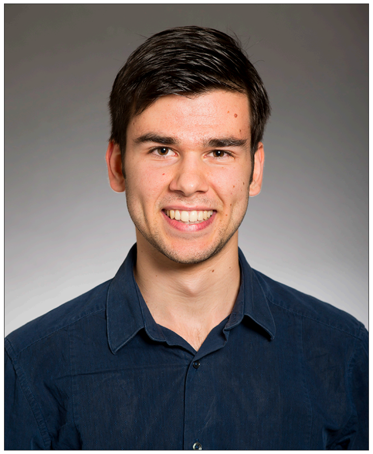
Sequential estimation of convex divergences using reverse submartingales and exchangeable filtrations

Aaditya Ramdas

Dept. of Statistics and Data Science
Machine Learning Dept.
Carnegie Mellon University
(Always looking for postdocs and funding)



Joint work with
Tudor Manole

Summary of the talk

1. The exchangeable filtration is an interesting object that arises naturally.

2. F-divergences, KL, TV, integral probability metrics, Wasserstein, etc. are "convex divergences". Entropy is a convex functional.

3. Convex divergences are reverse submartingales with respect to the exchangeable filtration.

4. A reverse Ville's inequality, converts fixed-time concentration into time-uniform (or stopping-time) concentration.

5. A sequence of random graphs, ignoring labeling (ordering), can play the role of the empirical distribution in the exchangeable filtration.

6. Are the above techniques useful for the study of random graphs?

Convex divergences

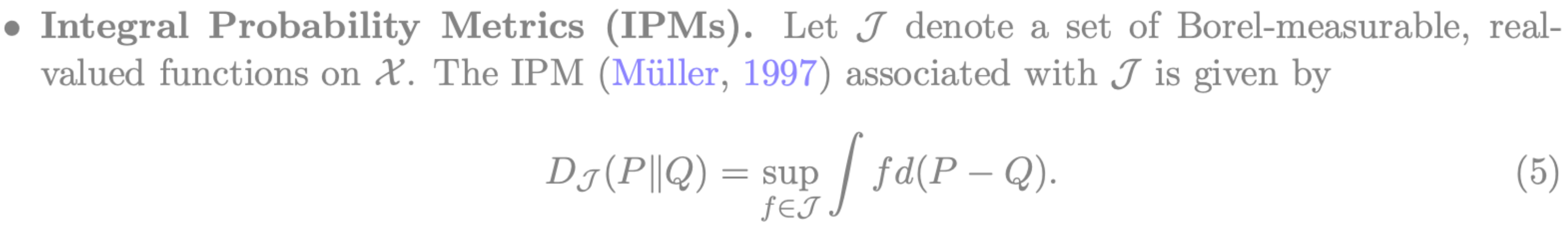
(Rockefeller'70s)

$$D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}_+$$

$\mathcal{P}(\mathcal{X})$ denotes the set of Borel probability measures over a set $\mathcal{X} \subseteq \mathbb{R}^d$.

$$\boxed{D\big(\lambda\mu_1 + (1-\lambda)\mu_2 \parallel \lambda\nu_1 + (1-\lambda)\nu_2\big) \leq \lambda D(\mu_1\|\nu_1) + (1-\lambda)D(\mu_2\|\nu_2).}$$

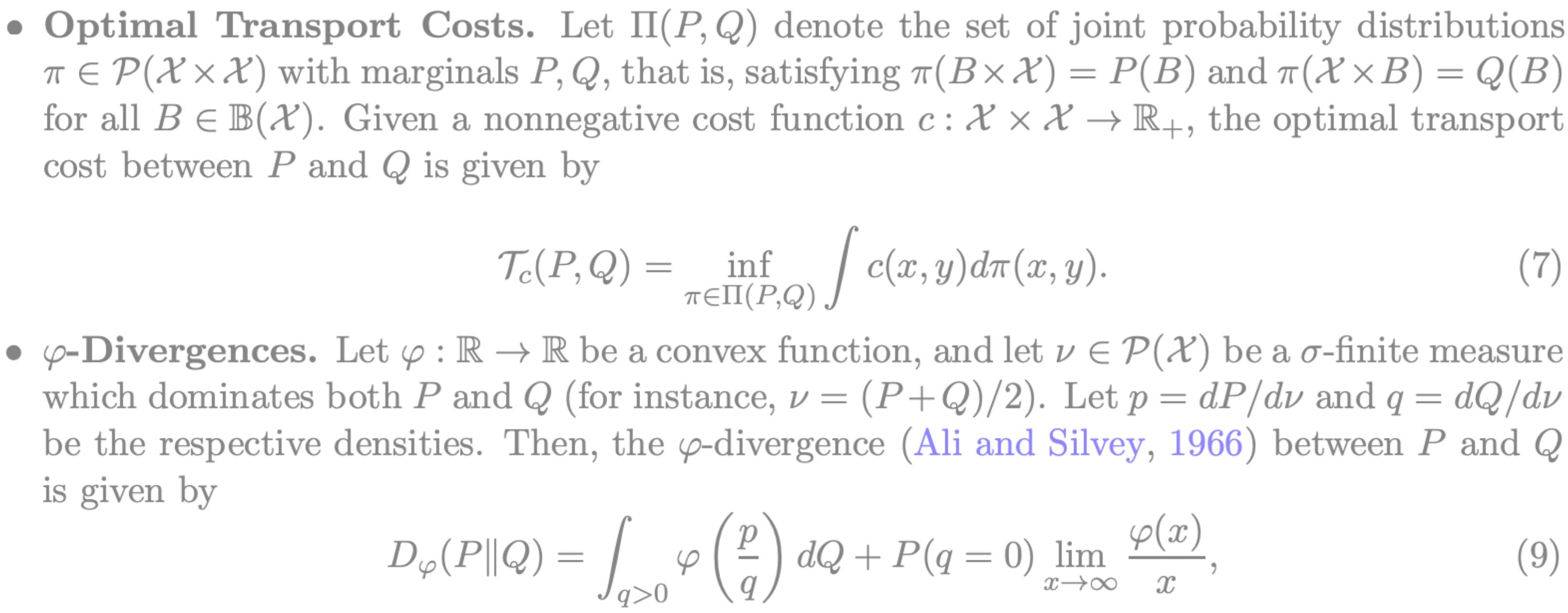
Examples:



- **Integral Probability Metrics (IPMs).** Let $\mathcal{J}$ denote a set of Borel-measurable, real-valued functions on $\mathcal{X}$. The IPM (Müller, 1997) associated with $\mathcal{J}$ is given by

$$D_{\mathcal{J}}(P\|Q) = \sup_{f\in\mathcal{J}} \int f d(P-Q). \tag{5}$$

(TV, KS, MMD)



- **Optimal Transport Costs.** Let $\Pi(P,Q)$ denote the set of joint probability distributions $\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{X})$ with marginals P, Q, that is, satisfying $\pi(B\times\mathcal{X}) = P(B)$ and $\pi(\mathcal{X}\times B) = Q(B)$ for all $B \in \mathbb{B}(\mathcal{X})$. Given a nonnegative cost function $c : \mathcal{X}\times\mathcal{X} \to \mathbb{R}_+$, the optimal transport cost between P and Q is given by

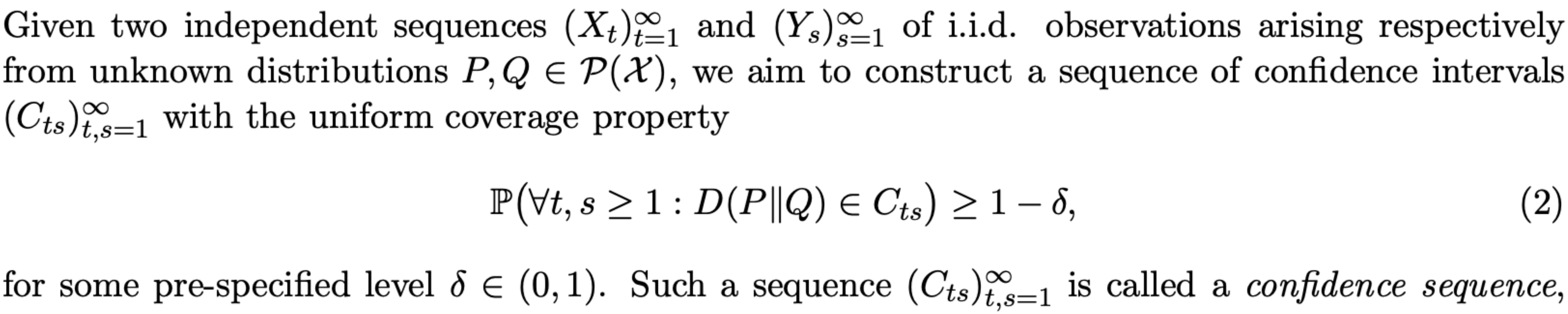
$$\mathcal{T}_c(P,Q) = \inf_{\pi\in\Pi(P,Q)} \int c(x,y)d\pi(x,y). \tag{7}$$

- **φ-Divergences.** Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function, and let $\nu \in \mathcal{P}(\mathcal{X})$ be a σ-finite measure which dominates both P and Q (for instance, $\nu = (P+Q)/2$). Let $p = dP/d\nu$ and $q = dQ/d\nu$ be the respective densities. Then, the φ-divergence (Ali and Silvey, 1966) between P and Q is given by

$$D_\varphi(P\|Q) = \int_{q>0} \varphi\left(\frac{p}{q}\right) dQ + P(q=0)\lim_{x\to\infty}\frac{\varphi(x)}{x}, \tag{9}$$

Confidence sequences

(Robbins'67)

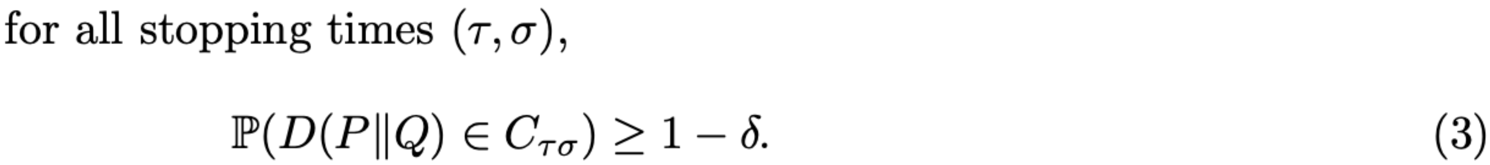


Given two independent sequences $(X_t)_{t=1}^{\infty}$ and $(Y_s)_{s=1}^{\infty}$ of i.i.d. observations arising respectively from unknown distributions $P, Q \in \mathcal{P}(\mathcal{X})$, we aim to construct a sequence of confidence intervals $(C_{ts})_{t,s=1}^{\infty}$ with the uniform coverage property

$$\mathbb{P}\big(\forall t, s \geq 1 : D(P\|Q) \in C_{ts}\big) \geq 1 - \delta, \tag{2}$$

for some pre-specified level $\delta \in (0,1)$. Such a sequence $(C_{ts})_{t,s=1}^{\infty}$ is called a *confidence sequence*,

or equivalently,



for all stopping times (τ, σ),

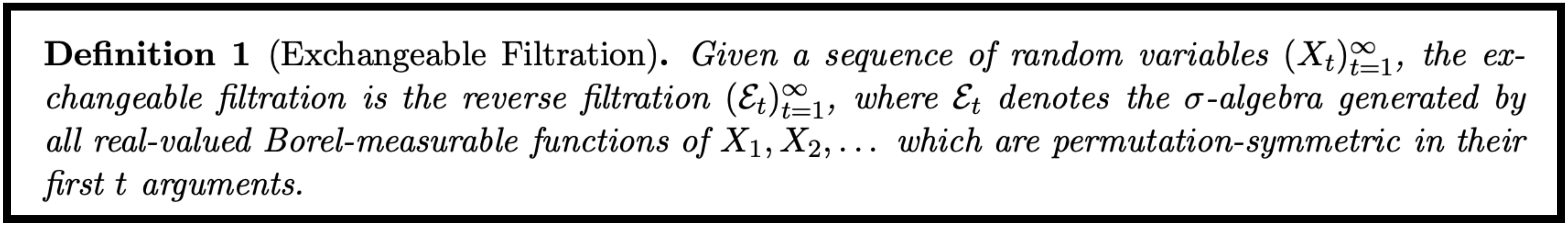
$$\mathbb{P}(D(P\|Q) \in C_{\tau\sigma}) \geq 1 - \delta. \tag{3}$$

Useful for sequential testing, estimation, bandits, etc.

Exchangeable filtration

(De-Finetti? Hewitt-Savage?)

$$X_1, X_2, \ldots \sim P$$



Definition 1 (Exchangeable Filtration). *Given a sequence of random variables $(X_t)_{t=1}^{\infty}$, the exchangeable filtration is the reverse filtration $(\mathcal{E}_t)_{t=1}^{\infty}$, where $\mathcal{E}_t$ denotes the σ-algebra generated by all real-valued Borel-measurable functions of $X_1, X_2, \ldots$ which are permutation-symmetric in their first t arguments.*

$$\mathscr{E}_1 \supseteq \mathscr{E}_2 \supseteq \ldots \mathscr{E}_\infty$$

$\mathscr{E}_t$ is the information known to an *amnesic oracle* (an oracle with amnesia).

Oracle = knows the entire future $X_t, X_{t+1}, \ldots$

Anmesia = forgotten order of events in the past $\sigma(X_1, \ldots, X_{t-1})$

Informally, $\mathscr{E}_t \approx \sigma(P_t, X_{t+1}, X_{t+1}, \ldots)$

Reverse submartingales

$X_1, X_2, \ldots \sim P$

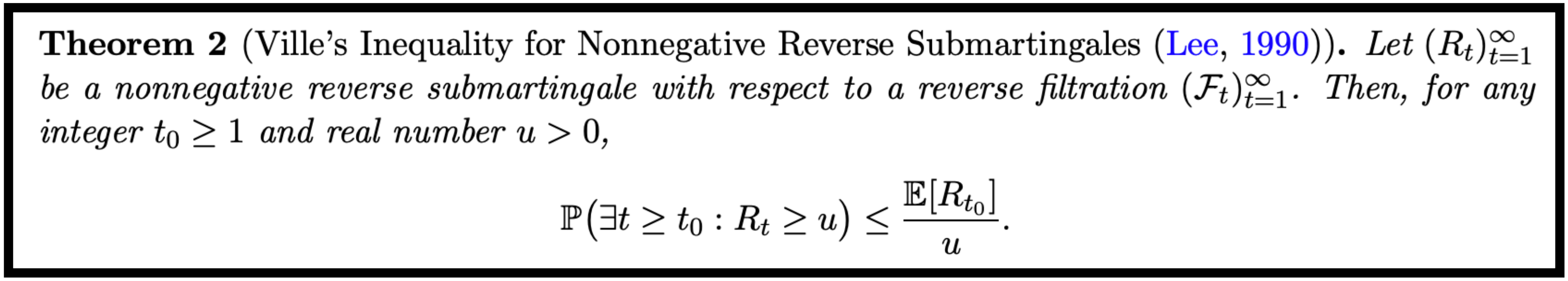
$\mathscr{E}_1 \supseteq \mathscr{E}_2 \supseteq \ldots \mathscr{E}_\infty$

An integrable process (M_t) is a reverse submartingale wrt $(\mathscr{E}_t)$ if

$$\mathbb{E}[M_t \,|\, \mathscr{E}_{t+1}] \geq M_{t+1}.$$

Eg: $M_t = (X_1 + \ldots + X_t)/t$ is a reverse martingale.

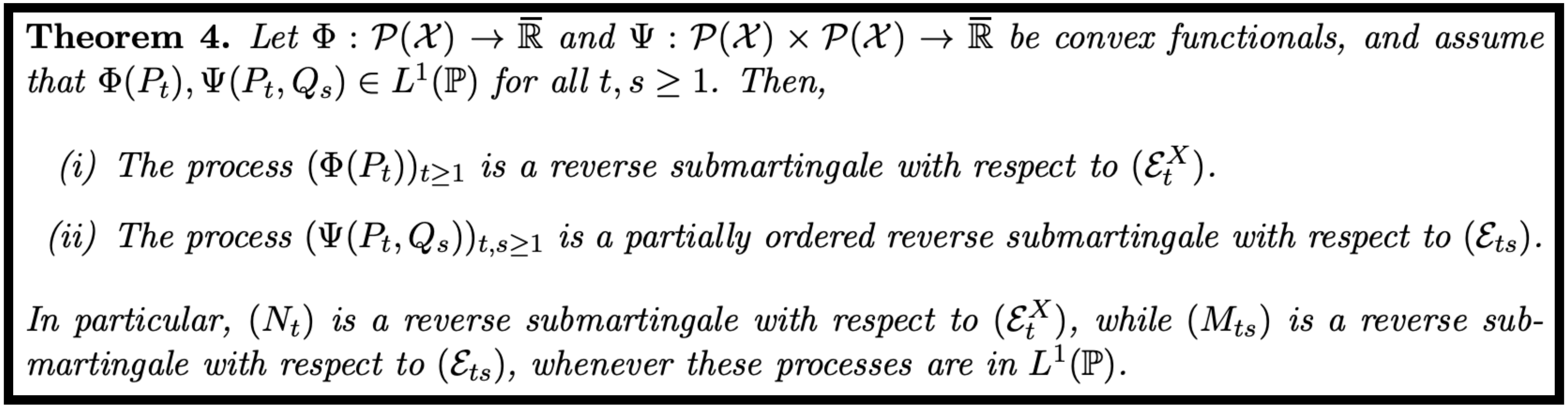
Eg: $P_t = (\delta_{X_1} + \ldots + \delta_{X_t})/t$ is a measure-valued reverse martingale.



Theorem 2 (Ville's Inequality for Nonnegative Reverse Submartingales (Lee, 1990)). *Let $(R_t)_{t=1}^{\infty}$ be a nonnegative reverse submartingale with respect to a reverse filtration $(\mathcal{F}_t)_{t=1}^{\infty}$. Then, for any integer $t_0 \geq 1$ and real number $u > 0$,*

$$\mathbb{P}\big(\exists t \geq t_0 : R_t \geq u\big) \leq \frac{\mathbb{E}[R_{t_0}]}{u}.$$

Define $N_t := D(P_t \| Q) - D(P \| Q), M_{ts} := D(P_t \| Q_s) - D(P \| Q)$



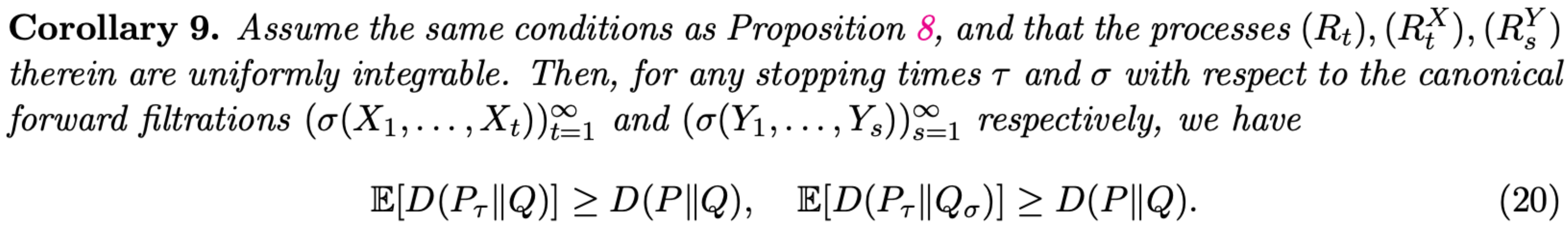
Theorem 4. *Let $\Phi : \mathcal{P}(\mathcal{X}) \to \overline{\mathbb{R}}$ and $\Psi : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \overline{\mathbb{R}}$ be convex functionals, and assume that $\Phi(P_t), \Psi(P_t, Q_s) \in L^1(\mathbb{P})$ for all $t, s \geq 1$. Then,*

(i) The process $(\Phi(P_t))_{t\geq 1}$ is a reverse submartingale with respect to $(\mathcal{E}_t^X)$.

(ii) The process $(\Psi(P_t, Q_s))_{t,s\geq 1}$ is a partially ordered reverse submartingale with respect to $(\mathcal{E}_{ts})$.

In particular, (N_t) is a reverse submartingale with respect to $(\mathcal{E}_t^X)$, while (M_{ts}) is a reverse submartingale with respect to $(\mathcal{E}_{ts})$, whenever these processes are in $L^1(\mathbb{P})$.

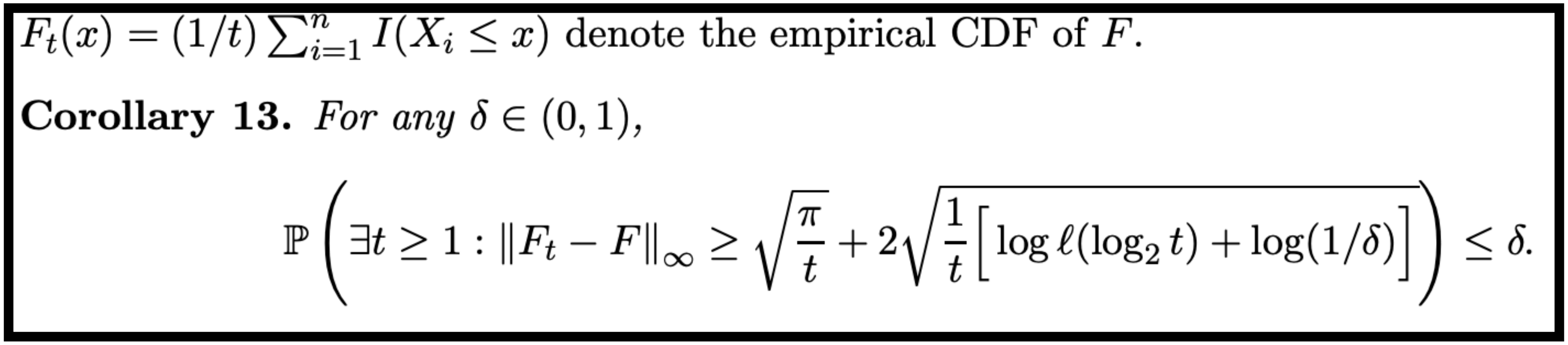
Reduces stopping-time concentration to fixed-time



Corollary 9. *Assume the same conditions as Proposition 8, and that the processes* $(R_t), (R_t^X), (R_s^Y)$ *therein are uniformly integrable. Then, for any stopping times* τ *and* σ *with respect to the canonical forward filtrations* $(\sigma(X_1, \ldots, X_t))_{t=1}^{\infty}$ *and* $(\sigma(Y_1, \ldots, Y_s))_{s=1}^{\infty}$ *respectively, we have*

$$\mathbb{E}[D(P_\tau \| Q)] \geq D(P \| Q), \quad \mathbb{E}[D(P_\tau \| Q_\sigma)] \geq D(P \| Q). \tag{20}$$

Eg: A time-uniform DKW inequality

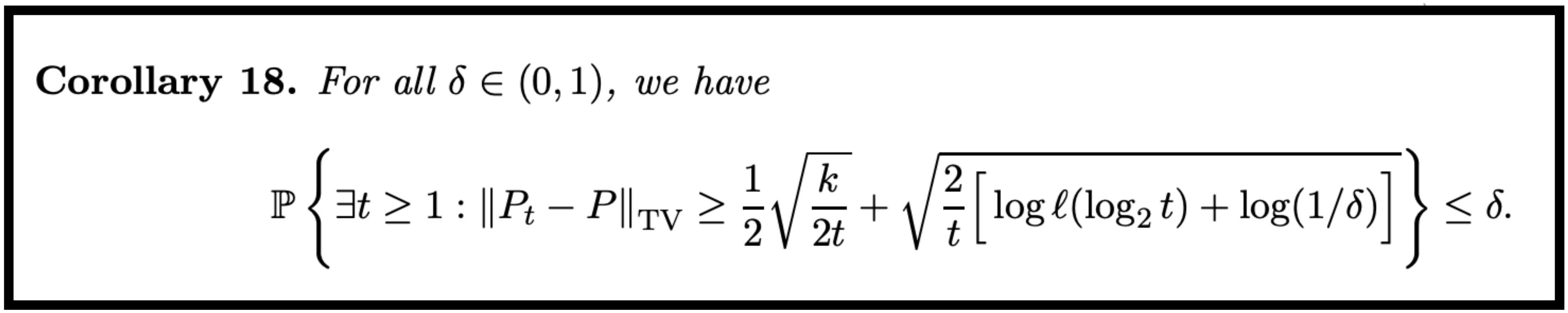


$F_t(x) = (1/t) \sum_{i=1}^{n} I(X_i \leq x)$ denote the empirical CDF of F.

Corollary 13. *For any* $\delta \in (0,1)$,

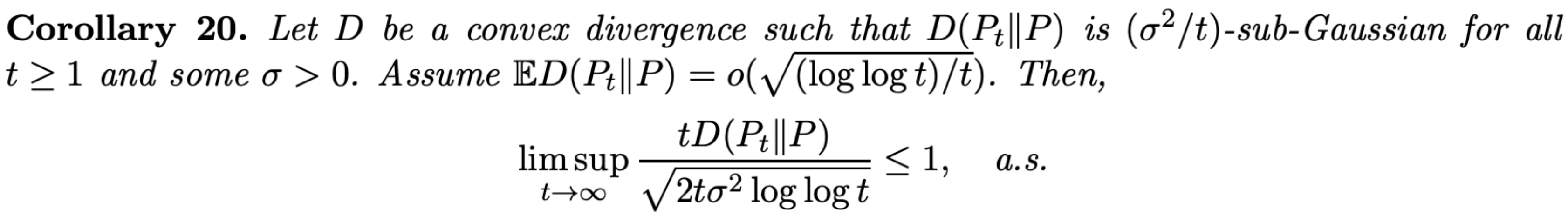
$$\mathbb{P}\left(\exists t \geq 1 : \|F_t - F\|_\infty \geq \sqrt{\frac{\pi}{t}} + 2\sqrt{\frac{1}{t}\left[\log \ell(\log_2 t) + \log(1/\delta)\right]}\right) \leq \delta.$$

Eg: A time-uniform TV inequality for alphabets of size k



Corollary 18. *For all* $\delta \in (0,1)$, *we have*

$$\mathbb{P}\left\{\exists t \geq 1 : \|P_t - P\|_{\mathrm{TV}} \geq \frac{1}{2}\sqrt{\frac{k}{2t}} + \sqrt{\frac{2}{t}\left[\log \ell(\log_2 t) + \log(1/\delta)\right]}\right\} \leq \delta.$$



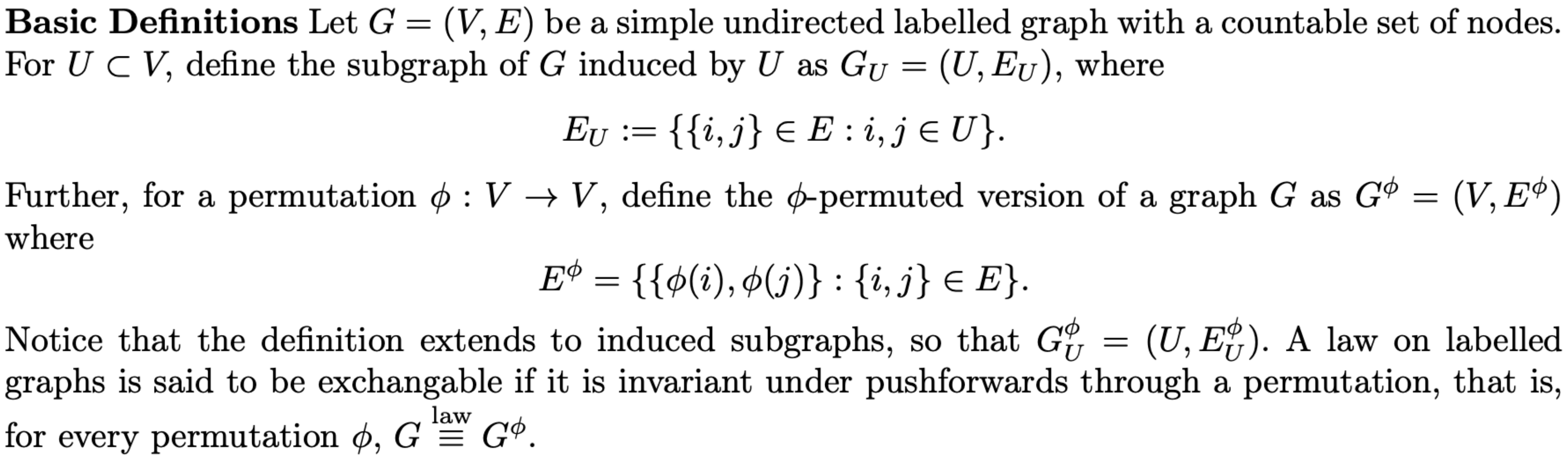
Corollary 20. *Let* D *be a convex divergence such that* $D(P_t \| P)$ *is* (σ^2/t)*-sub-Gaussian for all* $t \geq 1$ *and some* $\sigma > 0$. *Assume* $\mathbb{E}D(P_t \| P) = o(\sqrt{(\log\log t)/t})$. *Then,*

$$\limsup_{t\to\infty} \frac{tD(P_t \| P)}{\sqrt{2t\sigma^2 \log\log t}} \leq 1, \quad a.s.$$

The rest is ongoing and “shaky”
(work in progress, would love feedback)

Graphs are like "empirical distributions"

Given a sequence of random graphs (node exchangeable),
if we only kept the structure, but not the ordering in which nodes are introduced,
does the graph itself behave like an empirical distribution,
in the sense that it is a "graph-valued" reverse martingale?



Basic Definitions Let $G = (V, E)$ be a simple undirected labelled graph with a countable set of nodes. For $U \subset V$, define the subgraph of G induced by U as $G_U = (U, E_U)$, where

$$E_U := \{\{i, j\} \in E : i, j \in U\}.$$

Further, for a permutation $\phi : V \to V$, define the ϕ-permuted version of a graph G as $G^\phi = (V, E^\phi)$ where

$$E^\phi = \{\{\phi(i), \phi(j)\} : \{i, j\} \in E\}.$$

Notice that the definition extends to induced subgraphs, so that $G^\phi_U = (U, E^\phi_U)$. A law on labelled graphs is said to be exchangable if it is invariant under pushforwards through a permutation, that is, for every permutation ϕ, $G \overset{\text{law}}{\equiv} G^\phi$.

Submodularity? Subadditivity? Leave-one-out?

A function $f : \mathcal{G} \to \mathbb{R}$ is node-subadditive if for any $U, W \subset V$,

$$f(G_{U \cup W}) \leq f(G_U) + f(G_V).$$

Let $\mathcal{E}_t$ be the sigma algebra generated by node-symmetric functions of G_t. This yields the associated exchangable filtration $\{\mathcal{E}_t\}$. The basic property follows simply

Proposition 1. *Let G be an exchangable graph, and G_t its node-wise observation process. If f is a node-subadditive, node-symmetric, and trivial non-negative graph function, then $f(G_t)$ is a reverse submartingale with respect to $\{\mathcal{E}_t\}$.*

Weaker, sufficient condition:

$$f(G_{t+1}) \leq \frac{1}{t+1} \sum_{i=1}^{t+1} f(G_{V_{t+1} \setminus \{v_i\}})$$

Natural graph functions (counting motifs) are superadditive.

Is the cut metric between two graphons a convex divergence,
can it be sequentially tracked from two sequences of random graphs?

What's the right formalism? Is any of this interesting/useful?

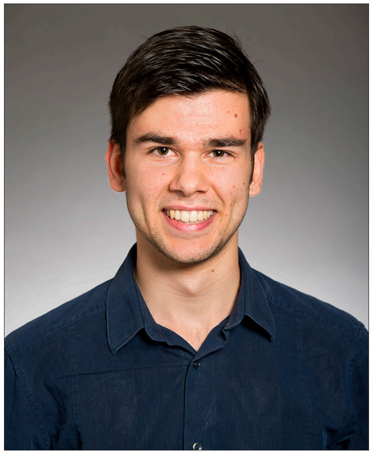
(Conversations with Aditya Gangrade)

Reverse martingales and exchangeable filtrations

Aaditya Ramdas

Dept. of Statistics and Data Science
Machine Learning Dept.
Carnegie Mellon University

(Always looking for postdocs and funding)



Joint work with

Tudor Manole