Concentration of Measure

Sudeep Kamath

PRINCETON UNIVERSITY

CIRM workshop, 26 Jan 2016

Goal

Goal

- Review tensorization of variance

Goal

- Review tensorization of variance
- Explore sub-Gaussian concentration

Goal

- Review tensorization of variance
- Explore sub-Gaussian concentration
- Develop basic information inequalities

What is concentration?

"A random variable that depends in a smooth way on many independent random variables (but not too much on any of them) is essentially constant."
- M. Talagrand, 1996.

If Z is a function of many independent variables $X_1, X_2, \ldots, X_n$, how large are typical deviations of Z?

$$\mathbb{P}[|Z - \mathbb{E}Z| \geq t] \leq \frac{\mathrm{Var}(Z)}{t^2}$$

Probability that Z deviates more than $10\sqrt{\mathrm{Var}(Z)}$ from $\mathbb{E}Z$ is at most 1%

Tensorization of variance

Let $Z = f(X_1, X_2, \ldots, X_n)$ where $X_1, X_2, \ldots, X_n$ are independent random variables.

$$X^{(i)} := (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n), \qquad \mathbb{E}^{(i)}[\,\cdot\,] := \mathbb{E}[\,\cdot\,|X^{(i)}]$$

$$\mathrm{Var}^{(i)}(Z) := \mathrm{Var}(Z|X^{(i)})$$

Tensorization of variance (Efron-Stein-Steele inequality)

$$\mathrm{Var}(Z) \leq \sum_{i=1}^{n} \mathbb{E}[\mathrm{Var}^{(i)}(Z)]$$

Recall if Z, Z' are i.i.d., then

Recall if Z, Z' are i.i.d., then

$$\mathrm{Var}(Z) = \frac{1}{2}\mathbb{E}[(Z - Z')^2]$$

Recall if Z, Z' are i.i.d., then

$$\mathrm{Var}(Z) = \frac{1}{2}\mathbb{E}[(Z - Z')^2]$$

If $Z = f(X_1, \ldots, X_i, \ldots, X_n)$, and $Z_i = f(X_1, \ldots, X_i', \ldots, X_n)$ where X_i' is an independent copy of X_i, then

Recall if Z, Z' are i.i.d., then

$$\mathrm{Var}(Z) = \frac{1}{2}\mathbb{E}[(Z - Z')^2]$$

If $Z = f(X_1, \ldots, X_i, \ldots, X_n)$, and $Z_i = f(X_1, \ldots, X_i', \ldots, X_n)$ where X_i' is an independent copy of X_i, then

$$\mathrm{Var}^{(i)}(Z) = \frac{1}{2}\mathbb{E}^{(i)}[(Z - Z_i)^2]$$

Recall if Z, Z' are i.i.d., then

$$\mathrm{Var}(Z) = \frac{1}{2}\mathbb{E}[(Z - Z')^2]$$

If $Z = f(X_1, \ldots, X_i, \ldots, X_n)$, and $Z_i = f(X_1, \ldots, X_i', \ldots, X_n)$ where X_i' is an independent copy of X_i, then

$$\mathrm{Var}^{(i)}(Z) = \frac{1}{2}\mathbb{E}^{(i)}[(Z - Z_i)^2]$$

Variant: "resampling coordinates"

Recall if Z, Z' are i.i.d., then

$$\mathrm{Var}(Z) = \frac{1}{2}\mathbb{E}[(Z - Z')^2]$$

If $Z = f(X_1, \ldots, X_i, \ldots, X_n)$, and $Z_i = f(X_1, \ldots, X_i', \ldots, X_n)$ where X_i' is an independent copy of X_i, then

$$\mathrm{Var}^{(i)}(Z) = \frac{1}{2}\mathbb{E}^{(i)}[(Z - Z_i)^2]$$

Variant: "resampling coordinates"

$$\mathrm{Var}(Z) \leq \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}[(Z - Z_i)^2]$$

Recall if Z, Z' are i.i.d., then

$$\mathrm{Var}(Z) = \frac{1}{2}\mathbb{E}[(Z - Z')^2] = \mathbb{E}[(Z - Z')_+^2], \quad (a)_+ = \max\{a, 0\}$$

If $Z = f(X_1, \ldots, X_i, \ldots, X_n)$, and $Z_i = f(X_1, \ldots, X_i', \ldots, X_n)$ where X_i' is an independent copy of X_i, then

$$\mathrm{Var}^{(i)}(Z) = \frac{1}{2}\mathbb{E}^{(i)}[(Z - Z_i)^2]$$

Variant: "resampling coordinates"

$$\mathrm{Var}(Z) \leq \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}[(Z - Z_i)^2]$$

Recall if Z, Z' are i.i.d., then

$$\mathrm{Var}(Z) = \frac{1}{2}\mathbb{E}[(Z - Z')^2] = \mathbb{E}[(Z - Z')_+^2], \quad (a)_+ = \max\{a, 0\}$$

If $Z = f(X_1, \ldots, X_i, \ldots, X_n)$, and $Z_i = f(X_1, \ldots, X_i', \ldots, X_n)$ where X_i' is an independent copy of X_i, then

$$\mathrm{Var}^{(i)}(Z) = \frac{1}{2}\mathbb{E}^{(i)}[(Z - Z_i)^2] = \mathbb{E}^{(i)}[(Z - Z_i)_+^2]$$

Variant: "resampling coordinates"

$$\mathrm{Var}(Z) \leq \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}[(Z - Z_i)^2]$$

Recall if Z, Z' are i.i.d., then

$$\mathrm{Var}(Z) = \frac{1}{2}\mathbb{E}[(Z - Z')^2] = \mathbb{E}[(Z - Z')_+^2], \quad (a)_+ = \max\{a, 0\}$$

If $Z = f(X_1, \ldots, X_i, \ldots, X_n)$, and $Z_i = f(X_1, \ldots, X_i', \ldots, X_n)$ where X_i' is an independent copy of X_i, then

$$\mathrm{Var}^{(i)}(Z) = \frac{1}{2}\mathbb{E}^{(i)}[(Z - Z_i)^2] = \mathbb{E}^{(i)}[(Z - Z_i)_+^2]$$

Variant: "resampling coordinates"

$$\mathrm{Var}(Z) \leq \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}[(Z - Z_i)^2] = \sum_{i=1}^{n}\mathbb{E}[(Z - Z_i)_+^2]$$

Largest eigenvalue of a random matrix

Let A be an $n \times n$ symmetric matrix with independent entries $X_{ij}, 1 \leq i \leq j \leq n$ independent, $-1 \leq X_{ij} \leq 1$.

Largest eigenvalue of a random matrix

Let A be an $n \times n$ symmetric matrix with independent entries $X_{ij}, 1 \leq i \leq j \leq n$ independent, $-1 \leq X_{ij} \leq 1$.

Let $Z = \lambda_{\max}(A) = \max_{\|w\|=1} w^{\mathrm{T}} A w = u^{\mathrm{T}} A u$

for some u that depends on the X_{ij}'s.

Largest eigenvalue of a random matrix

Let A be an $n \times n$ symmetric matrix with independent entries $X_{ij}, 1 \leq i \leq j \leq n$ independent, $-1 \leq X_{ij} \leq 1$.

Let $Z = \lambda_{\max}(A) = \max_{\|w\|=1} w^{\mathrm{T}} A w = u^{\mathrm{T}} A u$
for some u that depends on the X_{ij}'s.

$Z \in [-1, n]$ (exercise)

Largest eigenvalue of a random matrix

Let A be an $n \times n$ symmetric matrix with independent entries $X_{ij}, 1 \leq i \leq j \leq n$ independent, $-1 \leq X_{ij} \leq 1$.

Let $Z = \lambda_{\max}(A) = \max_{\|w\|=1} w^{\mathrm{T}} A w = u^{\mathrm{T}} A u$
for some u that depends on the X_{ij}'s.

$Z \in [-1, n]$ (exercise)

Let Z_{ij} denote $\lambda_{\max}$ for the matrix $\bar{A}^{ij}$ which is same as the matrix A except $X_{ij} = X_{ji}$ gets replaced by an independent copy $X'_{ij} = X'_{ji}$.

Largest eigenvalue of a random matrix

$$Z - Z_{ij} = u^{\mathrm{T}} A u - \max_{\|w\|=1} w^{\mathrm{T}} \bar{A}^{ij} w$$

Largest eigenvalue of a random matrix

$$
\begin{aligned}
Z - Z_{ij} &= u^{\mathrm{T}} A u - \max_{\|w\|=1} w^{\mathrm{T}} \bar{A}^{ij} w \\
&\le u^{\mathrm{T}} A u - u^{\mathrm{T}} \bar{A}^{ij} u
\end{aligned}
$$

Largest eigenvalue of a random matrix

$$
\begin{aligned}
Z - Z_{ij} &= u^{\mathrm{T}} A u - \max_{\|w\|=1} w^{\mathrm{T}} \bar{A}^{ij} w \\
&\le u^{\mathrm{T}} A u - u^{\mathrm{T}} \bar{A}^{ij} u \\
&\le 2 \cdot |u_i| \cdot |u_j| \cdot |X_{ij} - X'_{ij}|
\end{aligned}
$$

Largest eigenvalue of a random matrix

$$
\begin{aligned}
Z - Z_{ij} &= u^{\mathrm{T}} A u - \max_{\|w\|=1} w^{\mathrm{T}} \bar{A}^{ij} w \\
&\le u^{\mathrm{T}} A u - u^{\mathrm{T}} \bar{A}^{ij} u \\
&\le 2 \cdot |u_i| \cdot |u_j| \cdot |X_{ij} - X'_{ij}| \\
&\le 4 \cdot |u_i| \cdot |u_j|
\end{aligned}
$$

Largest eigenvalue of a random matrix

$$\begin{aligned} Z - Z_{ij} &= u^{\mathrm{T}} A u - \max_{\|w\|=1} w^{\mathrm{T}} \bar{A}^{ij} w \\ &\leq u^{\mathrm{T}} A u - u^{\mathrm{T}} \bar{A}^{ij} u \\ &\leq 2 \cdot |u_i| \cdot |u_j| \cdot |X_{ij} - X'_{ij}| \\ &\leq 4 \cdot |u_i| \cdot |u_j| \end{aligned}$$

$$\sum_{ij} (Z - Z_{ij})_+^2 \leq 16 \sum_{ij} u_i^2 u_j^2 = 16 \cdot \|u\|^2 \cdot \|u\|^2 = 16.$$

Largest eigenvalue of a random matrix

$$
\begin{aligned}
Z - Z_{ij} &= u^{\mathrm{T}} A u - \max_{\|w\|=1} w^{\mathrm{T}} \bar{A}^{ij} w \\
&\leq u^{\mathrm{T}} A u - u^{\mathrm{T}} \bar{A}^{ij} u \\
&\leq 2 \cdot |u_i| \cdot |u_j| \cdot |X_{ij} - X'_{ij}| \\
&\leq 4 \cdot |u_i| \cdot |u_j|
\end{aligned}
$$

$$
\sum_{ij} (Z - Z_{ij})_+^2 \leq 16 \sum_{ij} u_i^2 u_j^2 = 16 \cdot \|u\|^2 \cdot \|u\|^2 = 16.
$$

Thus, $\mathrm{Var}(Z) \leq 16$.

Suboptimality warning

Suboptimality warning

In fact, if X_{ij}'s are i.i.d. equiprobable on $\{-1, +1\}$, then

Suboptimality warning

In fact, if X_{ij}'s are i.i.d. equiprobable on $\{-1,+1\}$, then

$\mathrm{Var}(\lambda_{\max}(A)) = \Theta(n^{-1/3})$, i.e. *superconcentration*.

Suboptimality warning

In fact, if X_{ij}'s are i.i.d. equiprobable on $\{-1, +1\}$, then

$\mathrm{Var}(\lambda_{\max}(A)) = \Theta(n^{-1/3})$, i.e. *superconcentration*.

- We don't get the optimal bound by a general theorem

Suboptimality warning

In fact, if X_{ij}'s are i.i.d. equiprobable on $\{-1, +1\}$, then

$\mathrm{Var}(\lambda_{\max}(A)) = \Theta(n^{-1/3})$, i.e. *superconcentration*.

- We don't get the optimal bound by a general theorem
- But ...

Suboptimality warning

In fact, if X_{ij}'s are i.i.d. equiprobable on $\{-1, +1\}$, then

$\mathrm{Var}(\lambda_{\max}(A)) = \Theta(n^{-1/3})$, i.e. *superconcentration*.

- We don't get the optimal bound by a general theorem
- But ...
  - we didn't employ any random matrix theory

Suboptimality warning

In fact, if X_{ij}'s are i.i.d. equiprobable on $\{-1, +1\}$, then

$\mathrm{Var}(\lambda_{\max}(A)) = \Theta(n^{-1/3})$, i.e. *superconcentration*.

- We don't get the optimal bound by a general theorem
- But ...
  - we didn't employ any random matrix theory
  - we didn't carry out any detailed analysis

Suboptimality warning

In fact, if X_{ij}'s are i.i.d. equiprobable on $\{-1,+1\}$, then

$\mathrm{Var}(\lambda_{\max}(A)) = \Theta(n^{-1/3})$, i.e. *superconcentration*.

- We don't get the optimal bound by a general theorem
- But ...
  - we didn't employ any random matrix theory
  - we didn't carry out any detailed analysis
- Still, we obtained a genuinely non-trivial result

Suboptimality warning

In fact, if X_{ij}'s are i.i.d. equiprobable on $\{-1, +1\}$, then

$$\mathrm{Var}(\lambda_{\max}(A)) = \Theta(n^{-1/3}), \text{ i.e. } \textit{superconcentration}.$$

- We don't get the optimal bound by a general theorem
- But ...
  - we didn't employ any random matrix theory
  - we didn't carry out any detailed analysis
- Still, we obtained a genuinely non-trivial result
- In many cases, these are sufficient since they provide bounds of optimal order

Suboptimality warning

In fact, if X_{ij}'s are i.i.d. equiprobable on $\{-1, +1\}$, then

$$\mathrm{Var}(\lambda_{\max}(A)) = \Theta(n^{-1/3}),\text{ i.e. } \textit{superconcentration}.$$

- We don't get the optimal bound by a general theorem
- But ...
  - we didn't employ any random matrix theory
  - we didn't carry out any detailed analysis
- Still, we obtained a genuinely non-trivial result
- In many cases, these are sufficient since they provide bounds of optimal order
- E.g. here, $\mathrm{Var}(\lambda_{\max}(A))$ can be $1/4$ if entries are not identically distributed

Suboptimality warning

In fact, if X_{ij}'s are i.i.d. equiprobable on $\{-1, +1\}$, then

$\mathrm{Var}(\lambda_{\max}(A)) = \Theta(n^{-1/3})$, i.e. *superconcentration*.

- We don't get the optimal bound by a general theorem
- But ...
  - we didn't employ any random matrix theory
  - we didn't carry out any detailed analysis
- Still, we obtained a genuinely non-trivial result
- In many cases, these are sufficient since they provide bounds of optimal order
- E.g. here, $\mathrm{Var}(\lambda_{\max}(A))$ can be $1/4$ if entries are not identically distributed
- Can a general principle capture superconcentration? Active research area

Application: Gaussian Poincaré inequality

Application: Gaussian Poincaré inequality

- A Poincaré inequality is of the form

$$\text{“variance}(f) \lesssim \; c\, \mathbb{E}[\|\text{gradient}(f)\|^2]\text{”}$$

for a suitable notion of gradient.

Application: Gaussian Poincaré inequality

- A Poincaré inequality is of the form

$$\text{“variance}(f) \lesssim \ c\, \mathbb{E}[\|\text{gradient}(f)\|^2]\text{”}$$

  for a suitable notion of gradient.
- Such inequalities are closely associated with mixing in Markov processes

Application: Gaussian Poincaré inequality

- A Poincaré inequality is of the form

$$\text{“variance}(f) \lesssim c\, \mathbb{E}[\|\text{gradient}(f)\|^2]\text{”}$$

for a suitable notion of gradient.
- Such inequalities are closely associated with mixing in Markov processes

Gaussian Poincaré inequality

If $X \sim \mathcal{N}(0, I_n)$, and $f : \mathbb{R}^n \mapsto \mathbb{R}$ is continuously differentiable, then

Application: Gaussian Poincaré inequality

- A Poincaré inequality is of the form

$$\text{``variance}(f) \lesssim \; c\,\mathbb{E}[\|\text{gradient}(f)\|^2]\text{''}$$

for a suitable notion of gradient.
- Such inequalities are closely associated with mixing in Markov processes

Gaussian Poincaré inequality

If $X \sim \mathcal{N}(0, I_n)$, and $f : \mathbb{R}^n \mapsto \mathbb{R}$ is continuously differentiable, then

$$\mathrm{Var}(f(X)) \leq \mathbb{E}\left[\|\nabla f(X)\|^2\right]$$

Application: Gaussian Poincaré inequality

- A Poincaré inequality is of the form

$$\text{“variance}(f) \lesssim\ c\, \mathbb{E}[\|\text{gradient}(f)\|^2]\text{”}$$

  for a suitable notion of gradient.
- Such inequalities are closely associated with mixing in Markov processes

Gaussian Poincaré inequality

If $X \sim \mathcal{N}(0, I_n)$, and $f : \mathbb{R}^n \mapsto \mathbb{R}$ is continuously differentiable, then

$$\mathrm{Var}(f(X)) \leq \mathbb{E}\left[\|\nabla f(X)\|^2\right]$$

Note: Tight if f is linear!

Gaussian Poincaré: proof

Gaussian Poincaré: proof

First, consider a 1-dimensional Gaussian $X \sim \mathscr{N}(0, 1)$.

Gaussian Poincaré: proof

First, consider a 1-dimensional Gaussian $X \sim \mathcal{N}(0, 1)$.

Let $Y_1, Y_2, \ldots, Y_m$ be independent and equiprobable on $\{-1, +1\}$.

Gaussian Poincaré: proof

First, consider a 1-dimensional Gaussian $X \sim \mathcal{N}(0,1)$.

Let $Y_1, Y_2, \ldots, Y_m$ be independent and equiprobable on $\{-1,+1\}$.

Let $S_m = \frac{1}{\sqrt{m}}\left(Y_1 + Y_2 + \ldots + Y_m\right).$

Gaussian Poincaré: proof

First, consider a 1-dimensional Gaussian $X \sim \mathcal{N}(0,1)$.

Let $Y_1, Y_2, \ldots, Y_m$ be independent and equiprobable on $\{-1,+1\}$.

$$\text{Let } S_m = \frac{1}{\sqrt{m}}\left(Y_1 + Y_2 + \ldots + Y_m\right).$$

Let $f : \mathbb{R} \mapsto \mathbb{R}$ be twice continuously differentiable with compact support. Let $\sup_x |f''(x)| = K$.

Gaussian Poincaré: proof

First, consider a 1-dimensional Gaussian $X \sim \mathcal{N}(0,1)$.

Let $Y_1, Y_2, \ldots, Y_m$ be independent and equiprobable on $\{-1,+1\}$.

$$\text{Let } S_m = \frac{1}{\sqrt{m}}\left(Y_1 + Y_2 + \ldots + Y_m\right).$$

Let $f : \mathbb{R} \mapsto \mathbb{R}$ be twice continuously differentiable with compact support. Let $\sup_x |f''(x)| = K$.

$\mathrm{Var}^{(i)}(f(S_m))$

Gaussian Poincaré: proof

First, consider a 1-dimensional Gaussian $X \sim \mathcal{N}(0,1)$.

Let $Y_1, Y_2, \ldots, Y_m$ be independent and equiprobable on $\{-1, +1\}$.

$$\text{Let } S_m = \frac{1}{\sqrt{m}}\left(Y_1 + Y_2 + \ldots + Y_m\right).$$

Let $f : \mathbb{R} \mapsto \mathbb{R}$ be twice continuously differentiable with compact support. Let $\sup_x |f''(x)| = K$.

$$\mathrm{Var}^{(i)}(f(S_m)) = \frac{1}{4}\left(f\left(S_m - \frac{Y_i}{\sqrt{m}} + \frac{1}{\sqrt{m}}\right) - f\left(S_m - \frac{Y_i}{\sqrt{m}} - \frac{1}{\sqrt{m}}\right)\right)^2$$

Gaussian Poincaré: proof

First, consider a 1-dimensional Gaussian $X \sim \mathcal{N}(0,1)$.

Let $Y_1, Y_2, \ldots, Y_m$ be independent and equiprobable on $\{-1, +1\}$.

Let $S_m = \frac{1}{\sqrt{m}}\left(Y_1 + Y_2 + \ldots + Y_m\right)$.

Let $f : \mathbb{R} \mapsto \mathbb{R}$ be twice continuously differentiable with compact support. Let $\sup_x |f''(x)| = K$.

$$\begin{aligned}\mathrm{Var}^{(i)}(f(S_m)) &= \frac{1}{4}\left(f\left(S_m - \frac{Y_i}{\sqrt{m}} + \frac{1}{\sqrt{m}}\right) - f\left(S_m - \frac{Y_i}{\sqrt{m}} - \frac{1}{\sqrt{m}}\right)\right)^2 \\ &\leq \frac{1}{4}\left(\frac{2}{\sqrt{m}}|f'(S_m)| + \frac{2K}{m}\right)^2\end{aligned}$$

Gaussian Poincaré: proof

First, consider a 1-dimensional Gaussian $X \sim \mathcal{N}(0,1)$.

Let $Y_1, Y_2, \ldots, Y_m$ be independent and equiprobable on $\{-1,+1\}$.

Let $S_m = \frac{1}{\sqrt{m}}\left(Y_1 + Y_2 + \ldots + Y_m\right)$.

Let $f : \mathbb{R} \mapsto \mathbb{R}$ be twice continuously differentiable with compact support. Let $\sup_x |f''(x)| = K$.

$$\begin{aligned}\mathrm{Var}^{(i)}(f(S_m)) &= \frac{1}{4}\left(f\left(S_m - \frac{Y_i}{\sqrt{m}} + \frac{1}{\sqrt{m}}\right) - f\left(S_m - \frac{Y_i}{\sqrt{m}} - \frac{1}{\sqrt{m}}\right)\right)^2 \\ &\le \frac{1}{4}\left(\frac{2}{\sqrt{m}}|f'(S_m)| + \frac{2K}{m}\right)^2 = \frac{1}{m}\left(|f'(S_m)| + \frac{K}{\sqrt{m}}\right)^2\end{aligned}$$

Gaussian Poincaré: proof

Gaussian Poincaré: proof

$$\mathrm{Var}(f(S_m)) \le \sum_{i=1}^{m} \mathbb{E}\,\mathrm{Var}^{(i)}(f(S_m))$$

Gaussian Poincaré: proof

$$\mathrm{Var}(f(S_m)) \leq \sum_{i=1}^{m} \mathbb{E}\,\mathrm{Var}^{(i)}(f(S_m)) \leq \mathbb{E}\left[\left(|f'(S_m)| + \frac{K}{\sqrt{m}}\right)^2\right]$$

Gaussian Poincaré: proof

$$\mathrm{Var}(f(S_m)) \le \sum_{i=1}^{m} \mathbb{E}\,\mathrm{Var}^{(i)}(f(S_m)) \le \mathbb{E}\left[\left(|f'(S_m)| + \frac{K}{\sqrt{m}}\right)^2\right]$$

As $m \to \infty$, we have $S_m \to X \sim \mathcal{N}(0,1)$ in distribution by the Central Limit Theorem.

Gaussian Poincaré: proof

$$\mathrm{Var}(f(S_m)) \le \sum_{i=1}^{m} \mathbb{E}\,\mathrm{Var}^{(i)}(f(S_m)) \le \mathbb{E}\left[\left(|f'(S_m)| + \frac{K}{\sqrt{m}}\right)^2\right]$$

As $m \to \infty$, we have $S_m \to X \sim \mathscr{N}(0,1)$ in distribution by the Central Limit Theorem.

Since f and f' are continuous and bounded, we get $\mathrm{Var}(f(X)) \le \mathbb{E}\left[f'(X)^2\right]$

Gaussian Poincaré: proof

$$\mathrm{Var}(f(S_m)) \leq \sum_{i=1}^{m} \mathbb{E}\,\mathrm{Var}^{(i)}(f(S_m)) \leq \mathbb{E}\left[\left(|f'(S_m)| + \frac{K}{\sqrt{m}}\right)^2\right]$$

As $m \to \infty$, we have $S_m \to X \sim \mathcal{N}(0,1)$ in distribution by the Central Limit Theorem.

Since f and f' are continuous and bounded, we get

$$\mathrm{Var}(f(X)) \leq \mathbb{E}\left[f'(X)^2\right]$$

Extend to all continuously differentiable functions by

Gaussian Poincaré: proof

$$\mathrm{Var}(f(S_m)) \leq \sum_{i=1}^{m} \mathbb{E}\,\mathrm{Var}^{(i)}(f(S_m)) \leq \mathbb{E}\left[\left(|f'(S_m)| + \frac{K}{\sqrt{m}}\right)^2\right]$$

As $m \to \infty$, we have $S_m \to X \sim \mathcal{N}(0,1)$ in distribution by the Central Limit Theorem.

Since f and f' are continuous and bounded, we get $\mathrm{Var}(f(X)) \leq \mathbb{E}\left[f'(X)^2\right]$

Extend to all continuously differentiable functions by

- Truncation of f to $[-M, M]$ and apply dominated convergence theorem as $M \to \infty$

Gaussian Poincaré: proof

$$\mathrm{Var}(f(S_m)) \leq \sum_{i=1}^{m} \mathbb{E}\,\mathrm{Var}^{(i)}(f(S_m)) \leq \mathbb{E}\left[\left(|f'(S_m)| + \frac{K}{\sqrt{m}}\right)^2\right]$$

As $m \to \infty$, we have $S_m \to X \sim \mathcal{N}(0,1)$ in distribution by the Central Limit Theorem.

Since f and f' are continuous and bounded, we get
$$\mathrm{Var}(f(X)) \leq \mathbb{E}\left[f'(X)^2\right]$$

Extend to all continuously differentiable functions by

- Truncation of f to $[-M, M]$ and apply dominated convergence theorem as $M \to \infty$
- Smoothen truncated f by convolution with a sharply concentrated twice differentiable kernel with compact support

Gaussian Poincaré: proof

Gaussian Poincaré: proof

Now, if $X \sim \mathscr{N}(0, I_n)$ is an n-dimensional Gaussian vector

Gaussian Poincaré: proof

Now, if $X \sim \mathscr{N}(0, I_n)$ is an n-dimensional Gaussian vector

and $f : \mathbb{R}^n \mapsto \mathbb{R}$ is continuously differentiable,

Gaussian Poincaré: proof

Now, if $X \sim \mathscr{N}(0, I_n)$ is an n-dimensional Gaussian vector and $f : \mathbb{R}^n \mapsto \mathbb{R}$ is continuously differentiable, use tensorization of variance again.

Gaussian Poincaré: proof

Now, if $X \sim \mathcal{N}(0, I_n)$ is an n-dimensional Gaussian vector

and $f : \mathbb{R}^n \mapsto \mathbb{R}$ is continuously differentiable,

use tensorization of variance again.

$$\mathrm{Var}(f(X)) \leq \sum_{i=1}^{n} \mathbb{E}\left[\mathrm{Var}^{(i)}(f(X))\right]$$

Gaussian Poincaré: proof

Now, if $X \sim \mathscr{N}(0, I_n)$ is an n-dimensional Gaussian vector

and $f : \mathbb{R}^n \mapsto \mathbb{R}$ is continuously differentiable,

use tensorization of variance again.

$$\begin{aligned}
\mathrm{Var}(f(X)) &\leq \sum_{i=1}^{n} \mathbb{E}\left[\mathrm{Var}^{(i)}(f(X))\right] \\
&\leq \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}^{(i)}\left|\frac{\partial f}{\partial x_i}(X)\right|^2\right]
\end{aligned}$$

Gaussian Poincaré: proof

Now, if $X \sim \mathcal{N}(0, I_n)$ is an n-dimensional Gaussian vector

and $f : \mathbb{R}^n \mapsto \mathbb{R}$ is continuously differentiable,

use tensorization of variance again.

$$\begin{aligned} \mathrm{Var}(f(X)) &\leq \sum_{i=1}^{n} \mathbb{E}\left[\mathrm{Var}^{(i)}(f(X))\right] \\ &\leq \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}^{(i)}\left|\frac{\partial f}{\partial x_i}(X)\right|^2\right] \\ &= \mathbb{E}\left[\|\nabla f(X)\|^2\right] \end{aligned}$$

Revisiting trivial example

Let $Z = X_1 + X_2 + \ldots + X_n$ where $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) with finite variance. Then,

$$\mathbb{E}Z = n\mathbb{E}X_1 \qquad \mathrm{Var}(Z) = n\,\mathrm{Var}(X_1)$$

Revisiting trivial example

Let $Z = X_1 + X_2 + \ldots + X_n$ where $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) with finite variance. Then,

$$\mathbb{E}Z = n\mathbb{E}X_1 \qquad \mathrm{Var}(Z) = n\,\mathrm{Var}(X_1)$$

Mean $= \Theta(n)$, Standard Deviation $= O(\sqrt{n})$.

in-

ari-

Whatever spooky thing it is that makes 'Law of Large Numbers'-type bounds on variance work for general functions of independent random variables ...

in-

ari-

Whatever spooky thing it is that makes 'Law of Large Numbers'-type bounds on variance work for general functions of independent random variables ...

Whatever spooky thing it is that makes 'Law of Large Numbers'-type bounds on variance work for general functions of independent random variables ...

... who can say if that same apparition may not make the 'Central Limit Theorem'-type bounds work for general functions as well?

in-
ari-

Revisiting trivial example

Let $Z = X_1 + X_2 + \ldots + X_n$ where $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) with finite variance. Then,

$$\mathbb{E}Z = n\mathbb{E}X_1 \qquad \mathrm{Var}(Z) = n\,\mathrm{Var}(X_1)$$

Mean $= \Theta(n)$, Standard Deviation $= O(\sqrt{n})$.

Revisiting trivial example

Let $Z = X_1 + X_2 + \ldots + X_n$ where $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) with finite variance. Then,

$$\mathbb{E}Z = n\mathbb{E}X_1 \qquad \mathrm{Var}(Z) = n\,\mathrm{Var}(X_1)$$

Mean $= \Theta(n)$, Standard Deviation $= O(\sqrt{n})$.

But in fact, more: $\dfrac{Z - \mathbb{E}Z}{\sqrt{n}} \approx \mathscr{N}\left(0, \mathrm{Var}(X_1)\right)$

Revisiting trivial example

Let $Z = X_1 + X_2 + \ldots + X_n$ where $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) with finite variance. Then,

$$\mathbb{E}Z = n\mathbb{E}X_1 \qquad \mathrm{Var}(Z) = n\,\mathrm{Var}(X_1)$$

Mean $= \Theta(n)$, Standard Deviation $= O(\sqrt{n})$.

But in fact, more: $\dfrac{Z - \mathbb{E}Z}{\sqrt{n}} \approx \mathscr{N}\left(0, \mathrm{Var}(X_1)\right)$

So, $\mathbb{P}[Z - \mathbb{E}Z \geq t] \lesssim \exp\left(-\dfrac{t^2}{2n\,\mathrm{Var}(X_1)}\right)$

for $t = O(\text{typical deviation})$

Revisiting trivial example

Let $Z = X_1 + X_2 + \ldots + X_n$ where $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) with finite variance. Then,

$$\mathbb{E}Z = n\mathbb{E}X_1 \qquad \mathrm{Var}(Z) = n\,\mathrm{Var}(X_1)$$

Mean $= \Theta(n)$, Standard Deviation $= O(\sqrt{n})$.

But in fact, more: $\dfrac{Z - \mathbb{E}Z}{\sqrt{n}} \approx \mathcal{N}\left(0, \mathrm{Var}(X_1)\right)$

So, $\mathbb{P}[Z - \mathbb{E}Z \geq t] \lesssim \exp\left(-\dfrac{t^2}{2n\,\mathrm{Var}(X_1)}\right)$

for $t = O(\text{typical deviation})$

Such a sub-Gaussian tail inequality is also a manifestion of a general phenomenon that holds for a large family of functions.

How can we prove sub-Gaussian tail bounds?

How can we prove sub-Gaussian tail bounds?

The Chernoff bound

How can we prove sub-Gaussian tail bounds?

The Chernoff bound: for $\lambda > 0$,

$$\mathbb{P}[Z \geq t] = \mathbb{P}[e^{\lambda Z} \geq e^{\lambda t}] \leq \frac{\mathbb{E}e^{\lambda Z}}{e^{\lambda t}}$$

How can we prove sub-Gaussian tail bounds?

The Chernoff bound: for $\lambda > 0$,

$$\mathbb{P}[Z \geq t] = \mathbb{P}[e^{\lambda Z} \geq e^{\lambda t}] \leq \frac{\mathbb{E}e^{\lambda Z}}{e^{\lambda t}} = e^{-(\lambda t - \psi_Z(\lambda))}$$

How can we prove sub-Gaussian tail bounds?

The Chernoff bound: for $\lambda > 0$,

$$\mathbb{P}[Z \geq t] = \mathbb{P}[e^{\lambda Z} \geq e^{\lambda t}] \leq \frac{\mathbb{E}e^{\lambda Z}}{e^{\lambda t}} = e^{-(\lambda t - \psi_Z(\lambda))}$$

Log Moment Generating Functions (log m.g.f.)

For any random variable Z with $\mathbb{E}Z = 0$, define

$$\psi_Z(\lambda) = \psi(\lambda) := \log \mathbb{E}e^{\lambda Z}, \quad \lambda \in \mathbb{R}$$

How can we prove sub-Gaussian tail bounds?

The Chernoff bound: for $\lambda > 0$,

$$\mathbb{P}[Z \geq t] = \mathbb{P}[e^{\lambda Z} \geq e^{\lambda t}] \leq \frac{\mathbb{E}e^{\lambda Z}}{e^{\lambda t}} = e^{-(\lambda t - \psi_Z(\lambda))}$$

Log Moment Generating Functions (log m.g.f.)

For any random variable Z with $\mathbb{E}Z = 0$, define

$$\psi_Z(\lambda) = \psi(\lambda) := \log \mathbb{E}e^{\lambda Z}, \quad \lambda \in \mathbb{R}$$

Properties:

How can we prove sub-Gaussian tail bounds?

The Chernoff bound: for $\lambda > 0$,

$$\mathbb{P}[Z \geq t] = \mathbb{P}[e^{\lambda Z} \geq e^{\lambda t}] \leq \frac{\mathbb{E}e^{\lambda Z}}{e^{\lambda t}} = e^{-(\lambda t - \psi_Z(\lambda))}$$

Log Moment Generating Functions (log m.g.f.)

For any random variable Z with $\mathbb{E}Z = 0$, define

$$\psi_Z(\lambda) = \psi(\lambda) := \log \mathbb{E}e^{\lambda Z}, \quad \lambda \in \mathbb{R}$$

Properties:

- $\psi(0) = 0$

How can we prove sub-Gaussian tail bounds?

The Chernoff bound: for $\lambda > 0$,

$$\mathbb{P}[Z \geq t] = \mathbb{P}[e^{\lambda Z} \geq e^{\lambda t}] \leq \frac{\mathbb{E}e^{\lambda Z}}{e^{\lambda t}} = e^{-(\lambda t - \psi_Z(\lambda))}$$

Log Moment Generating Functions (log m.g.f.)

For any random variable Z with $\mathbb{E}Z = 0$, define

$$\psi_Z(\lambda) = \psi(\lambda) := \log \mathbb{E}e^{\lambda Z}, \quad \lambda \in \mathbb{R}$$

Properties:

- $\psi(0) = 0$
- $\psi(\lambda) = \log \mathbb{E}e^{\lambda Z} \geq \log e^{\lambda \mathbb{E}Z} = 0$ (Jensen)

How can we prove sub-Gaussian tail bounds?

The Chernoff bound: for $\lambda > 0$,

$$\mathbb{P}[Z \geq t] = \mathbb{P}[e^{\lambda Z} \geq e^{\lambda t}] \leq \frac{\mathbb{E}e^{\lambda Z}}{e^{\lambda t}} = e^{-(\lambda t - \psi_Z(\lambda))}$$

Log Moment Generating Functions (log m.g.f.)

For any random variable Z with $\mathbb{E}Z = 0$, define

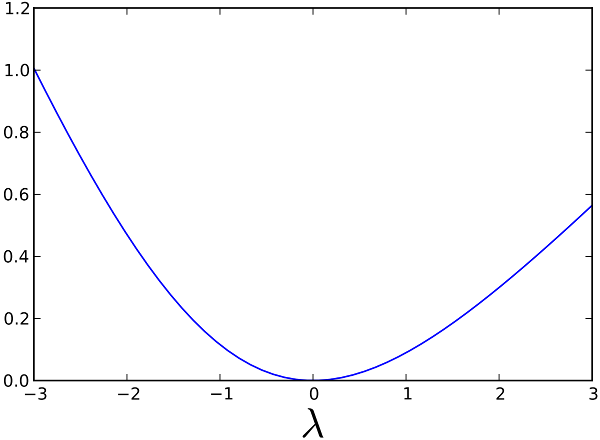
$$\psi_Z(\lambda) = \psi(\lambda) := \log \mathbb{E}e^{\lambda Z}, \quad \lambda \in \mathbb{R}$$

Properties:

- $\psi(0) = 0$
- $\psi(\lambda) = \log \mathbb{E}e^{\lambda Z} \geq \log e^{\lambda \mathbb{E}Z} = 0$ (Jensen)
- Whenever defined over (a, b), it can be differentiated under expectation infinitely many times

How can we prove sub-Gaussian tail bounds?

The Chernoff bound: for $\lambda > 0$,



under expectation infinitely many times

How can we prove sub-Gaussian tail bounds?

The Chernoff bound: for $\lambda > 0$,

$$\mathbb{P}[Z \geq t] = \mathbb{P}[e^{\lambda Z} \geq e^{\lambda t}] \leq \frac{\mathbb{E}e^{\lambda Z}}{e^{\lambda t}} = e^{-(\lambda t - \psi_Z(\lambda))}$$

Log Moment Generating Functions (log m.g.f.)

For any random variable Z with $\mathbb{E}Z = 0$, define

$$\psi_Z(\lambda) = \psi(\lambda) := \log \mathbb{E}e^{\lambda Z}, \quad \lambda \in \mathbb{R}$$

Properties:

- $\psi(0) = 0$
- $\psi(\lambda) = \log \mathbb{E}e^{\lambda Z} \geq \log e^{\lambda \mathbb{E}Z} = 0$ (Jensen)
- Whenever defined over (a, b), it can be differentiated under expectation infinitely many times

Example

If $Z \sim \mathscr{N}(0, \sigma^2)$, then $\psi_Z(\lambda) = \log \mathbb{E} e^{\lambda Z} =$

Example

If $Z \sim \mathcal{N}(0, \sigma^2)$, then $\psi_Z(\lambda) = \log \mathbb{E} e^{\lambda Z} = \dfrac{\lambda^2 \sigma^2}{2}$

Example

$$\text{If } Z \sim \mathscr{N}(0, \sigma^2), \text{ then } \psi_Z(\lambda) = \log \mathbb{E} e^{\lambda Z} = \frac{\lambda^2 \sigma^2}{2}$$

$$\mathbb{P}[Z \geq t] \leq e^{-\max_{\lambda > 0}(\lambda t - \lambda^2 \sigma^2 / 2)}$$

Example

$$\text{If } Z \sim \mathcal{N}(0, \sigma^2), \text{ then } \psi_Z(\lambda) = \log \mathbb{E} e^{\lambda Z} = \frac{\lambda^2 \sigma^2}{2}$$

$$\mathbb{P}[Z \geq t] \leq e^{-\max_{\lambda > 0}(\lambda t - \lambda^2 \sigma^2 / 2)} = e^{-t^2/2\sigma^2}, \ \forall t > 0$$

Example

If $Z \sim \mathcal{N}(0, \sigma^2)$, then $\psi_Z(\lambda) = \log \mathbb{E} e^{\lambda Z} = \dfrac{\lambda^2 \sigma^2}{2}$

$$\mathbb{P}[Z \geq t] \leq e^{-\max_{\lambda > 0}(\lambda t - \lambda^2 \sigma^2 / 2)} = e^{-t^2/2\sigma^2}, \ \forall t > 0$$

Sub-Gaussianity

Define Y as σ^2-sub-Gaussian if

$$\log \mathbb{E} e^{\lambda(Y - \mathbb{E}Y)} \leq \frac{\lambda^2 \sigma^2}{2}, \ \ \forall \ \ \lambda \in \mathbb{R}$$

Example

If $Z \sim \mathcal{N}(0, \sigma^2)$, then $\psi_Z(\lambda) = \log \mathbb{E} e^{\lambda Z} = \dfrac{\lambda^2 \sigma^2}{2}$

$$\mathbb{P}[Z \geq t] \leq e^{-\max_{\lambda > 0}(\lambda t - \lambda^2 \sigma^2 / 2)} = e^{-t^2/2\sigma^2}, \ \forall t > 0$$

Sub-Gaussianity

Define Y as σ^2-sub-Gaussian if

$$\log \mathbb{E} e^{\lambda(Y - \mathbb{E}Y)} \leq \frac{\lambda^2 \sigma^2}{2}, \ \ \forall \ \ \lambda \in \mathbb{R}$$

Then, $\mathbb{P}[Y - \mathbb{E}Y \geq t] \leq e^{-\frac{t^2}{2\sigma^2}}, \ \forall t > 0$

Example

If $Z \sim \mathcal{N}(0, \sigma^2)$, then $\psi_Z(\lambda) = \log \mathbb{E} e^{\lambda Z} = \dfrac{\lambda^2 \sigma^2}{2}$

$$\mathbb{P}[Z \geq t] \leq e^{-\max_{\lambda>0}(\lambda t - \lambda^2\sigma^2/2)} = e^{-t^2/2\sigma^2}, \ \forall t > 0$$

Sub-Gaussianity

Define Y as σ^2-sub-Gaussian if

$$\log \mathbb{E} e^{\lambda(Y - \mathbb{E}Y)} \leq \frac{\lambda^2\sigma^2}{2}, \ \ \forall \ \ \lambda \in \mathbb{R}$$

Then, $\mathbb{P}[Y - \mathbb{E}Y \geq t] \leq e^{-\frac{t^2}{2\sigma^2}}, \ \forall t > 0$

$$\mathbb{P}\left[|Y - \mathbb{E}Y| \geq 10\sqrt{\mathrm{Var}(Y)}\right] \leq 1\%$$ (Chebyshev)

Example

If $Z \sim \mathcal{N}(0, \sigma^2)$, then $\psi_Z(\lambda) = \log \mathbb{E} e^{\lambda Z} = \dfrac{\lambda^2 \sigma^2}{2}$

$$\mathbb{P}[Z \geq t] \leq e^{-\max_{\lambda > 0}(\lambda t - \lambda^2 \sigma^2 / 2)} = e^{-t^2/2\sigma^2}, \ \forall t > 0$$

Sub-Gaussianity

Define Y as σ^2-sub-Gaussian if

$$\log \mathbb{E} e^{\lambda(Y - \mathbb{E}Y)} \leq \frac{\lambda^2 \sigma^2}{2}, \quad \forall \quad \lambda \in \mathbb{R}$$

Then, $\mathbb{P}[Y - \mathbb{E}Y \geq t] \leq e^{-\frac{t^2}{2\sigma^2}}, \ \forall t > 0$

$\mathbb{P}\left[|Y - \mathbb{E}Y| \geq 10\sqrt{\mathrm{Var}(Y)}\right] \leq 1\%$ (Chebyshev)

$\mathbb{P}[|Y - \mathbb{E}Y| \geq 10\sigma] \leq$ (Sub-Gaussianity)

Example

If $Z \sim \mathcal{N}(0, \sigma^2)$, then $\psi_Z(\lambda) = \log \mathbb{E} e^{\lambda Z} = \frac{\lambda^2 \sigma^2}{2}$

$$\mathbb{P}[Z \geq t] \leq e^{-\max_{\lambda > 0}(\lambda t - \lambda^2 \sigma^2 / 2)} = e^{-t^2/2\sigma^2}, \ \forall t > 0$$

Sub-Gaussianity

Define Y as σ^2-sub-Gaussian if

$$\log \mathbb{E} e^{\lambda(Y - \mathbb{E}Y)} \leq \frac{\lambda^2 \sigma^2}{2}, \ \ \forall \ \ \lambda \in \mathbb{R}$$

Then, $\mathbb{P}[Y - \mathbb{E}Y \geq t] \leq e^{-\frac{t^2}{2\sigma^2}}, \ \forall t > 0$

$$\mathbb{P}\left[|Y - \mathbb{E}Y| \geq 10\sqrt{\mathrm{Var}(Y)}\right] \leq 1\% \text{ (Chebyshev)}$$

$$\mathbb{P}[|Y - \mathbb{E}Y| \geq 10\sigma] \leq 3.86 \times 10^{-22} \text{ (Sub-Gaussianity)}$$

Equivalent definitions of sub-Gaussianity

Suppose $\mathbb{E}Z = 0$. Each statement below implies the next:

Equivalent definitions of sub-Gaussianity

Suppose $\mathbb{E}Z = 0$. Each statement below implies the next:

$$\log \mathbb{E}e^{\lambda Z} \leq \frac{\lambda^2\sigma^2}{2}, \quad \forall\ \lambda \in \mathbb{R}$$

Equivalent definitions of sub-Gaussianity

Suppose $\mathbb{E}Z = 0$. Each statement below implies the next:

$$\log \mathbb{E} e^{\lambda Z} \leq \frac{\lambda^2 \sigma^2}{2}, \quad \forall\ \lambda \in \mathbb{R}$$

$$\mathbb{P}[Z \geq t], \mathbb{P}[Z \leq -t] \leq e^{-\frac{t^2}{2\sigma^2}}, \quad \forall\ t > 0$$

Equivalent definitions of sub-Gaussianity

Suppose $\mathbb{E}Z = 0$. Each statement below implies the next:

$$\log \mathbb{E} e^{\lambda Z} \leq \frac{\lambda^2 \sigma^2}{2}, \quad \forall\, \lambda \in \mathbb{R}$$

$$\mathbb{P}[Z \geq t], \mathbb{P}[Z \leq -t] \leq e^{-\frac{t^2}{2\sigma^2}}, \quad \forall\, t > 0$$

$$\mathbb{E}[Z^{2q}] \leq q!\,(4\sigma^2)^q, \quad q = 1, 2, 3, \ldots$$

Equivalent definitions of sub-Gaussianity

Suppose $\mathbb{E}Z = 0$. Each statement below implies the next:

$$\log \mathbb{E} e^{\lambda Z} \leq \frac{\lambda^2 \sigma^2}{2}, \quad \forall\, \lambda \in \mathbb{R}$$

$$\mathbb{P}[Z \geq t], \mathbb{P}[Z \leq -t] \leq e^{-\frac{t^2}{2\sigma^2}}, \quad \forall\, t > 0$$

$$\mathbb{E}[Z^{2q}] \leq q!\,(4\sigma^2)^q, \quad q = 1, 2, 3, \ldots$$

$$\mathbb{E}\left[e^{\frac{Z^2}{8\sigma^2}}\right] \leq 2$$

Equivalent definitions of sub-Gaussianity

Suppose $\mathbb{E}Z = 0$. Each statement below implies the next:

$$\log \mathbb{E}e^{\lambda Z} \leq \frac{\lambda^2\sigma^2}{2}, \;\; \forall\, \lambda \in \mathbb{R}$$

$$\mathbb{P}[Z \geq t], \mathbb{P}[Z \leq -t] \leq e^{-\frac{t^2}{2\sigma^2}}, \;\; \forall\, t > 0$$

$$\mathbb{E}[Z^{2q}] \leq q!\,(4\sigma^2)^q, \;\;\; q = 1, 2, 3, \ldots$$

$$\mathbb{E}\left[e^{\frac{Z^2}{8\sigma^2}}\right] \leq 2$$

$$\log \mathbb{E}e^{\lambda Z} \leq \frac{\lambda^2(24\sigma^2)}{2}, \;\; \forall\, \lambda \in \mathbb{R}$$

Tensorization

Tensorization

- Idea: prove a bound on a function of many independent variables by proving bounds on functions of one variable

Tensorization

- Idea: prove a bound on a function of many independent variables by proving bounds on functions of one variable
- Unlike variance, sub-Gaussianity does not tensorize naturally

Tensorization

- Idea: prove a bound on a function of many independent variables by proving bounds on functions of one variable
- Unlike variance, sub-Gaussianity does not tensorize naturally
- Need to develop further techniques

Tensorization

- Idea: prove a bound on a function of many independent variables by proving bounds on functions of one variable
- Unlike variance, sub-Gaussianity does not tensorize naturally
- Need to develop further techniques
- Two such general methodologies are:

Tensorization

- Idea: prove a bound on a function of many independent variables by proving bounds on functions of one variable
- Unlike variance, sub-Gaussianity does not tensorize naturally
- Need to develop further techniques
- Two such general methodologies are:
  - *Entropy method*: Ledoux (1996), Massart (1998), Lugosi et al. (1999, 2001)

Tensorization

- Idea: prove a bound on a function of many independent variables by proving bounds on functions of one variable
- Unlike variance, sub-Gaussianity does not tensorize naturally
- Need to develop further techniques
- Two such general methodologies are:
  - *Entropy method*: Ledoux (1996), Massart (1998), Lugosi et al. (1999, 2001)
  - *Transportation method*: Ahlswede, Gács and Körner (1976), Marton (1986, 1996, 1997), Dembo (1997), Villani (2003, 2008)

Tensorization

- Idea: prove a bound on a function of many independent variables by proving bounds on functions of one variable
- Unlike variance, sub-Gaussianity does not tensorize naturally
- Need to develop further techniques
- Two such general methodologies are:
  - *Entropy method*: Ledoux (1996), Massart (1998), Lugosi et al. (1999, 2001)
  - *Transportation method*: Ahlswede, Gács and Körner (1976), Marton (1986, 1996, 1997), Dembo (1997), Villani (2003, 2008)
- Both are fueled by basic information-theoretic tools

Information measures

For random variables Y, Z taking values in finite sets let Shannon entropy and conditional Shannon entropy be defined by

$$H(Y) := \sum_{y} p_Y(y) \log \frac{1}{p_Y(y)}$$

$$H(Y|Z) := \sum_{y,z} p_{Y,Z}(y,z) \log \frac{1}{p_{Y|Z}(y|z)}$$

Information measures

For random variables Y, Z taking values in finite sets let Shannon entropy and conditional Shannon entropy be defined by

$$H(Y) := \sum_{y} p_Y(y) \log \frac{1}{p_Y(y)}$$

$$H(Y|Z) := \sum_{y,z} p_{Y,Z}(y,z) \log \frac{1}{p_{Y|Z}(y|z)}$$

Properties:

For random variables Y, Z taking values in finite sets let Shannon entropy and conditional Shannon entropy be defined by

$$H(Y) := \sum_y p_Y(y) \log \frac{1}{p_Y(y)}$$

$$H(Y|Z) := \sum_{y,z} p_{Y,Z}(y,z) \log \frac{1}{p_{Y|Z}(y|z)}$$

Properties:

- $0 \leq H(Y|Z) \leq H(Y) \leq \log |\mathscr{Y}|$

For random variables Y, Z taking values in finite sets let Shannon entropy and conditional Shannon entropy be defined by

$$H(Y) := \sum_y p_Y(y) \log \frac{1}{p_Y(y)}$$
$$H(Y|Z) := \sum_{y,z} p_{Y,Z}(y,z) \log \frac{1}{p_{Y|Z}(y|z)}$$

Properties:

- $0 \leq H(Y|Z) \leq H(Y) \leq \log |\mathscr{Y}|$
- Conditioning reduces entropy: $H(Y|Z,W) \leq H(Y|Z)$

For random variables Y, Z taking values in finite sets let Shannon entropy and conditional Shannon entropy be defined by

$$H(Y) := \sum_{y} p_Y(y) \log \frac{1}{p_Y(y)}$$
$$H(Y|Z) := \sum_{y,z} p_{Y,Z}(y,z) \log \frac{1}{p_{Y|Z}(y|z)}$$

Properties:

- $0 \leq H(Y|Z) \leq H(Y) \leq \log |\mathscr{Y}|$
- Conditioning reduces entropy: $H(Y|Z,W) \leq H(Y|Z)$
- Chain rule: $H(Y,Z) = H(Y) + H(Z|Y)$, so $H(Y,Z,W) = H(Y) + H(Z|Y) + H(W|Z,Y)$

Han's inequality

For any n random variables $Y_1, Y_2, \ldots, Y_n$ with $Y = (Y_1, Y_2, \ldots, Y_n)$ and $Y^{(i)} = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)$, we have

$$H(Y) \leq \frac{1}{n-1} \sum_{i=1}^{n} H\left(Y^{(i)}\right)$$

Han's inequality

For any n random variables $Y_1, Y_2, \ldots, Y_n$ (not necessarily independent), with $Y = (Y_1, Y_2, \ldots, Y_n)$ and $Y^{(i)} = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)$, we have

$$H(Y) \le \frac{1}{n-1} \sum_{i=1}^{n} H\left(Y^{(i)}\right)$$

Han's inequality

For any n random variables $Y_1, Y_2, \ldots, Y_n$ (not necessarily independent), with $Y = (Y_1, Y_2, \ldots, Y_n)$ and $Y^{(i)} = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)$, we have

$$H(Y) \leq \frac{1}{n-1} \sum_{i=1}^{n} H\left(Y^{(i)}\right)$$

Proof of Han's inequality

$$H(Y) = H(Y_i | Y^{(i)}) + H(Y^{(i)})$$

Han's inequality

For any n random variables $Y_1, Y_2, \ldots, Y_n$ (not necessarily independent), with $Y = (Y_1, Y_2, \ldots, Y_n)$ and $Y^{(i)} = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)$, we have

$$H(Y) \leq \frac{1}{n-1} \sum_{i=1}^{n} H\left(Y^{(i)}\right)$$

Proof of Han's inequality

$$\begin{aligned} H(Y) &= H(Y_i | Y^{(i)}) + H(Y^{(i)}) \\ &\leq H(Y_i | Y_1, Y_2, \ldots, Y_{i-1}) + H(Y^{(i)}) \end{aligned}$$

Han's inequality

For any n random variables $Y_1, Y_2, \ldots, Y_n$ (not necessarily independent), with $Y = (Y_1, Y_2, \ldots, Y_n)$ and $Y^{(i)} = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)$, we have

$$H(Y) \leq \frac{1}{n-1} \sum_{i=1}^{n} H\left(Y^{(i)}\right)$$

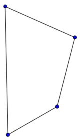
Proof of Han's inequality

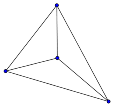
$$\begin{aligned} H(Y) &= H(Y_i | Y^{(i)}) + H(Y^{(i)}) \\ &\leq H(Y_i | Y_1, Y_2, \ldots, Y_{i-1}) + H(Y^{(i)}) \end{aligned}$$

Summing over i, we get $nH(Y) \leq H(Y) + \sum_{i=1}^{n} H\left(Y^{(i)}\right)$

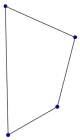
Application: growth rate of number of subsets in convex position

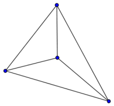
Application: growth rate of number of subsets in convex position





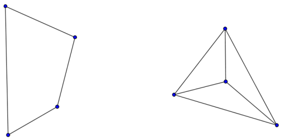
Application: growth rate of number of subsets in convex position





Points on left are in convex position

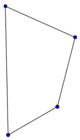
Application: growth rate of number of subsets in convex position

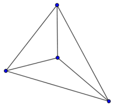


Points on left are in convex position

Points on right are not

Application: growth rate of number of subsets in convex position





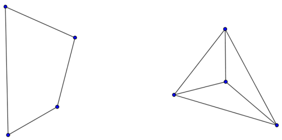
Points on left are in convex position

Points on right are not

Combinatorial entropy

Let $X_1, X_2, \dots, X_n$ be independent taking values in $\mathbb{R}^2$.
Let M be number of subsets of points in convex position.

Application: growth rate of number of subsets in convex position



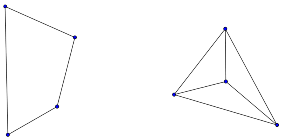
Points on left are in convex position

Points on right are not

Combinatorial entropy

Let $X_1, X_2, \ldots, X_n$ be independent taking values in $\mathbb{R}^2$.
Let M be number of subsets of points in convex position.
Let $Z = \log_2 M$ (called a combinatorial entropy)

Application: growth rate of number of subsets in convex position



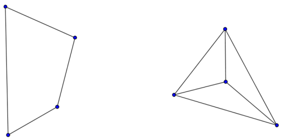
Points on left are in convex position

Points on right are not

Combinatorial entropy

Let $X_1, X_2, \ldots, X_n$ be independent taking values in $\mathbb{R}^2$.
Let M be number of subsets of points in convex position.
Let $Z = \log_2 M$ (called a combinatorial entropy)
Then, $\mathrm{Var}(Z) \leq \mathbb{E}Z$

Application: growth rate of number of subsets in convex position



Points on left are in convex position

Points on right are not

Combinatorial entropy

Let $X_1, X_2, \ldots, X_n$ be independent taking values in $\mathbb{R}^2$.
Let M be number of subsets of points in convex position.
Let $Z = \log_2 M$ (called a combinatorial entropy)
Then, $\mathrm{Var}(Z) \leq \mathbb{E}Z$
Thus, standard deviation $= O(\sqrt{\text{mean}})$

Relative entropy

Tilting of a measure

If P is a probability measure on Ω and $Y : \Omega \mapsto \mathbb{R}_{\geq 0}$ is any non-negative random variable such that $\mathbb{E}Y = 1$,

Tilting of a measure

If P is a probability measure on Ω and $Y : \Omega \mapsto \mathbb{R}_{\geq 0}$ is any non-negative random variable such that $\mathbb{E}Y = 1$, define a new probability measure $Q = P_Y$ by tilting of the original measure P with Y as:

Tilting of a measure

If P is a probability measure on Ω and $Y : \Omega \mapsto \mathbb{R}_{\geq 0}$ is any non-negative random variable such that $\mathbb{E}Y = 1$, define a new probability measure $Q = P_Y$ by tilting of the original measure P with Y as:

$$Q(A) = \mathbb{E}[Y 1_A]$$

Tilting of a measure

If P is a probability measure on Ω and $Y : \Omega \mapsto \mathbb{R}_{\geq 0}$ is any non-negative random variable such that $\mathbb{E}Y = 1$, define a new probability measure $Q = P_Y$ by tilting of the original measure P with Y as:

$$Q(A) = \mathbb{E}[Y 1_A], \qquad \mathbb{E}_Q[W] = \mathbb{E}[YW]$$

Tilting of a measure

If P is a probability measure on Ω and $Y : \Omega \mapsto \mathbb{R}_{\geq 0}$ is any non-negative random variable such that $\mathbb{E}Y = 1$, define a new probability measure $Q = P_Y$ by tilting of the original measure P with Y as:

$$Q(A) = \mathbb{E}[Y 1_A], \qquad \mathbb{E}_Q[W] = \mathbb{E}[YW]$$

We also write $\dfrac{dQ}{dP} = Y$ (called Radon-Nikodym derivative)

Tilting of a measure

If P is a probability measure on Ω and $Y : \Omega \mapsto \mathbb{R}_{\geq 0}$ is any non-negative random variable such that $\mathbb{E}Y = 1$, define a new probability measure $Q = P_Y$ by tilting of the original measure P with Y as:

$$Q(A) = \mathbb{E}[Y 1_A], \qquad \mathbb{E}_Q[W] = \mathbb{E}[YW]$$

We also write $\frac{dQ}{dP} = Y$ (called Radon-Nikodym derivative)

Say Q is absolutely continuous with respect to P (denoted $Q << P$),

Tilting of a measure

If P is a probability measure on Ω and $Y : \Omega \mapsto \mathbb{R}_{\geq 0}$ is any non-negative random variable such that $\mathbb{E}Y = 1$, define a new probability measure $Q = P_Y$ by tilting of the original measure P with Y as:

$$Q(A) = \mathbb{E}[Y 1_A], \qquad \mathbb{E}_Q[W] = \mathbb{E}[YW]$$

We also write $\dfrac{dQ}{dP} = Y$ (called Radon-Nikodym derivative)

Say Q is absolutely continuous with respect to P (denoted $Q << P$), if for any measurable A, $P(A) = 0 \implies Q(A) = 0$.

Tilting of a measure

If P is a probability measure on Ω and $Y : \Omega \mapsto \mathbb{R}_{\geq 0}$ is any non-negative random variable such that $\mathbb{E}Y = 1$, define a new probability measure $Q = P_Y$ by tilting of the original measure P with Y as:

$$Q(A) = \mathbb{E}[Y 1_A], \qquad \mathbb{E}_Q[W] = \mathbb{E}[YW]$$

We also write $\dfrac{dQ}{dP} = Y$ (called Radon-Nikodym derivative)

Say Q is absolutely continuous with respect to P (denoted $Q << P$), if for any measurable A, $P(A) = 0 \implies Q(A) = 0$.

Radon-Nikodym Theorem

If $Q << P$, then there exists Y such that $Q = P_Y$.

Relative entropy

Relative entropy

$$D(Q||P) = \mathbb{E}_Q\left[\log \frac{dQ}{dP}\right]$$

Relative entropy

$$D(Q||P) = \mathbb{E}_Q\left[\log\frac{dQ}{dP}\right] = \mathbb{E}\left[\frac{dQ}{dP}\log\frac{dQ}{dP}\right]$$

Relative entropy

$$D(Q||P) = \mathbb{E}_Q\left[\log\frac{dQ}{dP}\right] = \mathbb{E}\left[\frac{dQ}{dP}\log\frac{dQ}{dP}\right]$$

For Q, P on finite sets having mass functions q, p

$$D(Q||P) = \sum_x q(x)\log\frac{q(x)}{p(x)}$$

Relative entropy

$$D(Q||P) = \mathbb{E}_Q\left[\log\frac{dQ}{dP}\right] = \mathbb{E}\left[\frac{dQ}{dP}\log\frac{dQ}{dP}\right]$$

For Q, P on finite sets having mass functions q, p

$$D(Q||P) = \sum_x q(x)\log\frac{q(x)}{p(x)}$$

For Q, P having densities q, p respectively on the real line

$$D(Q||P) = \int_{-\infty}^{\infty} q(x)\log\frac{q(x)}{p(x)}\,dx$$

Relative entropy

$$D(Q||P) = \mathbb{E}_Q\left[\log \frac{dQ}{dP}\right] = \mathbb{E}\left[\frac{dQ}{dP}\log \frac{dQ}{dP}\right]$$

For Q, P on finite sets having mass functions q, p

$$D(Q||P) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

For Q, P having densities q, p respectively on the real line

$$D(Q||P) = \int_{-\infty}^{\infty} q(x) \log \frac{q(x)}{p(x)}\, dx$$

- $D(Q||P) \geq 0$, with equality if and only if $Q = P$

Relative entropy

$$D(Q||P) = \mathbb{E}_Q\left[\log \frac{dQ}{dP}\right] = \mathbb{E}\left[\frac{dQ}{dP}\log \frac{dQ}{dP}\right]$$

For Q, P on finite sets having mass functions q, p

$$D(Q||P) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

For Q, P having densities q, p respectively on the real line

$$D(Q||P) = \int_{-\infty}^{\infty} q(x) \log \frac{q(x)}{p(x)}\, dx$$

- $D(Q||P) \geq 0$, with equality if and only if $Q = P$
- $D(Q||P)$ not symmetric in Q and P

Relative entropy

$$D(Q||P) = \mathbb{E}_Q\left[\log \frac{dQ}{dP}\right] = \mathbb{E}\left[\frac{dQ}{dP}\log\frac{dQ}{dP}\right]$$

Very Important Notion in Information Theory

ne

$\int_{-\infty}$ $p(x)$

- $D(Q||P) \geq 0$, with equality if and only if $Q = P$
- $D(Q||P)$ not symmetric in Q and P

Relative entropy

$$D(Q||P) = \mathbb{E}_Q\left[\log \frac{dQ}{dP}\right] = \mathbb{E}\left[\frac{dQ}{dP} \log \frac{dQ}{dP}\right]$$

Very Important Notion in Information Theory

Shows up in the theory of concentration of measure in two distinct ways:

- entropy method
- transportation method

- $D(Q||P) \geq 0$, with equality if and only if $Q = P$
- $D(Q||P)$ not symmetric in Q and P

Relative entropy

$$D(Q||P) = \mathbb{E}_Q \left[\log \frac{dQ}{dP}\right] = \mathbb{E}\left[\frac{dQ}{dP} \log \frac{dQ}{dP}\right]$$

For Q, P on finite sets having mass functions q, p

$$D(Q||P) = \sum_x q(x) \log \frac{q(x)}{p(x)}$$

For Q, P having densities q, p respectively on the real line

$$D(Q||P) = \int_{-\infty}^{\infty} q(x) \log \frac{q(x)}{p(x)}\, dx$$

- $D(Q||P) \geq 0$, with equality if and only if $Q = P$
- $D(Q||P)$ not symmetric in Q and P

Han's inequality for relative entropy

Suppose $P = P_1 \times P_2 \times \ldots \times P_n$, and $Q << P$ are two probability measures on $\mathscr{X}^n$.

Han's inequality for relative entropy

Suppose $P = P_1 \times P_2 \times \ldots \times P_n$, and $Q << P$ are two probability measures on $\mathscr{X}^n$.
If $Q^{(i)}, P^{(i)}$ denote the marginals of Q and P on $\mathscr{X}^{(i)}$, then

Han's inequality for relative entropy

Suppose $P = P_1 \times P_2 \times \ldots \times P_n$, and $Q << P$ are two probability measures on $\mathscr{X}^n$.
If $Q^{(i)}, P^{(i)}$ denote the marginals of Q and P on $\mathscr{X}^{(i)}$, then

$$D(Q||P) \geq \frac{1}{n-1} \sum_{i=1}^{n} D(Q^{(i)}||P^{(i)})$$

Han's inequality for relative entropy

Suppose $P = P_1 \times P_2 \times \ldots \times P_n$, and $Q << P$ are two probability measures on $\mathscr{X}^n$.
If $Q^{(i)}, P^{(i)}$ denote the marginals of Q and P on $\mathscr{X}^{(i)}$, then

$$D(Q||P) \geq \frac{1}{n-1} \sum_{i=1}^{n} D(Q^{(i)}||P^{(i)})$$

LHS$= \sum_{x^n} q(x^n) \log q(x^n) - \sum_{x^n} q(x^n) \log p(x^n)$

Han's inequality for relative entropy

Suppose $P = P_1 \times P_2 \times \ldots \times P_n$, and $Q << P$ are two probability measures on $\mathscr{X}^n$.
If $Q^{(i)}, P^{(i)}$ denote the marginals of Q and P on $\mathscr{X}^{(i)}$, then

$$D(Q||P) \geq \frac{1}{n-1} \sum_{i=1}^{n} D(Q^{(i)}||P^{(i)})$$

LHS$= \sum_{x^n} q(x^n) \log q(x^n) - \sum_{x^n} q(x^n) \log p(x^n)$

RHS$= \frac{1}{n-1} \sum_{i=1}^{n} \sum_{x^{(i)}} q^{(i)}(x^{(i)}) \log q^{(i)}(x^{(i)})$

Han's inequality for relative entropy

Suppose $P = P_1 \times P_2 \times \ldots \times P_n$, and $Q << P$ are two probability measures on $\mathscr{X}^n$.
If $Q^{(i)}, P^{(i)}$ denote the marginals of Q and P on $\mathscr{X}^{(i)}$, then

$$D(Q||P) \geq \frac{1}{n-1} \sum_{i=1}^{n} D(Q^{(i)}||P^{(i)})$$

LHS$= \sum_{x^n} q(x^n) \log q(x^n) - \sum_{x^n} q(x^n) \log p(x^n)$

RHS$= \frac{1}{n-1} \sum_{i=1}^{n} \sum_{x^{(i)}} q^{(i)}(x^{(i)}) \log q^{(i)}(x^{(i)})$

$- \frac{1}{n-1} \sum_{i=1}^{n} \sum_{x^{(i)}} q^{(i)}(x^{(i)}) \log p^{(i)}(x^{(i)})$

Han's inequality for relative entropy

Suppose $P = P_1 \times P_2 \times \ldots \times P_n$, and $Q << P$ are two probability measures on $\mathscr{X}^n$.
If $Q^{(i)}, P^{(i)}$ denote the marginals of Q and P on $\mathscr{X}^{(i)}$, then

$$D(Q||P) \geq \frac{1}{n-1} \sum_{i=1}^{n} D(Q^{(i)}||P^{(i)})$$

LHS$= \sum_{x^n} q(x^n) \log q(x^n) - \sum_{x^n} q(x^n) \log p(x^n)$

RHS$= \frac{1}{n-1} \sum_{i=1}^{n} \sum_{x^{(i)}} q^{(i)}(x^{(i)}) \log q^{(i)}(x^{(i)})$

$-\frac{1}{n-1} \sum_{i=1}^{n} \sum_{x^{(i)}} q^{(i)}(x^{(i)}) \log p^{(i)}(x^{(i)})$

As $p(x^n) = \Pi_{i=1}^{n} p_i(x_i)$, second terms in red are equal.

Han's inequality for relative entropy

Suppose $P = P_1 \times P_2 \times \ldots \times P_n$, and $Q << P$ are two probability measures on $\mathscr{X}^n$.
If $Q^{(i)}, P^{(i)}$ denote the marginals of Q and P on $\mathscr{X}^{(i)}$, then

$$D(Q||P) \geq \frac{1}{n-1} \sum_{i=1}^{n} D(Q^{(i)}||P^{(i)})$$

LHS$= \sum_{x^n} q(x^n) \log q(x^n) - \sum_{x^n} q(x^n) \log p(x^n)$

RHS$= \frac{1}{n-1} \sum_{i=1}^{n} \sum_{x^{(i)}} q^{(i)}(x^{(i)}) \log q^{(i)}(x^{(i)})$

$- \frac{1}{n-1} \sum_{i=1}^{n} \sum_{x^{(i)}} q^{(i)}(x^{(i)}) \log p^{(i)}(x^{(i)})$

As $p(x^n) = \Pi_{i=1}^{n} p_i(x_i)$, second terms in red are equal.
Result follows from Han's inequality for Shannon entropy.

Entropy

If Z is a non-negative random variable, we define
$\mathrm{Ent}(Z) := \mathbb{E}[Z \log Z] - (\mathbb{E}Z) \log \mathbb{E}Z$

Entropy

If Z is a non-negative random variable, we define
$\mathrm{Ent}(Z) := \mathbb{E}[Z \log Z] - (\mathbb{E}Z) \log \mathbb{E}Z$

Properties:

- For convex ϕ, we have $\mathbb{E}\phi(X) - \phi(\mathbb{E}X) \geq 0$,

Entropy

If Z is a non-negative random variable, we define
$\mathrm{Ent}(Z) := \mathbb{E}[Z \log Z] - (\mathbb{E}Z) \log \mathbb{E}Z$

Properties:

- For convex ϕ, we have $\mathbb{E}\phi(X) - \phi(\mathbb{E}X) \geq 0$,

  $\phi(x) = x^2$ gives Var

Entropy

If Z is a non-negative random variable, we define
$\mathrm{Ent}(Z) := \mathbb{E}[Z \log Z] - (\mathbb{E}Z) \log \mathbb{E}Z$

Properties:

- For convex ϕ, we have $\mathbb{E}\phi(X) - \phi(\mathbb{E}X) \geq 0$,

  $\phi(x) = x^2$ gives Var

  $\phi(x) = x \log x$ gives Ent

Entropy

If Z is a non-negative random variable, we define
$\mathrm{Ent}(Z) := \mathbb{E}[Z \log Z] - (\mathbb{E}Z) \log \mathbb{E}Z$

Properties:

- For convex ϕ, we have $\mathbb{E}\phi(X) - \phi(\mathbb{E}X) \geq 0$,

  $\phi(x) = x^2$ gives Var

  $\phi(x) = x \log x$ gives Ent
- $\mathrm{Ent}(Z) \geq 0$

Entropy

If Z is a non-negative random variable, we define
$\mathrm{Ent}(Z) := \mathbb{E}[Z \log Z] - (\mathbb{E}Z) \log \mathbb{E}Z$

Properties:

- For convex ϕ, we have $\mathbb{E}\phi(X) - \phi(\mathbb{E}X) \geq 0$,

  $\phi(x) = x^2$ gives Var

  $\phi(x) = x \log x$ gives Ent
- $\mathrm{Ent}(Z) \geq 0$
- $\mathrm{Ent}(aZ) = \mathbb{E}[aZ \log(aZ)] - (\mathbb{E}aZ) \log(\mathbb{E}aZ) = a\,\mathrm{Ent}(Z)$

Entropy

If Z is a non-negative random variable, we define
$\mathrm{Ent}(Z) := \mathbb{E}[Z \log Z] - (\mathbb{E}Z) \log \mathbb{E}Z$

Properties:

- For convex ϕ, we have $\mathbb{E}\phi(X) - \phi(\mathbb{E}X) \geq 0$,

  $\phi(x) = x^2$ gives Var

  $\phi(x) = x \log x$ gives Ent
- $\mathrm{Ent}(Z) \geq 0$
- $\mathrm{Ent}(aZ) = \mathbb{E}[aZ \log(aZ)] - (\mathbb{E}aZ) \log(\mathbb{E}aZ) = a\,\mathrm{Ent}(Z)$
- If $\mathbb{E}Z = 1$, then $\mathrm{Ent}(Z) = \mathbb{E}[Z \log Z] = D(P_Z || P)$

Entropy

If Z is a non-negative random variable, we define
$\mathrm{Ent}(Z) := \mathbb{E}[Z \log Z] - (\mathbb{E}Z) \log \mathbb{E}Z$

Properties:

- For convex ϕ, we have $\mathbb{E}\phi(X) - \phi(\mathbb{E}X) \geq 0$,

  $\phi(x) = x^2$ gives Var

  $\phi(x) = x \log x$ gives Ent
- $\mathrm{Ent}(Z) \geq 0$
- $\mathrm{Ent}(aZ) = \mathbb{E}[aZ \log(aZ)] - (\mathbb{E}aZ) \log(\mathbb{E}aZ) = a\,\mathrm{Ent}(Z)$
- If $\mathbb{E}Z = 1$, then $\mathrm{Ent}(Z) = \mathbb{E}[Z \log Z] = D(P_Z || P)$
- Thus, $\dfrac{\mathrm{Ent}(Z)}{\mathbb{E}Z} = D\left(P_{\frac{Z}{\mathbb{E}Z}} || P\right)$

Entropy

If Z is a non-negative random variable, we define
$\mathrm{Ent}(Z) := \mathbb{E}[Z \log Z] - (\mathbb{E}Z) \log \mathbb{E}Z$

Properties:

- For convex ϕ, we have $\mathbb{E}\phi(X) - \phi(\mathbb{E}X) \geq 0$,

  $\phi(x) = x^2$ gives Var

  $\phi(x) = x \log x$ gives Ent
- $\mathrm{Ent}(Z) \geq 0$
- $\mathrm{Ent}(aZ) = \mathbb{E}[aZ \log(aZ)] - (\mathbb{E}aZ) \log(\mathbb{E}aZ) = a\,\mathrm{Ent}(Z)$
- If $\mathbb{E}Z = 1$, then $\mathrm{Ent}(Z) = \mathbb{E}[Z \log Z] = D(P_Z || P)$
- Thus, $\dfrac{\mathrm{Ent}(Z)}{\mathbb{E}Z} = D\left(P_{\frac{Z}{\mathbb{E}Z}} || P\right)$
- Crucial fact: Ent tensorizes!!

Tensorization of entropy

Tensorization of entropy

Let $Z = f(X_1, X_2, \ldots, X_n)$ where $X_1, X_2, \ldots, X_n$ are independent random variables.

Tensorization of entropy

Let $Z = f(X_1, X_2, \ldots, X_n)$ where $X_1, X_2, \ldots, X_n$ are independent random variables.

$X^{(i)} := (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$

Tensorization of entropy

Let $Z = f(X_1, X_2, \ldots, X_n)$ where $X_1, X_2, \ldots, X_n$ are independent random variables.

$$X^{(i)} := (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n), \qquad \mathbb{E}^{(i)}[\,\cdot\,] := \mathbb{E}[\,\cdot\,|X^{(i)}]$$

Tensorization of entropy

Let $Z = f(X_1, X_2, \ldots, X_n)$ where $X_1, X_2, \ldots, X_n$ are independent random variables.

$$X^{(i)} := (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n), \qquad \mathbb{E}^{(i)}[\,\cdot\,] := \mathbb{E}[\,\cdot\,|X^{(i)}]$$

$$\mathrm{Ent}^{(i)}(Z) := \mathrm{Ent}(Z|X^{(i)})$$

Tensorization of entropy

Let $Z = f(X_1, X_2, \ldots, X_n)$ where $X_1, X_2, \ldots, X_n$ are independent random variables.

$$X^{(i)} := (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n), \qquad \mathbb{E}^{(i)}[\,\cdot\,] := \mathbb{E}[\,\cdot\,|X^{(i)}]$$

$$\mathrm{Ent}^{(i)}(Z) := \mathrm{Ent}(Z|X^{(i)})$$

$$g_i(x^{(i)}) = \mathrm{Ent}(f(x_1, \ldots, X_i, \ldots, x_n))$$

Tensorization of entropy

Let $Z = f(X_1, X_2, \ldots, X_n)$ where $X_1, X_2, \ldots, X_n$ are independent random variables.

$$X^{(i)} := (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n), \qquad \mathbb{E}^{(i)}[\,\cdot\,] := \mathbb{E}[\,\cdot\,|X^{(i)}]$$

$$\mathrm{Ent}^{(i)}(Z) := \mathrm{Ent}(Z|X^{(i)})$$

$$g_i(x^{(i)}) = \mathrm{Ent}(f(x_1, \ldots, X_i, \ldots, x_n)) \implies \mathrm{Ent}^{(i)}(Z) = g_i(X^{(i)})$$

Tensorization of entropy

Let $Z = f(X_1, X_2, \dots, X_n)$ where $X_1, X_2, \dots, X_n$ are independent random variables.

$$X^{(i)} := (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n), \qquad \mathbb{E}^{(i)}[\,\cdot\,] := \mathbb{E}[\,\cdot\,|X^{(i)}]$$

$$\mathrm{Ent}^{(i)}(Z) := \mathrm{Ent}(Z|X^{(i)})$$

$$g_i(x^{(i)}) = \mathrm{Ent}(f(x_1, \dots, X_i, \dots, x_n)) \implies \mathrm{Ent}^{(i)}(Z) = g_i(X^{(i)})$$

Tensorization of entropy

$$\mathrm{Ent}(Z) \leq \sum_{i=1}^{n} \mathbb{E}[\mathrm{Ent}^{(i)}(Z)]$$

Proof

Proof

Since Ent is homogenous, assume $\mathbb{E}Z = \mathbb{E}f(X) = 1$.

Tensorization of entropy

Proof

Since Ent is homogenous, assume $\mathbb{E}Z = \mathbb{E}f(X) = 1$.

Let the distribution of $X = (X_1, X_2, \ldots, X_n)$ be $P = P_1 \times P_2 \times \ldots \times P_n$.

Proof

Since Ent is homogenous, assume $\mathbb{E}Z = \mathbb{E}f(X) = 1$.

Let the distribution of $X = (X_1, X_2, \ldots, X_n)$ be
$P = P_1 \times P_2 \times \ldots \times P_n$.

Let $q(x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n)p(x_1, x_2, \ldots, x_n)$

Proof

Since Ent is homogenous, assume $\mathbb{E}Z = \mathbb{E}f(X) = 1$.

Let the distribution of $X = (X_1, X_2, \ldots, X_n)$ be
$P = P_1 \times P_2 \times \ldots \times P_n$.

Let $q(x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n)p(x_1, x_2, \ldots, x_n)$

i.e. Q is obtained from P by tilting according to f.

Proof

Since Ent is homogenous, assume $\mathbb{E}Z = \mathbb{E}f(X) = 1$.

Let the distribution of $X = (X_1, X_2, \ldots, X_n)$ be
$P = P_1 \times P_2 \times \ldots \times P_n$.

Let $q(x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n)p(x_1, x_2, \ldots, x_n)$

i.e. Q is obtained from P by tilting according to f.

$\mathbb{E}[Z|X^{(i)} = x^{(i)}]$

Tensorization of entropy

Proof

Since Ent is homogenous, assume $\mathbb{E}Z = \mathbb{E}f(X) = 1$.

Let the distribution of $X = (X_1, X_2, \ldots, X_n)$ be $P = P_1 \times P_2 \times \ldots \times P_n$.

Let $q(x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n)p(x_1, x_2, \ldots, x_n)$

i.e. Q is obtained from P by tilting according to f.

$$\mathbb{E}[Z|X^{(i)} = x^{(i)}] = \frac{\sum_{x_i} p(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n)}{\sum_{x_i} p(x_1, x_2, \ldots, x_n)}$$

Proof

Since Ent is homogenous, assume $\mathbb{E}Z = \mathbb{E}f(X) = 1$.

Let the distribution of $X = (X_1, X_2, \ldots, X_n)$ be
$P = P_1 \times P_2 \times \ldots \times P_n$.

Let $q(x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n)p(x_1, x_2, \ldots, x_n)$

i.e. Q is obtained from P by tilting according to f.

$$\mathbb{E}[Z|X^{(i)} = x^{(i)}] = \frac{\sum_{x_i} p(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n)}{\sum_{x_i} p(x_1, x_2, \ldots, x_n)}$$

$$\mathbb{E}^{(i)}[Z]$$

Proof

Since Ent is homogenous, assume $\mathbb{E}Z = \mathbb{E}f(X) = 1$.

Let the distribution of $X = (X_1, X_2, \ldots, X_n)$ be
$P = P_1 \times P_2 \times \ldots \times P_n$.

Let $q(x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n)p(x_1, x_2, \ldots, x_n)$

i.e. Q is obtained from P by tilting according to f.

$$\mathbb{E}[Z|X^{(i)} = x^{(i)}] = \frac{\sum_{x_i} p(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n)}{\sum_{x_i} p(x_1, x_2, \ldots, x_n)}$$

$$\mathbb{E}^{(i)}[Z] = \mathbb{E}[Z|X^{(i)}] = \frac{q^{(i)}(X^{(i)})}{p^{(i)}(X^{(i)})}$$

Proof

Proof

$$\mathrm{Ent}(Z) = \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z]$$

Proof

$$\begin{aligned}\mathrm{Ent}(Z) &= \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z] \\ &= D(Q||P)\end{aligned}$$

Proof

$$
\begin{aligned}
\mathrm{Ent}(Z) &= \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z] \\
&= D(Q||P) \\
\mathbb{E}[\mathrm{Ent}^{(i)}(Z)] &= \mathbb{E}\left[\mathbb{E}^{(i)}[Z \log Z] - \mathbb{E}^{(i)}[Z] \log \mathbb{E}^{(i)}[Z]\right]
\end{aligned}
$$

Proof

$$
\begin{aligned}
\mathrm{Ent}(Z) &= \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z] \\
&= D(Q||P) \\
\mathbb{E}[\mathrm{Ent}^{(i)}(Z)] &= \mathbb{E}\left[\mathbb{E}^{(i)}[Z \log Z] - \mathbb{E}^{(i)}[Z] \log \mathbb{E}^{(i)}[Z]\right] \\
&= D(Q||P) - D(Q^{(i)}||P^{(i)})
\end{aligned}
$$

Proof

$$\begin{aligned}
\mathrm{Ent}(Z) &= \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z] \\
&= D(Q||P) \\
\mathbb{E}[\mathrm{Ent}^{(i)}(Z)] &= \mathbb{E}\left[\mathbb{E}^{(i)}[Z \log Z] - \mathbb{E}^{(i)}[Z] \log \mathbb{E}^{(i)}[Z]\right] \\
&= D(Q||P) - D(Q^{(i)}||P^{(i)})
\end{aligned}$$

We need to show

$$D(Q||P) \leq \sum_{i=1}^{n} \left[D(Q||P) - D(Q^{(i)}||P^{(i)})\right]$$

Proof

$$\begin{aligned}
\mathrm{Ent}(Z) &= \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z] \\
&= D(Q||P) \\
\mathbb{E}[\mathrm{Ent}^{(i)}(Z)] &= \mathbb{E}\left[\mathbb{E}^{(i)}[Z \log Z] - \mathbb{E}^{(i)}[Z] \log \mathbb{E}^{(i)}[Z]\right] \\
&= D(Q||P) - D(Q^{(i)}||P^{(i)})
\end{aligned}$$

We need to show

$$D(Q||P) \leq \sum_{i=1}^{n} \left[D(Q||P) - D(Q^{(i)}||P^{(i)})\right]$$

But this is exactly Han's inequality for relative entropy.

Summary

Summary

- Magical phenomenon causes functions of many independent variables to be concentrated in a way analogous to the law of large numbers

Summary

- Magical phenomenon causes functions of many independent variables to be concentrated in a way analogous to the law of large numbers
- Tensorization of variance can capture this phenomenon in its generality

Summary

- Magical phenomenon causes functions of many independent variables to be concentrated in a way analogous to the law of large numbers
- Tensorization of variance can capture this phenomenon in its generality
- But in fact, sub-Gaussian tail bounds also hold for functions of many independent variables

Summary

- Magical phenomenon causes functions of many independent variables to be concentrated in a way analogous to the law of large numbers
- Tensorization of variance can capture this phenomenon in its generality
- But in fact, sub-Gaussian tail bounds also hold for functions of many independent variables
- The entropy method and transportation method are two techniques to capture such behavior

Summary

- Magical phenomenon causes functions of many independent variables to be concentrated in a way analogous to the law of large numbers
- Tensorization of variance can capture this phenomenon in its generality
- But in fact, sub-Gaussian tail bounds also hold for functions of many independent variables
- The entropy method and transportation method are two techniques to capture such behavior
- We have discussed basic information inequalities and are in shape to talk about the entropy method next time