Self-study: Variational Inference via Divergence Minimization

Tianpei Xie

Nov. 9th., 2022

Contents

1 Statistical Divergence

1.1 Definitions

- **Definition** Given a *differentiable manifold* $\mathcal{M}$ of dimension n, a ***divergence*** on $\mathcal{M}$ is a C^2-function $\mathbb{D} : \mathcal{M} \times \mathcal{M} \to [0, \infty)$ satisfying:

  1. (***non-negativity***) $\mathbb{D}\,(p \,\|\, q) \ge 0$ for all $p, q \in \mathcal{M}$;
  2. (***positivity***) $\mathbb{D}\,(p \,\|\, q) = 0$ if and only if $p = q$;
  3. At every point $p \in \mathcal{M}$, $\mathbb{D}\,(p \,\|\, p + dp)$ is a ***positive-definite*** quadratic form for infinitesimal displacements dp from p.

  The last property means that divergence defines an *inner product* on the ***tangent space*** $T_p\mathcal{M}$ for every $p \in \mathcal{M}$. Since $\mathbb{D}$ is C^2 on $\mathcal{M}$, this defines a ***Riemannian metric*** g on $\mathcal{M}$.

- **Definition** Let p, q be $\mathbb{R}^d \supset \mathcal{M}_0 :\to \mathbb{R}$ density functions and let $\alpha \in \mathbb{R} \setminus \{1\}$. The ***Rényi divergence*** of order α or **α−divergence** of a distribution p from a distribution q is defined to be

$$\mathbb{D}^{\alpha}\,(p \,\|\, q) = \frac{1}{\alpha - 1} \log\left(\mathbb{E}_Q\left[\left(\frac{dP}{dQ}\right)^{\alpha}\right]\right) = \frac{1}{\alpha - 1} \log\left(\int_{\mathcal{M}_0} p^{\alpha}(x) q^{1-\alpha}(x)\; \mu(dx)\right) \tag{1}$$

- **Definition** Let P and Q be two probability distributions over a space Ω, such that $P \ll Q$, that is, P is ***absolutely continuous*** with respect to Q. Then, for a ***convex function*** $f : [0, +\infty) \to (-\infty, +\infty]$ such that $f(x)$ is finite for all $x > 0$, $f(1) = 0$, and $f(0) = \lim_{t \to 0^+} f(t)$ (which could be infinite), the ***f-divergence*** of P from Q is defined as

$$\mathbb{D}^{f}\,(P \,\|\, Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right] = \int_{\Omega} f\left(\frac{dP}{dQ}\right) dQ = \int_{\Omega} q(x) f\left(\frac{p(x)}{q(x)}\right) \mu(dx) \tag{2}$$

  The convex function f is referred as ***generator function***.

- **Definition** Let $F : \mathcal{X} \to \mathbb{R}$ be a *continuously-differentiable*, ***strictly convex*** function defined on a convex set $\mathcal{X}$. The ***Bregman divergence*** associated with F for points $p, q \in \mathcal{X}$ is the difference between the value of F at point p and the value of the *first-order Taylor expansion* of F around point q evaluated at point p:

$$\mathbb{D}^{F}\,(p \,\|\, q) = F(p) - F(q) - \langle \nabla F(q)\,,\, p - q \rangle \tag{3}$$

- **Definition** We suppose $\mathcal{X} = \mathcal{Y}$ and that for some $p \ge 1$, $c(x, y) = d(x, y)^p$, where d is a distance on $\mathcal{X}$, the **p-Wasserstein distance** between measures α, β on $\mathcal{X}$ is $\mathcal{W}_p(\alpha, \beta)$, where

$$(\mathcal{W}_p(\alpha, \beta))^p := \min_{\substack{(X,Y) \sim \pi; \\ X_{\#}\pi = \alpha, \\ Y_{\#}\pi = \beta}} \mathbb{E}_{(X,Y)}\left[d(X, Y)^p\right] \tag{4}$$

1.2 KL Divergence for Exponential Families

- The canonical representation of ***exponential famlity*** of distribution has the following form

$$\begin{aligned} p(x_1, \ldots, x_m) = p(\boldsymbol{x}; \boldsymbol{\eta}) &= \exp\left(\langle \boldsymbol{\eta}\,,\, \boldsymbol{\phi}(\boldsymbol{x})\rangle - A(\boldsymbol{\eta})\right) h(\boldsymbol{x})\nu(d\boldsymbol{x}) \\ &= \exp\left(\sum_{\alpha} \eta_\alpha \phi_\alpha(\boldsymbol{x}) - A(\boldsymbol{\eta})\right) \end{aligned} \tag{5}$$

  where ϕ is a feature map and $\boldsymbol{\phi}(\boldsymbol{x})$ defines a set of ***sufficient statistics*** (or ***potential functions***). The normalization factor is defined as

$$A(\boldsymbol{\eta}) := \log \int \exp\left(\langle \boldsymbol{\eta}\,,\, \boldsymbol{\phi}(\boldsymbol{x})\rangle\right) h(\boldsymbol{x})\nu(d\boldsymbol{x}) = \log Z(\boldsymbol{\eta})$$

  $A(\boldsymbol{\eta})$ is also referred as ***log-partition function*** or *cumulant function*. The parameters $\boldsymbol{\eta} = (\eta_\alpha)$ are called ***natural parameters*** or *canonical parameters*. The canonical parameter $\{\eta_\alpha\}$ forms a **natural (canonical) parameter space**

$$\Omega = \left\{\boldsymbol{\eta} \in \mathbb{R}^d : A(\boldsymbol{\eta}) < \infty\right\} \tag{6}$$

- The exponential family is the unique solution of ***maximum entropy estimation*** problem:

$$\min_{q \in \Delta} \quad \mathbb{KL}\left(q \,\|\, p_0\right) \tag{7}$$
$$\text{s.t.} \quad \mathbb{E}_q\left[\phi_\alpha(X)\right] = \mu_\alpha \quad \forall\, \alpha \in \mathcal{I} \tag{8}$$

  where $\mathbb{KL}\left(q \,\|\, p_0\right) = \int \log(\frac{q}{p_0}) q dx = \mathbb{E}_q\left[\log \frac{q}{p_0}\right]$ is the relative entropy or the Kullback-Leibler divergence of q w.r.t. p_0.

  Here $\boldsymbol{\mu} = (\mu_\alpha)_{\alpha \in \mathcal{I}}$ is a set of ***mean parameters***. The space of mean parameters $\mathcal{M}$ is a *convex polytope* spanned by potential functions $\{\phi_\alpha\}$.

$$\mathcal{M} := \left\{\boldsymbol{\mu} \in \mathbb{R}^d : \exists q \text{ s.t. } \mathbb{E}_q\left[\phi_\alpha(X)\right] = \mu_\alpha \quad \forall\, \alpha \in \mathcal{I}\right\} = \operatorname{conv}\left\{\phi_\alpha(x),\ x \in \mathcal{X},\ \alpha \in \mathcal{I}\right\} \tag{9}$$

- Moreover $A(\boldsymbol{\eta})$ has a variational form

$$A(\boldsymbol{\eta}) = \sup_{\boldsymbol{\mu} \in \mathcal{M}} \left\{\langle \boldsymbol{\eta}\,,\, \boldsymbol{\mu}\rangle - A^*(\boldsymbol{\mu})\right\} \tag{10}$$

  where $A^*(\boldsymbol{\mu})$ is the conjugate dual function of A and it is defined as

$$A^*(\boldsymbol{\mu}) := \sup_{\boldsymbol{\eta} \in \Omega} \left\{\langle \boldsymbol{\mu}\,,\, \boldsymbol{\eta}\rangle - A(\boldsymbol{\eta})\right\} \tag{11}$$

  It is shown that $A^*(\boldsymbol{\mu}) = -H(q_{\boldsymbol{\eta}(\boldsymbol{\mu})})$ for $\boldsymbol{\mu} \in \mathcal{M}^\circ$ which is the negative entropy. $A^*(\boldsymbol{\mu})$ is also the optimal value for the **maximum likelihood estimation** problem on p. The exponential family can be reparameterized according to its mean parameters $\boldsymbol{\mu}$ via backward mapping $(\nabla A)^{-1} : \mathcal{M}^\circ \to \Omega$, called **mean parameterization**.

- We can formulate the **KL divergence** between two distributions in exponential family Ω using its primal and dual form

– **Primal-form**: given $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \Omega$

$$\begin{aligned}\mathbb{KL}\left(p_{\boldsymbol{\eta}_1} \,\|\, p_{\boldsymbol{\eta}_2}\right) \equiv \mathbb{KL}\left(\boldsymbol{\eta}_1 \,\|\, \boldsymbol{\eta}_2\right) &= A(\boldsymbol{\eta}_2) - A(\boldsymbol{\eta}_1) - \langle \boldsymbol{\mu}_1\,, \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1 \rangle \\ &\equiv A(\boldsymbol{\eta}_2) - A(\boldsymbol{\eta}_1) - \langle \nabla A(\boldsymbol{\eta}_1)\,, \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1 \rangle\end{aligned} \tag{12}$$

– **Primal-dual form**: given $\boldsymbol{\mu}_1 \in \mathcal{M}, \boldsymbol{\eta}_2 \in \Omega$

$$\mathbb{KL}\left(\boldsymbol{\mu}_1 \,\|\, \boldsymbol{\eta}_2\right) = A(\boldsymbol{\eta}_2) + A^*(\boldsymbol{\mu}_1) - \langle \boldsymbol{\mu}_1\,, \boldsymbol{\eta}_2 \rangle \tag{13}$$

– **Dual-form**: given $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathcal{M}$

$$\begin{aligned}\mathbb{KL}\left(\boldsymbol{\mu}_1 \,\|\, \boldsymbol{\mu}_2\right) &= A^*(\boldsymbol{\mu}_1) - A^*(\boldsymbol{\mu}_2) - \langle \boldsymbol{\eta}_2\,, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \rangle \\ &\equiv A^*(\boldsymbol{\mu}_1) - A^*(\boldsymbol{\mu}_2) - \langle \nabla A^*(\boldsymbol{\mu}_2)\,, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \rangle\end{aligned} \tag{14}$$

- The dual form is related to the *Bregman divergence*, which induce the **projection operation**. We see that dual form $\mathbb{KL}\left(\boldsymbol{\mu}_1 \,\|\, \boldsymbol{\mu}_2\right) = \mathbb{D}^{A^*}\left(\boldsymbol{\mu}_1 \,\|\, \boldsymbol{\mu}_2\right)$, where $F = A^*$ is the negative entropy.

1.3 α-Divergence Properties

See papers in [Hero et al., 2001, Nielsen and Nock, 2011, Póczos and Schneider, 2011].

- $\mathbb{D}^{\alpha}\left(p \,\|\, q\right) = \mathbb{D}^{1-\alpha}\left(q \,\|\, p\right)$
- $\frac{\alpha}{1-\alpha}\mathbb{D}^{1-\alpha}\left(p \,\|\, q\right) = \mathbb{D}^{\alpha}\left(q \,\|\, p\right)$
- If $\alpha = -1$, $\mathbb{D}^{(-1)}\left(p \,\|\, q\right) = \mathbb{D}^{(1)}\left(q \,\|\, p\right) = \mathbb{KL}\left(p \,\|\, q\right) \equiv \int_x p(x) \log \frac{p(x)}{q(x)} dx$ is the **Kullback-Leibler divergence**.
- For $p_{\boldsymbol{\eta}_1}, q_{\boldsymbol{\eta}_2}$ exponential families, α-divergence has closed form expression:

$$\mathbb{D}^{\alpha}\left(p_{\boldsymbol{\eta}_1} \,\|\, q_{\boldsymbol{\eta}_2}\right) = \frac{1}{1-\alpha}\left(\alpha A(\boldsymbol{\eta}_1) + (1-\alpha)A(\boldsymbol{\eta}_2) - A(\alpha\boldsymbol{\eta}_1 + (1-\alpha)\boldsymbol{\eta}_2)\right) \tag{15}$$

  where $A(\boldsymbol{\eta})$ is the ***log-partition function*** or *cumulant function*.

1.4 f-Divergence Properties

For more details see tutorials in [Csiszár et al., 2004, Liese and Vajda, 2006] and see lecture notes in [Polyanskiy and Wu, 2014].

- $\mathbb{D}^{f_1+f_2}\left(p \,\|\, q\right) = \mathbb{D}^{f_1}\left(p \,\|\, q\right) + \mathbb{D}^{f_2}\left(p \,\|\, q\right)$
- $\mathbb{D}^{f}\left(p \,\|\, q\right) = \mathbb{D}^{g}\left(p \,\|\, q\right)$ if $f(x) = g(x) + c(x-1)$ for some $c \in \mathbb{R}$
- ***Reversal by convex inversion***: for any function f, its ***convex inversion*** is defined as $g(t) := tf(1/t)$. If f satisfies condition for f-divergence, then g satisfies the condition as well and $\mathbb{D}^{g}\left(Q \,\|\, P\right) = \mathbb{D}^{f}\left(P \,\|\, Q\right)$.
- ***Data processing inequality***: if κ is an arbitrary transition probability that transforms measures P and Q into P_κ and Q_κ correspondingly, then

$$\mathbb{D}^{f}\left(P \,\|\, Q\right) \geq \mathbb{D}^{f}\left(P_\kappa \,\|\, Q_\kappa\right). \tag{16}$$

The equality here holds if and only if the transition is induced from a ***sufficient statistic*** with respect to $\{P, Q\}$.

- ***Joint Convexity***: for any $0 \leq \lambda \leq 1$,

$$\mathbb{D}^f\left(\lambda P_1 + (1-\lambda)P_2 \,\|\, \lambda Q_1 + (1-\lambda)Q_2\right) \leq \lambda \mathbb{D}^f\left(P_1 \,\|\, Q_1\right) + (1-\lambda)\mathbb{D}^f\left((P_2 \,\|\, Q_2\right). \tag{17}$$

This follows from the convexity of the mapping $(p, q) \mapsto q\, f(p/q)$ on $\mathbb{R}_+^2$.

- **Theorem 1.1** ***(Variational representations)*** *[Polyanskiy and Wu, 2014, Wan et al., 2020]*
Let f^* be the **convex conjugate** *of f. Let* $\mathrm{effdom}(f^*)$ *be the effective domain of f^*, that is,* $\mathrm{effdom}(f^*) = \{y : f^*(y) < \infty\}$. *Then we have* two **variational representations** *of* $\mathbb{D}^f\left(p \,\|\, q\right)$*:*

$$\mathbb{D}^f\left(P \,\|\, Q\right) = \sup_{g:\, \Omega \to \mathrm{effdom}(f^*)} \mathbb{E}_P\left[g\right] - \mathbb{E}_Q\left[f^* \circ g\right] \tag{18}$$

- Special cases:
  1. ***KL divergence*** if $f(x) = x\log(x)$:

$$\mathbb{D}^f\left(P \,\|\, Q\right) = \int_{\Omega} dQ \frac{dP}{dQ}\log\left(\frac{dP}{dQ}\right) = \int_{\Omega} dP \log\left(\frac{dP}{dQ}\right) = \mathbb{E}_P\left[\log\left(\frac{dP}{dQ}\right)\right] = \mathbb{KL}\left(P \,\|\, Q\right)$$

  2. ***Total Variation divergence*** if $f(x) = \frac{1}{2}|x-1|$:

$$\mathbb{D}^f\left(P \,\|\, Q\right) = \frac{1}{2}\mathbb{E}_Q\left[\left|\left(\frac{dP}{dQ}\right) - 1\right|\right] = \frac{1}{2}\int |dP - dQ| := TV(P \| Q) \tag{19}$$

  It has *variational representation*

$$TV(P \| Q) = \sup_{f \in \mathrm{Lip}_1} \mathbb{E}_P\left[f(X)\right] - \mathbb{E}_Q\left[f(X)\right] = \mathcal{W}_1(P, Q) \tag{20}$$

  where $\mathrm{Lip}_1 := \{f : \mathcal{X} \to \mathbb{R} : \|f\|_\infty \leq 1\}$ is Lipschitz function. It is also equal to the Wasserstein-1 distance.

  3. ***χ^2-divergence*** if $f(x) = (x-1)^2$:

$$\mathbb{D}^f\left(P \,\|\, Q\right) = \mathbb{E}_Q\left[\left(\frac{dP}{dQ} - 1\right)^2\right] = \int_{\Omega} \frac{(dP - dQ)^2}{dQ} := \chi^2(P \| Q) \tag{21}$$

  4. ***Squared Hellinger distance***: $f(x) = (1 - \sqrt{x})^2$

$$\begin{aligned}\mathbb{D}^f\left(P \,\|\, Q\right) &= \mathbb{E}_Q\left[\left(1 - \sqrt{\frac{dP}{dQ}}\right)^2\right] \\ &= \int_{\Omega}\left(\sqrt{dP} - \sqrt{dQ}\right)^2 = 2 - 2\int \sqrt{dP\, dQ} := H^2(P \| Q)\end{aligned} \tag{22}$$

5. ***Jensen-Shannon divergence***: $f(x) = x\log(\frac{2x}{x+1}) + \log(\frac{2}{x+1})$,

$$\mathbb{D}^{f}\left(P \,\|\, Q\right) = \mathbb{KL}\left(P \,\|\, \frac{P+Q}{2}\right) + \mathbb{KL}\left(Q \,\|\, \frac{P+Q}{2}\right) := \mathbb{D}^{JS}\left(P \,\|\, Q\right) \tag{23}$$

6. ***Hellinger α-divergence*** $\mathbb{D}^{f_\alpha}\left(p \,\|\, q\right)$ is defined by generator

$$f^{(\alpha)}(x) := \begin{cases} \frac{4}{(1-\alpha^2)}\left\{1 - x^{\frac{(1+\alpha)}{2}}\right\} & \text{if } \alpha \neq \pm 1, \\ x\log(x), & \text{if } \alpha = 1, \\ -\log(x), & \text{if } \alpha = -1 \end{cases}.$$

For $\alpha = \pm 1$, it is the KL divergence. For $\alpha \neq \pm 1$, the corresponding divergence is

$$\mathbb{D}^{f^{(\alpha)}}\left(p \,\|\, q\right) = \frac{4}{(1-\alpha^2)}\left\{1 - \int_{\mathcal{X}} (p(x))^{\frac{1+\alpha}{2}} (q(x))^{\frac{1-\alpha}{2}} dx\right\} \tag{24}$$

The Rényi divergence and Hellinger α-divergence has one-to-one correspondence

$$\mathbb{D}^{\frac{\alpha+1}{2}}\left(p \,\|\, q\right) = \frac{2}{\alpha-1}\log\left(1 - \left(\frac{1-\alpha^2}{4}\right)\mathbb{D}^{f^{(\alpha)}}\left(P \,\|\, Q\right)\right).$$

Note that Rényi divergence itself is **not f-divergence**.

We can formulate the ***dual*** of Hellinger α-divergence using ***the conjugate dual*** of $(f^{(\alpha)})^* = f^{(-\alpha)}$. When $\alpha = 1$, it is the KL divergence.

7. ***Bregman divergence***: The only f-divergence that is also a Bregman divergences is the **KL divergence**.

- f-divergence is a **generalization** of KL divergence from ***information theorectial perspective*** [Cover and Thomas, 2006]. Bregman divergence is a generalization of KL divergence from the ***projection perspective*** as well as *Generalized Pythagorean Theorem.*

2 Divergence and Variational Inference

References

Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006. ISBN 978-0-471-24195-9.

Imre Csiszár, Paul C Shields, et al. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004.

Alfred O Hero, Bing Ma, Olivier Michel, and John Gorman. Alpha-divergence for classification, indexing and retrieval. In *University of Michigan*. Citeseer, 2001.

Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

Frank Nielsen and Richard Nock. On Rényi and Tsallis entropies and divergences for exponential families. *arXiv preprint arXiv:1105.3259*, 2011.

Barnabás Póczos and Jeff Schneider. On the estimation of alpha-divergences. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 609–617. JMLR Workshop and Conference Proceedings, 2011.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7, 2014.

Neng Wan, Dapeng Li, and Naira Hovakimyan. f-divergence variational inference. *Advances in neural information processing systems*, 33:17370–17379, 2020.