Random Projections and Ensemble Kalman Filter

BTP Stage-II Full Report

by

Raaz Dwivedi
Roll Number: 100070016
Department of Electrical Engineering
raaz10@iitb.ac.in

under the guidance of

Prof. Vivek Borkar
Department of Electrical Engineering
borkar@ee.iitb.ac.in

Indian Institute of Technology, Bombay
April 21, 2014

Department of Electrical Engineering

IIT Bombay

Acceptance Certificate

The BTP Stage-II Report titled Random Projections and Ensemble Kalman Filter submitted by Raaz Dwivedi [Roll No: 100070016] may be accepted for evaluation.

Signature of Guide: _____

Prof. Vivek S. Borkar

Department of Electrical Engineering

IIT Bombay

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Signature of Student: _____

Raaz Dwivedi
[Roll No: 100070016]

Acknowledgements

I am deeply thankful to my guide, Prof. Vivek S. Borkar for his untiring support and his expert guidance in this project. His role in this project has been very active. His motivation and patience helped me a lot. He took a personal interest in the problem and has been a wonderful guide. For that I am deeply indebted to him.
Finally, I wish to thank my parents and sister for their support and encouragement throughout my study.

Abstract

We review the celebrated Johnson Lindenstrauss Lemma and some recent advances in the understanding of probability measures with geometric characteristics on $\mathbb{R}^d$, for large d. These advances include the central limit theorem for convex sets, according to which the uniform measure on a high dimensional convex body[1] has marginals that are approximately Gaussian. We try to combine these two results to provide a theoretical justification to the successful heuristic methods implemented in Ensemble Kalman Filters for high dimensional data by oceanographars, meteorologists, etc.

[1] A convex body in $\mathbb{R}^d$ is a compact, convex set with a non-empty interior.

Contents

1. Introduction

1.1 Motivation

The Ensemble Kalman Filter is a Monte-Carlo implementation of the Bayesian Update. It uses ensemble mean and ensemble covariance for representing the distribution of the system state, but it assumes that all the involved probability distribution functions are Gaussian. Nevertheless, such ensemble methods have proven to work very well for high-dimensional data, even when the pdfs involved are not necessarily Gaussian. We use the recent advances in probability measures which imply that random projections on low dimensional subspaces are approximately gaussian, and a variant of the Johnson-Lindenstrauss Lemma (JL Lemma) which shows that for a set there exist norm preserving low-dimensional projections (with ϵ-distortion). We combine these two results to provide a theoretical support behind the success of the Ensemble Kalman Filter in the absence of Gaussianity.

1.2 Outline

The report has been organized into Six Sections. In next section, we look at the Kalman Filter and the Ensemble Kalman Filter (EnKF). We describe the mathematical model of these filters. And then we give details of a particular example from oceanography where Ensemble Kalman Filter is used. In Section 3, we discuss the Johnson-Lindenstrauss Lemma and mention several variants that appeal to the wide-applicability of the lemma. In the next section, results on Gaussian projections of high dimensional distributions are made precise followed by discussion of the strong variant of the JL Lemma. We combine these results in Section 5 to provide a mathematical justification behind the heuristics associated with the EnKF. Finally we conclude in Section 6 and outline a possible future line of work.

2. Ensemble Kalman Filter

In this chapter, we study in detail the working of a Ensemble Kalman Filter (EnKF). First we describe a Kalman Filter, followed by an overview of EnKF. After that we present the mathematical model of Kalman Filter (KF), Extended Kalman Filter (XKF) and the EnKF. After that we make a few remarks about the EnKF. Most of the content in the next two sections is a summary of [8] and [9].

2.1 Introduction

In the Ensemble Kalman Filter, given a probability density function (pdf) of the state of the modeled system (the prior, called often the forecast in geosciences) and the data likelihood, the Bayes theorem is used to to obtain pdf after the data likelihood has beed taken into account (the posterior, often called the analysis). This is called a Bayesian update. The Bayesian update is combined with advancing the model in time, incorporating new data from time to time. The original Kalman Filter [11] assumes that all pdfs are Gaussian (the Gaussian assumption) and provides algebraic formulas for the change of the mean and covariance by the Bayesian update, as well as a formula for advancing the covariance matrix in time provided the system is linear. However, maintaining the covariance matrix is not feasible computationally for high-dimensional systems. For this reason, EnKFs were developed [12, 13]. EnKFs represent the distribution of the system state using a random sample, called an ensemble, and replace the covariance matrix by the sample covariance computed from the ensemble. One advantage of EnKFs is that advancing the pdf in time is achieved by simply advancing each member of the ensemble. For a survey of EnKF and related data assimilation techniques, see [10].

2.1.1 Kalman Filter

Let us review first the Kalman Filter. Let x denote the n-dimensional state vector of a model, and assume that it has Gaussian probability distribution with mean μ and covariance Q, i.e., its pdf is ($\propto$ denotes proportionality)

$$p(x) \propto \exp\left(-\frac{1}{2}(x-\mu)^t Q^{-1}(x-\mu)\right). \tag{2.1}$$

This probability distribution, called the *prior*, was evolved in time by running the model and is to be updated to account for the new data. We assume that data has some associated error with it and the distribution of error is known. Here, the data d is assumed to be Gaussian with covariance R and mean Hx, where H is the so-called observation matrix. Naturally, Hx denotes the value of the data *if* the data was "error-free". Thus, the conditional density of d given x (known as data-likelihood) is given as

$$p(d|x) \propto \exp\left(-\frac{1}{2}(d-Hx)^t R^{-1}(d-Hx)\right). \tag{2.2}$$

The pdf of the state and the data-likelihood are combined to give the new probability density of the system state x conditional on the value of the data d (*the posterior*) by the Bayes Theorem,

$$p(x|d) \propto p(d|x)p(x). \tag{2.3}$$

The data d is fixed once it is received, so denote the posterior state by $\widehat{x}$ instead of $x|d$ and the posterior pdf by $p(\widehat{x})$. It can be shown by algebraic manipulations [14] that the posterior pdf is also Gaussian,

$$p(\widehat{x}) \propto \exp\left(-\frac{1}{2}(\widehat{x}-\widehat{\mu})^t \widehat{Q}^{-1}(\widehat{x}-\widehat{\mu})\right), \tag{2.4}$$

with the posterior mean $\widehat{\mu}$ and covariance $\widehat{Q}$ given by the KALMAN UPDATE FORMULAS

$$\widehat{\mu} = \mu + K(d-H\mu), \quad \widehat{Q} = (I-KH)Q, \tag{2.5}$$

where

$$K = QH^t(HQH^t+R)^{-1} \tag{2.6}$$

is the so-called KALMAN GAIN MATRIX.

2.1.2 Ensemble Kalman Filter (EnKF)

The EnKF is a Monte Carlo approximation of the Kalman Filter, which avoids evolving the covariance matrix of the pdf of the state vector x. Instead, the distribution is represented by a collection of realizations, called an ensemble. So, let

$$X = [x_1, \cdots, x_N] = [x_i] \tag{2.7}$$

be an $n \times N$ matrix whose columns are a sample from the prior distribution. The matrix X is called the *prior ensemble*. Replicate the data into an $m \times N$ matrix

$$D = [d_1, \cdots, d_N] = [d_i] \tag{2.8}$$

so that each column d_i consists of the data vector d plus a random vector from the n-dimensional normal distribution $\mathcal{N}(0, R)$. Then the columns of

$$\widehat{X} = X + K(D - HX) \tag{2.9}$$

form a random sample from the posterior distribution. The EnKF is now obtained simply by replacing the state covariance Q in Kalman Gain Matrix (2.5) by the sample covariance C computed from the ensemble members (called the *ensemble covariance*).

2.2 Mathematical Models

Consider a discrete-time nonlinear system with dynamics

$$x_{k+1} = f(x_k, u_k) + w_k \tag{2.10}$$

and measurements

$$y_k = h(x_k) + v_k, \tag{2.11}$$

where $x_k, w_k \in \mathbb{R}^n$; $u_k \in \mathbb{R}^m$; $y_k, v_k \in \mathbb{R}^p$. We assume that w_k and v_k are stationary zero-mean white noise processes with covariance matrices Q_k and R_k, respectively. Furthermore we assume that x_0, w_k and v_k are uncorrelated.The objective is to obtain estimates x_k^a of the state x_k using measurements y_k so that $tr\left(\mathbb{E}\left[e_k^a (e_k^a)^t\right]\right)$ is minimized, where $e_k^a \in \mathbb{R}^n$ denotes the error, and is defined by

$$e_k^a := x_k - x_k^a. \tag{2.12}$$

2.2.1 Kalman Filter

When the dynamics and measurement in (2.10) and (2.11) are linear, that is,

$$f(x_k, u_k) = A_k x_k + B_k u_k, \tag{2.13}$$

$$h(x_k) = C_k x_k \tag{2.14}$$

the Kalman Filter provides optimal estimates x_k^a of the state x_k. Define the analysis state error covariance $P_k^a \in \mathbb{R}^{n \times n}$ by $P_k^a := \mathbb{E}\left[e_k^a (e_k^a)^t\right]$. The Kalman Filter equations [11] are expressed in two steps, the Analysis Step, where information from measurements is used, and the Forecast Step, where information about the plant is used. These steps are

expressed as the analysis step:

$$K_k = P^f_{xy_k}\left(P^f_{yy_k}\right)^{-1}, \tag{2.15}$$
$$P^a_k = (I - K_k C_k)P^f_k, \tag{2.16}$$
$$x^a_k = x^f_k + K_k(y_k - C_k x^f_k), \quad \text{[Data Update]} \tag{2.17}$$

and the forecast step:

$$x^f_{k+1} = A_k x^a_k + B_k u_k, \quad \text{[Physics Update]} \tag{2.18}$$
$$P^f_{k+1} = A_k P^a_k A^t_k + Q_k, \tag{2.19}$$

where the forecast state error covariance $P^f_k \in \mathbb{R}^{n\times n}$ is defined by $P^f_k := \mathbb{E}\left[e^f_k\left(e^f_k\right)^t\right]$, and

$$P^f_{xy_k} := \mathbb{E}\left[e^f_k\left(y_k - y^f_k\right)^t\right] = P^f_k C^t_k, \tag{2.20}$$
$$P^f_{yy_k} := \mathbb{E}\left[(y_k - y^f_k)\left(y_k - y^f_k\right)^t\right] = C_k P^f_k C^t_k + R_k, \tag{2.21}$$

where $y^f_k := C_k x^f_k$ and $e^f_k := x_k - x^f_k$.
However when the dyanmics in (2.10) and (2.11) are non-linear, the discrete-time Riccati update equation (2.19) cannot be used to propogate the state error covariance matrix. Hence, some approximate techniques like Extended Kalman Filter (XKF) are used in which Physics update is modified as $x^f_{k+1} = f(x^a_k, u_k)$ and for other equations, A_k and C_k are approximated as Jacobians of $f(x,u)$ and $h(x)$ at current state estimate, respectively. That is

$$A_k := \left.\frac{\partial f(x,u)}{\partial x}\right|_{x=x^a_k} \tag{2.22}$$
$$C_k := \left.\frac{dh}{dx}\right|_{x=x^a_k} \tag{2.23}$$

2.2.2 Ensemble Kalman Filter

The EnKF is a suboptimal estimator, where the error statistics are predicted by using a Monte Carlo or ensemble integration to solve the Fokker-Planck equation. The Ensemble Kalman Filtering method is presented in three stages.
First, to represent the error statistics in the forecast step, we assume that at time k, we have an ensemble of q forecasted state estimates with random sample errors. We denote this ensemble as $X^f_k \in \mathbb{R}^{n\times q}$, where

$$X^f_k := (x^{f_1}_k, \cdots, x^{f_q}_k), \tag{2.24}$$

and the superscript f_i refers to the i-th forecast ensemble member. Then, the ensemble mean $\bar{x}_k^f \in \mathbb{R}^n$ is defined by

$$\bar{x}_k^f := \frac{1}{q}\sum_{i=1}^{q} x_k^{f_i}. \tag{2.25}$$

Similarly, we have $y_k^{f_i}$'s and we define their ensemble mean as $\bar{y}_k^f$. Since the true state x_k is not known, we approximate $e_k^f = x_k - x_k^f$ by using the ensemble members. The ensemble error matrix $E_k^f \in \mathbb{R}^{n\times q}$ around the ensemble mean is defined as

$$E_k^f := \left[x_k^{f_1} - \bar{x}_k^f, \cdots, x_k^{f_q} - \bar{x}_k^f\right] \tag{2.26}$$

and the ensemble of output errors $E_{y_k}^f \in \mathbb{R}^{p\times q}$ is defined as

$$E_{y_k}^a := \left[y_k^{f_1} - \bar{y}_k^f, \cdots, y_k^{f_q} - \bar{y}_k^f\right]. \tag{2.27}$$

And P_k^f, $P_{xy_k}^f$ and $P_{yy_k}^f$ are approximated by $\widehat{P}_k^f$, $\widehat{P}_{xy_k}^f$ and $\widehat{P}_{yy_k}^f$ respectively where,

$$\widehat{P}_k^f := \frac{1}{q-1}E_k^f(E_k^f)^t, \widehat{P}_{xy_k}^f := \frac{1}{q-1}E_k^f(E_{y_k}^f)^t, \text{ and } \widehat{P}_{yy_k}^f := \frac{1}{q-1}E_{y_k}^f(E_{y_k}^f)^t. \tag{2.28}$$

Thus, we interpret the forecast ensemble mean as the best forecast estimate of the state, and the spread of the ensemble members around the mean as the error between the best estimate and the actual state.
The second step is the analysis step: To obtain the analysis estimates of the state, the EnKF performs an ensemble of parallel data assimilation cycles, where for $i = 1, \cdots, q$

$$x_k^{a_i} = x_k^{f_i} + \widehat{K}_k\left(y_k^i - h(x_k^{f_i})\right). \tag{2.29}$$

The *perturbed observations* y_k^i are given by

$$y_k^i = y_k + v_k^i, \tag{2.30}$$

where v_k^i is distributed as $\mathcal{N}(0, R_k)$. The sample error covariance computed from v_k^i converges to R_k as $q \to \infty$. We approximate the analysis error covariance matrix P_k^a by $\widehat{P}_k^a$, where

$$\widehat{P}_k^a := \frac{1}{q-1}E_k^a(E_k^a)^t, \tag{2.31}$$

and E_k^a is defined by (2.26) with $x_k^{f_i}$ replaced by $x_k^{a_i}$ and $\bar{x}_k^f$ replaced by the ensemble mean of $x_k^{a_i}$. We use the classical Kalman filter gain expression and the approximations of the error covariances to determine the filter gain $\widehat{K}$ by

$$\widehat{K} = \widehat{P}_{xy_k}^f(\widehat{P}_{yy_k}^f)^{-1}. \tag{2.32}$$

The last step is the prediction of error statistics of error statistics in the forecast step:

$$x_{k+1}^{f_i} = f(x_k^{a_i}, u_k) + w_k^i, \tag{2.33}$$

where the values w_k^i are sampled from $\mathcal{N}(0, Q_k)$. The sample error covariance matrix computed from the w_k^i converges to Q_k as $q \to \infty$. Finally, we summarize the analysis and forecast steps.
Analysis Step:

$$\begin{aligned}
\widehat{K}_k &= \widehat{P}_{xy_k}^f (\widehat{P}_{yy_k}^f)^{-1}, && (2.34)\\
x_k^{a_i} &= x_k^{f_i} + \widehat{K}_k \left(y_k + v_k^i - h(x_k^{f_i}) \right) && (2.35)\\
\bar{x}_k^a &= \sum_{i=1}^{q} x_k^{a_i}/q. && (2.36)
\end{aligned}$$

Forecast Step:

$$\begin{aligned}
x_{k+1}^{f_i} &= f(x_k^{a_i}, w_k) + w_k^i, && (2.37)\\
\bar{x}_k^f &= \sum_{i=1}^{q} x_k^{f_i}/q, \quad \bar{y}_k^f = \sum_{i=1}^{q} y_k^{f_i}/q, && (2.38)\\
E_k^f &= \left[x_k^{f_1} - \bar{x}_k^f, \cdots, x_k^{f_q} - \bar{x}_k^f \right] && (2.39)\\
E_{y_k}^f &= \left[y_k^{f_1} - \bar{y}_k^f, \cdots, y_k^{f_q} - \bar{y}_k^f \right] && (2.40)\\
\widehat{P}_{xy_k}^f &= \frac{1}{q-1} E_k^f (E_{y_k}^f)^t && (2.41)\\
\widehat{P}_{yy_k}^f &= \frac{1}{q-1} E_{y_k}^f (E_{y_k}^f)^t. && (2.42)
\end{aligned}$$

[26] gives a beautiful pictorial view of all the steps. Unlike the XKF, the evaluation of the filter gain $\widehat{K}_k$ in the EnKF does not involve an approximation of the nonlinearity $f(x, u)$ and $h(x)$. Hence, the computational burden of evaluating the Jacobians is absent in the EnKF. Furthermore, evaluation of $\widehat{P}_{xy_k}^f \in \mathbb{R}^{n \times p}$ and $\widehat{P}_{yy_k}^f \in \mathbb{R}^{p \times p}$ is a $O(pqn)$ operation. While evaluation of P_k^f is an $O(n^3)$ operation in XKF. Hence, if $q \ll n$ then the computational burden of evaluating the approximate covariances in the EnKF is less when compared to XKF. However, in EnKF q parallel copies of the model have to be simulated, and, when q is large, the computational burden of the forecast step in the EnKF is large. Alternatively, in the XKF, only one copy of the model is simulated to obtain the state estimates. Hence, if n is very large and $q \ll n$, then the EnKF is computationally less intensive than the XKF. Usually, q is of the order of a few hundreds and n is of the order of a few millions and so we have $q \ll n$.

2.3 A Closer Look at EnKF

To summarize, the ensemble Kalman filter is a recursive filter suitable for problems with a large number of variables, such as discretizations of partial differential equations in geophysical models. The EnKF originated as a version of the Kalman filter for large problems (essentially, the covariance matrix is replaced by the sample covariance), and it is now an important data assimilation component of ensemble forecasting. EnKF is related to the particle filter (in this context, a particle is the same thing as an ensemble member) but the EnKF makes the assumption that *all probability distributions involved are "Gaussian"*.
In the limit of ensemble size becoming infinite, the KF and the EnKF are equivalent. For nonlinear dynamics the EnKF includes the full effect of the non-linear terms and there are no linearizations or closure assumptions used. In addition, there is no need for a tangent linear operator or its adjoint, and this makes the method very easy to implement for practical applications. This leads to an interpretation of the EnKF as a purely statistical Monte Carlo method where the ensemble of model states evolves in state space with the mean as the best estimate and the spreading of the ensemble as the error variance. At measurement times each observation is represented by another ensemble, where the mean is the actual measurement and the variance of the ensemble represents the measurement errors. Thus, we combine a stochastic prediction step with a stochastic analysis step. A few important observations are listed below:

1. The ensemble methods introduce an approximation by using only the mean and covariance of the prior joint pdf when computing the posterior ensemble update equation. Thus, it is effectively assumed that the prior joint pdf is Gaussian when computing the updates. This means that the *EnKF will not give the correct answer if the prior joint pdf has non-Gaussian contributions. However the ensemble methods have proven to work well with a large number of different nonlinear dynamical models.*

   Thus the EnKF analysis scheme is approximate in the sense that it does not properly take into account non-Gaussian contributions in the prior for x, the state of the system. In other words, it does not solve the Bayesian update equation for non-Gaussian pdfs. On the other hand, it is not a pure resampling of a Gaussian posterior distribution. Only the updates are linear and these are added to the prior non-Gaussian ensemble. Thus, the updated ensemble will inherit many of the non-Gaussian properties from the forecast ensemble. In summary, we have a very computational efficient analysis scheme where we avoid traditional resampling of the posterior, and the solution becomes something between a linear Gaussian update and a full Bayesian computation. It was also suggested that the sequential introduction of measurements, with Gaussian distributed errors, actually introduced "Gaussianity" to the ensemble representing the conditional joint density.

2. It is now known that EnKF can handle certain levels of nonlinearity in both the model

prediction and measurement functional. Even if the prior ensemble is non-Gaussian the ensemble methods will in many cases provide an updated ensemble having a realistic pdf. When the prior ensemble is non-Gaussian, the analyzed ensemble will inherit some of the non-Gaussian structures. On the other hand, it is also possible to make the EnKF fail completely; e.g., if the weight on the prior is low and a multimodal pdf develops, this may result in non-physical solutions.

3. It is seen that the residuals, as expected, are decreasing when the ensemble size is increased. In practical applications we are naturally limited by the number of ensemble members we can afford to run. However, from the central limit theorem, the accuracy in the EnKF estimate will improve proportionally to the square root of the ensemble size. In most published applications of the EnKF a typical ensemble size is around 100 members. This ensemble size is clearly much less than effective dimension of the solution space of many dynamical models, but in many cases a so-called localization or local analysis computation is often used to effectively increase the dimension of the space where the solution is searched for. We provide an overview of a practical example from [10].

    - The TOPAZ system consists of the HYCOM ocean model which has been coupled to two different sea-ice models, one is a simple model for ice-thickness and ice-concentration while the other is multi-category sea-ice model which represents ice-thickness distributions. Further, four ecosystem models of increasing complexity have been integrated in the system. The TOPAZ system has a huge state vector consisting of 79.6 million variables just for the physical ocean parameters. The inclusion of the marine ecosystem multiplies the number of unknowns by a factor 2 to 3, depending on the ecosystem model formulation used.
    - The system uses 100 members in the ensemble, thus the computational cost of running the system is 100 times the cost of running a single model. Clearly, it is a challenge to represent the solution search space for such a large state vector while assimilating huge number of measurements using only a limited ensemble size. To avoid the problems associated with a large number of measurements, many operational assimilation schemes have made the assumption that only measurements located within a certain distance from a grid point will impact the analysis in this grid point. This allows for an algorithm where the analysis is computed grid point by grid point and only a subset of observations, located near the current grid point, is then used in the analysis.
    - The analysis in the EnKF is computed in a space spanned by the ensemble members. This is a subspace which, in many cases, can be rather small compared to the total dimension of the model state. Computing the analysis grid point by

grid point implies that, for each grid point, a small model state is solved for in a relatively large ensemble space. The analysis will then result from a different combination of ensemble members for each grid point, and this allows the analysis scheme to reach solutions not originally represented by the ensemble. This algorithm is approximate and it does not solve the original problem posed. The local analysis is spatially discontinuous and the updated ensemble members may not represent solutions of the original model equations, but the deviation should not be too large as long as the range of influence is large enough. In addition the updated ensemble members are not represented in the space spanned by the predicted ensemble. In fact, the use of an update matrix which varies smoothly throughout the grid effectively reduces the dimension of the problem. That is, in an ocean model where we update the solution grid column by grid column, we are solving many small problems instead of one large. On the other hand, in the standard EnKF analysis we also introduce an approximation by using a limited ensemble size. With an infinite ensemble size there would be no need to use a local analysis scheme, since the whole solution space would be represented by the ensemble. The local analysis scheme will in many applications significantly reduce the impact of a limited ensemble size and allow for the use of the EnKF with high dimensional model systems

- The quality of the EnKF analysis is clearly connected to the ensemble size used. We expect that a larger ensemble is needed for the global analysis than the local analysis to achieve the same quality of the result. That is, in the global analysis a large ensemble is needed to properly explore the state space and to provide a consistent result for the global analysis. We expect this to be application dependent. In dynamical models with large state spaces, the local analysis allows for the computation of a realistic analysis result while still using a relatively small ensemble of model states.

3. Johnson-Lindenstrauss Lemma

3.1 Introduction

3.1.1 Dimensionality Reduction

Advancement in data collection and storage capabilities have enabled researchers in diverse domain to observe and collect huge amounts of data. These large data sets, however, present substantial challenges to existing data analysis tools. One major bottleneck in this regard is the large number of features or dimensions associated with some measured quantity, a problem frequently termed as the "curse of dimensionality". Existing algorithms usually scale very poorly with increase in number of dimensions of the data. This motivates mapping the data from the high-dimensional space to a lower dimensional space in a manner such that the mapping preserves (or almost preserves) the structure of the data. Substantial research efforts have been made (and are still being made) to overcome the aforementioned curse and tame high-dimensional data. *Dimensionality reduction* encompasses all such techniques which aim to reduce the number of random variables (dimensions or features) associated with some observable or measurable quantity with the hope that the data in lower dimensions would be much more amenable to efficient exploration and analysis.

3.1.2 The Johnson-Lindenstrauss Lemma

In the process of extending Lipschitz mappings to Hilbert spaces, Johnson and Lindenstrauss [15] formulated a key geometric lemma. This lemma (Lemma 1 of [15]) was thereafter referred to as Johnson-Lindenstrauss Lemma. The Johnson-Lindenstrauss Lemma states that a set of points in high-dimensional space can be mapped to a much lower dimension such that the pairwise distances of the points in the higher dimensional space are almost preserved. The cardinality of the lower dimension space depends on the number of input points and degree (approximation factor) to which the pairwise distances need to be preserved.

3.1.3 Applications

The lemma has uses in compressed sensing, manifold learning, dimensionality reduction, and graph embedding. Much of the data stored and manipulated on computers, including text and images, can be represented as points in a high-dimensional space. However, the essential algorithms for working with such data tend to get bogged down very quickly as dimension increases. It is therefore desirable to reduce the dimensionality of the data in a way that preserves its relevant structure. The Johnson-Lindenstrauss lemma is a classic result in this vein.

3.2 Random Projections

We first state JL Lemma formally and then discuss briefly about bounds of the dimension of subspace containing the ϵ-distortion embedding. We outline the key proof technique that is used in proving most of the results in various variants of JL Lemma. And then we state some variants and relaxation of the lemma.

3.2.1 Elementary Versions

The lemma as stated in [15] (throughout the paper, unless otherwise mentioned, $||.||$ denotes the ℓ_2 norm):

Theorem 3.2.1 (JL Lemma). *For any $0 < \epsilon < 1$ and any integer n, let k be a positive integer such that*

$$k \geqslant 8\frac{\ln n}{\epsilon^2}. \tag{3.1}$$

Then for any set V of n points in $\mathbb{R}^d$, there is a map $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $u, v \in V$,

$$(1-\epsilon)||u-v||^2 \leqslant ||f(u)-f(v)||^2 \leqslant (1+\epsilon)||u+v)||^2. \tag{3.2}$$

The bounds on k were improved later on, for instance, the statement of the theorem as given in [17]:

Theorem 3.2.2. *For any $0 < \epsilon < 1$ and any integer n, let k be a positive integer such that*

$$k \geqslant \frac{4\ln n}{(\epsilon^2/2 - \epsilon^3/3)}. \tag{3.3}$$

Then for any set V of n points in $\mathbb{R}^d$, there is a map $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $u, v \in V$,

$$(1-\epsilon)||u-v||^2 \leqslant ||f(u)-f(v)||^2 \leqslant (1+\epsilon)||u-v||^2. \tag{3.4}$$

Moreover, this map can be found in randomized polynomial time.

Frankl and Maehara in [18] had shown that dimension $k = \lceil 9 \ln n/(\epsilon^2 - 2\epsilon^3/3) \rceil + 1$ is sufficient; the papers of Indyk and Motwani [21] and Achlioptas [19] give essentially the same bounds for k as in (3.3). Before proceeding further, we state the following result Alon in by [24]:

Theorem 3.2.3. *For any $\epsilon > 0$, any set of A of n points in an Euclidean Space can be embedded in an Euclidean Space of dimension $k = c(\epsilon) \log n$ with distortion ϵ and $c(\epsilon) \leqslant O(1/\epsilon^2)$. Then*

$$c(\epsilon) \geqslant \Omega \left(\frac{1}{\epsilon^2 \log \frac{1}{\epsilon}} \right). \tag{3.5}$$

Thus one sees that the $k = O\left(\frac{\log n}{\epsilon^2}\right)$ is *nearly a tight bound* on k. The proof involves fairly simple concepts of linear algebra.

3.2.2 Key Proof Technique

All known proofs of the lemma proceed according to the following scheme: for given d and an appropriate k, one defines a suitable probability distribution $\mathcal{F}$ on the set of all linear maps $\mathbb{R}^d \to \mathbb{R}^k$. Then one proves the following lemma:

Lemma 3.2.4. *If $T : \mathbb{R}^d \to \mathbb{R}^k$ is a random linear mapping drawn from the distribution $\mathcal{F}$, then for every vector $x \in \mathbb{R}^d$ we have*

$$\mathbb{P}\left\{ (1-\epsilon)||x|| \leqslant ||T(x)|| \leqslant (1+\epsilon)||x|| \right\} \geqslant 1 - \frac{1}{n^2}. \tag{3.6}$$

Having established this statement for $\mathcal{F}$, the lemma follows easily:
Let $V = \{v_1, ..., v_n\} \subset \mathbb{R}^d$. We choose T at random according to $\mathcal{F}$. Then for every $i < j$, using linearity of T and Theorem 3.1 with $x = v_i - v_j$, we get that T fails to satisfy $(1-\epsilon)||v_i - v_j|| \leqslant ||T(v_i) - T(v_j)|| \leqslant (1+\epsilon)||v_i - v_j||$ with probability atmost $1/n^2$. Consequently the probability that any of the $\binom{n}{2}$ pairwise distances is distorted by more than $1 \pm \epsilon$ is at most $\binom{n}{2}/n^2 < 1/2$. Therefore a random T works with probability at least $1/2$. Repeating this projection $O(n)$ times can boost the success probability to the desired constant, giving us the claimed randomized polynomial time algorithm.

3.2.3 Variants of JL Lemma

From application point of view, it is important to be able to generate and evaluate the random linear map T fast. Now we state some different ways in which T can be generated.

- In the original paper by Johnson and Lindenstrauss [15], T is chosen as the orthogonal projection on a random k-dimensional subspace of $\mathbb{R}^d$ with a scaling factor of $\sqrt{d/k}$. Using the idea that the projection on random k-dimensional subspace is same as projection of a random vector on first k-coordinates, the proof in this case boils down to showing that if x is a random point on the unit sphere S^{d-1} in $\mathbb{R}^d$, then the length of its orthogonal projection on the first k co-ordinates (or in other words the quantity $\sqrt{x_1^2 + x_2^2 + ... + x_k^2}$) is sharply concentrated around $\sqrt{k/d}$. This is a simple consequence of measure concentration on S^{d-1}, see, e.g., [23] for a detailed presentation of such a proof.

- In [21], Indyk and Motwani use a matrix T where the entries of T are i.i.d. random variables distributed as $\mathcal{N}(0,1)$. Such an T is easier to generate. By simple properties of the normal distribution it follows that in this case, for every fixed unit vector $x \in \mathbb{R}^d$, the quantity $||T(x)||^2$ has the chi-square distribution with k degrees of freedom, and one can use known tail estimates for this distribution to prove Theorem 3.1.

- Achlioptas in [19], uses a database friendly projection T, by giving an even more tractable way of generating T_{ij}. He proved that T_{ij} can be chosen as independent ± 1 random variables taking either value with same probability. In another variant, he proved the lemma for T_{ij} taking values $\pm\sqrt{3}$ with probability $1/6$ each and 0 with probability $2/3$. This setting allows for computing $T(x)$ about 3 times faster than the former, since T is sparse - only about one third entries are non-zero.

- In [16], Matousek generalizes the result to random variables with subgaussian tails[1]. He proved that for an integer n, $\epsilon \in (0, 1/2]$, $\delta \in (0, 1]$, $k = C\epsilon^{-2} \log \frac{2}{\delta}$ and

$$T(x)_i = \frac{1}{\sqrt{k}} \sum_{j=1}^{n} R_{ij} x_j, i = 1, 2, ..., k \tag{3.7}$$

  where R_{ij} are independent random variables with zero mean and unit variance and a uniform subgaussian tail[2] (and C depends on the constant in the subgaussian tail), we have for every $x \in \mathbb{R}^n$

$$\mathbb{P}\left\{(1-\epsilon)||x|| \leqslant ||T(x)|| \leqslant (1+\epsilon)||x||\right\} \geqslant 1 - \delta. \tag{3.8}$$

3.2.4 Variants involving Sparse Projections

To fasten the evaluation of the map Tx, we need the associated matrix to be sparse. Now we take a look at the results which involve sparse T. A significant obstacle for a sparse

[1] A random variable X is said to have a subgaussian tail if there exists a constant $a > 0$ such that for all $c > 0$, $\mathbb{P}[X > c] \leqslant e^{-ac^2}$ and $\mathbb{P}[X < -c] \leqslant e^{-ac^2}$.

[2] A sequence of random variables $X_1, X_2, .., X_n$ are said to have a uniform subgaussian tail if all of them have subgaussian tails with the same constant a.

matrix T is that once T becomes significantly sparse with the fraction of nonzero entries tending to 0, the length of the image $||T(x)||$ is not sufficiently concentrated for some vectors, for example, for $x = [1, 0, 0, ..., 0]^T$. In [20], it was proved that the concentration is sufficient even for sparse T provided that the vector x is "well-spread", which can be quantified as follows: Assuming $||x|| = 1$, we require that $||x||_\infty = \max_j |x_j|$ be close to $1/\sqrt{d}$. This means that the mass of x has to be distributed over many components. These authors were, however, successful in dealing with vectors x that are not well-spread by introducing the Fourier transform. Another variant with the constraint in place was proved by Matousek in [16].

- Ailon and Chazelle in [20] use a central concept from harmonic analysis known as the Heisenberg principle: A signal and its spectrum cannot be both concentrated. With this in mind, they precondition the random projection with a Fourier transform (via an FFT) in order to isometrically enlarge the support of any sparse vector. To prevent the inverse effect, i.e., the sparsification of dense vectors, they randomize the Fourier transform. The result is the Fast-Johnson-Lindenstrauss-Transform: a randomized FFT followed by a sparse projection. The FJLT shares the low-distortion characteristics of a random projection but with a lower complexity. The overall linear map is a combination of three matrices given by $\Phi = PHD$ where P is $k \times d$ matrix with a large number of entries as 0 and the rest of them being distributed as normal (more precisely, $P_{ij} \sim \mathcal{N}(0, q^{-1})$ with probability q, and $P_{ij} = 0$ with probability $1 - q$, where q can be taken as $\log^2 n/d$). H is $d \times d$ normalized Hadamard matrix[3](deterministic) which calculates the *Walsh Fourier Transform* (WFT), and D is $d \times d$ diagonal matrix with diagonal entries drawn randomly from $\{1, -1\}$ with probability $1/2$, and this D does the job of randomizing the WFT.

- In [16], Matousek proved (3.8) for a linear map given by equation (3.7) with the constraint that x must be "well-spread", that is the result holds for $||x||_\infty \leqslant \alpha$ where $\alpha \geqslant 1/\sqrt{d}$. R_{ij} are independent random variables taking values 0 with probability $1 - q$ and $\pm 1/\sqrt{q}$ with probability $q/2$ each where $q = C_0 \alpha^2 \log(d/\epsilon\delta)$. Thus $q = O(\log d/d)$ is small for $\alpha = 1/\sqrt{d}$. The proof relied on showing that the $T(x)_i$ had subgaussian tails for the "well-spread" vectors.

[3]Hadamard Matrix, denoted by $H_d \in \mathbb{R}^{d \times d}$ where d is a power of 2, is a matrix with entries as ± 1 and it is usually expressed in a recursive way with $H_0 = 1$ and $H_{2^k} = \begin{pmatrix} H_{2^{k-1}} & H_{2^{k-1}} \\ H_{2^{k-1}} & -H_{2^{k-1}} \end{pmatrix}$.

4. Random Low Dimensional Projections

A general study of probability distributions in high dimension is likely to be hopeless, as such distributions may exhibit a wide range of entirely unrelated phenomena. There seem to exist, nevertheless, some large classes of distributions which obey some interesting, non-trivial principles. One of the earliest such examples is provided by the classical Central Limit Theorem. Suppose we are given a probability density $f : \mathbb{R}^n \to [0, 1)$ which is a product density, i.e.,

$$f(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f_i(x_i) \tag{4.1}$$

for some density functions $f_1, ..., f_n$. Then f is the joint density of n independent random variables $X_1, ..., X_n$. Assume that the dimension n is large. Then under mild integrability assumptions on the $f_i's$, it is guaranteed that for appropriate coefficients $b, \theta_1^n, \theta_2^n, ..., \theta_n^n$ we have

$$\mathbb{P}\left(\sum_{i=1}^{n} \theta_i X_i \leqslant t\right) \approx \int_{-\infty}^{t} \exp\left(-(s-b)^2/2\right) ds, \ \forall t \in \mathbb{R} \tag{4.2}$$

When the density f is properly normalized (such that $X_1, ..., X_n$ have mean zero and variance one), the gaussian approximation (4.2) actually holds for "most" choices of $\theta_1^n, \theta_2^n, ..., \theta_n^n \in \mathbb{R}$ with $\sum_{i=1}^{n} (\theta_i^n)^2 = 1$. By "most" we mean that the coefficients $\theta_1^n, \theta_2^n, ..., \theta_n^n$ may be chosen randomly, uniformly on the unit sphere S^{n-1} in $\mathbb{R}^n$. As we shall later see, this has been generalized to a great extent.

In the next two sections, we discuss two key aspects of low dimensional projections of random distributions in high dimensions. First one is on random distributions with convexity assumption. It is observed that random distributions in high dimension have Gaussian marginals when projected to low dimensional subspaces, a generalisation of (4.2). Another result is a strong variant of the JL Lemma (Theorem 3.2.1) which says that for a given metric space, suitable random projections "almost" preserve the norm of the points. All the results are easy to understand but the proofs require heavy machinery and as a consequence have been skipped.

4.1 Low Dimensional Projections with Gaussian Densities

In this section, we mention several results by Klartag ([1], [2], [3], [4]; overview in [5]) which state that uniform measure on high dimensional convex body has Gaussian marginals. We also state some intermediate results to show that these results are indeed strong.

A function $f : \mathbb{R}^n \to [0, \infty)$ is log-concave if $\forall x, y \in \mathbb{R}^n$ and $0 < \lambda < 1$,

$$f(\lambda x + (1-\lambda)y) \geqslant f(x)^{\lambda} f(y)^{1-\lambda} \tag{4.3}$$

That is, f is log-concave when $\log f$ is concave on the support of f. Examples of interest for log-concave functions include characteristic functions of convex sets, the Gaussian density, and several densities from statistical mechanics. In this section, random vectors in $\mathbb{R}^n$ that are distributed according to a log-concave density, are considered. And so, it also includes as a special case the uniform distribution on an arbitrary compact, convex set with a non-empty interior.
We say that $f : \mathbb{R}^n \to [0, \infty)$ is *isotropic* if it is the density function of some random variable with zero mean and identity covariance matrix. That is, f is isotropic when

$$\int_{\mathbb{R}^n} f(x)dx = 1, \int_{\mathbb{R}^n} xf(x)dx = 0 \text{ and } \int_{\mathbb{R}^n} \langle x, \theta \rangle^2 f(x)dx = |\theta|^2, \ \forall \theta \in \mathbb{R}^n. \tag{4.4}$$

Any log concave function with $0 < \int f < \infty$ can be brought to an isotropic position via an affine map, that is, $f \circ T$ is isotropic for some affine map $T : \mathbb{R}^n \to \mathbb{R}^n$ (see, e.g., [6]).
We denote the standard Euclidean norm on $\mathbb{R}^n$ by $||.||$, and for $x \in \mathbb{R}$, $|x|$ denotes its magnitude.
We begin with one of the main results as given in [2]:

Result 4.1.1. *There exists a sequence $\epsilon_n \searrow 0$ for which the following holds. Let $K \subset \mathbb{R}^n$ be a compact, convex set with a non-empty interior. Let X be a random vector that is distributed uniformly in K. Then there exist a unit vector θ in $\mathbb{R}^n, t_0 \in \mathbb{R}$ and $\sigma > 0$ such that*

$$\sup_{A \subset \mathbb{R}} \left| \mathbb{P}\left\{ \langle X, \theta \rangle \in A \right\} - \int_A \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t-t_0)^2}{2\sigma^2} \right) dt \right| \leqslant \epsilon_n \tag{4.5}$$

where the supremum runs over all measurable sets $A \subset \mathbb{R}$, and where $<,>$ denotes the usual scalar product in $\mathbb{R}^n$.

Furthermore, under the additional assumptions that the expectation of X is zero and that the covariance matrix of X is the identity matrix, we may assert that most (in a sense to be made precise later) unit vectors θ satisfy (4.5), with $t_0 = 0$ and $\sigma = 1$. Corresponding principles also hold for multidimensional marginal distributions of convex sets. The following lemma played an important role in proving Result 4.1.1:

Lemma 4.1.2. *Let $n \geqslant 1$ be an integer and let X be a random vector with an isotropic, log-concave density in $\mathbb{R}^n$. Then for all $0 \leqslant \epsilon \leqslant 1$,*

$$\mathbb{P}\left\{\left|\frac{||X||}{\sqrt{n}} - 1\right| \geqslant \epsilon\right\} \leqslant Cn^{-c\epsilon^2}, \tag{4.6}$$

where $c, C > 0$ are universal constants.

It was later improved in [3] by replacing RHS of (4.6) by $C\,\exp\left(-c\epsilon^{3.33}n^{0.33}\right)$. Consequently we have a tightened bound as :

$$\mathbb{P}\left\{\left|\frac{||X||}{\sqrt{n}} - 1\right| \geqslant \frac{1}{n^{1/15}}\right\} \leqslant C\,\exp(-n^{1/15}) \tag{4.7}$$

For the results that we mention next, we introduce a few more notations: Suppose that X and Y are two random variables attaining values in some measure space (here Ω will always be $\mathbb{R}$ or $\mathbb{R}^n$ or a subspace $E \subset \mathbb{R}^n$). We define their total variation distance as

$$d_{TV}(X, Y) = 2\sup_{A \subset \Omega} |\mathbb{P}\{X \in A\} - \mathbb{P}\{Y \in A\}| \tag{4.8}$$

where the supremum runs over all measurable sets $A \subset \Omega$. Note that $d_{TV}(X, Y)$ equals the L^1-distance between the densities of X and Y, when these densities exist. Let σ_{n-1} stand for the unique rotationally invariant probability measure on S^{n-1}, also referred to as the uniform probability measure on the sphere S^{n-1}. Then we have the following result [2]-

Theorem 4.1.3. *There exist sequences $\epsilon_n \searrow 0, \delta_n \searrow 0$ for which the following holds: Let $n \geqslant 1$, and let X be a random vector in $\mathbb{R}^n$ with an isotropic log-concave density. Then there exists a subset $\Theta \subset S^{n-1}$ with $\sigma_{n-1}(\Theta) \geqslant 1 - \delta_n$, such that for all $\theta \in \Theta$,*

$$d_{TV}\left(\langle X, \theta\rangle, Z\right) \leqslant \epsilon_n \tag{4.9}$$

where $Z \sim N(0, 1)$ is the standard normal random variable.

The bounds given were $\epsilon_n \leqslant C\left(\dfrac{\log\log(n+2)}{\log(n+1)}\right)^{1/2}$ and $\delta_n \leqslant \exp(-cn^{0.99})$ with $c, C > 0$ as universal constants.
The bounds were later improved in [3] by Klartag and the improved result is stated below:

Theorem 4.1.4. *Let $n \geqslant 1$, and let X be a random vector in $\mathbb{R}^n$ with an isotropic log-concave density. Then there exists a subset $\Theta \subset S^{n-1}$ with $\sigma_{n-1}(\Theta) \geqslant 1 - Ce^{-\sqrt{n}}$, such that for all $\theta \in \Theta$, the real-valued random variable $\langle X, \theta\rangle$ has a density $f_\theta : \mathbb{R} \to [0, \infty)$ with the following two properties:*
(i) $\int_{-\infty}^{\infty} |f_\theta(t) - \zeta(t)|\, dt \leqslant \dfrac{1}{n^k}$,
(ii) For all, $|t| \leqslant n^k$ *we have* $\left|\dfrac{f_\theta(t)}{\zeta(t)} - 1\right| \leqslant \dfrac{1}{n^k}$.

where $\zeta(t)$ denotes the denisty of zero mean, unit variance Gaussian random variable, and $C, k > 0$ are universal constants.

For the special case where X is uniformly distributed over a convex body $K \subset \mathbb{R}^n$ we have that any random projection of X over an ℓ-dimensional subspace is close to an ℓ-dimensional gaussian random variable in the total variation sense. To formalise the result in [3], let us denote by $G_{n,\ell}$ the grassmanian[1] of all ℓ-dimensional subspaces of $\mathbb{R}^n$, and let $\sigma_{n,\ell}$ stand for the unique rotationally invariant probability measure (for details, refer to e.g., [22]) on $G_{n,\ell}$. For a subspace $E \subset \mathbb{R}^n$ and a point $x \in \mathbb{R}^n$ we write $Proj_E(x)$ for the orthogonal projection of x onto E. Then we have:

Theorem 4.1.5. *Let $1 \leqslant \ell \leqslant n$ be an integer and let $K \subset \mathbb{R}^n$ be a convex body. Let X be a random vector that is distributed uniformly in K, and suppose that X has zero mean and identity covariance matrix. Assume that $\ell \leqslant cn^k$. Then there exists a subset $\mathcal{E} \subset G_{n,\ell}$ with $\sigma_{n,\ell}(\mathcal{E}) \geqslant 1 - e^{-c\sqrt{n}}$ such that for any $E \in \mathcal{E}$,*

$$\sup_{A \subset E} \left| \mathbb{P}\left\{ Proj_E(X) \in A \right\} - \int_A \zeta^\ell(x) dx \right| \leqslant \frac{1}{n^k}, \tag{4.10}$$

where the supremum runs over all measurable sets $A \subset E$. Here $\zeta^\ell(x)$ denotes the density of standard multivariate normal distribution in $\mathbb{R}^\ell$ and $c, k > 0$ are universal constants.

In [4] Klartag strengthens the previous result for projections of random variables with support on convex bodies. The total variation estimate (4.10) implies that density of $Proj_E(X)$ is close to the density of a certain Gaussian random vector Γ in L^1-norm. Consequenctly, we might deduce that the ratio between the density of $Proj_E(X)$ and the density of Γ deviates from 1 by no more than $1/n^k$, in the significant parts of the subspace E. And this deduction led to a convergence result in pointwise sense [4].

Theorem 4.1.6. *Let X be an isotropic random vector in $\mathbb{R}^n$ with a log-concave density. Let $1 \leqslant \ell \leqslant n^{c_1}$ be an integer. Then there exists a subset $\mathcal{E} \subset G_{n,\ell}$ with $\sigma_{n,\ell}(\mathcal{E}) \geqslant 1 - \exp(-n^{c_2})$ such that for any $E \in \mathcal{E}$, the following holds. Denote by f_E the density of the random vector $Proj_E(X)$. Then,*

$$\left| \frac{f_E(x)}{\zeta_E^\ell(x)} - 1 \right| \leqslant \frac{C}{n^{c_3}} \tag{4.11}$$

for all $x \in E$ with $|x| \leqslant n^{c_4}$. Here $\zeta_E^\ell(x)$ is the standard ℓ-dimensional Gaussian density in E, and $C, c_1, c_2, c_3, c_4 > 0$ are universal constants.

[1] Grassmanian $\mathrm{Gr}(k, V)$ is a space which parametrizes all linear subspaces of a vector space V of given dimension k. For example, the Grassmanian $\mathrm{Gr}(1, \mathbb{R}^n)$ is the space of all lines through the origin in $\mathbb{R}^n$.

4.2 Norm Preserving Random Projections

In [1], Klartag generalizes the notion of JL Lemma (3.2) to a general set and attempts to reduce the dependence of k on n, in (3.3).

Definition 4.2.1. *For a metric space (T, d) define*

$$\gamma_\alpha(T,d) = \underset{\substack{T_s \subset T \\ |T_s| \leqslant 2^{2^s}, |T_0| = 1}}{\inf} \left\{ \sup_{t \in T} \sum_{s=0}^{\infty} 2^{s/\alpha} d(t, T_s) \right\}. \tag{4.12}$$

Note that, by the celebrated "majorizing measures" theorem in [7], if $\{X_t : t \in T\}$ is a gaussian process, and $d_2^2(s,t) = \mathbb{E}|X_s - X_t|^2$ is its covariance structure, then with $c_1, c_2 > 0$ as absolute constants, we have

$$c_1 \gamma_2(T, d_2) \leqslant \mathbb{E}\left[\sup_{t \in T} X_t\right] \leqslant c_2 \gamma_2(T, d_2). \tag{4.13}$$

Definition 4.2.2 (Orlicz Norms). *Let X be a random variable. Define the ψ_p norm of X as*

$$||X||_{\psi_p} = \inf_{C>0} \left\{ \mathbb{E}\left[\exp\left(\frac{|X|^p}{C^p} \right) \right] \leqslant 2 \right\}. \tag{4.14}$$

A standard argument shows that if X has a bounded ψ_p norm then the tail of $|X|$ decays faster than $2\exp(-u^p/||X||_{\psi_p}^p)$. In particular, for $p = 2$, it means that X has a subgaussian tail. Next, we mention few useful results from Klartag and Mendelson [1]. In the paper, using these results, they also establish some bounds on the supremum of certain empirical processes indexed by set of functions with the same L_2 norm. Here, we state only the geometric applications of the result, the most important of which (Theorem 4.2.4) is a sharpening of JL Lemma (3.2). The key feature of the results is the wide-range applicability, as these results only require that the matrix entries have subgaussian tails. Recall, in Section 3.2, we had discussed a few results by Matousek [16] for matrix entries with subgaussian tails.

Theorem 4.2.3. *Let (Ω, μ) be a probability space and let $X, X_1, X_2, ..., X_k$ be independent random variables distriuted according to μ. Set T to be a collection of functions, such that for every $f \in T$, $\mathbb{E}[f^2(X)] = ||f||_{L_2}^2 = 1$ and $f(X)$ be subgaussian with $||f||_{\psi_2} \leqslant \beta$. Define the random variable*

$$Z_f^k = \frac{1}{k} \sum_{i=1}^{k} f^2(X_i) - ||f||_{L_2}^2. \tag{4.15}$$

Then for any $e^{-c'\gamma_2^2(T, ||.||_{\psi_2})} < \delta < 1$, with probability larger than $1 - \delta$,

$$\sup_{f \in T} |Z_f^k| \leqslant \frac{c(\delta, \beta)}{\sqrt{k}} \gamma_2(T, ||.||_{\psi_2}) \tag{4.16}$$

where $c' > 0$ is an absolute constant and $c(\delta, \beta)$ depends on solely on δ, β.

As an application we have the following result (analogous to JL Lemma) -

Theorem 4.2.4. *For every $\beta > 0$ there exists a constant $c(\beta)$ for which the following holds. Let $T \subset S^{n-1}$ be a set and let $\Gamma : \mathbb{R}^n \to \mathbb{R}^k$ be a random matrix whose entries are independent, identically distributed random variable with zero mean, variance 1, and are subgaussian with $||\Gamma_{i,j}||_{\psi_2} \leqslant \beta$. Then, with probability larger than $1/2$, for any $x \in T$ and $\epsilon \geqslant \frac{c(\beta)}{\sqrt{k}}\gamma_2(T, ||.||_2)$,*

$$1 - \epsilon \leqslant \frac{1}{\sqrt{k}}||\Gamma x||_{\ell_2^k} < 1 + \epsilon. \tag{4.17}$$

REMARK: The JL Lemma and its variants mentioned in Chapter 3 were valid for a finite set. In contrast, all the results in this section are applicable to a general set. Also, for a set $T \subset S^{d-1}$ of cardinality n, $\gamma_2(T, ||.||_2) \leqslant c\sqrt{\log n}$, so another novelty in this theorem compared to the JL Lemma is that $\log n$ can be slightly improved to $\gamma_2^2(T, ||.||_2)$, however the order dependency on n remains the same. And this observation is consistent with Theorem 3.2.3 by Alon which shows that $k = O(\epsilon^{-2} \log n)$ is nearly a tight bound.
Using Theorem 4.2.4, Klartag extended the result in [1] as follows:

Theorem 4.2.5. *For every $\beta > 0$ there exists a constant $c(\beta)$ for which of the following holds. Let $T \subset S^{n-1}$ be a set, and let $\Gamma : \mathbb{R}^n \to \mathbb{R}^k$ be a random operator whose rows are independent vectors $\Gamma_1, \Gamma_2, ..., \Gamma_k \in \mathbb{R}^n$. Assume that for any $x \in \mathbb{R}^n$ and $1 \leqslant i \leqslant k$, $\mathbb{E}\langle \Gamma_i, x\rangle^2 = \frac{1}{k}||x||^2$ and $||\langle \Gamma_i, x\rangle||_{\psi_2} \leqslant \beta||\langle \Gamma_i, x\rangle||_{L_2}$. Then, with probability larger than $1/2$, for any $x \in T$ and any $\epsilon \leqslant \frac{c(\beta)}{\sqrt{k}}\gamma_2(T, ||.||_2)$,*

$$(1 - \epsilon) \leqslant ||\Gamma x||_{\ell_2^k} < 1 + \epsilon. \tag{4.18}$$

Result 4.2.6. *For the random projection Γ, where the elements of the matrix Γ are independent random variables with zero mean, variance $1/k$ and $||\Gamma_{i,j}||_{\psi_2} \leqslant \beta||\Gamma_{i,j}||_{L_2}$, the conditions required by Theorem 4.2.5 are met, and hence such a choice of random projection also works fine.*

Result 4.2.7. *Another example from [1] is when Γ is an orthogonal projection on random k-dimensional subspaces of $\mathbb{R}^n$ (note that this choice of Γ is exactly same as the choice of random projection in [15] for proving JL Lemma, for details refer to Section 3.2). Recall $G_{n,k}$ is the grassmanian of k-dimesional subspaces of $\mathbb{R}^n$ and that there exists a unique rotation invariant probability measure $\sigma_{n,k}$ on $G_{n,k}$. Assume that P is an orthogonal projection on a random k-dimensional subspace in $\mathbb{R}^n$. Then for $k \geqslant C\gamma_2^2(T, ||.||_2)/\epsilon^2$, Theorem 4.2.5 also holds for $\Gamma = \sqrt{n}P$ where the random k-dimensional subspace is drawn from $G_{n,k}$ as per $\sigma_{n,k}$.*

5. Our Attempts

5.1 Some Results for Gaussian Random Variables and Vectors

In this section, we state some results from [29] and [30] for the conditional density of jointly gaussian random variables and vectors.

Result 5.1.1. *If $X \sim \mathcal{N}_n(\mu, C)$ and Q is $n \times n$ non-singular matrix then for any $\eta \in \mathbb{R}^n$,*

$$QX + \eta \sim \mathcal{N}_n(Q\mu + \eta, QCQ^t). \tag{5.1}$$

For details, refer to Lemma 7, Chapter 1 of [30].

Jointly Gaussian Random Variables

It is well known that for $X_1, X_2, ..., X_n$ jointly Gaussian, the conditional mean of X_n given $X_1, ..., X_{n-1}$ is an affine map in $X_1, ..., X_{n-1}$. Let $\mathbf{X} = (X_1, X_2, ..., X_n) \sim \mathcal{N}(\mu, \mathbf{K})$ where $\mu = \mathbb{E}(\mathbf{X}) = [\mathbb{E}(X_1), ..., \mathbb{E}(X_n)]^t$ denotes the mean vector and $\mathbf{K} = \mathbb{E}\left[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^t\right]$ is the covariance matrix with $\mathbf{K^{-1}} = [q_{ij}]$. Then X_n is normal given $X_1, X_2, ..., X_{n-1}$ with

$$\mathbb{E}(X_n | X_1, ..., X_{n-1}) = \mu_n - \frac{1}{q_{nn}} \sum_{j=1}^{n-1} q_{nj}(X_j - \mu_j) \tag{5.2}$$

and

$$\mathrm{Var}(X_n | X_1, ..., X_{n-1}) = \frac{1}{q_{nn}}. \tag{5.3}$$

Jointly Gaussian Random Vectors

Let $\mathbf{X}$ be $p + q$ dimensional gaussian random vector whose density is given by $\mathcal{N}_{p+q}(\mu, \Sigma)$ where $\mathbf{X} = \begin{pmatrix} \mathbf{X}_{(1)} \\ \mathbf{X}_{(2)} \end{pmatrix}, \mu = \begin{pmatrix} \mu_{(1)} \\ \mu_{(2)} \end{pmatrix}$ and

$$\Sigma = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array} \right)$$

where $\mathbf{X}_{(1)}$ and $\mu_{(1)}$ are p-dimensional and Σ_{11} is $p \times p$. Define $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. Then the conditional density of $\mathbf{X}_{(2)}$, given $\mathbf{X}_{(1)} = \mathbf{x}_{(1)}$, given by

$$\left(\mathbf{X}_{(2)} | \mathbf{X}_{(1)} = \mathbf{x}_{(1)}\right) \sim \mathcal{N}_q\left(\mu_{(2)} + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_{(1)} - \mu_{(1)}),\ \Sigma_{22.1}\right). \tag{5.4}$$

5.2 Projections in EnKF: Why do they work?

Let (X, Y) be randomly distributed in $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ with a log concave density f, where n_1, n_2 are large. Then there exists an affine map $A : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ such that $f \circ A$ is isotropic log-concave density (look at discussion in Section 4.1). Let

$$(\tilde{X}, \tilde{Y}) := A\left((X, Y)\right), \tag{5.5}$$

then clearly $(\tilde{X}, \tilde{Y}) \sim f \circ A$ which is isotropic and log-concave.
Suppose we are given a random sample of N points from the joint distribution of $(\tilde{X}, \tilde{Y})$ (indexed by first N natural numbers). For a given $\epsilon > 0$, let k_1, k_2 be integers such that

$$O(\log N/\epsilon^2) \leqslant k_i \leqslant (2n_i)^{c_1}, i = 1, 2. \tag{5.6}$$

Consider an orthogonal projection of $(\tilde{X}, \tilde{Y})$ onto a random $k_1 \times k_2$-dimensional subspace E of $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, that is to say, $\tilde{X}$ gets projected from $\mathbb{R}^{n_1}$ to a k_1-dimensional subspace and $\tilde{Y}$ gets projected from $\mathbb{R}^{n_2}$ to a k_2-dimensional subspace. Then from Result 4.2.7, it follows that with high probability norms are preserved for $(\tilde{X}_i, \tilde{Y}_i)$ where $i = 1, 2, ..., N$, with distortion atmost ϵ. Let us denote $Proj_E(\tilde{X}, \tilde{Y}) = (\widehat{X}, \widehat{Y})$. Then with high probability (obtained after $O(N)$ projections)

$$(1 - \epsilon)||(\widehat{x}_i, \widehat{y}_i)||_2 \leqslant ||(\tilde{x}_i, \tilde{y}_i)||_2 \leqslant (1 + \epsilon)||(\widehat{x}_i, \widehat{y}_i)||_2, \quad i \in \{1, ..., N\} \tag{5.7}$$

where $||.||_2$ denotes 2-norm on $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$.
Let G denote the grassmanian of all $S_1 \times S_2$ subspaces of $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ where for $i = 1, 2$, $S_i \subset \mathbb{R}^{n_i}$ and $\dim(S_i) = k_i$ and let σ denote the unique rotationally invariant probability measure on G. Then from Result 4.1.6 we have that, there exists $\mathcal{E} \subset G$ with $\sigma(\mathcal{E}) \geqslant 1 - e^{-(n_1+n_2)c_2}$ such that for all $(\widehat{x}, \widehat{y}) \in E$ with $||(\widehat{x}, \widehat{y})||_2 \leqslant (n_1 + n_2)^{c_4}$

$$\left| \frac{f_E(\widehat{x}, \widehat{y})}{\zeta_E(\widehat{x}, \widehat{y})} - 1 \right| \leqslant \frac{C}{(n_1 + n_2)^{c_3}} \tag{5.8}$$

where f_E denotes the density of $Proj_E(X, Y)$ and $\zeta_E(.)$ is the standard Gaussian density in E. As a consequence, we have that with high probability ($\geqslant 1 - e^{-(n_1+n_2)c_2}$), $(\widehat{X}, \widehat{Y})$ are approximately gaussian, where approximation is in the sense of (5.8). That is we have

$$(\widehat{X}, \widehat{Y}) \approx \mathcal{N}_{k_1+k_2}(\mathbf{0}, I_{k_1+k_2}). \tag{5.9}$$

Let us *assume* that A is an invertible map.

$$(\tilde{X}, \tilde{Y}) = A(X, Y) \Rightarrow (X, Y) = A^{-1}(\tilde{X}, \tilde{Y}). \tag{5.10}$$

Let's denote $Proj_E$ by Γ. Then we have,

$$\Gamma(X,Y) = \Gamma A^{-1}(\tilde{X},\tilde{Y}) \Rightarrow A\Gamma(X,Y) = A\Gamma A^{-1}(\tilde{X},\tilde{Y}). \tag{5.11}$$

As $(A\Gamma A^{-1})^2 = A\Gamma^2 A^{-1} = A\Gamma A^{-1}$, we have that $A\Gamma A^{-1}$ is also a projection, consequently results analogus to (5.7) and (5.8) hold for $A\Gamma A^{-1}$ as well. Thus, we have with high probability, $A\Gamma A^{-1}(\tilde{X},\tilde{Y})$ is approximately gaussian. That is

$$A\Gamma(X,Y) \approx \mathcal{N}_{k_1+k_2}(\mathbf{0}, I_{k_1+k_2}) \tag{5.12}$$

Note that a Gaussian distribution in a smaller dimension can be extended to a Gaussian distribution in higher dimensions by zero padding, i.e.

$$\mathcal{N}_{k_1+k_2}(\mathbf{0}, I_{k_1+k_2}) = \mathcal{N}_{n_1+n_2}(\widehat{\mathbf{0}}, C) \tag{5.13}$$

where $\widehat{\mathbf{0}} \in \mathbb{R}^{n_1+n_2}$ and C is defined as

$$C := \left(\begin{array}{cc|cc} I_{k_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & [O]_{n_1-k_1} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & I_{k_2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & [O]_{n_2-k_2} \end{array}\right) \tag{5.14}$$

Here $[O]_p$ denotes the zero matrix belonging to $\mathbb{R}^{p\times p}$.
We make an abuse of notation. Let ΓX denote the first k_1 components of $\Gamma(X,Y)$ padded with (n_1-k_1) zeros and similarly let ΓY denote the last k_2 components of $\Gamma(X,Y)$ padded with (n_2-k_2) zeros. That is for

$$\Gamma(X,Y) = (g_1, ..., g_{k_1}, ..., g_{k_1+k_2}) \in \mathbb{R}^{k_1+k_2}, \tag{5.15}$$

we have

$$\Gamma X = (g_1, ..., g_{k_1}, 0, ..., 0) \in \mathbb{R}^{n_1}, \tag{5.16}$$

$$\Gamma Y = (g_{k_1+1}, ..., g_{k_1+k_2}, 0, ..., 0) \in \mathbb{R}^{n_2}. \tag{5.17}$$

As A is affine, we have $A(x,y) = B\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ where $B \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$ and $b_1 \in \mathbb{R}^{n_1}$ and $b_2 \in \mathbb{R}^{n_2}$. Note that A is invertible iff B is invertible. We *assume* the invertibility of A and hence of B. With the notation introduced in (5.16) and (5.17), using (5.12), (5.13) and Result 5.1.1 we have,

$$(\Gamma X, \Gamma Y) \approx A^{-1}\mathcal{N}_{n_1+n_2}(\mathbf{0}, C) = \mathcal{N}_{n_1+n_2}\left(\begin{bmatrix} -b_1 \\ -b_2 \end{bmatrix}, B^{-1}C(B^{-1})^t\right). \tag{5.18}$$

In other words, with high probability a random projection of (X, Y) is approximately gaussian (though the projection is not centered), where approximation is in the sense of (5.8).
Now, we consider the random variable $\mathbb{E}(\Gamma Y|\Gamma X)$. Define the square matrix R of size $n_1 + n_2$ as

$$R = \left(\begin{array}{c|c} R_{11} & R_{12} \\ \hline R_{21} & R_{22} \end{array}\right) := B^{-1}C(B^{-1})^t \tag{5.19}$$

where $R_{11} \in \mathbb{R}^{n_1 \times n_1}$. Then for $\Gamma X = z$, (5.4) implies that $\Gamma Y|\Gamma X = z$ is approximately gaussian.
Using (5.4), (5.18) and (5.19), we get

$$\begin{aligned} & \mathbb{E}(\Gamma Y|\Gamma X)|_{\Gamma X=z} \approx -b_2 + R_{21}R_{11}^{-1}(z + b_1) \\ i.e., \quad & \mathbb{E}(\Gamma Y|\Gamma X)|_{\Gamma X=z} = -b_2 + R_{21}R_{11}^{-1}(z + b_1) + o(\epsilon) \end{aligned}$$

Let $z = \Gamma x$, where x is fixed. Based on the fact $||\Gamma x|| \approx ||x||$ (that follows from Theorem 4.2.5 and Result 4.2.7), if we are able to show that

$$\mathbb{E}(Y|\Gamma X)|_{\Gamma X=\Gamma x} \approx \mathbb{E}(Y|X)|_{X=x} \tag{5.20}$$

Then applying Theorem 4.2.5 and Result 4.2.7 on ΓY we have $||\Gamma y|| \approx ||y||$ and that gives,

$$\mathbb{E}(Y|\Gamma X)|_{\Gamma X=x} \approx \mathbb{E}(\Gamma Y|\Gamma X)|_{\Gamma X=\Gamma x}, \tag{5.21}$$

and consequently, we have

$$\mathbb{E}(Y|X)|_{X=x} \approx \mathbb{E}(Y|\Gamma X)|_{X=x} \approx \mathbb{E}(\Gamma Y|\Gamma X)|_{\Gamma X=\Gamma x}. \tag{5.22}$$

That is,

$$\mathbb{E}(Y|X)|_{X=x} = -b_2 + R_{21}R_{11}^{-1}(\Gamma x + b_1) + o(\epsilon). \tag{5.23}$$

Now we use martingale theory to prove (5.20). Consider an orthonormal basis $\{e_i\}$ of $\mathbb{R}^\infty$. Let $X = \sum_{i=1}^{n_1} X_i e_i$. And let $\mathcal{F}_n = \sigma(X_i, i \leqslant n)$ be a filtration. If we consider the sequence of projection Γ_n, as the projection on first n components, then we have,

$$\mathbb{E}[Y|\Gamma_n X] = \mathbb{E}[Y|\sigma(X_i, i \leqslant n)] = \mathbb{E}[Y|\mathcal{F}_n]. \tag{5.24}$$

Using martingale convergence and concentration inequality (see, e.g, Chap. 3 of [27]), we can show that

$$\mathbb{E}[Y|\Gamma_n X] \approx \mathbb{E}[Y|\Gamma_\infty X] = \mathbb{E}[Y|X]. \tag{5.25}$$

And thus we get 5.23. Note that, the ensemble mean of the projections and its ensemble covariance matrix are the Maximum likelihood estimates of actual mean and the covariance matrix of the projections (for detailed discussion, see, Chap. 4 of Murphy [28]).

6. Conclusion and Future Work

A theoretical justification behind the "magical success" of the heuristical methods associated with Ensemble Kalman Filter was provided, when the data involved was "non-Gaussian" and the dimension of the data involved was huge. With the results mentioned in Section 4, with some work in Section 5, we used the fact that EnkF and alike methods involve projections from very high dimensions to low dimensions which introduce Gaussianity into the data. Also, the projections are successful in preserving norms of the originall points with small distortion, leading to the ensemble mean of the projected data being close to the ensemble mean of the original data, which in turn is close to the original mean of the original data (by strong law of large numbers). Till now, we didn't find a good reason in literature as to why EnKF works with the non-Gaussian data. However, it is still popular because of its appeal to easy implementation and surprising success as compared to other methods. We believe, that the novelty of this work lies in its first attempt to justify the success of EnKF.
An obvious future line of work would be get exact error bounds and associated probability in all the approximations we made. Also, the assumptions made by us seem *intuitively* "reasonable", but practical validation of those assumptions might provide better support to our arguments.

References

[1] Klartag, B., and S. Mendelson. "Empirical processes and random projections." *Journal of Functional Analysis 225.1 (2005): 229-245*.

[2] Klartag, Bo'az. "A central limit theorem for convex sets." *Inventiones mathematicae* 168.1 (2007): 91-131.

[3] Klartag, B. "Power-law estimates for the central limit theorem for convex sets." *Journal of Functional Analysis* 245.1 (2007): 284-310.

[4] Eldan, Ronen, and Bo'az Klartag. "Pointwise estimates for marginals of convex bodies." *Journal of Functional Analysis* 254.8 (2008): 2275-2293.

[5] Klartag, Bo'az. "High-dimensional distributions with convexity properties." *A text based on a talk given by the author at the fifth European Congress of Mathematics*, 2010. Available online at `http://www.math.tau.ac.il/∼ klartagb/index.html`.

[6] Milman, Vitali D., and Alain Pajor. "Isotropic position and inertia ellipsoids and zonoids of the unit ball of a normed n-dimensional space." *Geometric aspects of functional analysis*. Springer Berlin Heidelberg, 1989. 64-104.

[7] Talagrand, M. "A simple proof of the majorizing measure theorem." *Geometric & Functional Analysis GAFA* 2.1 (1992): 118-125.

[8] Mandel, Jan. "A brief Tutorial on the ensemble Kalman filter." *arXiv preprint arXiv*:0901.3725 (2009).

[9] Gillijns, Steven, et al. "What is the ensemble Kalman filter and how well does it work?." *American Control Conference*, 2006. IEEE, 2006.

[10] Evensen, Geir. *Data assimilation*. New York: Springer, 2007.

[11] Kalman, Rudolph Emil. "A new approach to linear filtering and prediction problems." *Journal of basic Engineering 82.1 (1960): 35-45*.

[12] Houtekamer, Peter L., and Herschel L. Mitchell. "Data assimilation using an ensemble Kalman filter technique." *Monthly Weather Review 126.3 (1998)*.

[13] Evensen, Geir. "Sequential data assimilation with a nonlinear quasigeostrophic model using Monte Carlo methods to forecast error statistics." *Journal of Geophysical Research: Oceans(1978-2012) 99.C5 (1994)*: 10143-10162.

[14] Anderson, Brian DO, and John B. Moore. *Optimal filtering*. Courier Dover Publications, 2012.

[15] Johnson, William B., and Joram Lindenstrauss. "Extensions of Lipschitz mappings into a Hilbert space." *Contemporary Mathematics* 26.189-206 (1984): 1.

[16] Matousek, Jiri. "On variants of the Johnson-Lindenstrauss lemma." *Random Structures & Algorithms* 33.2 (2008): 142-156.

[17] Dasgupta, Sanjoy, and Anupam Gupta. "An elementary proof of a theorem of Johnson and Lindenstrauss." *Random Structures & Algorithms* 22.1 (2003): 60-65.

[18] Frankl, Peter, and Hiroshi Maehara. "The Johnson-Lindenstrauss lemma and the sphericity of some graphs." *Journal of Combinatorial Theory, Series B* 44.3 (1988): 355-362.

[19] Achlioptas, Dimitris. "Database-friendly random projections: Johnson-Lindenstrauss with binary coins." *Journal of computer and System Sciences* 66.4 (2003): 671-687.

[20] Ailon, Nir, and Bernard Chazelle. "Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform." *Proceedings of the thirty-eighth annual ACM Symposium on Theory of Computing*. ACM, 2006.

[21] Indyk, Piotr, and Rajeev Motwani. "Approximate nearest neighbors: towards removing the curse of dimensionality." *Proceedings of the thirtieth annual ACM Symposium on Theory of Computing*. ACM, 1998.

[22] Sullivan, Dennis. "For $n > 3$ there is only one finitely additive rotationally invariant measure on the n-sphere defined on all Lebesgue measurable subsets." *Bulletin (New Series) of the American Mathematical Society* 4.1 (1981): 121-123.

[23] Matousek, Jiri. *Lectures on discrete geometry*. Vol. 212. Springer, 2002.

[24] Alon, Noga. "Problems and results in extremal combinatorics-I." *Discrete Mathematics* 273.1 (2003): 31-53.

[25] Jeffrey Anderson, Tim Hoar and Nancy Collins. "Introduction to Ensemble Kalman Filters and the Data Assimilation Research Testbed." `http://icap.atmos.und.edu/EnsembleForecastsDataAssimilation/MeetingPDFs/May11/07_Anderson_icap.pdf`

[26] "Data Assimilation Research Testbed Tutorial" by DART.
http://www2.cscamm.umd.edu/programs/das13/presentations/DART_section4.pdf

[27] Borkar, Vivek S. *Probability theory: an advanced course*. Springer, 1995.

[28] Murphy, Kevin P. *Machine learning: a probabilistic perspective*. MIT Press, 2012.

[29] Ash, R. B. "Lectures on Statistics."
http://www.math.uiuc.edu/∼r-ash/Stat/StatLec21-25.pdf

[30] Tucker, H. "Lectures on Multivariate Analysis." http://www.math.uci.edu/∼
htucker/LectureNotes/ MultivariateAnalysis.PDF