Bregman divergences and exponential families

Jongkyoung Kim

Department of Computer Science
POSTECH, Korea
blkimjk@postech.ac.kr
Last updated: 15 Dec. 2005

Outline

- Bregman Divergence
- Measure Theory
- Probability Theory
- Exponential Families

Bregman Divergence: Definition

Definition 1. *Let $\mathcal{S}$ be a nonempty, open, convex set such that $\overline{\mathcal{S}} \subseteq dom(\phi)$, where a function $\phi : dom(\phi) \subseteq \mathbb{R}^n \mapsto \mathbb{R}$ is strictly convex on $\overline{\mathcal{S}}$. Assume that ϕ is differentiable on $\mathcal{S}$. The Bregman divergence $d_\phi : \overline{\mathcal{S}} \times \mathcal{S} \mapsto [0, \infty)$ is defined as*

$$d_\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla\phi(\mathbf{y}) \rangle$$

where $\nabla\phi(\mathbf{y})$ represents the gradient vector of ϕ evaluated at $\mathbf{y}$.

Remark 1. [Geometric Interpretation] *The Bregman divergence $d_\phi(\mathbf{x}, \mathbf{y})$ may be interpreted as the difference $\phi(\mathbf{x}) - h(\mathbf{x})$ where $h(\mathbf{z})$ represents the hyperplane, which is tangent to the epigraph of ϕ at the point $(\mathbf{y}, \phi(\mathbf{y}))$ in $\mathbb{R}^{n+1}$. The epigraph of ϕ is defined as*

$$\mathbf{epi}(\phi) = \{(\mathbf{x}, t) | \mathbf{x} \in dom(\phi),\ \phi(\mathbf{x}) \leq t\}$$

Consider the first-order condition for convexity:

$$\phi(\mathbf{x}) \geq \phi(\mathbf{y}) + \nabla\phi(\mathbf{y})^T(\mathbf{x} - \mathbf{y})$$

If $(\mathbf{x}, t) \in \mathbf{epi}(\phi)$*, then*

$$t \geq \phi(\mathbf{x}) \geq \phi(\mathbf{y}) + \nabla\phi(\mathbf{y})^T(\mathbf{x} - \mathbf{y})$$

It can be expressed as:

$$\begin{bmatrix} \nabla\phi(\mathbf{y}) \\ -1 \end{bmatrix}^T \left(\begin{bmatrix} \mathbf{x} \\ t \end{bmatrix} - \begin{bmatrix} \mathbf{y} \\ \phi(\mathbf{y}) \end{bmatrix} \right) \leq 0$$

This means that the hyperplane $h(\mathbf{z})$ *defined by* $(\nabla\phi(\mathbf{y}), -1)$ *supports* $\mathbf{epi}(\phi)$ *at the point* $(\mathbf{y}, \phi(\mathbf{y}))$*.*

Remark 2. [Relationship with exponential families] *The log-likelihood of the density of an exponential family* $p(\mathbf{x}; \theta)$ *can be written as*

$$\log(p(\mathbf{x}; \theta)) = -d_\phi(T(\mathbf{x}), \mu(\theta)) + \log(b_\phi(\mathbf{x}))$$

where ϕ *is the conjugate function of* $A(\theta)$ *and* $\mu(\theta) = E_\theta[T(\mathbf{x})] = \nabla A(\theta)$*. The mappings between* θ *and* μ *are given by the Legendre transformation*

$$\mu(\theta) = \nabla A(\theta) \text{ and } \theta(\mu) = \nabla\phi(\mu).$$

The conjugate function ϕ can be expressed as

$$\begin{aligned}\phi(\mu) &= \sup_{\theta\in\Theta}\{\langle\mu,\theta\rangle - A(\theta)\} \\ &= \langle\nabla A(\theta),\theta\rangle - A(\theta) \\ &= \langle\mu,\theta(\mu)\rangle - A(\theta(\mu))\end{aligned}$$

For a given $\mathbf{x}$, we obtain the relationship by using the above Legendre transformation.

$$\begin{aligned}\log(p(\mathbf{x};\theta)) &= \langle\theta,T(\mathbf{x})\rangle - A(\theta) + \log h(\mathbf{x}) \\ &= \langle\theta,\mu\rangle - A(\theta) + \langle T(\mathbf{x})-\mu,\theta\rangle + \log h(\mathbf{x}) \\ &= \phi(\mu) + \langle T(\mathbf{x})-\mu,\nabla\phi(\mu)\rangle + \log h(\mathbf{x}) \\ &= -d_\phi(T(\mathbf{x}),\mu) + \phi(T(\mathbf{x})) + \log h(\mathbf{x}) \\ &= -d_\phi(T(\mathbf{x}),\mu) + \log(b_\phi(\mathbf{x}))\end{aligned}$$

where $b_\phi(\mathbf{x}) = \exp\{\phi(T(\mathbf{x}))\}h(\mathbf{x})$.

Bregman Divergence: Properties

Property 1. [Non-negativity] *For all* $\mathbf{x} \in \overline{\mathcal{S}}$ *and all* $\mathbf{y} \in \mathcal{S}$*, we have* $d_\phi(\mathbf{x}, \mathbf{y}) \geq 0$*, and equality holds if and only if* $\mathbf{x} = \mathbf{y}$*.*

Proof: From the first-order condition for strict convexity, we have

$$\phi(\mathbf{x}) > \phi(\mathbf{y}) + \nabla\phi(\mathbf{y})^T(\mathbf{x} - \mathbf{y})$$

Therefore, $d_\phi(\mathbf{x}, \mathbf{y}) > 0$. $d_\phi(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$. ■

Property 2. [Non-symmetry] *In general,* $d_\phi(\mathbf{x}, \mathbf{y}) \neq d_\phi(\mathbf{y}, \mathbf{x})$

Property 3. [Three-point] *For* $\mathbf{x}_1 \in \overline{\mathcal{S}}$ *and* $\mathbf{x}_2, \mathbf{x}_3 \in \mathcal{S}$*,*

$$d_\phi(\mathbf{x}_1, \mathbf{x}_3) = d_\phi(\mathbf{x}_1, \mathbf{x}_2) + d_\phi(\mathbf{x}_2, \mathbf{x}_3) - \langle \mathbf{x}_1 - \mathbf{x}_2, \nabla\phi(\mathbf{x}_3) - \nabla\phi(\mathbf{x}_2) \rangle$$

Property 4. [Convexity] d_ϕ *is always convex in the first argument, but not necessarily convex in the second argument.*

Proof: Since the nonnegative weighted sum of convex functions is convex, we can easily see that d_ϕ is convex in the first argument. ■

Definition 2. *Given a closed convex set $\Omega \subseteq \mathbb{R}^n$ and $\mathbf{y} \in \mathcal{S}$, a point $\mathbf{x}^* \in \Omega \cap \overline{\mathcal{S}}$ for which*

$$\mathbf{x}^* = \arg\min_{\mathbf{z} \in \Omega \cap \overline{\mathcal{S}}} d_\phi(\mathbf{z}, \mathbf{y})$$

is denoted by $P_\Omega(\mathbf{y})$ and is called a Bregman projection of the point $\mathbf{y}$ onto the set Ω.

Property 5. [Generalized Pythagoras theorem: Part I] *Let $\Omega \subseteq \mathbb{R}^n$ be a closed convex set such that $\Omega \cap \overline{\mathcal{S}} \neq \emptyset$. Assume that $\mathbf{y}, P_\Omega(\mathbf{y}) \in \mathcal{S}$. Let $\mathbf{z} \in \Omega \cap \overline{\mathcal{S}}$, then the following inequality holds*

$$d_\phi(P_\Omega(\mathbf{y}), \mathbf{y}) + d_\phi(\mathbf{z}, P_\Omega(\mathbf{y})) \leq d_\phi(\mathbf{z}, \mathbf{y})$$

Proof: Define the function

$$\begin{aligned} G(\mathbf{u}) &\equiv d_\phi(\mathbf{u}, \mathbf{y}) - d_\phi(\mathbf{u}, P_\Omega(\mathbf{y})) \\ &= \langle \mathbf{u}, \nabla\phi(P_\Omega(\mathbf{y})) - \nabla\phi(\mathbf{y}) \rangle + \alpha \end{aligned}$$

where α is a real number independent of $\mathbf{u}$. We can see that $G(\mathbf{u})$ is convex. For any

λ with $0 \leq \lambda \leq 1$, we denote $\mathbf{u}_\lambda = \lambda \mathbf{z} + (1-\lambda) P_\Omega(\mathbf{y})$ and obtain that

$$d_\phi(\mathbf{u}_\lambda, \mathbf{y}) - d_\phi(\mathbf{u}_\lambda, P_\Omega(\mathbf{y})) \leq \lambda [d_\phi(\mathbf{z}, \mathbf{y}) - d_\phi(\mathbf{z}, P_\Omega(\mathbf{y}))] + (1-\lambda) d_\phi(P_\Omega(\mathbf{y}), \mathbf{y})$$

For $\lambda > 0$, this leads to

$$d_\phi(\mathbf{z}, \mathbf{y}) - d_\phi(\mathbf{z}, P_\Omega(\mathbf{y})) - d_\phi(P_\Omega(\mathbf{y}), \mathbf{y}) \geq \frac{1}{\lambda} [d_\phi(\mathbf{u}_\lambda, \mathbf{y}) - d_\phi(P_\Omega(\mathbf{y}), \mathbf{y})] - \frac{1}{\lambda} d_\phi(\mathbf{u}_\lambda, P_\Omega(\mathbf{y}))$$

The first term is nonnegative and the second term tends to zero as $\lambda \to 0$. To see this, we use the directional derivative.

$$\begin{aligned} \lim_{\lambda \to 0+} \frac{d_\phi(\mathbf{u}_\lambda, P_\Omega(\mathbf{y}))}{\lambda} &= \lim_{\lambda \to 0+} \frac{d_\phi(P_\Omega(\mathbf{y}) + \lambda(\mathbf{z} - P_\Omega(\mathbf{y})), P_\Omega(\mathbf{y})) - d_\phi(P_\Omega(\mathbf{y}), P_\Omega(\mathbf{y}))}{\lambda} \\ &= \langle \nabla d_\phi(\mathbf{x}, P_\Omega(\mathbf{y}))|_{\mathbf{x}=P_\Omega(\mathbf{y})}, \mathbf{z} - P_\Omega(\mathbf{y}) \rangle \\ &= 0 \end{aligned}$$

The last equality holds since $\nabla d_\phi(\mathbf{x}, P_\Omega(\mathbf{y}))|_{\mathbf{x}=P_\Omega(\mathbf{y})} = 0$. ■

Remark 3. *From the property 2 and 5, we can see that Bregman divergences are not a metric.*

Property 6. [Generalized Pythagoras theorem: Part II] *In the same conditions of property 5, if $\Omega \cap \overline{\mathcal{S}}$ is an affine set, then the following equality holds*

$$d_\phi(P_\Omega(\mathbf{y}), \mathbf{y}) + d_\phi(\mathbf{z}, P_\Omega(\mathbf{y})) = d_\phi(\mathbf{z}, \mathbf{y})$$

Proof: We know that every affine subset of $\mathbb{R}^n$ is an intersection of a finite collection of hyperplanes. Denote those hyperplanes by $\langle \mathbf{a}_i, \mathbf{x} \rangle = b_i$, $i = 1, \ldots, N$. Since $\mathbf{z}, P_\Omega(\mathbf{y}) \in \Omega \cap \overline{\mathcal{S}}$,

$$\langle \mathbf{a}_i, \mathbf{z} - P_\Omega(\mathbf{y}) \rangle = 0, \ i = 1, \ldots, N$$

To show the orthogonality between $\mathbf{z} - P_\Omega(\mathbf{y})$ and $\nabla\phi(\mathbf{y}) - \nabla\phi(P_\Omega(\mathbf{y}))$, consider the following constrained optimization problem

$$\begin{aligned} &\textit{minimize} && d_\phi(\mathbf{z}, \mathbf{y}) \\ &\textit{subject to} && \langle \mathbf{a}_i, \mathbf{z} \rangle = b_i, \ i = 1, \ldots, N \end{aligned}$$

Because this problem is a convex programming, we can use the KKT conditions which are sufficient and necessary. The Lagrangian of this problem is

$$\mathcal{L}(\mathbf{z}, \lambda) = \phi(\mathbf{z}) - \phi(\mathbf{y}) - \langle \mathbf{z} - \mathbf{y}, \nabla\phi(\mathbf{y}) \rangle - \sum_i^N \lambda_i (\langle \mathbf{a}_i, \mathbf{z} \rangle - b)$$

and the KKT conditions are then

$$\nabla_{\mathbf{z}} \mathcal{L}(\mathbf{z}, \lambda) = \nabla\phi(\mathbf{z}^*) - \nabla\phi(\mathbf{y}) - \sum_i^N \lambda_i \mathbf{a}_i = 0$$

$$\langle \mathbf{a}_i, \mathbf{z}^* \rangle = b_i, \; i = 1, \ldots, N$$

Here $\mathbf{z}^*$ *is equal to* $P_\Omega(\mathbf{y})$. *Therefore, we can see that*

$$\begin{aligned} \langle \mathbf{z} - P_\Omega(\mathbf{y}), \nabla\phi(\mathbf{y}) - \nabla\phi(P_\Omega(\mathbf{y})) \rangle &= \langle \mathbf{z} - P_\Omega(\mathbf{y}), - \sum_i^N \lambda_i \mathbf{a}_i \rangle \\ &= 0 \end{aligned}$$

From property 3, it follows that

$$d_\phi(P_\Omega(\mathbf{y}), \mathbf{y}) + d_\phi(\mathbf{z}, P_\Omega(\mathbf{y})) = d_\phi(\mathbf{z}, \mathbf{y}) \; \blacksquare$$

Bregman Divergence: Examples

Example 1. [Squared Euclidean distance] *The underlying function* $\phi(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x} \rangle$ *is strictly convex, differentiable on* $\mathbb{R}^n$.

$$\begin{aligned} d_\phi(\mathbf{x}, \mathbf{y}) &= \langle \mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{y}, \mathbf{y} \rangle - \langle \mathbf{x} - \mathbf{y}, \nabla\phi(\mathbf{y}) \rangle \\ &= \langle \mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{y}, \mathbf{y} \rangle - \langle \mathbf{x} - \mathbf{y}, 2\mathbf{y} \rangle \\ &= \langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle \\ &= \|\mathbf{x} - \mathbf{y}\|^2 \end{aligned}$$

Example 2. [KL divergence] *If* $\mathbf{p}$ *is a positive vector, the negative entropy* $\phi(\mathbf{p}) =$

$\sum_{i=1}^{n} p_i \ln p_i$ *is strictly convex, differentiable on* $\mathbb{R}^n_{++}$.

$$
\begin{aligned}
d_\phi(\mathbf{p}, \mathbf{q}) &= \sum_{i=1}^{n} p_i \ln p_i - \sum_{i=1}^{n} q_i \ln q_i - \langle \mathbf{p} - \mathbf{q}, \nabla\phi(\mathbf{q}) \rangle \\
&= \sum_{i=1}^{n} p_i \ln p_i - \sum_{i=1}^{n} q_i \ln q_i - \sum_{i=1}^{n} (p_i - q_i)(\ln q_i + 1) \\
&= \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i} - \sum_{i=1}^{n} (p_i - q_i) \\
&= KL(\mathbf{p} \| \mathbf{q})
\end{aligned}
$$

Measure Theory: σ-field

Definition 3. [σ-field] *Let Ω be an arbitrary space or nonempty set of points ω. A class $\mathcal{F}_0$ of subsets of Ω is called a field if it satisfies these conditions:*

1. $\Omega \in \mathcal{F}_0$
2. $A \in \mathcal{F}_0$ *implies* $A^c \in \mathcal{F}_0$
3. $A, B \in \mathcal{F}_0$ *implies* $A \cup B \in \mathcal{F}_0$

A field $\mathcal{F}$ is called a σ-field if, in addition, $A_n \in \mathcal{F}, n = 1, 2, \ldots,$ implies $\bigcup_1^\infty A_n \in \mathcal{F}$.

Remark 4. *A field is closed under the finite set-theoretic operations, and a σ-field is closed also under the countable ones.*

Remark 5. *The largest σ-field in Ω is the power class 2^Ω, consisting of all the subsets of Ω; the smallest σ-field consists only of the two sets $\{\emptyset, \Omega\}$.*

Example 3. [Field but not σ-field] *Consider the class of the finite disjoint unions of subintervals of $\Omega = (0, 1]$. Augmented by the empty set, this class is a field $\mathcal{B}_0$. Suppose that $A = (a_1, a_1'] \bigcup \cdots \bigcup (a_m, a_m']$, where $(a_i, a_i']$ are disjoint and $a_1 \leq \cdots \leq a_m$.*

Then $A^c = (a_1, a_1']^c \bigcap \cdots \bigcap (a_m, a_m']^c = (0, a_1] \bigcup (a_1', a_2] \bigcup \cdots \bigcup (a_{m-1}', a_m]$ *and so lies in* $\mathcal{B}_0$. *If* $B = (b_1, b_1'] \bigcup \cdots \bigcup (b_n, b_n']$, *then* $A \bigcap B = \bigcup_{i=1}^{m} \bigcup_{j=1}^{n} \left\{(a_i, a_i'] \bigcap (b_j, b_j']\right\}$, *and so* $A \bigcap B$ *is in* $\mathcal{B}_0$. *Thus* $\mathcal{B}_0$ *is a field. Although* $\mathcal{B}_0$ *is a field, it is not a* σ*-field. It does not contain the singletons* $\{x\}$, *even though each is a countable intersection* $\bigcap_n (x - \frac{1}{n}, x]$.

Proposition 1. *Given any class* $\mathcal{C}$ *of subsets of* Ω, *there is a smallest* σ*-field* $\mathcal{F}(\mathcal{C})$ *containing* $\mathcal{C}$. *The smallest* σ*-field is the intersection of all the* σ*-field containing* $\mathcal{C}$.

Proof: There exists at least one σ-field containing $\mathcal{C}$, the power class 2^{Ω}. Moreover, an arbitrary intersection of σ-fields is itself a σ-field. Suppose that $\mathcal{F} = \bigcap_i \mathcal{F}_i$, where $\mathcal{F}_i$ is a σ-field. Then $\Omega \in \mathcal{F}_i$ for all i, so that $\Omega \in \mathcal{F}$. And $A \in \mathcal{F}$ implies that $A \in \mathcal{F}_i$ and $A^c \in \mathcal{F}_i$ for each i, so that $A^c \in \mathcal{F}$. If $A_n \in \mathcal{F}$ for each n, then $A_n \in \mathcal{F}_i$ for each n and i, so that $\bigcup_n A_n$ lies in each F_i and hence in $\mathcal{F}$. Thus $\mathcal{F}(\mathcal{C})$ is a σ-field containing $\mathcal{C}$. It is smallest in the sense that it is contained in every σ-field that contains $\mathcal{C}$. ■

Definition 4. [Measurable space] *A pair* $(\Omega, \mathcal{F})$ *is a measurable space if* $\mathcal{F}$ *is a* σ*-field in* Ω.

Example 4. [Borel σ-field] *Some important measurable spaces are:*

1. $\Omega = (a, b] \subset \mathbb{R}$, $\mathcal{C}$: *the class of all sub-intervals of* Ω.
   $\mathcal{B}_I = \mathcal{F}(\mathcal{C})$

2. $\Omega = \mathbb{R}$, C : *the class of all sub-intervals of* Ω.
   $\mathcal{B} = \mathcal{F}(C)$
3. $\Omega = \mathbb{R}^k$, C : *the class of all sub-rectangles of* Ω.
   $\mathcal{B}_k = \mathcal{F}(C)$

Remark 6. *The Borel* σ*-field contains the open and closed sets. It contains all the subsets in* Ω *that actually arise in ordinary analysis and probability theory.*

Measure Theory: Measures

Definition 5. [Finitely additive measure] *A set function μ_0 on a field $\mathcal{F}_0$ is a finitely additive measure if it satisfies these conditions:*

1. $\mu_0(A) \in [0, \infty]$ *for* $A \in \mathcal{F}_0$
2. $\mu_0(\emptyset) = 0$
3. *for disjoint* $A, B \in \mathcal{F}_0$,

$$\mu_0(A \cup B) = \mu_0(A) + \mu_0(B)$$

Definition 6. [Measure] *A set function μ on a σ-field $\mathcal{F}$ is a (σ-additive) measure if it satisfies these conditions:*

1. $\mu(A) \in [0, \infty]$ *for* $A \in \mathcal{F}$
2. $\mu(\emptyset) = 0$
3. *for disjoint* $A_1, A_2, \ldots \in \mathcal{F}$,

$$\mu\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mu(A_k)$$

The measure μ is finite or infinite as $\mu(\Omega) < \infty$ or $\mu(\Omega) = \infty$. It is a probability measure if $\mu(\Omega) = 1$. If there exists a countable sequence $A_1, A_2, \ldots$ such that $\Omega = \bigcup_{k=1}^{\infty} A_k$ and $\mu(A_k) < \infty$, then μ is σ-finite.

Definition 7. [Measure space] *The triple $(\Omega, \mathcal{F}, \mu)$ is a measure space if μ is a measure on a σ-field $\mathcal{F}$ in Ω.*

Proposition 2. [Continuity from above and below] *Let μ be a measure on the σ-field $\mathcal{F}$.*

1. *If $A_n \in \mathcal{F}, A_n \downarrow A$, and if $\mu(A_n) < \infty$ for some n, then $\mu(A_n) \downarrow \mu(A)$.*
2. *If $A_n \in \mathcal{F}, A_n \uparrow A$, then $\mu(A_n) \uparrow \mu(A)$.*

Remark 7. *By $A_n \uparrow A$ is meant $A_1 \subset A_2 \subset \cdots$ and $A = \bigcup_n A_n$; by $A_n \downarrow A$ is meant $A_1 \supset A_2 \supset \cdots$ and $A = \bigcap_n A_n$*

Theorem 1. [Carathéodory's extension theorem] *If a finitely additive measure μ_0 on a field $\mathcal{F}_0$ is continuous from above at $\emptyset$, then there exists a unique measure μ on $\mathcal{F}$ such that $\mu(A) = \mu_0(A)$ for every $A \in \mathcal{F}_0$.*

Remark 8. [How to construct measure spaces] *Carathéodory's extension theorem will almost always allow one to extend a finitely additive measure to a measure. In*

fact, it is the tool that is generally applied to construct measure spaces. Typically, to define a measure space, we first construct a finitely additive measure on a field by an explicit formula. Then we check continuity from above at $\emptyset$ and invoke Carathéodory's extension theorem.

Example 5. [Counting measure] *Let $\mathcal{F}$ be the σ-field of all subsets of Ω, and let $\mu(A)$ be the number of points in A, where $\mu(A) = \infty$ if A is not finite. This μ is counting measure. It is finite iff Ω is finite, and is σ-finite iff Ω is countable. Even if $\mathcal{F}$ does not contain every subset of Ω, counting measure is well defined on $\mathcal{F}$.*

Example 6. [Lebesgue measure] *Let $\mathcal{F}$ be the Borel σ-field of all sub-intervals of $\mathbb{R}$, and consider $\mu((a, b]) = b - a$. This μ is Lebesgue measure. For k-dimensional Borel σ-field, the Lebesgue measure is defined as*

$$\mu(A) = \prod_{i=1}^{k}(b_i - a_i),\ A = \{\mathbf{x} \in \mathbb{R}^k | a_i < x_i \leq b_i, i = 1, \ldots, k\}.$$

The Lebesgue measure is σ-finite.

Remark 9. *The class of all finite union of disjoint sub-intervals in $\mathbb{R}$ is a field. Take μ_0 on this field to be length. Then μ_0 is continuous from above at $\emptyset$. Therefore, Lebesgue measure is an unique extension.*

Definition 8. [Almost everywhere (a.e.)] *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. We say a property P holds almost everywhere with respect to the measure μ if the set $A = \{\omega \in \Omega | P$ does not hold at $\omega\} \in \mathcal{F}$ and $\mu(A) = 0$. We write this as P $\mu-$a.e.*

Measure Theory: Measurable Functions

Definition 9. [Inverse image] *Given two spaces Ω, $\mathbb{R}$, and a function $X : \Omega \mapsto \mathbb{R}$, the inverse image of a set $B \subset \mathbb{R}$ is defined as $X^{-1}B = \{\omega \in \Omega | X(\omega) \in B\}$. Denote this by $\{\omega | X(\omega) \in B\}$.*

Definition 10. [(Borel) Measurable function] *Given two measurable spaces $(\Omega, \mathcal{F})$, $(\mathbb{R}, \mathcal{B})$, a function $X : \Omega \mapsto \mathbb{R}$ is called (Borel) measurable if the inverse image of every set in $\mathcal{B}$ is in $\mathcal{F}$. Refer measurable functions on $(\Omega, \mathcal{F})$ as $\mathcal{F}$-measurable. Denote the σ-field of inverse images of sets in $\mathcal{B}$ by $\mathcal{F}(X)$.*

Proposition 3. *Let $\mathcal{A} \subset \mathcal{B}$ such that $\mathcal{F}(\mathcal{A}) = \mathcal{B}$. Then $X : \Omega \mapsto \mathbb{R}$ is measurable if the inverse of every set in $\mathcal{A}$ is in $\mathcal{F}$.*

Remark 10. *Hence, if X is a real function such that $\{\omega | X(\omega) \leq x\} \in \mathcal{F}$ for all x, then X is $\mathcal{F}$-measurable.*

Proposition 4. *If $X : \mathbb{R} \mapsto \mathbb{R}$ is continuous, then it is measurable.*

Proposition 5. *If X is a measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B})$ and ϕ is a $\mathcal{B}$-measurable function, then $\phi(X)$ is an $\mathcal{F}$-measurable function.*

Remark 11. *If ϕ is a continuous function, $\phi(X)$ is $\mathcal{F}$-measurable.*

Proposition 6. *Suppose that $X_1, X_2, \ldots$ are $\mathcal{F}$-measurable.*

1. *The functions $\sup_n X_n$, $\inf_n X_n$, $\limsup_n X_n$, and $\liminf_n X_n$ are $\mathcal{F}$-measurable.*
2. *If $\lim_n X_n$ exists everywhere (pointwise convergence), then it is $\mathcal{F}$-measurable.*

Definition 11. [Simple function] *The set indicator of a subset $A \in \mathcal{F}$ is the function*

$$1_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

A simple function is any finite linear combination of set indicators,

$$g(\omega) = \sum_{i=1}^{n} \alpha_i 1_{A_i}(\omega), \; A_i \in \mathcal{F}$$

Remark 12. *A set indicator and simple function are $\mathcal{F}$-measurable.*

Proposition 7. *If X is $\mathcal{F}$-measurable, there exists a sequence $\{X_n\}$ of simple*

functions such that

$$0 \leq X_n(\omega) \uparrow X(\omega) \text{ if } X(\omega) \geq 0$$

$$0 \geq X_n(\omega) \downarrow X(\omega) \text{ if } X(\omega) \leq 0$$

for every $\omega \in \Omega$.

Proposition 8. [Transformation of measures] *Let* $(\Omega, \mathcal{F})$ *and* $(\mathbb{R}^n, \mathcal{B}_n)$ *be measurable spaces, and suppose that the mapping* $X : \Omega \mapsto \mathbb{R}^n$ *is* $\mathcal{F}$*-measurable. Given a measure* μ *on* $\mathcal{F}$*, define a set function* $\tilde{\mu}$ *on* $\mathcal{B}_n$ *by*

$$\tilde{\mu}(B) = \mu(X^{-1}B), \ B \in \mathcal{B}_n$$

Then $\tilde{\mu}$ *is a measure on* $(\mathbb{R}^n, \mathcal{B}_n)$.

Measure Theory: Integration

Definition 12. [Definite integral] *Let f be a nonnegative $\mathcal{F}$-measurable functions on a measure space $(\Omega, \mathcal{F}, \mu)$. The integral of f is defined as*

$$\begin{aligned} \int f d\mu &= \int_\Omega f(\omega) d\mu(\omega) = \int_\Omega f(\omega) \mu(d\omega) \\ &= \sup \left\{ \sum_i \left[\inf_{\omega \in A_i} f(\omega) \right] \mu(A_i) \right\} \end{aligned}$$

where $A_i \in \mathcal{F}$; $A_i \cap A_j = \emptyset$ for $i \neq j$; and $\bigcup_i A_i = \Omega$. The supremum extends over all finite decomposition $\{A_i\}$ of Ω. If $\int f d\mu$ is finite, then f is integrable.

Remark 13. *For general f, the integral is defined by*

$$\int f d\mu = \int f^+ d\mu - \int f^- d\mu$$

where

$$f^+(\omega) = \begin{cases} f(\omega) & \text{if } 0 \leq f(\omega) \leq \infty \\ 0 & \text{if } -\infty \leq f(\omega) \leq 0 \end{cases}$$

and

$$f^{-}(\omega) = \begin{cases} -f(\omega) & \text{if } -\infty \le f(\omega) \le 0 \\ 0 & \text{if } 0 \le f(\omega) \le \infty \end{cases}$$

Proposition 9. *If $f(\omega) = \sum_i \alpha_i 1_{A_i}(\omega)$ is a nonnegative simple function, $\{A_i\}$ being a finite decomposition of Ω, then $\int f d\mu = \sum_i \alpha_i \mu(A_i)$.*

Proof: Let $\{B_j\}$ be a finite decomposition of Ω and β_j be the infimum of f over B_j. If $A_i \bigcap B_j \neq \emptyset$, then $\beta_j \le \alpha_i$. Therefore, $\sum_j \beta_j \mu(B_j) = \sum_{ij} \beta_j \mu(A_i \bigcap B_j) \le \sum_{ij} \alpha_i \mu(A_i \bigcap B_j) = \sum_i \alpha_i \mu(A_i)$. The equality holds if $\{B_j\}$ coincides with $\{A_i\}$. ■

Theorem 2. [Monotone convergence theorem] *If $0 \le f_n(\omega) \uparrow f(\omega)$ for all ω, then $0 \le \int f_n d\mu \uparrow \int f d\mu$.*

Remark 14. [Uniqueness of the integral] *Although there are various ways to define the integral, they are all equivalent if they have properties of proposition 9 and theorem 2. For f nonnegative, there exist simple functions f_n such that $0 \le f_n \uparrow f$. The integral $\int f d\mu$ must be $\lim_n \int f_n d\mu$, and the value of each $\int f_n d\mu$ is determined.*

Proposition 10. [Monotonicity] *If f and g are integrable and $f \le g$ a.e., then $\int f d\mu \le \int g d\mu$.*

Proposition 11. [Linearity] *If f and g are integrable and α, β are finite real numbers, then $\alpha f + \beta g$ is integrable and $\int \alpha f + \beta g d\mu = \alpha \int f d\mu + \beta \int g d\mu$.*

Theorem 3. [Fatou's lemma] *If $\{f_n\}$ is a sequence of nonnegative measurable functions, then*

$$\int \liminf_n f_n d\mu \leq \liminf_n \int f_n d\mu.$$

Proof: If $g_n = \inf_{k \geq n} f_k$, then $0 \leq g_n \uparrow g = \liminf_n f_n$. Thus by the theorem 2, $\int g_n d\mu \uparrow \int g d\mu$. Since $\int f_n d\mu \geq \int g_n d\mu$, we have

$$\begin{aligned} \int \liminf_n f_n d\mu &= \lim_n \int g_n d\mu \\ &= \liminf_n \int g_n d\mu \\ &\leq \liminf_n \int f_n d\mu. \blacksquare \end{aligned}$$

Theorem 4. [Lebesgue's dominated convergence theorem] *If $|f_n| \leq g$ a.e., where g is integrable, and if $f_n \to f$ a.e., then f and f_n are integrable and $\int f_n d\mu \to \int f d\mu$.*

Theorem 5. [Interchanging integration and differentiation] *Suppose that* $f(\omega, t)$ *is a measurable and integrable function of* ω *for each* t *in* (a, b). *Let* $\phi(t) = \int f(\omega, t)\mu(d\omega)$. *Suppose that for* $\omega \in A$, *where* $A \in \mathcal{F}, \mu(\Omega - A) = 0$, $f(\omega, t)$ *has a derivative* $f'(\omega, t)$ *in* (a, b) *and that* $|f'(\omega, t)| \leq g(\omega)$ *for* $\omega \in A$ *and* $t \in (a, b)$, *where* g *is integrable. Then* $\phi(t)$ *has derivative* $\int f'(\omega, t)\mu(d\omega)$ *on* (a, b).

Proof: By Lebesgue's dominated convergence theorem and mean value theorem,

$$\begin{aligned} \phi'(t) = \lim_{h \to 0} \frac{\phi(t+h) - \phi(t)}{h} &= \lim_{h \to 0} \int \frac{f(\omega, t+h) - f(\omega, t)}{h} \mu(d\omega) \\ &= \int f'(\omega, t)\mu(d\omega). \blacksquare \end{aligned}$$

Definition 13. [Integration over sets] *The integral of* f *over a set* A *in* $\mathcal{F}$ *is defined by*

$$\int_A f d\mu = \int 1_A f d\mu$$

Notice that $\int_A f d\mu = 0$ *if* $\mu(A) = 0$.

Proposition 12. *If* $A_1, A_2, \ldots$ *are disjoint, and if* f *is either nonnegative or integrable, then* $\int_{\cup_n A_n} f d\mu = \sum_n \int_{A_n} f d\mu$.

Proposition 13. [Hölder's inequality] *Let f and g be $\mathcal{F}$-measurable functions on a measure space $(\Omega, \mathcal{F}, \mu)$ for which $|f|^p$ and $|g|^q$ are integrable. For $p, q > 0$ such that $\frac{1}{p} + \frac{1}{q} = 1$,*

$$\left| \int f g d\mu \right| \leq \int |fg| d\mu \leq \left[\int |f|^p d\mu \right]^{\frac{1}{p}} \left[\int |g|^q d\mu \right]^{\frac{1}{q}}$$

Definition 14. [Density] *Suppose that f is a nonnegative measurable function. Define a measure ν by*

$$\nu(A) = \int_A f d\mu, \ A \in \mathcal{F}$$

The measure ν is said to have density f with respect to μ. For formal substitution, we write $d\nu = f d\mu$.

Remark 15. *This is one of the methods of constructing a new measure from a given measure. Note that $\mu(A) = 0$ implies that $\nu(A) = 0$. Clearly, ν is finite iff f is integrable.*

Theorem 6. *If ν has density f with respect to μ, then*

$$\int g d\nu = \int g f d\mu$$

holds for nonnegative g. Moreover, g is integrable with respect to ν iff gf is integrable with respect to μ, in which case

$$\int g d\nu = \int gf d\mu, \quad \int_A g d\nu = \int_A gf d\mu$$

both hold.

Measure Theory: Lebesgue Integral

Definition 15. [Lebesgue integral] *A $\mathcal{B}$-measurable function on $\mathbb{R}$ is Lebesgue integrable if it is integrable with respect to Lebesgue measure λ, and its Lebesgue integral $\int f d\lambda$ is denoted by $\int f(x)dx$.*

Definition 16. [Riemann integral] *Let $f : [a, b] \to \mathbb{R}$ be a bounded function. The function f is said to be Riemann integrable on $[a, b]$ if*

$$\overline{\int_a^b} f(x)dx = \underline{\int_a^b} f(x)dx$$

where

$$\underline{\int_a^b} f(x)dx = \sup\{L(P, f) = \sum_{i=1}^{n} m_i(x_i - x_{i-1}) | P \text{ is a partition of } f\}$$

and

$$\overline{\int_a^b} f(x)dx = \inf\{U(P, f) = \sum_{i=1}^{n} M_i(x_i - x_{i-1}) | P \text{ is a partition of } f\}.$$

Here, $m_i = \inf\{f(x)|x_{i-1} \leq x \leq x_i\}$ and $M_i = \sup\{f(x)|x_{i-1} \leq x \leq x_i\}$. The common value in that case is called the Riemann integral of f over $[a, b]$ and is denoted by $\int_a^b f(x)dx$

Remark 16. *The Riemann integral when it exists coincides with the Lebesgue integral.*

Remark 17. [Riemann integrable] *Let $\mathcal{R}[a, b]$ denote the set of all functions $f : [a, b] \mapsto \mathbb{R}$ which are Riemann integrable. $\mathcal{R}[a, b]$ includes the class of all monotone functions, the class of all continuous functions, and the class of all functions with finite discontinuity points. A bounded function on a bounded interval is Riemann integrable iff the set of its discontinuities has Lebesgue measure 0.*

Example 7. [Dirichlet function] *Let $f : (a, b] \to \mathbb{R}$ be defined by*

$$f(x) = \begin{cases} 1 & \text{if } x \text{ is a rational} \\ 0 & \text{if } x \text{ is a irrational} \end{cases}$$

For any partition P of $(a, b]$, $U(P, f) = (b - a)$ and $L(P, f) = 0$. Thus f is not Riemann integrable. But the Lebesgue integral is 0 because $f = 0$ a.e.

Remark 18. *It is because of its extreme oscillations that the above function fails to be Riemann integrable. This cannot happen in the case of the Lebesgue integral of a*

measurable function. If f fails to be Lebesgue integrable, it is because of its positive part or its negative part is too large, not because of one or the other is too irregular.

Measure Theory: Absolute Continuity

Definition 17. [Absolutely continuous] *Let μ and ν be two measures on a measurable space $(\Omega, \mathcal{F})$. We say that ν is absolutely continuous with respect μ, denoted by $\nu \ll \mu$, if, for $A \in \mathcal{F}$, $\mu(A) = 0$ implies $\nu(A) = 0$.*

Remark 19. *If $\nu(A) = \int_A f d\mu$, then certainly $\nu \ll \mu$. The Radon-Nikodym theorem goes in the opposite direction.*

Theorem 7. [Radon-Nikodym theorem] *If μ and ν are σ-finite measures such that $\nu \ll \mu$, then there exists a nonnegative f, a density, such that $\nu(A) = \int_A f d\mu$ for all $A \in \mathcal{F}$. For two such densities f and g, $\mu[f \neq g] = 0$.*

Probability Theory: Probability Spaces

Definition 18. [Probability space] *A measure space $(\Omega, \mathcal{F}, P)$ is called a probability space, where P is a probability measure on $\mathcal{F}$. A point $\omega \in \Omega$ is called the sample point. A set $A \in \mathcal{F}$ is called an event. The event N with $P(N) = 0$ is called a null event. A property which is true except for null events is said to hold almost surely (a.s.) or with probability 1.*

Example 8. [Tossing coins] *Suppose we toss a coin three times. The sample space is given by*

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

Each of the followings is a σ-field

$\mathcal{F}_0 = \{\emptyset, \Omega\}$

$\mathcal{F}_1 = \{\emptyset, \{HHH, HHT, HTH, HTT\}, \{THH, THT, TTH, TTT\}, \Omega\}$

$\mathcal{F}_2 =$ $\{\emptyset, \{HHH, HHT\}, \{HTH, HTT\}, \{THH, THT\}, \{TTH, TTT\}, \Omega,$ *and their unions*$\}$

$\mathcal{F}_3 = \mathcal{F} =$ *The power set of* Ω

Define $P(A) = \frac{1}{8}$(the number of elements in A). Then each $(\Omega, \mathcal{F}_k, P), k = 0, 1, 2, 3,$ is a probability space.

Remark 20. *In probability, the σ-field is interpreted as the degree of information we have about the result of an experiment.*

Example 9. [Counting measure] *Let $\Omega = \{\omega_1, \omega_2, \ldots\}$ and $\mathcal{F}$ be the power set of Ω. Here, the counting measure is σ-finite. Let $p_1, p_2, \ldots$ be nonnegative numbers such that $\sum_i p_i = 1$. For each $A = \{\omega_{k_1}, \omega_{k_2}, \ldots\}$, define $P(A) = \sum_j p_{k_j}$. Then P is a probability measure, and is absolutely continuous with respect to the counting measure.*

Example 10. [Lebesgue measure] *Let $\Omega = (0, 1]$. Then $\mathcal{F} = \{\emptyset, (0, c], (c, 1], \Omega\}, 0 < c < 1$, is a σ-field. The Lebesgue measure is then a probability.*

Example 11. [Density] *Consider a measurable space $(\mathbb{R}^n, \mathcal{B}_n)$ and the Lebesgue measure μ. Assume that f is a measurable nonnegative function such that $\int f d\mu = 1$, and define $P(A) = \int_A f d\mu$ for each $A \in \mathcal{B}_n$. Then $(\mathbb{R}^n, \mathcal{B}_n, P)$ is a probability space. We call f the density of P with respect to μ.*

Probability Theory: Random Variables

Definition 19. [Random variable] *Let $(\Omega, \mathcal{F}, P)$ be a probability space. A $\mathcal{F}$-measurable function $X : \Omega \mapsto \mathbb{R}^n$ is called a random variable or random vector. We usually denote $P(X^{-1}(B)) = P(X \in B)$, the probability that X is in B.*

Remark 21. *The probability space is essentially a mathematical construct, which may be quite abstract. We therefore introduce mappings X from Ω to $\mathbb{R}^n$, the space we understand easily. With a random variable X properly defined, we may consider $(\mathbb{R}^n, \mathcal{B}_n, \tilde{P})$ as a proper probability space to work in, the probability space induced by X, where $\tilde{P}(B) = P(X \in B)$ for $B \in \mathcal{B}_n$. (see proposition 8.)*

Probability Theory: Expectation

Definition 20. [Expected value and variance] *Let* $\mathbf{X} = (X_1, \ldots, X_n)^\top$ *be a random vector. We define*

$$\mu = E(\mathbf{X}) = \int \mathbf{X} dP = (\int X_1 dP, \ldots, \int X_n dP)^\top$$

and call it the expected value (or mean) of $\mathbf{X}$*. The* n *by* n *matrix*

$$\mathbf{\Sigma} = Var(\mathbf{X}) = \int (\mathbf{X} - \mu)(\mathbf{X} - \mu)^\top dP$$

where the integral is also taken componentwise, is called the variance of $\mathbf{X}$*.*

Definition 21. [Distribution and density] *The distribution function of* $\mathbf{X}$ *is the function* $F = F_{\mathbf{X}} : \mathbb{R}^n \mapsto [0, 1]$ *defined by*

$$F(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x}) \text{ for all } \mathbf{x} \in \mathbb{R}^n$$

If there is a nonnegative integrable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ *such that*

$$F(\mathbf{x}) = \int_{\{\mathbf{y} \leq \mathbf{x}\}} f(\mathbf{y}) \mu(d\mathbf{y})$$

then it is called the density function of $\mathbf{X}$ *with respect to* μ*. It follows that*

$$P(\mathbf{X} \in B) = \int_B f(\mathbf{x})\mu(d\mathbf{x}), \text{ for all } B \in \mathcal{B}_n$$

Remark 22. *The distribution* $P(\mathbf{X} \in B)$ *is the probability measure on* $(\mathbb{R}^n, \mathcal{B}_n)$*.*

Remark 23. *The integral is an ordinary sum (discrete random variables), when* μ *is the counting measure, or an ordinary integral (continuous random variables) when* μ *is the Lebesgue measure.*

Theorem 8. *Let* $\mathbf{X} : \Omega \mapsto \mathbb{R}^n$ *be a random variable with distribution function* F *and density* f *with respect to* μ*. Suppose* $g : \mathbb{R}^n \mapsto \mathbb{R}$*, and* $Y = g(\mathbf{X})$ *is integrable. Then*

$$E(Y) = \int_{\mathbb{R}^n} g(\mathbf{x}) f(\mathbf{x})\mu(d\mathbf{x}).$$

In particular,

$$E(\mathbf{X}) = \int_{\mathbb{R}^n} \mathbf{x} f(\mathbf{x})\mu(d\mathbf{x}), \; Var(\mathbf{x}) = \int_{\mathbb{R}^n} (\mathbf{x} - E(\mathbf{X}))(\mathbf{x} - E(\mathbf{X}))^\top f(\mathbf{x})\mu(d\mathbf{x}).$$

Exponential Families: Definition

Definition 22. [Exponential family] *Let ν be a σ-finite measure on $(\mathbb{R}^n, \mathcal{B}_n)$. Let*

$$\Theta = \{\theta \in \mathbb{R}^d | \int_{\mathbb{R}^n} h(\mathbf{x}) \exp\langle \theta, T(\mathbf{x})\rangle \nu(d\mathbf{x}) < \infty\}.$$

The quantity A, known as the log partition function, is defined by the integral

$$A(\theta) = \log \int_{\mathbb{R}^n} h(\mathbf{x}) \exp\langle \theta, T(\mathbf{x})\rangle \nu(d\mathbf{x})$$

and define

$$p(\mathbf{x}; \theta) = h(\mathbf{x}) \exp\{\langle \theta, T(\mathbf{x})\rangle - A(\theta)\}$$

The family of probability densities $\{p(\mathbf{x}; \theta) | \theta \in \Theta\}$ is called a d-dimensional exponential family of probability densities. Θ is called the natural parameter space. $\theta \in \Theta$ is referred to as a canonical parameter, and $T(\mathbf{x})$ is known as a sufficient statistic.

Definition 23. [Regular exponential family] *An exponential family for which the natural parameter space Θ is an open set is called regular. Herein, we restrict our attention to regular exponential families.*

Definition 24. [Minimal representation] *If there is no vector* $\mathbf{a} \in \mathbb{R}^d$ *and constant* $b \in \mathbb{R}$ *such that* $\langle \mathbf{a}, T(\mathbf{x}) \rangle = b$ *holds* ν*-a.e., the representation is said to be minimal.*

Definition 25. [Overcomplete representation] *If the representation is not minimal, it is called overcomplete.*

Exponential Families: Basic Convexity

Proposition 14. *Θ is a convex set.*

Proof: Let $\theta_1, \theta_2 \in \Theta, 0 < \alpha < 1$. Then, by Hölder's inequality,

$$\begin{aligned}
\int h(\mathbf{x}) \exp\langle \alpha\theta_1 + (1-\alpha)\theta_2, T(\mathbf{x})\rangle \nu(d\mathbf{x}) &= \int h(\mathbf{x})^{\alpha} (\exp\langle \theta_1, T(\mathbf{x})\rangle)^{\alpha} h(\mathbf{x})^{1-\alpha} (\exp\langle \theta_2, T(\mathbf{x})\rangle)^{1-\alpha} \nu(d\mathbf{x}) \\
&\leq \left[\int h(\mathbf{x}) \exp\langle \theta_1, T(\mathbf{x})\rangle \nu(d\mathbf{x})\right]^{\alpha} \left[\int h(\mathbf{x}) \exp\langle \theta_1, T(\mathbf{x})\rangle \nu(d\mathbf{x})\right]^{1-\alpha} \\
&< \infty. \blacksquare
\end{aligned}$$

Definition 26. [Lower semi-continuous] *A function $f : dom(f) \subset \mathbb{R}^n \mapsto \mathbb{R}$ is called lower semi-continuous at $\mathbf{x} \in dom(f)$ if $f(\mathbf{x}) \leq \liminf_n f(\mathbf{x}_n)$ for every sequence $\{\mathbf{x}_n\} \subset dom(f)$ that converges to $\mathbf{x}$.*

Proposition 15. *$A(\theta)$ is lower semi-continuous on $\mathbb{R}^d$, and C^{∞} on Θ.*

Proof: By Fatou's lemma,

$$\begin{aligned}\int h(\mathbf{x}) \exp\langle \theta, T(\mathbf{x})\rangle \nu(d\mathbf{x}) &= \int h(\mathbf{x}) \liminf_n \exp\langle \theta_n, T(\mathbf{x})\rangle \nu(d\mathbf{x}) \\ &\leq \liminf_n \int h(\mathbf{x}) \exp\langle \theta_n, T(\mathbf{x})\rangle \nu(d\mathbf{x})\end{aligned}$$

Hence, $A(\theta) \leq \liminf_n A(\theta_n)$. ■

Proposition 16. *The derivatives of $A(\theta)$ are the cumulants of the random vector $T(\mathbf{x})$. In particular*

$$\begin{aligned}\frac{\partial A(\theta)}{\partial \theta_\alpha} &= E_\theta[T_\alpha(\mathbf{x})] \\ \frac{\partial^2 A(\theta)}{\partial \theta_\alpha \partial_\beta} &= E_\theta[T_\alpha(\mathbf{x}) T_\beta(\mathbf{x})] - E_\theta[T_\alpha(\mathbf{x})] E_\theta[T_\beta(\mathbf{x})]\end{aligned}$$

Proof: We can interchange the order of differentiation and integration.

$$\begin{aligned}
\frac{\partial A(\theta)}{\partial \theta_\alpha} &= \frac{\partial}{\partial \theta_\alpha}\left\{\log \int h(\mathbf{x}) \exp\langle \theta, T(\mathbf{x})\rangle \nu(d\mathbf{x})\right\} \\
&= \frac{\int T_\alpha(\mathbf{x}) h(\mathbf{x}) \exp\langle \theta, T(\mathbf{x})\rangle \nu(d\mathbf{x})}{\int h(\mathbf{x}) \exp\langle \theta, T(\mathbf{x})\rangle \nu(d\mathbf{x})} \\
&= \int T_\alpha(\mathbf{x}) h(\mathbf{x}) \{\exp\langle \theta, T(\mathbf{x})\rangle - A(\theta)\} \nu(d\mathbf{x}) \\
&= E_\theta[T_\alpha(\mathbf{x})]. \blacksquare
\end{aligned}$$

Remark 24. *A is also called the cumulant generating function of the random vector $T(\mathbf{x})$.*

Proposition 17. *$A(\theta)$ is a convex function, and strictly convex if the representation is minimal.*

Proof: Since

$$\nabla^2 A(\theta) = E\left(\begin{bmatrix} T_1(\mathbf{x}) - ET_1(\mathbf{x}) \\ \vdots \\ T_d(\mathbf{x}) - ET_d(\mathbf{x}) \end{bmatrix} [T_1(\mathbf{x}) - ET_1(\mathbf{x}), \ \cdots, \ T_d(\mathbf{x}) - ET_d(\mathbf{x})]^\top\right)$$

the hessian $\nabla^2 A(\theta)$ is positive semi-definite on the open set Θ, which ensures convexity. If the representation is minimal, there is no vector $\mathbf{a} \in \mathbb{R}^d$ and constant $b \in \mathbb{R}$ such that $\langle \mathbf{a}, T(\mathbf{x}) \rangle = b$ holds ν-a.e. Then

$$\begin{aligned} var_\theta \left[\langle \mathbf{a}, T(\mathbf{x}) \rangle \right] &= E \left[\mathbf{a}^\top T(\mathbf{x}) - E(\mathbf{a}^\top T(\mathbf{x})) \right]^2 \\ &= \mathbf{a}^\top E[T(\mathbf{x}) T(\mathbf{x})^\top] \mathbf{a} - \mathbf{a}^\top E(T(\mathbf{x})) E(T(\mathbf{x})^\top) \mathbf{a} \\ &= \mathbf{a}^\top \nabla^2 A(\theta) \mathbf{a} > 0 \end{aligned}$$

for all $\mathbf{a} \neq 0 \in \mathbb{R}^d$ and $\theta \in \Theta$. This strict positive definiteness of the Hessian on the open set Θ implies strict convexity. ■

Proposition 18. $\|\nabla^2 A(\theta_t)\| \to \infty$ *for any sequence* $\{\theta_t\} \subset \Theta$ *approaching the boundary.*

Proof: Let θ_b be a boundary point, and let $\theta_0 \in \Theta$ be arbitrary. By the convexity and openness of Θ, the line segment $\theta_t = \alpha \theta_b + (1 - \alpha)\theta_0$ is contained in Θ for all $\alpha \in [0, 1)$. Since $A(\theta)$ is differentiable and convex on Θ,

$$A(\theta_0) \geq A(\theta_t) + \langle \nabla A(\theta_t), \theta_0 - \theta_t \rangle.$$

Then,

$$\begin{aligned} A(\theta_t) - A(\theta_0) &\leq \langle \nabla A(\theta_t), \theta_t - \theta_0 \rangle \\ &\leq |\langle \nabla A(\theta_t), \theta_t - \theta_0 \rangle| \\ &\leq ||\theta_t - \theta_0||\, ||\nabla A(\theta_t)||. \end{aligned}$$

As $\alpha \to 1-$, $A(\theta_t) \to \infty$ by the lower semi-continuity of A. Since $||\theta_t - \theta_0||$ is bounded, $||\nabla^2 A(\theta_t)|| \to \infty$. ■

Exponential Families: Legendre Duality

Definition 27. [Conjugate] *Let* $f : dom(f) \subset \mathbb{R}^n \mapsto \mathbb{R}$. *The function* $f^* : \mathbb{R}^n \mapsto \mathbb{R}$, *defined as*

$$f^*(\mathbf{y}) = \sup_{\mathbf{x} \in dom(f)} \{\langle \mathbf{y}, \mathbf{x} \rangle - f(\mathbf{x})\}$$

is called the conjugate of the function f. *The domain of the conjugate function consists of* $\mathbf{y} \in \mathbb{R}^n$ *for which the supremum is finite. The conjugate of a differentiable function is also called the Legendre transform of* f.

Remark 25. f^* *is a convex function since it is the pointwise supremum of a family of convex functions of* $\mathbf{y}$.

References

[Topology & Real Analysis]

[1] W. Rudin. Principles of mathematical analysis. 1976.

[2] H. L. Royden. Real analysis. 1988.

[Convex Analysis]

[3] R. T. Rockafellar. Convex analysis. 1970.

[4] D. P. Bertsekas, A. Nedic, and A. E. Ozdaglar. Convex analysis and optimization. 2003.

[5] S. Boyd and L. Vandenberghe. Convex optimization. 2004.

[Measure Theory]

[6] I. K. Rana. An introduction to measure and integration. 2002.

[7] S. Choi. Lecture notes on introduction to stochastic calculus and mathematical finance. POSTECH. 2005.

[Probability Theory]

[8] P. Bilingsley. Probability and measure. 1995

[Exponential Family]

[9] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and

variational inference. Technical Report 649, Department of Statistics, University of California, Berkeley, 2003.
[10] L. D. Brown. Fundamentals of statistical exponential families. 1986.

[Bregman Divergence]
[11] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. Journal of machine learning research 6. 1705-1749. 2005.
[12] Y. Censor and S. A. Zenios. Parallel optimization. 1997.