Math Refresher

Uri Shaham

Outline

- Sample and population statistics

Outline

- Sample and population statistics
- Multivariate Gaussians

Outline

- Sample and population statistics
- Multivariate Gaussians
- Linear transformation of random variables

Outline

- Sample and population statistics
- Multivariate Gaussians
- Linear transformation of random variables
- Bayes rule, entropy

Outline

- Sample and population statistics
- Multivariate Gaussians
- Linear transformation of random variables
- Bayes rule, entropy
- Multivariate calculus

Outline

- Sample and population statistics
- Multivariate Gaussians
- Linear transformation of random variables
- Bayes rule, entropy
- Multivariate calculus
- Eigendecomposition

Outline

- Sample and population statistics
- Multivariate Gaussians
- Linear transformation of random variables
- Bayes rule, entropy
- Multivariate calculus
- Eigendecomposition
- KKT conditions and Lagrangians

Sample and population statistics

	Population	Sample
	X	x
	$\mathbb{E}[X]$	$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$
Univariate	$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$	$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$
Multivariate	$\mathrm{Cov}(X) = \mathbb{E}[(X - \mathbb{E}[X])^T(X - \mathbb{E}[X])]$	$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^T(x_i - \bar{x}_n)$

Multivariate Gaussians

- Univariate normal density: $f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.

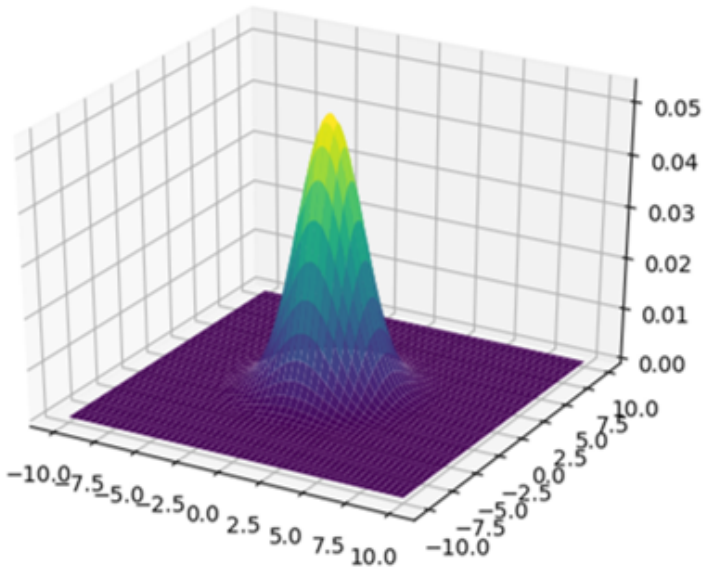

Figure: Multivariate normal density, credit:https://web.stanford.edu/class/archive/cs/cs109/cs109.1218/files/student_drive/5.9.pdf

Multivariate Gaussians

- Univariate normal density: $f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.
- Multivariate normal density:

$$f(x;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp\left((x-\mu)^T \Sigma^{-1} (x-\mu)\right).$$

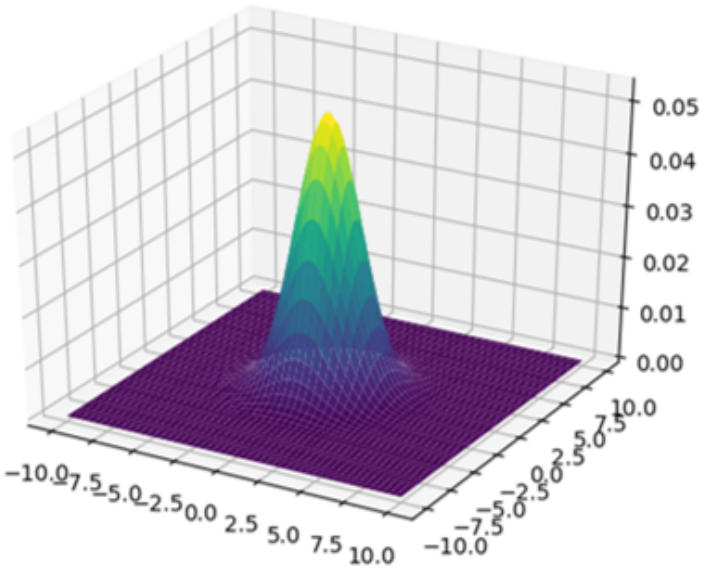

Figure: Multivariate normal density, credit:https://web.stanford.edu/class/archive/cs/cs109/cs109.1218/files/student_drive/5.9.pdf

Multivariate Gaussians

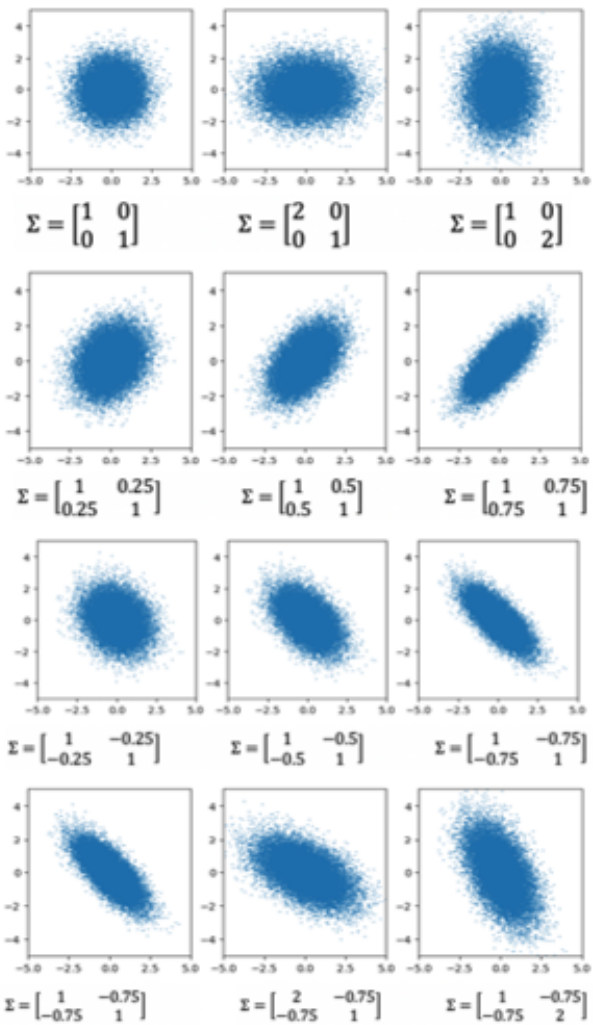

Figure: Effect of covariance, credit:https://web.stanford.edu/class/archive/cs/cs109/cs109.1218/files/student_drive/5.9.pdf

Linear transformations of random vectors

- Let $X \in \mathbb{R}^d$ be a random vector, $a \in \mathbb{R}^d$, $b \in \mathbb{R}$ be constants.

Linear transformations of random vectors

- Let $X \in \mathbb{R}^d$ be a random vector, $a \in \mathbb{R}^d$, $b \in \mathbb{R}$ be constants.
- $\mathbb{E}[a^T X + b] = a^T \mathbb{E}[X] + b$.

Linear transformations of random vectors

- Let $X \in \mathbb{R}^d$ be a random vector, $a \in \mathbb{R}^d$, $b \in \mathbb{R}$ be constants.
- $\mathbb{E}[a^T X + b] = a^T \mathbb{E}[X] + b$.
- $\text{Var}[a^T X + b] = a^T \text{Cov}(X) a$.

Linear transformations of random vectors

- Let $X \in \mathbb{R}^d$ be a random vector, $a \in \mathbb{R}^d$, $b \in \mathbb{R}$ be constants.
- $\mathbb{E}[a^T X + b] = a^T \mathbb{E}[X] + b$.
- $\text{Var}[a^T X + b] = a^T \text{Cov}(X) a$.
- If $X \in \mathbb{R}^d$ is a multivariate normal vector $X \sim \mathcal{N}(\mu, \Sigma)$, then for any constant $a \in \mathbb{R}^d$, $a^T x \in \mathbb{R}$ is a normal random variable $a^T X \sim \mathcal{N}(a^T \mu, a^T \Sigma a)$.

Probability basics: Bayes rule, and MLE

- Bayes rule: Let A, B be events such that $\Pr(B) > 0$. Then

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}.$$

Probability basics: Bayes rule, and MLE

- Bayes rule: Let A, B be events such that $\Pr(B) > 0$. Then

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}.$$

- Let p_θ be a probabilistic model, assigning probability $p_\theta(x)$ to data point (x). The maximum likelihood estimate $\hat{\theta}_{\text{MLE}}$ is

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta p_\theta(x).$$

Probability basics: Bayes rule, and MLE

- Bayes rule: Let A, B be events such that $\Pr(B) > 0$. Then

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}.$$

- Let p_θ be a probabilistic model, assigning probability $p_\theta(x)$ to data point (x). The maximum likelihood estimate $\hat{\theta}_{\text{MLE}}$ is

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta p_\theta(x).$$

- If $x_1, \ldots x_n$ are independent, then the MLE can be found by

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta \sum_{i=1}^{n} \log p_\theta(x).$$

Probability basics: Entropy

- Let X be a multinomial random variable with parameters $p_1, \ldots p_k$ such that $\forall_i p_i \geq 0$ and $\sum_{i=1}^{k} p_i = 1$. The entropy of X is

$$H(X) = -\sum_{i=1}^{k} p_i \log(p_i).$$

Probability basics: Entropy

- Let X be a multinomial random variable with parameters $p_1, \ldots p_k$ such that $\forall_i p_i \geq 0$ and $\sum_{i=1}^{k} p_i = 1$. The entropy of X is

$$H(X) = -\sum_{i=1}^{k} p_i \log(p_i).$$

- $H(X)$ quantifies the amount of uncertainty about the value of X, in a similar fashion to variance with numeric random variables.

A Bit of Multivariate calculus

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function.

A Bit of Multivariate calculus

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function.
- The gradient of f is the (d-dimensional) vector of partial derivatives $\nabla f = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d})^T$.

A Bit of Multivariate calculus

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function.
- The gradient of f is the (d-dimensional) vector of partial derivatives $\nabla f = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d})^T$.
- Dot product generalize standard products for scalars: for $x, y \in \mathbb{R}^d$ $\langle x, y \rangle = x^T y = \sum_{i=1}^{d} x_i y_i$.

A Bit of Multivariate calculus

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function.
- The gradient of f is the (d-dimensional) vector of partial derivatives $\nabla f = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d})^T$.
- Dot product generalize standard products for scalars: for $x, y \in \mathbb{R}^d$ $\langle x, y \rangle = x^T y = \sum_{i=1}^{d} x_i y_i$.
- 1st order Taylor series for multivariate functions:

$$f(x + \epsilon) = f(x) + \langle \nabla f(x), \epsilon \rangle + O(\epsilon^2).$$

A Bit of Multivariate calculus

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function.
- The gradient of f is the (d-dimensional) vector of partial derivatives $\nabla f = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d})^T$.
- Dot product generalize standard products for scalars: for $x, y \in \mathbb{R}^d$ $\langle x, y \rangle = x^T y = \sum_{i=1}^{d} x_i y_i$.
- 1st order Taylor series for multivariate functions:

$$f(x + \epsilon) = f(x) + \langle \nabla f(x), \epsilon \rangle + O(\epsilon^2).$$

- This means that the gradient is the direction of the steepest ascent of f.

A Bit of Multivariate calculus

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function.
- The gradient of f is the (d-dimensional) vector of partial derivatives $\nabla f = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d})^T$.
- Dot product generalize standard products for scalars: for $x, y \in \mathbb{R}^d$ $\langle x, y \rangle = x^T y = \sum_{i=1}^{d} x_i y_i$.
- 1st order Taylor series for multivariate functions:

$$f(x + \epsilon) = f(x) + \langle \nabla f(x), \epsilon \rangle + O(\epsilon^2).$$

- This means that the gradient is the direction of the steepest ascent of f.
- The notion of norm generalizes the absolute value of scalars, and expresses the length of a vector: for $x \in \mathbb{R}^d$, $\|x\|^2 = \langle x, x \rangle = \sum_{i=1}^{d} x_i^2$.

A Bit of Multivariate calculus

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function.
- The gradient of f is the (d-dimensional) vector of partial derivatives $\nabla f = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d})^T$.
- Dot product generalize standard products for scalars: for $x, y \in \mathbb{R}^d$ $\langle x, y \rangle = x^T y = \sum_{i=1}^{d} x_i y_i$.
- 1st order Taylor series for multivariate functions:

$$f(x + \epsilon) = f(x) + \langle \nabla f(x), \epsilon \rangle + O(\epsilon^2).$$

- This means that the gradient is the direction of the steepest ascent of f.
- The notion of norm generalizes the absolute value of scalars, and expresses the length of a vector: for $x \in \mathbb{R}^d$, $\|x\|^2 = \langle x, x \rangle = \sum_{i=1}^{d} x_i^2$.
- $f(x) = a^T x \Rightarrow \nabla f = a$.

A Bit of Multivariate calculus

- Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function.
- The gradient of f is the (d-dimensional) vector of partial derivatives $\nabla f = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d})^T$.
- Dot product generalize standard products for scalars: for $x, y \in \mathbb{R}^d$ $\langle x, y \rangle = x^T y = \sum_{i=1}^{d} x_i y_i$.
- 1st order Taylor series for multivariate functions:

$$f(x + \epsilon) = f(x) + \langle \nabla f(x), \epsilon \rangle + O(\epsilon^2).$$

- This means that the gradient is the direction of the steepest ascent of f.
- The notion of norm generalizes the absolute value of scalars, and expresses the length of a vector: for $x \in \mathbb{R}^d$, $\|x\|^2 = \langle x, x \rangle = \sum_{i=1}^{d} x_i^2$.
- $f(x) = a^T x \Rightarrow \nabla f = a$.
- $f(x) = x^t A x$, where A is symmetric $\Rightarrow \nabla f = 2Ax$.

Eigendecomposition

- Let A be a $d \times d$ matrix.

Eigendecomposition

- Let A be a $d \times d$ matrix.
- A vector v is an eigenvector of A with eigenvalue λ if $Av = \lambda v$.

Eigendecomposition

- Let A be a $d \times d$ matrix.
- A vector v is an eigenvector of A with eigenvalue λ if $Av = \lambda v$.
- This means that when viewing A as an operator on $\mathbb{R}^d$, A only streches (or shrinks) v, but does not rotate it.

Eigendecomposition

- Let A be a $d \times d$ matrix.
- A vector v is an eigenvector of A with eigenvalue λ if $Av = \lambda v$.
- This means that when viewing A as an operator on $\mathbb{R}^d$, A only streches (or shrinks) v, but does not rotate it.
- Note: if v is an eigenvector of A, then so is av for any constant a. Hence we usually take v to be a unit vector.

Eigendecomposition

- Let A be a $d \times d$ matrix.
- A vector v is an eigenvector of A with eigenvalue λ if $Av = \lambda v$.
- This means that when viewing A as an operator on $\mathbb{R}^d$, A only streches (or shrinks) v, but does not rotate it.
- Note: if v is an eigenvector of A, then so is av for any constant a. Hence we usually take v to be a unit vector.
- The spectral theorem: if A is symmetric, there exists an orthonornal basis of $\mathbb{R}^d$, consisting of eigenvectors of A. In addition, all eigenvalues are real.

Eigendecomposition

- Let A be a $d \times d$ matrix.
- A vector v is an eigenvector of A with eigenvalue λ if $Av = \lambda v$.
- This means that when viewing A as an operator on $\mathbb{R}^d$, A only streches (or shrinks) v, but does not rotate it.
- Note: if v is an eigenvector of A, then so is av for any constant a. Hence we usually take v to be a unit vector.
- The spectral theorem: if A is symmetric, there exists an orthonornal basis of $\mathbb{R}^d$, consisting of eigenvectors of A. In addition, all eigenvalues are real.
- In such cases we can write $A = U\Lambda U^T$, where U is an orthonormal matrix whose columns are the eigenvectors of A, and Λ is a diagonal matrix consisting of the corresponding eigenvalues of A.

Constrained Optimization in a Nutshell

- Unconstrained optimization: $\min_x f(x)$.

Constrained Optimization in a Nutshell

- Unconstrained optimization: $\min_x f(x)$.
- Constrained optimization: $\min_x f(x)$, subject to $g_i(x) \leq 0$, $i = 1, \ldots, k$.

Constrained Optimization in a Nutshell

- Unconstrained optimization: $\min_x f(x)$.
- Constrained optimization: $\min_x f(x)$, subject to $g_i(x) \leq 0$, $i = 1, \ldots, k$.
- Lagrangian: $\mathcal{L}(x, \lambda) = f(x) + \sum_{i=1}^{k} \lambda_i g_i(x)$.

Constrained Optimization in a Nutshell

- Unconstrained optimization: $\min_x f(x)$.
- Constrained optimization: $\min_x f(x)$, subject to $g_i(x) \leq 0$, $i = 1, \ldots, k$.
- Lagrangian: $\mathcal{L}(x, \lambda) = f(x) + \sum_{i=1}^{k} \lambda_i g_i(x)$.
- The Lagrangian bounds the optimal constrained solution from below:

$$\min_x \max_{\forall i \lambda_i \geq 0} \mathcal{L}(x, \lambda) \leq \min_{x \text{ is feasible}} f(x).$$

Constrained Optimization in a Nutshell- cont'

- Strong duality: if the problem f and the constraints $g_i, i = 1, \ldots, k$ are convex, and there is at least one x for which all constraints are fully satisifed (i.e., $g_i(x) < 0$), then the duality gap is zero.

Constrained Optimization in a Nutshell- cont'

- Strong duality: if the problem f and the constraints $g_i, i = 1, \ldots, k$ are convex, and there is at least one x for which all constraints are fully satisifed (i.e., $g_i(x) < 0$), then the duality gap is zero.
- In this case, it is sufficient to solve for the Lagrangian.

Constrained Optimization in a Nutshell- cont'

- Strong duality: if the problem f and the constraints $g_i, i = 1, \ldots, k$ are convex, and there is at least one x for which all constraints are fully satisifed (i.e., $g_i(x) < 0$), then the duality gap is zero.
- In this case, it is sufficient to solve for the Lagrangian.
- The solution of which is called "the dual problem", and is sometimes easier to solve.

Constrained Optimization in a Nutshell- cont'

- Strong duality: if the problem f and the constraints $g_i, i = 1, \ldots, k$ are convex, and there is at least one x for which all constraints are fully satisifed (i.e., $g_i(x) < 0$), then the duality gap is zero.
- In this case, it is sufficient to solve for the Lagrangian.
- The solution of which is called "the dual problem", and is sometimes easier to solve.
- This is done by fixing λ, and finding $x = \arg\min \mathcal{L}(x, \lambda)$, and then solving for λ.

Constrained Optimization in a Nutshell- cont'

- Strong duality: if the problem f and the constraints $g_i, i = 1, \ldots, k$ are convex, and there is at least one x for which all constraints are fully satisifed (i.e., $g_i(x) < 0$), then the duality gap is zero.
- In this case, it is sufficient to solve for the Lagrangian.
- The solution of which is called "the dual problem", and is sometimes easier to solve.
- This is done by fixing λ, and finding $x = \arg\min \mathcal{L}(x, \lambda)$, and then solving for λ.
- If strong duality holds, KKT conditions imply that at the optimum x^*, for each i either $\lambda_i = 0$ or $g_i(x^*) < 0$.

Further Reading

- Multivariate Gaussians

Thank You!