Learning Goals

- Concept of dimensenality reduction
- Correctly state the procedure and guarantee of Johnson-Lindenstrauss transform
- Proof idea of JL-transform

Dimensionality Reduction

- Data points can often live in very high dimensions

Dimensionality Reduction

- Data points can often live in very high dimensions
  - Images

Dimensionality Reduction

- Data points can often live in very high dimensions
  - Images
  - Vector representation of articles

Dimensionality Reduction

- Data points can often live in very high dimensions
  - Images
  - Vector representation of articles
  - Vector representation of words

Dimensionality Reduction

- Data points can often live in very high dimensions
  - Images
  - Vector representation of articles
  - Vector representation of words
- Many algorithms are very slow when run on high dimensional input
  - *Curse of dimensionality*

Dimensionality Reduction

- Data points can often live in very high dimensions
  - Images
  - Vector representation of articles
  - Vector representation of words
- Many algorithms are very slow when run on high dimensional input
  - *Curse of dimensionality*
- *Dimensionality reduction*: Transform data to lower dimensions while preserving information useful for analysis/application

Johnson-Lindenstrauss

- Distances between data points are often meaningful

Johnson-Lindenstrauss

- Distances between data points are often meaningful
- For $x \in \mathbb{R}^d$, the *ℓ_2 norm* of x is

$$||x|| = \left(\sum_{i=1}^{d} x_i^2 \right)^{1/2}.$$

For $x, y \in \mathbb{R}^d$, $||x - y||$ is their ℓ_2-distance, or *Euclidean distance*.

Johnson-Lindenstrauss

- Distances between data points are often meaningful
- For $x \in \mathbb{R}^d$, the *ℓ_2 norm* of x is

$$||x|| = \left(\sum_{i=1}^{d} x_i^2 \right)^{1/2}.$$

  For $x, y \in \mathbb{R}^d$, $||x - y||$ is their ℓ_2-distance, or *Euclidean distance*.
- The *Johnson-Lindenstrauss* transform is a *randomized* dimensionality reduction algorithm that *approximately* preserves Euclidean distances.

JL Statement

Theorem (Johnson-Lindenstrauss)

For arbitrary $x_1, \ldots, x_n \in \mathbb{R}^d$, and any $\epsilon \in (0,1)$, there is $t = O(\log n/\epsilon^2)$ such that there are $y_1, \ldots, y_n \in \mathbb{R}^t$ with

$$(1-\epsilon)||x_j|| \leq ||y_j|| \leq (1+\epsilon)||x_j||, \quad \forall j$$
$$(1-\epsilon)||x_j - x_{j'}|| \leq ||y_j - y_{j'}|| \leq (1+\epsilon)||x_j - x_{j'}||, \quad \forall j, j'.$$

Moreover, $y_1, \ldots, y_n$ can be computed in polynomial time.

Main Lemma

Lemma

Distributional JL For any $\epsilon, \delta \in (0, 1]$, there is a $t = O(\log(1/\delta)/\epsilon^2)$ and a random linear map $f : \mathbb{R}^d \to \mathbb{R}^t$, such that, for any $v \in \mathbb{R}^d$ with $||v|| = 1$,

$$\mathbf{Pr}\left[1 - \epsilon \leq \frac{||f(v)||}{\sqrt{t}} \leq 1 + \epsilon\right] \geq 1 - 2\delta.$$

Main Lemma

Lemma

Distributional JL For any $\epsilon, \delta \in (0, 1]$, there is a $t = O(\log(1/\delta)/\epsilon^2)$ and a random linear map $f : \mathbb{R}^d \to \mathbb{R}^t$, such that, for any $v \in \mathbb{R}^d$ with $||v|| = 1$,

$$\mathbf{Pr}\left[1 - \epsilon \leq \frac{||f(v)||}{\sqrt{t}} \leq 1 + \epsilon\right] \geq 1 - 2\delta.$$

Proof of Theorem using Lemma.

Consider $W = \{x_1, \ldots, x_n\} \cup \{x_i - x_j : i < j\}$.

Main Lemma

Lemma

Distributional JL For any $\epsilon, \delta \in (0, 1]$, there is a $t = O(\log(1/\delta)/\epsilon^2)$ and a random linear map $f : \mathbb{R}^d \to \mathbb{R}^t$, such that, for any $v \in \mathbb{R}^d$ with $||v|| = 1$,

$$\mathbf{Pr}\left[1 - \epsilon \leq \frac{||f(v)||}{\sqrt{t}} \leq 1 + \epsilon\right] \geq 1 - 2\delta.$$

Proof of Theorem using Lemma.

Consider $W = \{x_1, \ldots, x_n\} \cup \{x_i - x_j : i < j\}$. Note $|W| \leq n^2$. Take $\delta = 1/n^3$.

Main Lemma

Lemma

Distributional JL For any $\epsilon, \delta \in (0, 1]$, there is a $t = O(\log(1/\delta)/\epsilon^2)$ and a random linear map $f : \mathbb{R}^d \to \mathbb{R}^t$, such that, for any $v \in \mathbb{R}^d$ with $||v|| = 1$,

$$\mathbf{Pr}\left[1 - \epsilon \leq \frac{||f(v)||}{\sqrt{t}} \leq 1 + \epsilon\right] \geq 1 - 2\delta.$$

Proof of Theorem using Lemma.

Consider $W = \{x_1, \ldots, x_n\} \cup \{x_i - x_j : i < j\}$. Note $|W| \leq n^2$. Take $\delta = 1/n^3$. For each $w \in W$, consider $v = \frac{w}{||w||}$.

Main Lemma

Lemma

Distributional JL For any $\epsilon, \delta \in (0, 1]$, there is a $t = O(\log(1/\delta)/\epsilon^2)$ and a random linear map $f : \mathbb{R}^d \to \mathbb{R}^t$, such that, for any $v \in \mathbb{R}^d$ with $||v|| = 1$,

$$\mathbf{Pr}\left[1 - \epsilon \leq \frac{||f(v)||}{\sqrt{t}} \leq 1 + \epsilon\right] \geq 1 - 2\delta.$$

Proof of Theorem using Lemma.

Consider $W = \{x_1, \ldots, x_n\} \cup \{x_i - x_j : i < j\}$. Note $|W| \leq n^2$. Take $\delta = 1/n^3$. For each $w \in W$, consider $v = \frac{w}{||w||}$. Consider the event

$$\mathcal{E}_w := \left\{\frac{||f(w)||}{\sqrt{t}} \notin [1 - \epsilon, 1 + \epsilon] \cdot ||w||\right\} = \left\{\frac{||f(v)||}{\sqrt{t}} \notin [1 - \epsilon, 1 + \epsilon]\right\}.$$

Main Lemma

Lemma

Distributional JL For any $\epsilon, \delta \in (0, 1]$, there is a $t = O(\log(1/\delta)/\epsilon^2)$ and a random linear map $f : \mathbb{R}^d \to \mathbb{R}^t$, such that, for any $v \in \mathbb{R}^d$ with $||v|| = 1$,

$$\mathbf{Pr}\left[1 - \epsilon \leq \frac{||f(v)||}{\sqrt{t}} \leq 1 + \epsilon\right] \geq 1 - 2\delta.$$

Proof of Theorem using Lemma.

Consider $W = \{x_1, \ldots, x_n\} \cup \{x_i - x_j : i < j\}$. Note $|W| \leq n^2$. Take $\delta = 1/n^3$. For each $w \in W$, consider $v = \frac{w}{||w||}$. Consider the event

$$\mathcal{E}_w := \left\{\frac{||f(w)||}{\sqrt{t}} \notin [1 - \epsilon, 1 + \epsilon] \cdot ||w||\right\} = \left\{\frac{||f(v)||}{\sqrt{t}} \notin [1 - \epsilon, 1 + \epsilon]\right\}.$$

Each such event occurs w.p. $\leq 2\delta$. By the union bound, the probability that none of these happen is $\geq 1|W| \cdot 2\delta \geq 1 - \frac{2}{n}$. □

Gaussian Distribution

- For a random variable X, its *cumulative distribution function* (CDF) is $F_X(x) := \mathbf{Pr}[X \leq x]$.

Gaussian Distribution

- For a random variable X, its *cumulative distribution function* (CDF) is $F_X(x) := \mathbf{Pr}[X \leq x]$.
  - Example: For X uniformly distributed on $[0, 1]$, $F(x) = x$, for $x \in [0, 1]$.

Gaussian Distribution

- For a random variable X, its *cumulative distribution function* (CDF) is $F_X(x) := \mathbf{Pr}[X \leq x]$.
  - Example: For X uniformly distributed on $[0, 1]$, $F(x) = x$, for $x \in [0, 1]$.
- For a continuous random variable, the *probability density function* (PDF) is $f_X(x) := \frac{\mathrm{d}}{\mathrm{d}x} F_X(x)$.

Gaussian Distribution

- For a random variable X, its *cumulative distribution function* (CDF) is $F_X(x) := \mathbf{Pr}[X \leq x]$.
  - Example: For X uniformly distributed on $[0, 1]$, $F(x) = x$, for $x \in [0, 1]$.
- For a continuous random variable, the *probability density function* (PDF) is $f_X(x) := \frac{\mathrm{d}}{\mathrm{d}x} F_X(x)$.
  - For X uniformly distributed on $[0, 1]$, $f(x) = 1$ for $x \in [0, 1]$.

Gaussian Distribution

- For a random variable X, its *cumulative distribution function* (CDF) is $F_X(x) := \mathbf{Pr}[X \leq x]$.
  - Example: For X uniformly distributed on $[0, 1]$, $F(x) = x$, for $x \in [0, 1]$.
- For a continuous random variable, the *probability density function* (PDF) is $f_X(x) := \frac{\mathrm{d}}{\mathrm{d}x} F_X(x)$.
  - For X uniformly distributed on $[0, 1]$, $f(x) = 1$ for $x \in [0, 1]$.
- A random variable is drawn from *Gaussian distribution* (or *Normal distribution*) $\mathcal{N}(\mu, \sigma^2)$ if its PDF is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Gaussian Distribution

- For a random variable X, its *cumulative distribution function* (CDF) is $F_X(x) := \mathbf{Pr}[X \leq x]$.
  - Example: For X uniformly distributed on $[0, 1]$, $F(x) = x$, for $x \in [0, 1]$.
- For a continuous random variable, the *probability density function* (PDF) is $f_X(x) := \frac{\mathrm{d}}{\mathrm{d}x} F_X(x)$.
  - For X uniformly distributed on $[0, 1]$, $f(x) = 1$ for $x \in [0, 1]$.
- A random variable is drawn from *Gaussian distribution* (or *Normal distribution*) $\mathcal{N}(\mu, \sigma^2)$ if its PDF is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}.$$

  - In particular, the *standard normal distribution* has PDF $\varphi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$.

Gaussian Distribution

- For a random variable X, its *cumulative distribution function* (CDF) is $F_X(x) := \mathbf{Pr}[X \leq x]$.
  - Example: For X uniformly distributed on $[0, 1]$, $F(x) = x$, for $x \in [0, 1]$.
- For a continuous random variable, the *probability density function* (PDF) is $f_X(x) := \frac{\mathrm{d}}{\mathrm{d}x} F_X(x)$.
  - For X uniformly distributed on $[0, 1]$, $f(x) = 1$ for $x \in [0, 1]$.
- A random variable is drawn from *Gaussian distribution* (or *Normal distribution*) $\mathcal{N}(\mu, \sigma^2)$ if its PDF is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}.$$

  - In particular, the *standard normal distribution* has PDF $\varphi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$.
  - If $X \sim \mathcal{N}(0, 1)$, then $\sigma X + \mu \sim \mathcal{N}(\mu, \sigma^2)$.

Properties of Gaussian Distribution

Theorem

Linear combinations of independent Gaussian variables are still Gaussian.

Properties of Gaussian Distribution

Theorem

Linear combinations of independent Gaussian variables are still Gaussian.

Fact: The moment generating function $\mathbf{E}[e^{\lambda X}]$ of a random variable X uniquely determines its CDF.

Properties of Gaussian Distribution

Theorem

Linear combinations of independent Gaussian variables are still Gaussian.

Fact: The moment generating function $\mathbf{E}[e^{\lambda X}]$ of a random variable X uniquely determines its CDF.

Proof of Theorem.

We show only the zero mean case. For $X \sim \mathcal{N}(0, \sigma^2)$,

$$\begin{aligned}\mathbf{E}\left[e^{\lambda X}\right] &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2} + \lambda x\right) \mathrm{d}x \\ &= \frac{e^{\sigma^2\lambda^2/2}}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(\frac{x}{\sigma} - \sigma\lambda)^2} \mathrm{d}x = e^{\frac{\sigma^2\lambda^2}{2}}.\end{aligned}$$

So for independent $X \sim \mathcal{N}(0, \sigma_1^2)$, $Y \sim \mathcal{N}(0, \sigma_2^2)$,
$\mathbf{E}[e^{\lambda(X+Y)}] = \mathbf{E}[e^{\lambda X}] \cdot \mathbf{E}[e^{\lambda Y}] = e^{(\sigma_1^2+\sigma_2^2)\lambda^2/2}$. □

Idea of JL

- For $x \in \mathbb{R}^d$ with $||x|| = 1$, let $G_1, \cdots, G_d$ be i.i.d. from $\mathcal{N}(0, 1)$, then $\sum_i G_i x_i \sim \mathcal{N}(0, ||x||^2) = \mathcal{G}(0, 1)$.

Idea of JL

- For $x \in \mathbb{R}^d$ with $||x|| = 1$, let $G_1, \cdots, G_d$ be i.i.d. from $\mathcal{N}(0, 1)$, then $\sum_i G_i x_i \sim \mathcal{N}(0, ||x||^2) = \mathcal{G}(0, 1)$.
  - $\sum_i G_i x_i$ is obviously zero mean.

Idea of JL

- For $x \in \mathbb{R}^d$ with $||x|| = 1$, let $G_1, \cdots, G_d$ be i.i.d. from $\mathcal{N}(0,1)$, then $\sum_i G_i x_i \sim \mathcal{N}(0, ||x||^2) = \mathcal{G}(0,1)$.
  - $\sum_i G_i x_i$ is obviously zero mean.
  - Its variance is: $\mathbf{E}[(\sum_i G_i x_i)^2] = \mathbf{E}[\sum_i x_i^2] = ||x||^2 = 1$.

Idea of JL

- For $x \in \mathbb{R}^d$ with $||x|| = 1$, let $G_1, \cdots, G_d$ be i.i.d. from $\mathcal{N}(0, 1)$, then $\sum_i G_i x_i \sim \mathcal{N}(0, ||x||^2) = \mathcal{G}(0, 1)$.
  - $\sum_i G_i x_i$ is obviously zero mean.
  - Its variance is: $\mathbf{E}[(\sum_i G_i x_i)^2] = \mathbf{E}[\sum_i x_i^2] = ||x||^2 = 1$.
- By sampling $(\sum_i G_i x_i)^2$ multiple times, with good probability the average should be around the mean.

Idea of JL

- For $x \in \mathbb{R}^d$ with $||x|| = 1$, let $G_1, \cdots, G_d$ be i.i.d. from $\mathcal{N}(0, 1)$, then $\sum_i G_i x_i \sim \mathcal{N}(0, ||x||^2) = \mathcal{G}(0, 1)$.
  - $\sum_i G_i x_i$ is obviously zero mean.
  - Its variance is: $\mathbf{E}[(\sum_i G_i x_i)^2] = \mathbf{E}[\sum_i x_i^2] = ||x||^2 = 1$.
- By sampling $(\sum_i G_i x_i)^2$ multiple times, with good probability the average should be around the mean.
- If we multiply x by a $t \times d$ matrix A, whose entries are i.i.d. standard Gaussian variables

Idea of JL

- For $x \in \mathbb{R}^d$ with $||x|| = 1$, let $G_1, \cdots, G_d$ be i.i.d. from $\mathcal{N}(0, 1)$, then $\sum_i G_i x_i \sim \mathcal{N}(0, ||x||^2) = \mathcal{G}(0, 1)$.
  - $\sum_i G_i x_i$ is obviously zero mean.
  - Its variance is: $\mathbf{E}[(\sum_i G_i x_i)^2] = \mathbf{E}[\sum_i x_i^2] = ||x||^2 = 1$.
- By sampling $(\sum_i G_i x_i)^2$ multiple times, with good probability the average should be around the mean.
- If we multiply x by a $t \times d$ matrix A, whose entries are i.i.d. standard Gaussian variables
  - The resulting random vector $Ax \in \mathbb{R}^t$ has each coordinate drawn from $\mathcal{N}(0, 1)$.

Idea of JL

- For $x \in \mathbb{R}^d$ with $||x|| = 1$, let $G_1, \cdots, G_d$ be i.i.d. from $\mathcal{N}(0, 1)$, then $\sum_i G_i x_i \sim \mathcal{N}(0, ||x||^2) = \mathcal{G}(0, 1)$.
  - $\sum_i G_i x_i$ is obviously zero mean.
  - Its variance is: $\mathbf{E}[(\sum_i G_i x_i)^2] = \mathbf{E}[\sum_i x_i^2] = ||x||^2 = 1$.
- By sampling $(\sum_i G_i x_i)^2$ multiple times, with good probability the average should be around the mean.
- If we multiply x by a $t \times d$ matrix A, whose entries are i.i.d. standard Gaussian variables
  - The resulting random vector $Ax \in \mathbb{R}^t$ has each coordinate drawn from $\mathcal{N}(0, 1)$.
  - The expectation of $||Ax||^2$ is t.

Idea of JL

- For $x \in \mathbb{R}^d$ with $||x|| = 1$, let $G_1, \cdots, G_d$ be i.i.d. from $\mathcal{N}(0,1)$, then $\sum_i G_i x_i \sim \mathcal{N}(0, ||x||^2) = \mathcal{G}(0,1)$.
  - $\sum_i G_i x_i$ is obviously zero mean.
  - Its variance is: $\mathbf{E}[(\sum_i G_i x_i)^2] = \mathbf{E}[\sum_i x_i^2] = ||x||^2 = 1$.
- By sampling $(\sum_i G_i x_i)^2$ multiple times, with good probability the average should be around the mean.
- If we multiply x by a $t \times d$ matrix A, whose entries are i.i.d. standard Gaussian variables
  - The resulting random vector $Ax \in \mathbb{R}^t$ has each coordinate drawn from $\mathcal{N}(0,1)$.
  - The expectation of $||Ax||^2$ is t.
  - Let $A' = \frac{1}{\sqrt{t}} A$, then $\mathbf{E}[||A'x||^2] = 1$.

Idea of JL

- For $x \in \mathbb{R}^d$ with $||x|| = 1$, let $G_1, \cdots, G_d$ be i.i.d. from $\mathcal{N}(0,1)$, then $\sum_i G_i x_i \sim \mathcal{N}(0, ||x||^2) = \mathcal{G}(0,1)$.
  - $\sum_i G_i x_i$ is obviously zero mean.
  - Its variance is: $\mathbf{E}[(\sum_i G_i x_i)^2] = \mathbf{E}[\sum_i x_i^2] = ||x||^2 = 1$.
- By sampling $(\sum_i G_i x_i)^2$ multiple times, with good probability the average should be around the mean.
- If we multiply x by a $t \times d$ matrix A, whose entries are i.i.d. standard Gaussian variables
  - The resulting random vector $Ax \in \mathbb{R}^t$ has each coordinate drawn from $\mathcal{N}(0,1)$.
  - The expectation of $||Ax||^2$ is t.
  - Let $A' = \frac{1}{\sqrt{t}} A$, then $\mathbf{E}[||A'x||^2] = 1$.
- We just need to show that the empirical average converges to the expectation fast enough with t.

Proof of Lemma

- For each $i = 1, \ldots, t$, let Y_i be the i^{th} coordinate of AX.

Proof of Lemma

- For each $i = 1, \ldots, t$, let Y_i be the i^{th} coordinate of AX.
- Then Y_i is Gaussian from $\mathcal{N}(0, 1)$.

Proof of Lemma

- For each $i = 1, \ldots, t$, let Y_i be the i^{th} coordinate of AX.
- Then Y_i is Gaussian from $\mathcal{N}(0, 1)$.
- Let Y be $\sum_i Y_i^2$, then $\mathbf{E}[Y] = t$.

Proof of Lemma

- For each $i = 1, \ldots, t$, let Y_i be the i^{th} coordinate of AX.
- Then Y_i is Gaussian from $\mathcal{N}(0, 1)$.
- Let Y be $\sum_i Y_i^2$, then $\mathbf{E}[Y] = t$.

$$\begin{aligned}\mathbf{Pr}\left[\frac{||Ax||}{\sqrt{t}} \geq (1+\epsilon)\right] &= \mathbf{Pr}\left[Y \geq (1+\epsilon)^2 t\right] \\ &= \mathbf{Pr}\left[Y \geq (1+\epsilon)^2 \mathbf{E}\left[Y\right]\right].\end{aligned}$$

Proof of Lemma

- For each $i = 1, \ldots, t$, let Y_i be the i^{th} coordinate of AX.
- Then Y_i is Gaussian from $\mathcal{N}(0, 1)$.
- Let Y be $\sum_i Y_i^2$, then $\mathbf{E}[Y] = t$.

$$\begin{aligned}\mathbf{Pr}\left[\frac{||Ax||}{\sqrt{t}} \geq (1+\epsilon)\right] &= \mathbf{Pr}\left[Y \geq (1+\epsilon)^2 t\right] \\ &= \mathbf{Pr}\left[Y \geq (1+\epsilon)^2 \mathbf{E}\left[Y\right]\right].\end{aligned}$$

Let's bound $\mathbf{Pr}[Y > \alpha]$ for any α. For any $\lambda > 0$, we have

$$\begin{aligned}\mathbf{Pr}\left[Y \geq \alpha\right] &= \mathbf{Pr}\left[e^{\lambda Y} \geq e^{\lambda\alpha}\right] \\ &\leq \frac{\mathbf{E}[e^{\lambda Y}]}{e^{\lambda\alpha}} = \frac{\prod_i \mathbf{E}[e^{\lambda Y_i^2}]}{e^{\lambda\alpha}}.\end{aligned}$$

Moment Generating Function of χ^2-distributions

If $X_1, \cdot, X_k$ are independent standard normal random variables, then $Q = \sum_i X_i$ is said to be distributed according to the χ^2-distribution with k degrees of freedom.

Moment Generating Function of χ^2-distributions

If $X_1, \cdot, X_k$ are independent standard normal random variables, then $Q = \sum_i X_i$ is said to be distributed according to the χ^2-distribution with k degrees of freedom.

Claim

If $X \sim \mathcal{N}(0,1)$, then $\mathbf{E}[e^{\lambda X^2}] = 1/\sqrt{1-2\lambda}$, for $-\infty < \lambda < \frac{1}{2}$.

Moment Generating Function of χ^2-distributions

If $X_1, \cdot, X_k$ are independent standard normal random variables, then $Q = \sum_i X_i$ is said to be distributed according to the χ^2-distribution with k degrees of freedom.

Claim

If $X \sim \mathcal{N}(0,1)$, then $\mathbf{E}[e^{\lambda X^2}] = 1/\sqrt{1-2\lambda}$, for $-\infty < \lambda < \frac{1}{2}$.

Proof of Claim.

$$\begin{aligned}\mathbf{E}\left[e^{\lambda X^2}\right] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda x^2 - \frac{x^2}{2}}\, \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{1-2\lambda}} \int e^{-y^2/2}\, \mathrm{d}y = \frac{1}{\sqrt{1-2\lambda}}.\end{aligned}$$

where we substituted $y = \sqrt{1-2\lambda}x$. □

Finishing Proof of Lemma

$$\mathbf{Pr}\left[Y \geq \alpha\right] \leq \frac{\prod_i \mathbf{E}[e^{\lambda Y_i^2}]}{e^{\lambda\alpha}} = (1-2\lambda)^{-t/2} e^{-\lambda\alpha}.$$

Finishing Proof of Lemma

$$\mathbf{Pr}\left[Y \geq \alpha\right] \leq \frac{\prod_i \mathbf{E}[e^{\lambda Y_i^2}]}{e^{\lambda\alpha}} = (1-2\lambda)^{-t/2} e^{-\lambda\alpha}.$$

Now minimize the RHS by setting $\lambda = \frac{1}{2}(1 - \frac{t}{\alpha})$, we obtain
$\mathbf{Pr}[Y \geq \alpha] \leq e^{(t-\alpha)/2}(t/\alpha)^{-t/2}$.
Now let α be $(1+\epsilon)^2 t$, we get

$$\mathbf{Pr}\left[Y \geq (1+\epsilon)^2 t\right] \leq \exp\left(-t(\epsilon + \frac{\epsilon^2}{2} - \ln(1+\epsilon))\right).$$

Finishing Proof of Lemma

$$\mathbf{Pr}\left[Y \geq \alpha\right] \leq \frac{\prod_i \mathbf{E}[e^{\lambda Y_i^2}]}{e^{\lambda\alpha}} = (1-2\lambda)^{-t/2} e^{-\lambda\alpha}.$$

Now minimize the RHS by setting $\lambda = \frac{1}{2}(1 - \frac{t}{\alpha})$, we obtain
$\mathbf{Pr}[Y \geq \alpha] \leq e^{(t-\alpha)/2}(t/\alpha)^{-t/2}$.
Now let α be $(1+\epsilon)^2 t$, we get

$$\mathbf{Pr}\left[Y \geq (1+\epsilon)^2 t\right] \leq \exp\left(-t(\epsilon + \frac{\epsilon^2}{2} - \ln(1+\epsilon))\right).$$

Using basic calculus, we can show $\ln(1+\epsilon) \leq \epsilon - \frac{\epsilon^2}{4}$ for $\epsilon \in [0, 1]$, so we have

$$\mathbf{Pr}\left[Y \geq (1+\epsilon)^2 t\right] \leq e^{-\frac{3}{4}\epsilon^2 t}.$$

Remarks

We performed the proof of Chernoff bound for Gaussian distributions.

Remarks

We performed the proof of Chernoff bound for Gaussian distributions.

- Gaussian variables are not bounded, so Chernoff bound does not directly apply.

Remarks

We performed the proof of Chernoff bound for Gaussian distributions.

- Gaussian variables are not bounded, so Chernoff bound does not directly apply.
- The proof we gave though is almost identical to that of Chernoff bound, except for the calculation of the expectation.

Remarks

We performed the proof of Chernoff bound for Gaussian distributions.

- Gaussian variables are not bounded, so Chernoff bound does not directly apply.
- The proof we gave though is almost identical to that of Chernoff bound, except for the calculation of the expectation.

In general, "Chernoff like" bounds hold for distributions with tails smaller than Gaussian. They are known as *sub-Gaussian distributions*.

Remarks

We performed the proof of Chernoff bound for Gaussian distributions.

- Gaussian variables are not bounded, so Chernoff bound does not directly apply.
- The proof we gave though is almost identical to that of Chernoff bound, except for the calculation of the expectation.

In general, "Chernoff like" bounds hold for distributions with tails smaller than Gaussian. They are known as *sub-Gaussian distributions*.