.

Likelihood Distribution	**Conjugate Prior**
Bernoulli / Binomial	Beta
Geometric	Beta
Categorical	Dirichlet
Normal	Normal

Example: Beta-Binomial Conjugacy

Consider a binomial likelihood function with parameters n (number of trials) and θ (probability of success in each trial). If we choose a Beta distribution as the prior for θ, then the posterior distribution after observing data D will also be a Beta distribution.

The Beta distribution is the conjugate prior for the binomial likelihood, and the posterior is given by:

$$P(\theta|D) \propto P(D|\theta) \cdot P(\theta) \quad \text{where} \quad P(\theta|D) \sim \text{Beta}(\alpha + k, \beta + n - k)$$

Here, α and β are the parameters of the Beta prior, k is the number of successes observed in the data, and n is the total number of trials.

Calculating θ_{MAP} for the Previous Coin Problem

We want to estimate the bias (θ) of a coin based on k heads observed in n tosses. The prior distribution is Beta(α, β).

The posterior distribution is given by Bayes' theorem:

$$P(\theta|D) \propto P(D|\theta) \cdot P(\theta)$$

Assuming a binomial distribution for the likelihood (coin tosses):

$$P(D|\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

The prior distribution is Beta(α, β):

$$P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

$$P(\theta|D) \propto \binom{n}{k} \theta^k (1-\theta)^{n-k} \cdot \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

$$\log P(\theta|D) \propto (k + \alpha - 1)\log(\theta) + (n - k + \beta - 1)\log(1-\theta)$$

Set $\frac{d}{d\theta} \log P(\theta|D) = 0$ and solve for θ:

$$\frac{k + \alpha - 1}{\theta} - \frac{n - k + \beta - 1}{1 - \theta} = 0$$

$$\theta(n + \alpha + \beta - 2) = k + \alpha - 1$$

$$\boxed{\theta_{MAP} = \frac{k + \alpha - 1}{n + \alpha + \beta - 2}} \tag{5.12}$$

Note that when $\alpha = \beta = 1$, i.e., the beta distribution is an uniform one, the MAP estimate coincides with the MLE we calculated (Eq. 5.11).

5.8 MAP estimate for Univariate Gaussian with known Varience

The likelihood for a normal distribution is given by:

$$P(x_1, x_2, \ldots, x_n | \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

The prior is a normal distribution:

$$P(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right)$$

The logarithm of the posterior distribution is:

$$\begin{aligned}
\log P(\mu | x_1, x_2, \ldots, x_n) &\propto -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 + \text{constant} \\
&\propto -\frac{1}{2} \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right) \mu^2 + \left(\frac{\sum_{i=1}^{n} x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) \mu + \text{constant}
\end{aligned}$$

This expression is proportional to a Gaussian distribution. Factoring out common terms, we get:

$$\log P(\mu | x_1, x_2, \ldots, x_n) \propto -\frac{1}{2} \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right) (\mu - \mu_{\text{post}})^2 + \text{constant}$$

Where:

$$\boxed{\mu_{\text{post}} = \frac{\frac{1}{\sigma^2} \sum_{i=1}^{n} x_i + \frac{1}{\sigma_0^2} \mu_0}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}}$$

The variance of the posterior is given by:

$$\boxed{\sigma^2_{\text{post}} = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}}$$

Therefore, the posterior distribution $P(\mu | x_1, x_2, \ldots, x_n)$ is a Gaussian distribution with mean μ_{post} and variance σ^2_{post}.

For complete proof refer to Conjugate Bayesian analysis of the Gaussian distribution.

5.9 MAP Estimate for Linear Regression

The probability density function of a multivariate normal distribution is given by:

$$f(\mathbf{x};\mu,\mathbf{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)$$

$$\mathbf{x} \in \mathbb{R}^d$$

$\mathbf{x}$ is the column vector of the random variables,
$\boldsymbol{\mu}$ is the mean vector,
$\mathbf{\Sigma}$ is the covariance matrix,
$|\mathbf{\Sigma}|$ is the determinant of the covariance matrix,
d is the number of dimensions.

Assuming a linear regression model:

$$y_i = w^\top x_i + \epsilon_i$$

where

$$\epsilon_i \sim \mathcal{N}(0,\sigma^2)$$

Therefore,

$$y_i \sim \mathcal{N}(w^\top x_i, \sigma^2)$$

The likelihood function

$$P(D|w) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w^\top x_i)^2\right\}$$

$$\theta_{\text{MLE}} = \arg\min_w \left(\sum_{i=1}^{n}(y_i - w^\top x_i)^2\right)$$

A Gaussian prior is represented by

$$P(w) \sim \mathcal{N}(\mu, \Sigma),$$

where μ is the mean vector and Σ is the covariance matrix (symmetric) such that each of its entries are

$$\sigma_{ij} = \mathbf{cov}(x_i, x_j) = \mathbb{E}\left((x_i - \mathbb{E}(x_i))(x_j - \mathbb{E}(x_j))\right)$$

Suppose the prior of w is such that

$$P(w) \sim \mathcal{N}(\mathbf{0}, \frac{1}{\lambda}I)$$

$$P(w) = \frac{1}{\sqrt{(2\pi)^k \det\left(\frac{1}{\lambda}\mathbf{I}\right)}} \exp\left(-\frac{1}{2}w^\top \left(\frac{1}{\lambda}\mathbf{I}\right)^{-1} w\right)$$

$$P(w) = (\frac{\lambda}{2\pi})^{\frac{k}{2}} \exp\left(-\frac{\lambda}{2}w^\top w\right)$$

The posterior distribution is proportional to the likelihood times the prior:

$$P(w|D) \propto P(D|w) \cdot P(w)$$

$$\arg\max_w \log(P(w|D)) = \arg\max_w \log(P(D|w)) + \log(P(w))$$

$$= \arg\max_w -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w^\top x_i)^2 + -\frac{\lambda}{2}\|w\|_2^2$$

$$= \arg\min_w \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w^\top x_i)^2 + \frac{\lambda}{2}\|w\|_2^2$$

$$= \arg\min_w \frac{1}{2\sigma^2}\|\mathbf{X}w - y\|_2^2 + \boxed{\frac{\lambda}{2}\|w\|_2^2}$$

The term $\left(\frac{\lambda}{2}\|w\|_2^2\right)$ is called the regularizer.

Note : In the above equations λ is a *hyper-parameter* which is decided beforehand and is a design choice.