Discrepancies Between Measures

Alessandro Barp

December 4, 2020

Imperial College London, Alan Turing Institute

Discrepancies between measures

Aim

Quantify the "difference" between two measures.

- Let $\mathcal{P}(\mathcal{X})$ be the set of probability measures on a sample space $\mathcal{X}$.
- We need a map
$$D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}.$$
- Desiderata
  1. Tractability: need to be able to implement $D(\mathbb{P}, \mathbb{Q})$
  2. Meaningful: the output of $D(\mathbb{P}, \mathbb{Q})$ should be consistent with my application. E.g.
  $$D(\mathbb{P}, \mathbb{P}) = 0,$$
  also
  $$D(\mathbb{P}, \mathbb{Q}) = 0 \iff \mathbb{Q} = \mathbb{P},$$
  and if $D(\mathbb{P}_1, \mathbb{Q}) \leq D(\mathbb{P}_2, \mathbb{Q})$, then $\mathbb{P}_1$ is closer to $\mathbb{Q}$ than $\mathbb{P}_2$.
  3. Sampling Approximation: if $\mathbb{Q}$ is replaced by an empirical measure $\mathbb{Q}_n$, then $D(\mathbb{P}, \mathbb{Q}_n)$ should be defined

Three main families:

1. If $\mathcal{F}$ is a space of bounded functions, set

$$D(\mathbb{P}, \mathbb{Q}) \equiv \sup_{f \in \mathcal{F}} \left| \int f \mathrm{d}\mathbb{P} - \int f \mathrm{d}\mathbb{Q} \right|.$$

   This is a "worst-case error" in expectation. Note $D(\mathbb{P}, \mathbb{P}) = 0$, and we can easily replace $\int f \mathrm{d}\mathbb{Q}$ with a U-statistic.

2. If d is a metric on some metric space $\mathcal{H}$, and $\Phi : \mathcal{P}(\mathcal{X}) \to \mathcal{H}$, then

$$D(\mathbb{P}, \mathbb{Q}) \equiv d(\Phi(\mathbb{P}), \Phi(\mathbb{Q})).$$

   This is a pseudo-metric. However $\Phi(\mathbb{Q}_n)$ might not be defined.

3. Statistical divergences are such that $D(\mathbb{P} \| \mathbb{Q}) = 0$ iff $\mathbb{P} = \mathbb{Q}$. Divergence $\sim$ discrete Lagrangian, further require that information tensor

$$g^D_{ij}(\theta) \equiv -\partial_{\theta^i} \partial_{\alpha^j} D(\mathbb{P}_\theta, \mathbb{P}_\alpha)|_{\alpha=\theta}$$

   is Riemannian metric. They generate gradient flows.

Maximum Mean Discrepancies

- An inner product space allows us to measure projections

$$\langle u, v \rangle .$$

  A Hilbert space $\mathcal{H}$ is one for which sequences that are getting closer and closer converge.

- We then obtain a metric $\|u - v\| \equiv \sqrt{\langle u - v, u - v \rangle}$ which measure distances.
- If $\mathcal{H}$ is a Hilbert space of functions, measures can act by integration $\mathbb{P} : \mathcal{H} \to \mathbb{R}$. If $\mathbb{P}$ is continuous, then we can define a map $\Phi : \mathbb{P} \mapsto \Phi(\mathbb{P})$ by Riesz representation.
- We obtain a pseudo-metric

$$D(\mathbb{P}, \mathbb{Q}) \equiv \|\Phi(\mathbb{P}) - \Phi(\mathbb{Q})\|.$$

- If $\mathcal{H}$ is a RKHS (i.e., δ_x continuous), this is the MMD.

KL statistical divergence

- KL divergence

$$\mathsf{KL}(\mathbb{Q}\|\mathbb{P}) \equiv \int \log \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \mathrm{d}\mathbb{Q}.$$

- Information metric is the Fisher Matrix.
- $$\mathsf{KL}(q\mathrm{d}x, p\mathrm{d}x) = \int \log q \mathrm{d}\mathbb{Q} - \int \log p \mathrm{d}\mathbb{Q}.$$
- Ignore q term, so we can use U-statistics

$$\widehat{\mathsf{KL}}(\{X_i\}, p\mathrm{d}x) = -\sum \log p(X_i),$$

$X_i \sim \mathbb{Q}$, which defines maximum likelihood estimators.

- Fisher "divergence"

$$\begin{aligned}\mathrm{SM}(\mathbb{Q}\|\mathbb{P}_\theta) &\equiv \int_{\mathcal{X}} \|\nabla \log p_\theta - \nabla \log q\|_2^2 \,\mathrm{d}\mathbb{Q} \\ &= \int_{\mathcal{X}} \left(\|\nabla \log q\|_2^2 + \|\nabla \log p_\theta\|_2^2 + 2\Delta \log p_\theta\right) \mathrm{d}\mathbb{Q}\end{aligned}$$

  SM estimator is defined as $\hat{\theta}_n^{\mathrm{SM}} \equiv \mathrm{argmin}_{\theta\in\Theta} \widehat{\mathrm{SM}}(\{X_i\}_{i=1}^n\|\mathbb{P}_\theta)$ where

$$\widehat{\mathrm{SM}}(\{X_i\}_{i=1}^n\|\mathbb{P}_\theta) \equiv \frac{1}{n}\sum_{i=1}^n \Delta \log p_\theta(X_i) + \frac{1}{2}\|\nabla \log p_\theta(X_i)\|_2^2$$

- SM breaks down for non-smooth models or for models in which the second derivative grows very rapidly, inefficient for heavy-tailed distributions, non-robust for light-tailed distributions

A Stein operator $\mathcal{S}_{\mathbb{P}} : \mathcal{G} \to \Gamma(\mathbb{R})$ for $\mathbb{P}$ with Stein class $\mathcal{G}$, in this context means:

$$\int_{\mathcal{X}} \mathcal{S}_{\mathbb{P}}[f] \mathrm{d}\mathbb{P} = 0 \quad \forall f \in \mathcal{G}.$$

Used to construct integral probability discrepancies with no $\mathbb{P}$-integration: the *Stein discrepancy* (SD) $\mathcal{F} \equiv \mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]$

$$\mathrm{SD}_{\mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]}(\mathbb{Q}\|\mathbb{P}_\theta) \equiv \sup_{f \in \mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]} \left| \int_{\mathcal{X}} f \mathrm{d}\mathbb{P}_\theta - \int_{\mathcal{X}} f \mathrm{d}\mathbb{Q} \right| = \sup_{g \in \mathcal{G}} \left| \int_{\mathcal{X}} \mathcal{S}_{\mathbb{P}_\theta}[g] \mathrm{d}\mathbb{Q} \right|.$$

Langevin-Stein discrepancy $\mathcal{T}_p[g] = \langle \nabla \log p, g \rangle + \nabla \cdot g$. More generally given $m \in \Gamma(\mathbb{R}^{d \times d})$

$$\mathcal{S}_p^m[g] \equiv \frac{1}{p} \nabla \cdot (pmg), \quad \mathcal{S}_p^m[A] \equiv \frac{1}{p} \nabla \cdot (pmA).$$

Hence the learning task consists on obtaining the *minimum Stein discrepancy estimators*

$$\hat{\theta}_n^{\mathrm{Stein}} \equiv \mathrm{argmin}_{\theta \in \Theta} \widehat{\mathrm{SD}}_{\mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]}(\{X_i\}_i^n \| \mathbb{P}_\theta).$$

Diffusion Score Matching

For $\mathcal{S}_p^m[g]$ and $\mathcal{G} \equiv \{g \in C^1(\mathcal{X}, \mathbb{R}^d) \cap L^2(\mathcal{X}; \mathbb{Q}) : \|g\|_{L^2(\mathcal{X};\mathbb{Q})} \leq 1\}$:

$$\mathrm{DSM}_m(\mathbb{Q}\|\mathbb{P}) \equiv \sup_{f \in \mathcal{S}_p[\mathcal{G}]} \left| \int_{\mathcal{X}} f \mathrm{d}\mathbb{Q} - \int_{\mathcal{X}} f \mathrm{d}\mathbb{P} \right|^2 = \int_{\mathcal{X}} \left\| m^\top \left(\nabla \log q - \nabla \log p \right) \right\|_2^2 \mathrm{d}\mathbb{Q}.$$

- $\mathrm{DSM}_m(\mathbb{Q}\|\mathbb{P}) = 0$ iff $\mathbb{Q} = \mathbb{P}$ when $m(x)$ is invertible
- Recovers SM for $m(x)m^\top(x) = I$.
- Under appropriate assumptions

$$\mathrm{DSM}_m(\mathbb{Q}\|\mathbb{P}) = \int_{\mathcal{X}} \left(\|m^\top \nabla_x \log p\|_2^2 + \|m^\top \nabla \log q\|_2^2 + 2\nabla \cdot \left(m m^\top \nabla \log p \right) \right) \mathrm{d}\mathbb{Q}.$$

- If m is θ-independent

$$\widehat{\mathrm{DSM}}_m(\{X_i\}_{i=1}^n \|\mathbb{P}_\theta) \equiv \frac{1}{n} \sum_{i=1}^n \left(\|m^\top \nabla_x \log p_\theta\|_2^2 + 2\nabla \cdot \left(m m^\top \nabla \log p_\theta \right) \right)(X_i)$$

- For $\mathcal{G}$ unit ball vector-valued RKHS with matrix kernel K

$$\mathrm{DKSD}_{K,m}(\mathbb{Q}\|\mathbb{P})^2 = \int k^0(x,y)\mathrm{d}\mathbb{Q}\otimes\mathrm{d}\mathbb{Q}$$

$$k^0(x,y) \equiv \frac{1}{p(y)p(x)}\nabla_y\cdot\nabla_x\cdot\Big(p(x)m(x)K(x,y)m(y)^\top p(y)\Big)$$

- U-statistic approximation leads to DKSD estimators

$$\widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n\|\mathbb{P}_\theta)^2 = \frac{1}{n(n-1)}\sum_{i\neq j} k^0_\theta(X_i,X_j)$$

- m can depend on θ
- $K = kI$, $m = I$ then DKSD is KSD
- DKSD recovers DSM as a limit
- Statistical divergence when m invertible and K integrally positive definite
- Other examples recover contrastive divergences and minimum probability flow

Gradient Descent

- It is often stated that an advantage of Wasserstein-based estimators is that they take into account the geometry of the sample space
- In order to reflect the geometry of the statistical model in learning tasks you can follow a stochastic gradient flow generated by the information metric

$$g_{\mathrm{DKSD}}(\theta)_{ij} = \int_{\mathcal{X}^2} \left(\nabla_x \partial_{\theta^j} \log p_\theta\right)^\top m_\theta(x) K(x,y) m_\theta^\top(y) \nabla_y \partial_{\theta^i} \log p_\theta \, \mathrm{d}\mathbb{P}_\theta(x) \mathrm{d}\mathbb{P}_\theta(y),$$

$$g_{\mathrm{DSM}}(\theta)_{ij} = \int_{\mathcal{X}} \left\langle m^\top \nabla \partial_{\theta^i} \log p_\theta, m^\top \nabla \partial_{\theta^j} \log p_\theta \right\rangle \mathrm{d}\mathbb{P}_\theta.$$