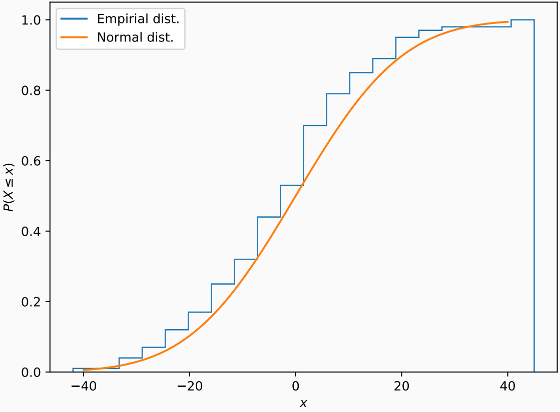
Optimal Bounds between f-Divergences and Integral Probability Metrics

Rohit Agrawal (Harvard)
Thibaut Horel (MIT)

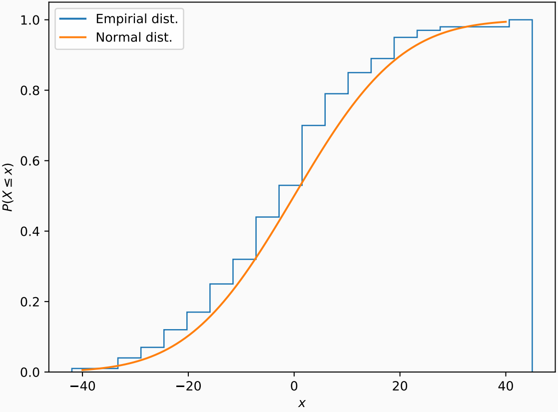
Motivation



Is the empirical distribution approximately normal?

What is the normal distribution best approximating it?

Motivation



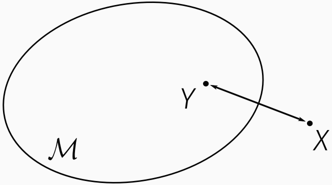
Is the empirical distribution approximately normal?

What is the normal distribution best approximating it?

Motivation

Typical learning procedure:

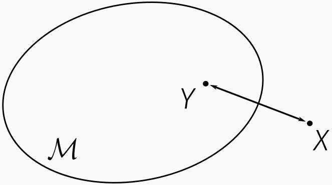
$$\text{given} \quad \left\{ \begin{array}{l} \text{observations } X \\ \text{model class } \mathcal{M} \\ \text{cost function } \mathrm{D}(\cdot \| \cdot) \end{array} \right\} \quad \text{solve} \quad \min_{Y \in \mathcal{M}} \mathrm{D}(X \| Y)$$



Motivation

Typical learning procedure:

$$\text{given} \quad \left\{ \begin{array}{l} \text{observations } X \\ \text{model class } \mathcal{M} \\ \text{cost function } \; \mathrm{D}(\cdot \| \cdot) \end{array} \right\} \quad \text{solve} \quad \min_{Y \in \mathcal{M}} \mathrm{D}(X \| Y)$$

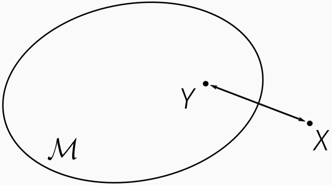


Example: if $\mathrm{D}(X \| Y)$ is the Kullback–Leibler divergence
$\Rightarrow$ maximum likelihood estimation

Motivation

Typical learning procedure:

$$\text{given} \quad \left\{\begin{array}{l} \text{observations } X \\ \text{model class } \mathcal{M} \\ \text{cost function } \mathrm{D}(\cdot\|\cdot) \end{array}\right\} \quad \text{solve} \quad \min_{Y \in \mathcal{M}} \mathrm{D}(X\|Y)$$



Example: if $\mathrm{D}(X\|Y)$ is the Kullback–Leibler divergence
$\Rightarrow$ maximum likelihood estimation

Problem: what statistical guarantees are implied by $\mathrm{D}(X\|Y) \leq \varepsilon$?

Measures of similarity for random variables

How "close" to each other are X and Y?

ϕ-divergences

$$\mathrm{D}_\phi(X \| Y) = \mathbb{E}_{y \sim Y}\left[\phi\left(\frac{\mathbb{P}[X = y]}{\mathbb{P}[Y = y]}\right)\right]$$

for convex ϕ with $\phi(1) = 0$

Ex: Kullback–Leibler (KL) div., χ^2-div., Hellinger dist., α-div., etc.

integral probability metrics

$$\mathrm{d}_{\mathcal{F}}(X, Y) = \sup_{f \in \mathcal{F}} \left|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]\right|$$

class $\mathcal{F}$ of "test" functions

Ex: total variation dist., max. mean discrepancy, etc.

What is the best lower bound of $\mathrm{D}_\phi(X\|Y)$
in terms of $\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]$*?*

Result

Theorem (Informal)

There exists an explicit function $K_{f(Y)} : \mathbb{R} \to \mathbb{R}$ *associated with* $f(Y)$ *inducing a correspondence between*

1. *lower bounds* $\mathrm{D}_\phi(X \| Y) \geq L\big(\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]\big)$ *for all* X

and

2. *upper bounds* $K_{f(Y)}(t) \leq B(t)$ *for all* $t \in \mathbb{R}$

Result

Theorem (Informal)

There exists an explicit function $K_{f(Y)} : \mathbb{R} \to \mathbb{R}$ associated with $f(Y)$ inducing a correspondence between

1. *lower bounds* $\mathrm{D}_\phi(X \| Y) \geq L\big(\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]\big)$ *for all X*

and

2. *upper bounds* $K_{f(Y)}(t) \leq B(t)$ *for all $t \in \mathbb{R}$*

Ex: for the KL divergence, $K_{f(Y)}$ is the log moment-generating function

Cumulant-generating function

For a given ϕ-divergence, define:

- the convex conjugate $\phi^\star(y) = \sup_{x \geq 0} \left\{ x \cdot y - \phi(x) \right\}$
- the ϕ-cumulant-generating function of $f(Y)$

$$K_{f(Y)}(t) = \inf_{\lambda \in \mathbb{R}} \mathbb{E}\left[\phi^\star(t \cdot f(Y) + \lambda) - t \cdot f(Y) - \lambda\right]$$

Cumulant-generating function

For a given ϕ-divergence, define:

- the convex conjugate $\phi^\star(y) = \sup_{x \geq 0} \{x \cdot y - \phi(x)\}$
- the ϕ-cumulant-generating function of $f(Y)$

$$K_{f(Y)}(t) = \inf_{\lambda \in \mathbb{R}} \mathbb{E}\big[\phi^\star(t \cdot f(Y) + \lambda) - t \cdot f(Y) - \lambda\big]$$

Example: for the KL divergence, $\phi(x) = x \log x$ and:

- $\phi^\star(y) = e^{y-1}$
- we recover the (centered) cumulant-generating function

$$K_{f(Y)}(t) = \log \mathbb{E}\left[e^{t \cdot f(Y) - t \cdot \mathbb{E}[f(Y)]}\right]$$

Result

Theorem

The following are equivalent:

1. $K_{f(Y)}(t) \leq B(t)$ *for all* $t \in \mathbb{R}$
2. $\mathrm{D}_\phi(X \| Y) \geq B^\star\big(\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]\big)$ *for all* X

where

$$K_{f(Y)}(t) = \inf_{\lambda \in \mathbb{R}} \mathbb{E}\big[\phi^\star(t \cdot f(Y) + \lambda) - t \cdot f(Y) - \lambda\big]$$

and $\star$ *denotes the convex conjugate*

Result

Theorem

The following are equivalent:

1. $K_{f(Y)}(t) \leq B(t)$ *for all* $t \in \mathbb{R}$
2. $\mathrm{D}_\phi(X\|Y) \geq B^\star\big(\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]\big)$ *for all* X

where

$$K_{f(Y)}(t) = \inf_{\lambda\in\mathbb{R}} \mathbb{E}\big[\phi^\star(t\cdot f(Y)+\lambda) - t\cdot f(Y) - \lambda\big]$$

and $\star$ *denotes the convex conjugate*

Key technique: use convex analysis to obtain variational representations of $\mathrm{D}_\phi(X\|Y)$

Applications and examples

1. for the KL divergence, if f takes values in $[-1, 1]$:

$$K_{f(Y)}(t) = \log \mathbb{E}\left[e^{t \cdot f(Y) - t \cdot \mathbb{E}[f(Y)]}\right] \leq \frac{t^2}{2} \quad \text{(Hoeffding's lemma)}$$

$\Rightarrow \mathrm{D}(X \| Y) \geq \frac{1}{2}\big(\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]\big)^2$ (Pinsker's inequality)

Holds more generally if $f(Y)$ is subgaussian

Applications and examples

1. for the KL divergence, if f takes values in $[-1,1]$:

$$K_{f(Y)}(t) = \log \mathbb{E}\left[e^{t\cdot f(Y) - t\cdot \mathbb{E}[f(Y)]}\right] \leq \frac{t^2}{2} \quad \text{(Hoeffding's lemma)}$$

$\Rightarrow \mathrm{D}(X\|Y) \geq \frac{1}{2}\left(\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]\right)^2$ (Pinsker's inequality)

Holds more generally if $f(Y)$ is subgaussian

2. "Pinkser's type" inequality for all α-divergences (Rényi divergences)

Applications and examples

1. for the KL divergence, if f takes values in $[-1, 1]$:

   $$K_{f(Y)}(t) = \log \mathbb{E}\left[e^{t \cdot f(Y) - t \cdot \mathbb{E}[f(Y)]}\right] \leq \frac{t^2}{2} \quad \text{(Hoeffding's lemma)}$$

   $\Rightarrow \mathrm{D}(X \| Y) \geq \frac{1}{2}\left(\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]\right)^2$ (Pinsker's inequality)
   Holds more generally if $f(Y)$ is subgaussian

2. "Pinkser's type" inequality for all α-divergences (Rényi divergences)

3. Negative result, when $\lim_{x \to \infty} \phi(x)/x < \infty$:
   $f(Y)$ unbounded $\Rightarrow$ no nontrivial lower bound

Conclusion

- complete description of optimal lower bounds of ϕ-divergences in terms of IPMs

Conclusion

- complete description of optimal lower bounds of ϕ-divergences in terms of IPMs
- results of independent interest on topological properties of ϕ-divergences

Conclusion

- complete description of optimal lower bounds of ϕ-divergences in terms of IPMs
- results of independent interest on topological properties of ϕ-divergences
- tools and techniques could be more broadly applied

Conclusion

- complete description of optimal lower bounds of ϕ-divergences in terms of IPMs
- results of independent interest on topological properties of ϕ-divergences
- tools and techniques could be more broadly applied

Thanks!