1 Just some Regular Ol' Definitions

1.1 Regularization

1. ***Why Regularize?***: Machine learning models tend to overfit if they memorize the training data too closely, failing to generalize well to unseen examples. Regularization techniques tackle this by penalizing complex models, pushing them towards simpler solutions that perform better on unseen data. Norms, like L0, L1, and L2, offer different ways to quantify and penalize model complexity.

2. ***L2-norm***: Also known as the Euclidean norm, is a measure of the magnitude of a vector in Euclidean space. It is calculated as:

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{N} x_i^2} \tag{1}$$

   where N is the dimensionality of the vector $\mathbf{x}$.

3. ***L1-norm***: a measure of the absolute magnitude of a vector. It is calculated as:

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{N} |x_i| \tag{2}$$

   again, where N is the dimensionality of the vector $\mathbf{x}$.

4. ***L0-norm***: also known as the "counting norm", is a measure of the number of non-zero elements in a vector. It is calculated as:

$$\|x\|_0 = \sum_{i=1}^{N} \mathbb{I}(x_i \neq 0) \tag{3}$$

   where $\mathbb{I}(x_i \neq 0)$ is the indicator function that is equal to 1 if $x_i \neq 0$ and 0 otherwise.

Properties:

1. Sparsity: L0 and L1 encourage sparsity (many zero parameters), while L2 prefers smaller but not necessarily zero values.

2. Optimization: L1 and L2 are differentiable, facilitating optimization, while L0 is not.

2 Regularization

Think back to linear regression. We would calculate our objective function that we want to minimize as follows: $J(\boldsymbol{\theta}) = \frac{1}{N}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$.

Consider we add L2 regularization to this objective function.
This would give us: $J(\boldsymbol{\theta}) = \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2$. Determine the closed-form solution of this objective function.

We can expand our objective function as follows:

$$\begin{aligned} J(\boldsymbol{\theta}) &= \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2 \\ &= (\mathbf{y} - X\boldsymbol{\theta})^\top(\mathbf{y} - X\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_2^2 \\ &= (\mathbf{y} - X\boldsymbol{\theta})^\top(\mathbf{y} - X\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^\top\boldsymbol{\theta} \end{aligned}$$

Because we want to optimize this, we take the derivative of $J(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and set it to 0, to get our closed-form solution

$$\begin{aligned} \nabla J(\boldsymbol{\theta}) &= \frac{\partial}{\partial\boldsymbol{\theta}}(\mathbf{y} - X\boldsymbol{\theta})^\top(\mathbf{y} - X\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}^\top\boldsymbol{\theta} \\ &= \frac{\partial}{\partial\boldsymbol{\theta}}(\mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top X\boldsymbol{\theta} - \boldsymbol{\theta}^\top X^\top\mathbf{y} + \boldsymbol{\theta}^\top X^\top X\boldsymbol{\theta} + \lambda\boldsymbol{\theta}^\top\boldsymbol{\theta}) \\ &= \frac{\partial}{\partial\boldsymbol{\theta}}(\mathbf{y}^\top\mathbf{y} - 2\boldsymbol{\theta}^\top X^\top\mathbf{y} + \boldsymbol{\theta}^\top X^\top X\boldsymbol{\theta} + \lambda\boldsymbol{\theta}^\top\boldsymbol{\theta}) \\ &= 0 - 2X^\top\mathbf{y} + 2X^\top X\boldsymbol{\theta} + \frac{\partial}{\partial\boldsymbol{\theta}}(\lambda\boldsymbol{\theta}^\top\boldsymbol{\theta}) \\ &= -2X^\top(\mathbf{y} - X\boldsymbol{\theta}) + 2\lambda\boldsymbol{\theta} \end{aligned}$$

$$\begin{aligned} -2X^\top(\mathbf{y} - X\boldsymbol{\theta}) + 2\lambda\boldsymbol{\theta} &= 0 \\ -X^\top\mathbf{y} + X^\top X\boldsymbol{\theta} + \lambda\boldsymbol{\theta} &= 0 \\ X^\top X\boldsymbol{\theta} + \lambda\boldsymbol{\theta} &= X^\top\mathbf{y} \\ \cancel{(X^\top X + \lambda)}\boldsymbol{\theta} &= X^\top\mathbf{y} \\ (X^\top X + \lambda I)\boldsymbol{\theta} &= X^\top\mathbf{y} \\ \boldsymbol{\theta} &= (X^\top X + \lambda I)^{-1}X^\top\mathbf{y} \end{aligned}$$

Hence the closed form solution of our objective function with L2 regularization included is $\boldsymbol{\theta} = (X^\top X + \lambda I)^{-1}X^\top\mathbf{y}$

3 Multivariate Gaussian Distribution

3.1 Covariance Matrix

You may be familiar with Gaussian distributions for scalar random variables, i.e., one-dimensional, but we can also have N-dimensional random vector, $X = [X_1, X_2, \ldots, X_N]^T$. For this, we use multivariate Gaussian distribution. Notice we cannot use the normal variance, σ^2, anymore, so we use a covariance matrix. The covariance matrix, Σ, has the dimensions $N \times N$ and $\Sigma_{i,j} = cov[X_i, X_j] = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$

What do the diagonal elements represent?

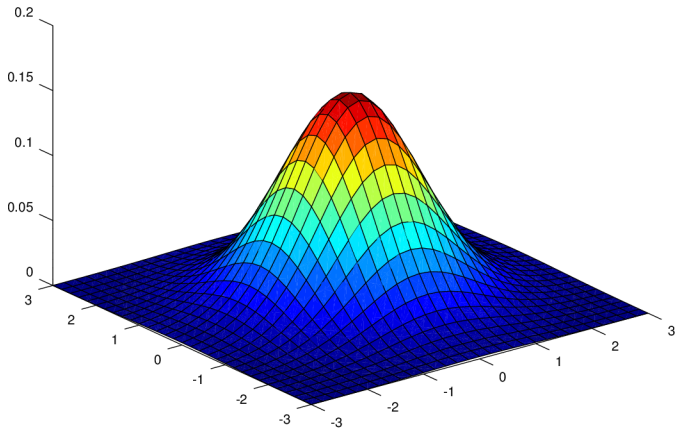
$\Sigma_{i,i} = Cov(X_i, X_i) = \mathbb{E}[X_i X_i] - \mathbb{E}[X_i]\mathbb{E}[X_i] = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 = Var(X_i)$
So, they represent the variance of the respective element.

What are some special properties of the covariance matrix?

It's always both symmetric (i.e., $\Sigma_{i,j} = \Sigma_{j,i}\ \forall i, j$) and positive semi-definite (i.e., $\mathbf{z}^\top \Sigma \mathbf{z} \geq 0 \forall \mathbf{z} \in \mathbb{R}^N$).

3.2 pdf

A multivariate gaussian distribution would look something like this:



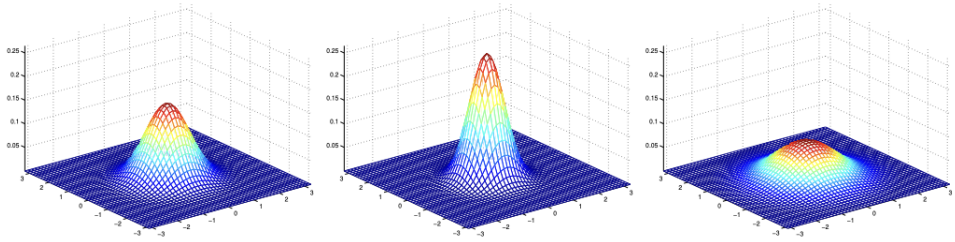
where the probability density function (pdf) is given by:

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{N/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}))$$

$\boldsymbol{\mu} \in \mathbb{R}^N$ is the mean and $\Sigma \in \mathbb{R}^{N \times N}$ is the covariance matrix. $|\Sigma|$ refers to the determinant of Σ.

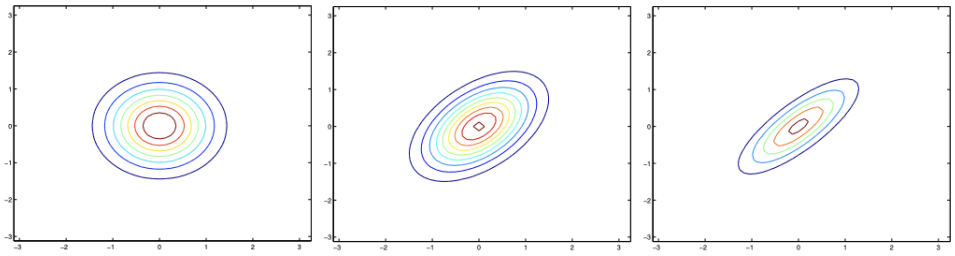
The following $\boldsymbol{\mu}$ and Σ values correspond to these multivariate Gaussian examples:

$$\boldsymbol{\mu} = \begin{bmatrix}0\\0\end{bmatrix} \Sigma = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix}0\\0\end{bmatrix} \Sigma = \begin{bmatrix}0.6 & 0\\0 & 0.6\end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix}0\\0\end{bmatrix} \Sigma = \begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$$



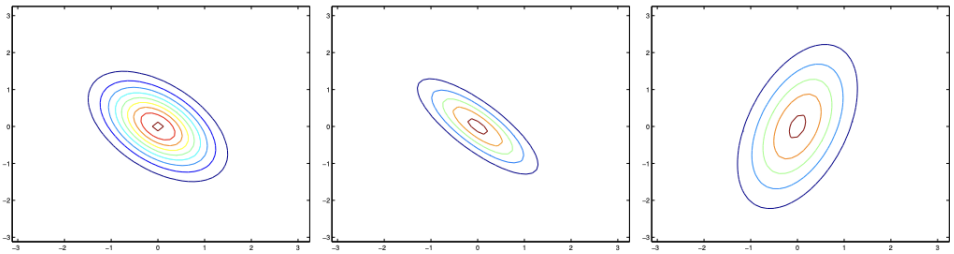
We can look at the contour lines to analyze Σ.

$$\Sigma = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}, \qquad \Sigma = \begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}, \qquad \Sigma = \begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}:$$



3.2.1 Covariance Matching

Given the gaussian plots below, use this colab notebook to determine what covariance matrix results in each plot. (Hint: Try starting off with 1's on the diagonal and try varying the other sliders.)



$$\Sigma = \begin{bmatrix}1 & -0.5\\-0.5 & 1\end{bmatrix}, \qquad \Sigma = \begin{bmatrix}1 & -0.8\\-0.8 & 1\end{bmatrix}, \qquad \Sigma = \begin{bmatrix}3 & 0.8\\0.8 & 1\end{bmatrix}$$

3.3 Multivariate Gaussian: What affects the "spread" of the distribution across the diagonal?

The covariance between variables X_1 and X_2 affects the spread of the distribution across the diagonal. A distribution that is more concentrated along the diagonal would have larger covariance between the two variables A negative covariance between two variables would reverse the direction of the diagonal.

4 Maximum Likelihood Estimation (MLE)

Maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model by maximizing the likelihood function.

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\theta; \mathcal{D})$$

$\mathcal{D}$ is our data sample that we observe, $\mathcal{D} = \{y^{(i)}\}_{i=1}^{N}$.

$\mathcal{L}(\theta)$ is the likelihood of observing our sampled data points for a specific parameter θ, i.e., it is the joint density of the observed data sample given the parameters, $p(y^{(1)}, y^{(2)}, \ldots, y^{(N)} \mid \theta)$. $\mathcal{L}$ is a real-valued function of θ so we want to pick the values for our parameters θ that maximizes the likelihood function. With more data points (larger N), we can arrive at parameters that are closer to their true values.

Oftentimes, we take the log of the $\mathcal{L}$ because maximizing the log function is easier. And, maximizing the log function would also maximize the likelihood function.

Very Important!!!: We assume that the data is i.i.d. here

4.1 Multivariate Gaussian

Assume you observe $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$, where each $\mathbf{x}^{(i)} \in \mathbb{R}^M$ is drawn from $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$

Derive the log-likelihood function first.

$\mathbf{x}^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and $\boldsymbol{\mu}, \Sigma$ are unknown. We want to estimate them by maximizing $\mathcal{L}$. p is the pdf function. Note that $\mathcal{L}$ is the likelihood function while ℓ is the log-likelihood function.

$$\mathcal{L}(\boldsymbol{\mu}, \Sigma; \mathcal{D}) = \prod_{i=1}^{N} p(\mathbf{x}^{(i)} | \boldsymbol{\mu}, \Sigma)$$

$$\ell(\boldsymbol{\mu}, \Sigma; \mathbf{x}) = \log \prod_{i=1}^{N} \frac{1}{(2\pi)^{M/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu})\right)$$

$$= \sum_{i=1}^{N} \left(-\frac{M}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu})\right)$$

$$\ell(\boldsymbol{\mu}, \Sigma; \mathcal{D}) = -\frac{NM}{2}\log(2\pi) - \frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}^{(i)} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}^{(i)} - \boldsymbol{\mu})$$

Now, using this result, derive $\hat{\boldsymbol{\mu}}$. Note that $\frac{\partial}{\partial \mathbf{v}} \mathbf{v}^T A \mathbf{v} = 2A\mathbf{v}$ if A is symmetric

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{N} \Sigma^{-1}(\mathbf{x}^{(i)} - \boldsymbol{\mu}) = 0$$

$$\text{Since } \Sigma \text{ is positive definite, } 0 = N\boldsymbol{\mu} - \sum_{i=1}^{N} \mathbf{x}^{(i)}$$

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)} = \bar{\mathbf{x}} \text{ as expected}$$

5 Colab Demo Link