ECE253/CSE208 Introduction to Information Theory

Lecture 4: Convexity and Inequalities

Dr. Yu Zhang

ECE Department
University of California, Santa Cruz

- Chap 2 of *Elements of Information Theory (2nd Edition) by Thomas Cover & Joy Thomas*

Convex Functions

Definition (Convexity of functions)

- A function $f(x)$ is *convex* over an interval (a, b) if and only if

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

holds for any $x_1, x_2 \in (a, b)$ and $\lambda \in [0, 1]$.

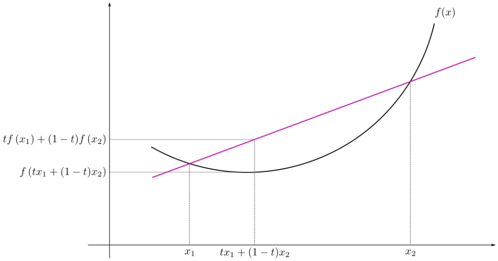


Figure: Any *chord* of a convex function is always above the function itself: The 0th-order condition of convexity.

Convex Functions

Definition (Convexity of functions)

- A function $f(\cdot)$ that is *differentiable everywhere* in (a, b) is *convex* if and only if

$$f(y) \geq f(x) + f'(x)(y - x)$$

for any $x, y \in (a, b)$.

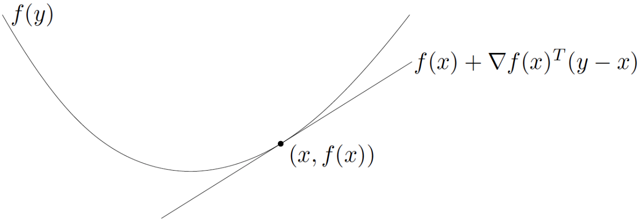


Figure: Tangent lines are always global under-estimator of the function: The 1st-order condition of convexity.

Convex Functions

Definition (Convexity of functions)

- A function $f(x)$ that is *twice differentiable* over (a, b) is *convex* if and only if

$$f''(x) \geq 0$$

for any $x \in (a, b)$.

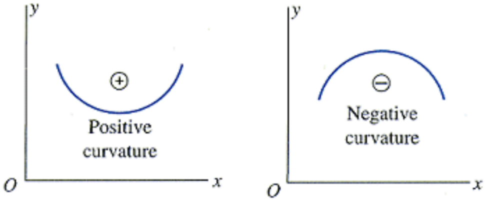


Figure: Convex functions always curve upward (positive curvature).: The 2nd-order condition of convexity.

Examples. Convex functions: $ax + b, |x|, x^2, x^4, e^{\pm x}, x \log x$.

If $f(x)$ is *convex*, then $-f(x)$ is *concave*. Affine functions are both convex and concave.

Convexity in High-dimensional Spaces

Definition (Convex function)

A function $f(\mathbf{x}) : \mathbb{R}^n \mapsto \mathbb{R}$ is *convex* iif for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have one of the following:

- 0th-order condition: $f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$ for any $\lambda \in [0,1]$.
- 1st-order condition: $f(\mathbf{y}) \geq f(\mathbf{x}) + \bigtriangledown f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$
- 2nd-order condition: $\mathbf{H}_f := \bigtriangledown^2 f(\mathbf{x}) \succeq \mathbf{0}$; i.e., the *Hessian* matrix is positive semi-definite (all eigenvalues are nonnegative), where $\mathbf{H}_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$.

- Strictly convex: if strict inequality always holds when $\mathbf{x} \neq \mathbf{y}$ and $\lambda \in (0,1)$.
- Strongly convex: $\mathbf{H}_f \succeq a\mathbf{I}$ for some constant $a > 0$ (the Hessian is positive definite).
- Geometrically, strict/strong convexity implies that the function has no flat part and curves upward everywhere $\implies$ unique minimizer.

Convex Sets

Definition (Convex sets)

A set $S \subseteq \mathbb{R}^n$ is convex if and only if $\lambda \mathbf{x}_1 + (1-\lambda)\mathbf{x}_2 \in S, \forall\ \mathbf{x}_1, \mathbf{x}_2 \in S, \lambda \in [0,1]$.

Geometrically, a convex set contains line segment between any two points in the set.

Convex sets are solid body without holes and curve outward.

examples (one convex, two nonconvex sets)

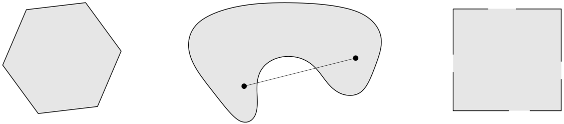


Figure: Convex and nonconvex sets (source: Stephen Boyd, Stanford).

Epigraph and Sublevel Set

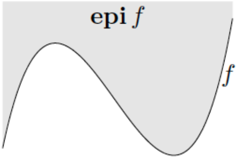
- α-**sublevel set** of $f : \mathbb{R}^n \to \mathbb{R}$:

$$C_\alpha = \{\mathbf{x} \in \mathrm{dom} f : f(\mathbf{x}) \leq \alpha\}$$

  sublevel sets of convex functions are convex (converse is false)

- **epigraph** of $f : \mathbb{R}^n \to \mathbb{R}$:

$$\mathrm{epi} f = \{(\mathbf{x}, t) \in \mathbb{R}^{n+1} : \mathbf{x} \in \mathrm{dom} f,\ f(\mathbf{x}) \leq t\}$$



f is a convex function $\iff$ $\mathrm{epi} f$ is a convex set

Convex Optimization Problem

Convex optimization problem in standard form:

$$\begin{aligned} \min_{\mathbf{x}} \quad & f_0(\mathbf{x}) \\ \text{s.to} \quad & f_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m \\ & \mathbf{a}_i^\top \mathbf{x} = b_i, \quad i = 1, \ldots, p \end{aligned}$$

- $f_0, f_1, \ldots, f_m$ are convex
- equality constraints are affine (alternatively $\mathbf{Ax} = \mathbf{b}$)

Important properties:

1. For convex problems, any local solution is also global.
2. If $f_0()$ is strictly convex, the minimizer is unique.
3. The optimal set X_{opt} is convex.

Jensen's Inequality

Lemma (Jenson's Inequality)

If X is a random variable and $f(\cdot)$ is a convex function, then

$$\boxed{E(f(X)) \geq f(E(X))}$$

Moreover, if $f(X)$ is strictly convex, equality implies $X = \mathrm{E}(X)$ with probability 1.

Jensen's Inequality (Cont'd)

Proof:

1) For a *two-point distribution* $X \in \{x_1, x_2\}$, the convexity of $f(\cdot) \implies$

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2).$$

2) Proof by *induction*: Assume Jenson's inequality holds for a $(k-1)$-mass point distribution. To show the inequality holds for a k-mass point distribution, define $p_i' = \frac{p_i}{1-p_k}$ for all $i = 1, 2, ..., k-1$:

$$\begin{aligned}
\sum_{i=1}^{k} p_i f(x_i) &= p_k f(x_k) + (1-p_k) \sum_{i=1}^{k-1} p_i' f(x_i) \\
&\geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i\right) \\
&\geq f\left(p_k x_k + (1-p_k) \sum_{i=1}^{k-1} p_i' x_i\right) = f\left(\sum_{i=1}^{k} p_i x_i\right)
\end{aligned}$$

where the 1st inequality is from the induction while the 2nd inequality is due to the convexity of $f(\cdot)$.

Gibbs' Inequality (Information Inequality)

Theorem (Gibbs' Inequality)

Let $p(x)$ and $q(x)$ be two probability mass functions. Then, $D(p\|q) \geq 0$ with equality if and only if $p(x) = q(x)$ for all x.

Proof: $-D(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{q(x)}{p(x)} \leq \log \left(\sum_{x \in \mathcal{X}} p(x) \frac{q(x)}{p(x)} \right) = \log \left(\sum_{x \in \mathcal{X}} q(x) \right) = 0.$

Corollary (Nonnegativity of mutual information)

For any two random variables X and Y, we have $I(X;Y) \geq 0$ with equality iff X and Y are independent.

Corollary (Conditional mutual Information)

$$I(X;Y|Z) \geq 0$$

with equality iff X and Y are conditionally independent given Z, which is denoted as $(X \perp\!\!\!\perp Y) \mid Z$.

Gibbs' Inequality (Cont'd)

Theorem (Conditioning reduces entropy (information cannot hurt))

$H(X|Y) \leq H(X)$ *with equality iif* $X \perp\!\!\!\perp Y$.

Intuitively, knowing Y can only reduce the uncertainty in X. Note that this is true only on the average (expectation) sense. That is, $H(X|Y = y) > H(X)$ can happen.

Theorem (Uniform distribution has the maximum entropy)

$$H(X) \leq \log |\mathcal{X}|$$

where $|\mathcal{X}|$ *is the cardinality of the set* $\mathcal{X}$ *(i.e., the number of elements in the set) with equality iff* X *has a uniform distribution over* $\mathcal{X}$.

Proof: Let $u(x) = \frac{1}{|\mathcal{X}|}$ be the uniform PMF, and $p(x)$ be the PMF of X over $\mathcal{X}$, respectively. Then, $D(p\|u) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{u(x)} = \log |\mathcal{X}| - H(X) \geq 0$.

Gibbs' Inequality (Cont'd)

Theorem (Independence bound on entropy)

Let $X_1, X_2, \ldots, X_n$ *be drawn from* $p(x_1, x_2, ..., x_n)$. *Then,*

$$H\left(X_1, X_2, ..., X_n\right) \leq \sum_{i=1}^{n} H\left(X_i\right)$$

with equality iff the $\{X_i\}_{i=1}^{n}$ *are independent.*

Proof: By the chain rule: $H\left(X_1, X_2, \ldots, X_n\right) = \sum_{i=1}^{n} H\left(X_i | X_{i-1}, \ldots, X_1\right) \leq \sum_{i=1}^{n} H(X_i)$.

Theorem (Convexity of relative entropy)

$D(p||q)$ *is convex in the pair* (p, q). *That is, if* (p_1, q_1) *and* (p_2, q_2) *are two pairs of PMFs. Then,*

$$D\left(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2\right) \leq \lambda D\left(p_1 \| q_1\right) + (1-\lambda) D\left(p_2 \| q_2\right)$$

for all $0 \leq \lambda \leq 1$.

Proof: By the log-sum inequality; see details on page 31–32 of the textbook.

Entropy and Probability of Errors

Lemma

Let X, X' be independent with $X \sim p(x), X' \sim r(x'), x, x' \in \mathcal{X}$. Then,

$$\textit{Prob}(X = X') \geq \max\left\{2^{-H(p,r)}, 2^{-H(r,p)}\right\}.$$

Proof: $2^{-H(p,r)} = 2^{\mathrm{E}_p(\log_2 r(X))} \leq \mathrm{E}_p\left(2^{\log_2 r(X)}\right) = \sum_{x \in \mathcal{X}} p(x)r(x) = \mathrm{Prob}(X = X')$.

Corollary

Let X, X' are i.i.d. with entropy $H(X)$. Then,

$$\textit{Prob}(X = X') \geq 2^{-H(X)},$$

with equality iif X has a uniform distribution.

Data Processing Inequality (DPI)

Definition

Random variables X, Y, Z are said to form a Markov chain in the $X \to Y \to Z$ if the joint PMF can be written as $p(x, y, z) = p(x)p(y|x)p(z|y)$.

Note: $p(x, y, z) = p(x)p(y, z|x) = p(x)p(y|x)p(z|y, x) = p(x)p(y|x)p(z|y)$.

Corollary

$X \to Y \to Z$ iff X and Z are conditionally independent given Y [i.e., $(X \perp\!\!\!\perp Z) \mid Y$]. Markovity implies conditional independence since

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x, y)p(z|y)}{p(y)} = p(x|y)p(z|y).$$

Corollary ($X \leftrightarrow Y \leftrightarrow Z$)

$X \to Y \to Z \Leftrightarrow Z \to Y \to X$.

DPI

Theorem (Data Processing Inequality (DPI))

If $X \to Y \to Z$ then $I(X;Y) \geq I(X;Z)$ with equality iif $I(X;Y|Z) = 0$ (i.e., $X \to Z \to Y$).

Proof: By the chain rule, we can expand mutual information in two different ways:

$$\begin{aligned} I(X;Y,Z) &= I(X;Y) + I(X;Z|Y) \\ &= I(X;Z) + I(X;Y|Z) \end{aligned}$$

Since $(X \perp\!\!\!\perp Z) \mid Y \implies I(X;Z|Y) = 0$. Consider the fact that $I(X;Y|Z) \geq 0$, we have $I(X;Y) \geq I(X;Z)$.

Corollary

$I(X;Y) \geq I(X;g(Y))$ for any function $g(\cdot)$. Thus, functions (post-processing) of Y cannot increase the information about X.

Proof: $X \to Y \to g(Y)$.

DPI (Cont'd)

Corollary

If $X \to Y \to Z$, then $I(X;Y|Z) \le I(X;Y)$.

Note that it is possible to have $I(X;Y|Z) > I(X;Y)$; see the following example when X, Y, Z do not form a Markov chain.

Example

Consider $Z = X + Y$ for two i.i.d. $X, Y \sim \mathrm{Bern}(0.5)$, Find $I(X;Y|Z)$.

$$
\begin{aligned}
I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) \\
&= \Pr(Z=0)H(X|Z=0) + \Pr(Z=2)H(X|Z=2) + \Pr(Z=1)H(X|Z=1) \\
&= [\Pr(\mathrm{X}=1, \mathrm{Y}=0) + \Pr(\mathrm{Y}=1, \mathrm{X}=0)] \times H(\mathrm{X}|\mathrm{Z}=1) \\
&= 2 \times \frac{1}{4} \times H(1/2, 1/2) \\
&= 0.5 > I(X;Y) = 0
\end{aligned}
$$

Sufficient Statistics

Given a family of distributions $\{f_\theta(x)\}$. Let X be a sample drawn from a distribution in this family, and $T(X)$ be any statistic (function of the sample such as sample mean or variance). Thus, we have $\theta \to X \to T(X) \Rightarrow I(\theta; X) \geq I(\theta; T(X))$.

Definition (Sufficient Statistic)

A function $T(X)$ is said to be a sufficient statistic relative to the family $\{f_\theta(x)\}$ if X is independent of θ given $T(X)$ for any distribution on θ; i.e., $\theta \to T(X) \to X$ forms a Markov chain. This is the same as the condition for equality in the DPI:

$$\boxed{I(\theta; X) = I(\theta; T(X))}$$

Implication

A statistics is sufficient for θ if it contains all information in X about θ: Once we know $T(X)$, the remaining randomness in X does not depend on θ.

Sufficient Statistics (Cont'd)

Example

Given i.i.d. $X_1, \ldots, X_n \sim \mathrm{Bern}(\theta)$. Let $X^n := (X_1, \ldots, X_n)$, a sufficient statistic of θ is $T(X^n) = \sum_{i=1}^n X_i$.

Proof: Need to prove: $\theta \to T(X^n) \to X^n$; i.e., $P(X^n)$ is independent of θ given $T(X^n)$:

$$P\left((X_1, ..., X_n) = (x_1, ..., x_n) \middle| \sum_{i=1}^n X_i = k\right) = \begin{cases} 0, & \text{if } \sum_{i=1}^n x_i \neq k \\ 1/\binom{n}{k}, & \text{if } \sum_{i=1}^n x_i = k \end{cases}$$

Another example: For $f_\theta = \mathrm{Uniform}(\theta, \theta+1)$, a sufficient statistic for θ is $T(X_1, ..., X_n) = (\max\{X_i\}, \min\{X_i\})$.

Definition (Minimal sufficient statistic)

A statistic $T(X)$ is a *minimal* sufficient statistic if it is a function of every other sufficient statistics $U(X)$, which implies that $\theta \to T(X) \to U(X) \to X$.

Minimal sufficient statistic maximally compresses the information about θ in the sample; see more discussions in the lecture notes on minimal sufficient statistics by Yukai Sun.

Fano's Inequality

Consider a Markov chain $X \to Y \to \hat{X}$. In the context of communications:

- Send symbol X via a noisy channel.
- The received symbol $Y \neq X$ due to the noise.
- Try to recover X by post-processing Y; i.e., $\hat{X} = g(Y)$ for some function g.
- The probability of error is defined as $P_e := P(\hat{X} \neq X)$.

Fano's inequality: We may estimate X with small P_e when $H(X|Y)$ is small.

For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$, define $P_e = \Pr(X \neq \hat{X})$, we have

$$H\left(P_e\right) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y) \implies P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}.$$

If $\hat{X} \in \mathcal{X}$, we then have a slightly stronger inequality:

$$H\left(P_e\right) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y).$$

Fano's Inequality (Cont'd)

Proof: Define $E = \mathbb{1}_{\{\hat{X} \neq X\}}$. By the chain rule, we have

$$\begin{aligned} H(E, X|\hat{X}) &= H(X|\hat{X}) + H(E|X, \hat{X}) \\ &= H(E|\hat{X}) + H(X|E, \hat{X}) \\ &\leq H(P_e) + P(E=0)H(X|\hat{X}, E=0) + P(E=1)H(X|\hat{X}, E=1) \\ &\leq H(P_e) + P_e \log|\mathcal{X}| \end{aligned}$$

$\implies H(P_e) + P_e \log|\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y)$, where the 2nd inequality is due to the DPI: $I(X;Y) \geq I(X;\hat{X})$.

Furthermore, given $E = 1$, the range of possible X outcomes is $|\mathcal{X}| - 1$
$\implies H(X|E, \hat{X}) \leq P_e \log(|\mathcal{X}| - 1)$.

Corollary

For any two random variables X, Y, let $p = P(X \neq Y)$. We have

$$H(p) + p \log|\mathcal{X}| \geq H(X|Y).$$

Proof: Let $\hat{X} = Y$ in Fano's inequality.

Fano's Inequality (Cont'd)

Fano's inequality establishes the fundamental limits of data compression and transmission. It can be used to characterize when a perfect reconstruction of sent code is not possible, i.e. P_e is bounded away from zero.

Example (Fano's inequality is sharp)

Let $X \in \{1, 2, \ldots, m\}$ and $p_1 \geq p_2 \geq \cdots \geq p_m$. Then the best guess of X is $\hat{X} = 1$ and the resulting probability of error is $P_e = 1 - p_1$. Fano's inequality becomes

$$H(P_e) + P_e \log(m-1) \geq H(X).$$

The PMF $(p_1, p_2, \ldots, p_m) = \left(1 - P_e, \frac{P_e}{m-1}, \ldots, \frac{P_e}{m-1}\right)$ achieves the lower bound with equality. To see this,

$$\begin{aligned} H(X) &= -(1-P_e)\log(1-P_e) - (m-1) \times \frac{P_e}{m-1} \log \frac{P_e}{m-1} \\ &= -(1-P_e)\log(1-P_e) - P_e \log P_e + P_e \log(m-1) \\ &= H(P_e) + P_e \log(m-1). \end{aligned}$$

Thank You!

Email: <zhangy@ucsc.edu>
Homepage: https://people.ucsc.edu/~yzhan419/