Recap

- Transition Kernel: $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$ are measurable spaces.

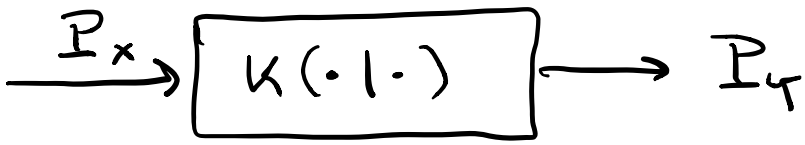
$K(\cdot|\cdot)$ is a transition kernel from space one to space two if:

(1) $K(\cdot|x) \in P_{\mathcal{G}}(\mathcal{Y})$ ← set of all probability measures over space $\mathcal{Y}$ wrt $\mathcal{G}$ $\quad \forall x \in \mathcal{X}$

(2) $K(B|\cdot) : \mathcal{X} \to \mathbb{R} \quad \forall B \in \mathcal{G}$

is a rv wrt $(\mathcal{X}, \mathcal{F})$ space

Interpretation: A transition kernel transforms a $P_X \in P_{\mathcal{F}}(\mathcal{X})$ into $P_Y \in P_{\mathcal{G}}(\mathcal{Y})$



$$P_Y(B) := \mathbb{E}_{P_X}[K(B|X)] = \int_{\mathcal{X}} K(B|x)\, dP_X(x) \quad \forall B \in \mathcal{G}$$

There is a close relation between transition kernels and joint distributions that will give rise to decompositions such as $P_{XY} = P_X P_{Y|X}$

Proposition: Let $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$ be measurable spaces.

$P_x \in \mathcal{P}_{\mathcal{F}}(\mathcal{X})$ and $K(\cdot|\cdot)$ be a transition kernel from $(\mathcal{X}, \mathcal{F})$ to $(\mathcal{Y}, \mathcal{G})$.
Then, $\exists!$ probability measure P_{xy} on $(\mathcal{X} \times \mathcal{Y}, \mathcal{F} \otimes \mathcal{G})$ such that

$$P_{xy}((A,B)) = \int_A K(B|x)\, dP_x(x) \quad \forall (A,B) \in \mathcal{F} \otimes \mathcal{G}$$

Conversely, given a probability space $(\mathcal{X} \times \mathcal{Y}, \mathcal{F} \otimes \mathcal{G}, P_{xy})$ $\exists!$ pair comprising $P_x \in \mathcal{P}_{\mathcal{F}}(\mathcal{X})$ and a transition kernel $K(\cdot|\cdot)$ s.t. the previous equation holds.

comment

Henceforth, we will write $P_{xy} = P_x P_{y|x} = P_y P_{x|y}$ while understanding this equality wrt the previous proposition.

Conditional Expectations

Have $(X, Y) \sim \mathbb{P}_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$

Definition:

$$\mathbb{E}[Y \mid X = x_0] := \int_{\mathcal{Y}} y \, d\mathbb{P}_{Y|X}(y|x_0)$$

↙ a number, deterministic, function of x_0

$$\mathbb{E}[Y|X] := \int_{\mathcal{Y}} y \, d\mathbb{P}_{Y|X}(y|X)$$

↙ a random variable

Proposition: Let $(X, Y) \sim \mathbb{P}_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. $\exists!$ measurable function $h: \mathcal{X} \to \mathcal{Y}$ such that

(i) $h(x) = \mathbb{E}[Y|X = x]$ $\quad \forall x \in \mathcal{X}$

(ii) $h(X) = h \circ X = \mathbb{E}[Y|X]$

Example/Exercise: $X = \begin{cases} 1 & \text{w.p. } 1/2 \\ -1 & \text{w.p. } 1/2 \end{cases}$ $\perp\!\!\!\perp$ $Z \sim N(0, \sigma^2)$

$Y \overset{d}{=} X + Z$. $\quad \mathbb{E}[X|Y] \overset{?}{=} \ldots = \tanh\left(\frac{Y}{\sigma^2}\right) = h(Y)$

Theorem: Law of Total Expectation

Let $(X, Y) \sim P_{XY} = P_X P_{Y|X} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. Then

$$\mathbb{E}_{P_Y}[Y] = \mathbb{E}_{P_X}\left[\mathbb{E}_{P_{Y|X}}[Y|X]\right]$$

$$= \int_{\mathcal{X}} h(x)\, dP_X(x)$$

$$= \int_{\mathcal{X}} \mathbb{E}[Y|X=x]\, dP_X(x)$$

$$\left(\mathbb{E}_{P_Y}[Y] = \mathbb{E}_{P_X}\left[\mathbb{E}_{P_{Y|X}}[Y|X]\right] = \sum_{x \in \mathcal{X}} P_X(x)\, \mathbb{E}_{P_{Y|X=x}}[Y|X=x]\right.$$

$$\left. = \sum_{x \in \mathcal{X}} P_X(x) \left(\sum_{y \in \mathcal{Y}} y \cdot P_{Y|X}(y|x)\right)\right)$$

f-divergences

Goal: We want to develop a way to measure a reasonable notion of "distance" between probability distributions.

Definition (Divergence): Let δ be a functional

$$\delta : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \longrightarrow \mathbb{R}_{\geq 0} \cup \{\infty\}$$

We will call δ a divergence if $\delta(P, Q) = 0 \Leftrightarrow P = Q$

Definition (Metric): If a divergence δ also satisfies

(i) $\delta(P, Q) = \delta(Q, P)$ — Symmetry

(ii) $\delta(P, Q) \leq \delta(P, R) + \delta(R, Q)$ $\forall P, Q, R \in \mathcal{P}(\mathcal{X})$ — Triangle inequality

then δ is a metric

Why do f-divergences?

Aside: Generative Modeling

Target P

Model Class $\{Q_\theta\}_{\theta \in \Theta}$

"Learn a Q_θ that is a good proxy of P..."

Formulate: Pick δ divergence

$$\inf_{\theta \in \Theta} \delta(Q_\theta, P)$$

AMAZING

Motivates Generative adversarial networks!

A large class of divergences falls under the framework of f-divergences. This includes

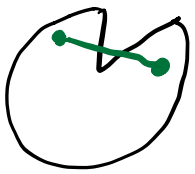
- Kullback Leibler distance
- Total Variation distance
- Hellinger distance
- χ^2 divergence
- Lecoin distance

Primer: Convexity

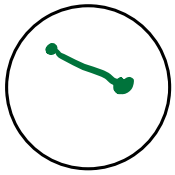
A subset K of a vector space V is convex if

$$\alpha x + (1-\alpha) y \in K \quad \forall x, y \in K \; ; \; \alpha \in [0,1]$$

Non-Convex



Convex



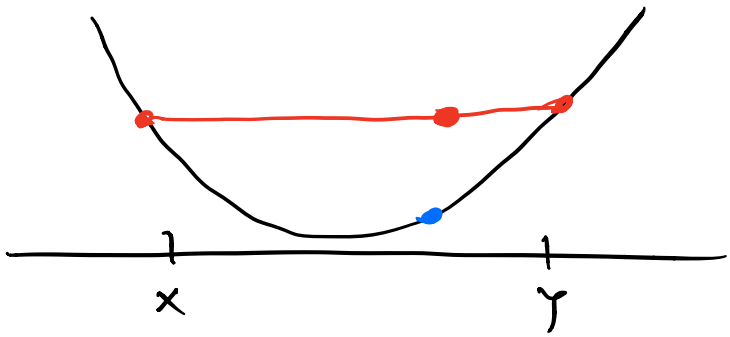
Example: $\mathcal{P}(X)$ is convex

$$(\alpha P_1 + (1-\alpha) P_2)(A) \stackrel{*}{=} \alpha P_1(A) + (1-\alpha) P_2(A)$$

Definition: A function $f: K \to \mathbb{R}$, where $K \subseteq \mathbb{R}^d$ is convex is convex if it satisfies

$$f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y) \quad \forall x, y \in K, \ \alpha \in [0,1]$$

Illustration



Comments: ① f is strictly convex if the above inequality is STRICT

② f is concave if $-f$ is convex

Def (epigraph):
The epigraph of a function $f: K \to \mathbb{R}$ is

$$\text{epi}(f) := \{(x, y) \in K \times \mathbb{R} \mid y \geq f(x)\}$$

Prop: $f: K \to \mathbb{R}$ is convex $\Leftrightarrow$ epi(f) a convex set

Prop (Operations that conserve convexity): Let $f_1, f_2 : \mathbb{R}^d \to \mathbb{R}$ be convex. Then:

1. $f_1 + f_2$ is convex
2. $\max\{f_1, f_2\}$ is convex
3. $A \in \mathbb{R}^{d \times n}$, $g : \mathbb{R}^n \to \mathbb{R}$, $g(x) = f_1(Ax)$ is convex

Proposition: $f : K \to \mathbb{R}$, $K \subseteq \mathbb{R}^d$ is convex iff its Hessian is Positive Semi Definite.

$$\text{Hess}(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix}$$

Note: $A \in \mathbb{R}^{n \times n}$ ($A = A^T$) is psd if $x^T A x \geq 0 \quad \forall x \in \mathbb{R}^n$

Corollary: $f : \mathbb{R} \to \mathbb{R}$

1. f convex iff $\frac{\partial^2 f}{\partial x^2} \geq 0$
2. f strictly convex iff $\frac{\partial^2 f}{\partial x^2} > 0$

Theorem (Jensen's Inequality)

Let $X \sim P_X \in \mathcal{P}(\mathcal{X})$ and $f: \mathcal{X} \to \mathbb{R}$. Then

① If f is convex

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

② If f is strictly convex

$$\mathbb{E}[f(X)] > f(\mathbb{E}[X])$$

unless X is almost surely constant

③ If f is concave

$$\mathbb{E}[f(X)] < f(\mathbb{E}[X])$$