Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2015

11 Useful Inequalities

Here we prove two simple but extremely useful inequalities. We focus on discrete random variables in the proofs. The statements are also true about continuous random variables, but the proofs are a bit more elaborate.

Definition 1 (Convex set). *A set $\mathcal{V}$ is convex if for any $v_1, v_2 \in \mathcal{V}$ and any $\lambda \in [0,1]$ we have $\lambda v_1 + (1-\lambda)v_2 \in \mathcal{V}$.*

Definition 2 (Convex function). *Let $\mathcal{V}$ be a convex set. Function $\phi(\cdot) : \mathcal{V} \mapsto \mathcal{R}$ is* convex *if for any $v_1, v_2 \in \mathcal{V}$ and any $\lambda \in [0,1]$,*

$$\phi(\lambda v_1 + (1-\lambda)v_2) \leq \lambda\phi(v_1) + (1-\lambda)\phi(v_2). \tag{1}$$

A function that satisfies the definition with strict inequality for all $\lambda \neq 0, 1$ and $v_1 \neq v_2$ is called *strictly convex*. As additional terminology, a function $\phi(\cdot)$ is *concave* if $-\phi(\cdot)$ is convex.

Now we are ready to state and prove the inequalities.

Theorem 1 (Jensen's Inequality). *If $\phi(\cdot)$ is a concave function and v is a random variable defined over alphabet $\mathcal{V}$, then*

$$\mathbb{E}\left[\phi(\mathsf{v})\right] \leq \phi\left(\mathbb{E}\left[\mathsf{v}\right]\right). \tag{2}$$

If $\phi(\cdot)$ is strictly concave, (2) *holds with equality if and only if v is a deterministic constant.*

Proof. We prove this inequality by induction on the size of the alphabet $\mathcal{V}$.

First, we consider $|\mathcal{V}| = 2$ and let v_1 and v_2 be the two elements in $\mathcal{V}$. The definition of concavity implies

$$\mathbb{E}\left[\phi(\mathsf{v})\right] = p_{\mathsf{v}}(v_1)\phi(v_1) + p_{\mathsf{v}}(v_2)\phi(v_2) \leq \phi(p_{\mathsf{v}}(v_1)v_1 + p_{\mathsf{v}}(v_2)v_2) = \phi\left(\mathbb{E}\left[\mathsf{v}\right]\right). \tag{3}$$

If $\phi(\cdot)$ is strictly concave, (3) is satisfied with equality if and only if $p_{\mathsf{v}}(v_1) = 0$ or $p_{\mathsf{v}}(v_2) = 0$, i.e., v is a deterministic constant.

We now consider $\mathcal{V} = \{v_1, \ldots, v_M\}$, $M > 2$, and assume that (2) holds for all random variables defined over alphabets smaller than M elements. Suppose v is not deterministic, i.e., there exists v_i such that $p_{\mathsf{v}}(v_i)$ is neither 0 nor 1. In this case, we can expand the left-hand side of (2) as follows:

$$\begin{aligned}
\mathbb{E}\left[\phi(\mathsf{v})\right] &= \sum_{m=1}^{M} p_{\mathsf{v}}(v_m)\,\phi(v_m) = p_{\mathsf{v}}(v_i)\,\phi(v_i) + \sum_{m\neq i} p_{\mathsf{v}}(v_m)\,\phi(v_m) \\
&= p_{\mathsf{v}}(v_i)\,\phi(v_i) + (1 - p_{\mathsf{v}}(v_i)) \sum_{m\neq i} \frac{p_{\mathsf{v}}(v_m)}{1 - p_{\mathsf{v}}(v_i)} \phi(v_m).
\end{aligned} \tag{4}$$

It is easy to see that the summation in (4) is $\mathbb{E}\left[\phi(\mathsf{v})|\mathsf{v} \neq v_i\right]$. By induction, we have $\mathbb{E}\left[\phi(\mathsf{v})|\mathsf{v} \neq v_i\right] \leq \phi\left(\mathbb{E}\left[\mathsf{v}|\mathsf{v} \neq v_i\right]\right)$ and

$$\begin{aligned}
\mathbb{E}\left[\phi(\mathsf{v})\right] &= p_{\mathsf{v}}(v_i)\,\phi(v_i) + (1 - p_{\mathsf{v}}(v_i)) \sum_{m \neq i} \frac{p_{\mathsf{v}}(v_m)}{1 - p_{\mathsf{v}}(v_i)} \phi(v_m) \\
&\leq p_{\mathsf{v}}(v_i)\,\phi(v_i) + (1 - p_{\mathsf{v}}(v_i))\phi\left(\sum_{m \neq i} \frac{p_{\mathsf{v}}(v_m)}{1 - p_{\mathsf{v}}(v_i)} v_m\right) && (5) \\
&\leq \phi\left(p_{\mathsf{v}}(v_i)\,v_i + (1 - p_{\mathsf{v}}(v_i)) \sum_{m \neq i} \frac{p_{\mathsf{v}}(v_m)}{1 - p_{\mathsf{v}}(v_i)} v_m\right) && (6) \\
&= \phi\left(\sum_{m=1}^{M} p_{\mathsf{v}}(v_m)\,v_m\right) = \phi\left(\mathbb{E}\left[\mathsf{v}\right]\right), && (7)
\end{aligned}$$

where (5) follows from the induction step for $|\mathcal{V}| = M - 1$, and (6) follows from the induction step for $|\mathcal{V}| = 2$.

If $\phi(\cdot)$ is strictly concave, the only way we can get equalities in the derivation above is to make sure that $p_{\mathsf{v}}(v_i) = 0$ or $p_{\mathsf{v}}(v_i) = 1$, and furthermore, conditioned on $\mathsf{v} \neq v_i$, v is deterministic. These conditions are satisfied if and only if v is deterministic. □

Theorem 2 (Gibbs' Inequality). *Let v be a random variable distributed according to distribution $p(\cdot)$. Then for any distribution $q(\cdot)$,*

$$\mathbb{E}_p\left[\log p(\mathsf{v})\right] \geq \mathbb{E}_p\left[\log q(\mathsf{v})\right], \tag{8}$$

with equality if and only if $q \equiv p$.

Before proving this result, we note that we use the notation $\mathbb{E}_p\left[\cdot\right]$ to emphasize the distribution with respect to which the expectation is being taken. Sometimes, we will similarly use a random variable as the subscript (for example, in this case $\mathbb{E}_p\left[\cdot\right] = \mathbb{E}_{\mathsf{v}}\left[\cdot\right]$). In general, we will use such subscripts only when the relevant distribution is not clear from context.

Proof of Gibbs' inequality. By concavity of the log function,

$$\begin{aligned}
\mathbb{E}_p\left[\log q(\mathsf{v})\right] - \mathbb{E}_p\left[\log p(\mathsf{v})\right] &= \mathbb{E}_p\left[\log \frac{q(\mathsf{v})}{p(\mathsf{v})}\right] \\
&\leq \log \mathbb{E}_p\left[\frac{q(\mathsf{v})}{p(\mathsf{v})}\right] && (9) \\
&= \log\left(\sum_v p(v) \cdot \frac{q(v)}{p(v)}\right) = 0, && (10)
\end{aligned}$$

where the inequality in (9) follows from the Jensen's inequality. Note that since the log function is strictly concave, equality in (9) holds if and only if $q(\mathsf{v})/p(\mathsf{v})$ is a constant, which must be 1 since distributions are normalized. □