

A NON-EXPONENTIAL EXTENSION OF SANOV'S THEOREM VIA CONVEX DUALITY

DANIEL LACKER,* *Columbia University*

Abstract

This work is devoted to a vast extension of Sanov's theorem, in Laplace principle form, based on alternatives to the classical convex dual pair of relative entropy and cumulant generating functional. The abstract results give rise to a number of probabilistic limit theorems and asymptotics. For instance, widely applicable non-exponential large deviation upper bounds are derived for empirical distributions and averages of independent and identically distributed samples under minimal integrability assumptions, notably accommodating heavy-tailed distributions. Other interesting manifestations of the abstract results include new results on the rate of convergence of empirical measures in Wasserstein distance, uniform large deviation bounds, and variational problems involving optimal transport costs, as well as an application to error estimates for approximate solutions of stochastic optimization problems. The proofs build on the Dupuis–Ellis weak convergence approach to large deviations as well as the duality theory for convex risk measures.

Keywords: Sanov's theorem; large deviations; convex duality; risk measures; weak convergence; empirical measures; heavy tails; stochastic optimization

2010 Mathematics Subject Classification: Primary 60F10
Secondary 46N10

1. Introduction

An original goal of this paper was to extend the weak convergence methodology of Dupuis and Ellis [22] to the context of non-exponential (e.g. heavy-tailed) large deviations. While we claim only modest success in this regard, we do find some general-purpose large deviation upper bounds which can be seen as polynomial-rate analogs of the upper bounds in the classical theorems of Sanov and Cramér. At least as interesting, however, are the abstract principles behind these bounds, which have broad implications beyond the realm of large deviations. Let us first describe these abstract principles before specializing them in various ways.

Let E be a Polish space, and let $\mathcal{P}(E)$ denote the set of Borel probability measures on E endowed with the topology of weak convergence. Let $B(E)$ (resp. $C_b(E)$) denote the set of measurable (resp. continuous) and bounded real-valued functions on E . For $n \geq 1$ and $\nu \in \mathcal{P}(E^n)$, define $\nu_{0,1} \in \mathcal{P}(E)$ and measurable maps $\nu_{k-1,k}: E^{k-1} \rightarrow \mathcal{P}(E)$ for $k = 2, \dots, n$ via the disintegration

$$\nu(dx_1, \dots, dx_n) = \nu_{0,1}(dx_1) \prod_{k=2}^n \nu_{k-1,k}(x_1, \dots, x_{k-1})(dx_k).$$

Received 1 January 2018; revision received 5 September 2019.

* Postal address: 306 Mudd, 500 West 120th St, New York, NY 10027, USA.

Email address: daniel.lacker@columbia.edu

In other words, if (X_1, \dots, X_n) is an E^n -valued random variable with law ν , then $\nu_{0,1}$ is the law of X_1 , and $\nu_{k-1,k}(X_1, \dots, X_{k-1})$ is the conditional law of X_k given (X_1, \dots, X_{k-1}) . Of course, $\nu_{k-1,k}$ are uniquely defined up to ν -almost sure equality.

The protagonist of the paper is a proper (i.e. not identically ∞) convex function $\alpha: \mathcal{P}(E) \rightarrow (-\infty, \infty]$ with compact sub-level sets; that is, $\{\nu \in \mathcal{P}(E): \alpha(\nu) \leq c\}$ is compact for every $c \in \mathbb{R}$. For $n \geq 1$ define $\alpha_n: \mathcal{P}(E^n) \rightarrow (-\infty, \infty]$ by

$$\alpha_n(\nu) = \int_{E^n} \sum_{k=1}^n \alpha(\nu_{k-1,k}(x_1, \dots, x_{k-1})) \nu(dx_1, \dots, dx_n),$$

and note that $\alpha_1 \equiv \alpha$. Define the convex conjugate $\rho_n: B(E^n) \rightarrow \mathbb{R}$ by

$$\rho_n(f) = \sup_{\nu \in \mathcal{P}(E^n)} \left(\int_{E^n} f \, d\nu - \alpha_n(\nu) \right) \quad \text{and} \quad \rho \equiv \rho_1. \tag{1.1}$$

Our main interest is in evaluating ρ_n at functions of the *empirical measure* $L_n: E^n \rightarrow \mathcal{P}(E)$ defined by

$$L_n(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}.$$

The main abstract result of the paper is the following extension of Sanov’s theorem, proved in more generality in Section 2.2 by adapting the weak convergence techniques of Dupuis and Ellis [22].

Theorem 1.1. *For $F \in C_b(\mathcal{P}(E))$,*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \rho_n(nF \circ L_n) = \sup_{\nu \in \mathcal{P}(E)} (F(\nu) - \alpha(\nu)).$$

The guiding example is the relative entropy, $\alpha(\cdot) = H(\cdot | \mu)$, where $\mu \in \mathcal{P}(E)$ is a fixed reference measure, and H is defined by

$$H(\nu | \mu) = \int_E \log(d\nu/d\mu) \, d\nu \quad \text{for } \nu \ll \mu, \quad H(\nu | \mu) = \infty \quad \text{otherwise,} \tag{1.2}$$

Letting μ^n denote the n -fold product measure, it turns out that $\alpha_n(\cdot) = H(\cdot | \mu^n)$, by the so-called *chain rule* of relative entropy [22, Theorem B.2.1]. The dual ρ_n is well known to be $\rho_n(f) = \log \int_{E^n} e^f \, d\mu^n$, and the duality formulas relating ρ_n and α_n are often known as the Gibbs variational principle or the Donsker–Varadhan formula [22, Proposition 1.4.2 and Lemma 1.4.3]. In this case Theorem 1.1 reduces to the Laplace principle form of Sanov’s theorem:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \int_{E^n} e^{nF \circ L_n} \, d\mu^n = \sup_{\nu \in \mathcal{P}(E)} (F(\nu) - H(\nu | \mu)).$$

Well-known theorems of Varadhan and of Dupuis and Ellis (see [22, Theorems 1.2.1 and 1.2.3]) assert the equivalence of this form of Sanov’s theorem with the more common form: for every Borel set $A \subset \mathcal{P}(E)$ with closure \bar{A} and interior A° ,

$$\begin{aligned} - \inf_{\nu \in A^\circ} H(\nu | \mu) &\leq \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mu^n(L_n \in A) \\ &\leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log \mu^n(L_n \in A) \leq - \inf_{\nu \in \bar{A}} H(\nu | \mu). \end{aligned} \tag{1.3}$$

To derive this heuristically, apply Theorem 1.1 to the function

$$F(v) = \begin{cases} 0 & \text{if } v \in A, \\ -\infty & \text{otherwise.} \end{cases}$$

For general α , Theorem 1.1 does not permit an analogous *equivalent* formulation in terms of deviation probabilities. In fact, for many α , Theorem 1.1 has nothing to do with large deviations (see Sections 1.3 and 1.4 below). Nonetheless, for certain α , Theorem 1.1 *implies* interesting large deviation upper bounds, which we prove by formalizing the aforementioned heuristic. While many α admit fairly explicit known formulas for the dual ρ , the recurring challenge in applying Theorem 1.1 is finding a useful expression for ρ_n , and herein lies but one of many instances of the wonderful tractability of relative entropy. The examples to follow do admit good expressions for ρ_n , or at least workable one-sided bounds, but we also catalog in Section 1.5 some natural alternative choices of α for which we did not find useful bounds or expressions for ρ_n .

The functional ρ is (up to a sign change) a *convex risk measure*, in the language of Föllmer and Schied [28]. A rich duality theory for convex risk measures has emerged over the past two decades, primarily geared toward applications in financial mathematics and optimization. We take advantage of this theory in Section 2 to demonstrate how α can be reconstructed from ρ , which shows that ρ could be taken as the starting point instead of α . Additionally, the theory of risk measures provides insight into how to deal with the subtleties that arise in extending the domain of ρ (and Theorem 1.1) to accommodate unbounded functions or stronger topologies on $\mathcal{P}(E)$. Section 1.6 briefly reinterprets Theorem 1.1 in a language more consistent with the risk measure literature. The reader familiar with risk measures may notice a *time-consistent dynamic risk measure* (see [1] for definitions and survey) hidden in the definition of ρ_n above.

We will make no use of the interpretation in terms of dynamic risk measures, but it did inspire a recursive formula for ρ_n (similar to a result of [14]). To state it loosely, if $f \in B(E^n)$ then we may write

$$\rho_n(f) = \rho_{n-1}(g), \quad \text{where } g(x_1, \dots, x_{n-1}) := \rho(f(x_1, \dots, x_{n-1}, \cdot)). \tag{1.4}$$

To make rigorous sense of this, we must note that $g: E^{n-1} \rightarrow \mathbb{R}$ is merely upper semianalytic and not Borel-measurable in general, and argue that ρ is well-defined for such functions. We make this precise in Proposition A.1. This recursive formula is not essential for any of the main arguments but is convenient for some calculations.

1.1. Non-exponential large deviations

Our first application, and the one we discuss in the most detail, comes from applying (an extension of) Theorem 1.1 with

$$\alpha(v) = \|\mathrm{d}v/\mathrm{d}\mu\|_{L^p(\mu)} - 1 \quad \text{for } v \ll \mu, \quad \alpha(v) = \infty \quad \text{otherwise,} \tag{1.5}$$

where $\mu \in \mathcal{P}(E)$ is fixed. We state the abstract result first. For a continuous function $\psi: E \rightarrow \mathbb{R}_+ := [0, \infty)$, let $\mathcal{P}_\psi(E)$ denote the set of $v \in \mathcal{P}(E)$ satisfying $\int \psi \, \mathrm{d}v < \infty$. Equip $\mathcal{P}_\psi(E)$ with the topology induced by the linear maps $v \mapsto \int f \, \mathrm{d}v$, where $f: E \rightarrow \mathbb{R}$ is continuous and $|f| \leq 1 + \psi$. Recall in the following that μ^n denotes the n -fold product measure.

Theorem 1.2. *Let $q \in (1, \infty)$, and let $p = q/(q - 1)$ denote the conjugate exponent. Let $\mu \in \mathcal{P}(E)$, and suppose $\int \psi^q d\mu < \infty$ for some continuous $\psi : E \rightarrow \mathbb{R}_+$. Then, for every closed set $A \subset \mathcal{P}_\psi(E)$,*

$$\limsup_{n \rightarrow \infty} n^{q-1} \mu^n(L_n \in A) \leq \left(\inf_{\nu \in A} \|d\nu/d\mu\|_{L^p(\mu)} - 1 \right)^{-q}.$$

We view Theorem 1.2 as a non-exponential version of the upper bound of Sanov’s theorem, and the proof is given in Section 4.1. At this level of generality, there cannot be a matching lower bound for open sets as in the classical case (1.3), as will be explained more in Section 1.1.2. Of course, Sanov’s theorem applies without any moment assumptions, but the upper bound provides no information in many heavy-tailed contexts. We illustrate this with three applications below, all of which take advantage of the crucial fact that Theorem 1.2 applies to arbitrary closed sets A , which enables a natural contraction principle (i.e. continuous mapping). The first example gives new results on the rate of convergence of empirical measures in Wasserstein distance. Second, we derive non-exponential upper bounds analogous to Cramér’s theorem for sums of independent and identically distributed (i.i.d.) random variables with values in Banach spaces. Lastly, we derive error bounds for the usual Monte Carlo scheme in stochastic optimization, essentially providing a heavy-tailed analog of the results of [37].

1.1.1. *Rate of convergence of empirical measures in Wasserstein distance.* First, some terminology: a *compatible metric* on E is any metric on E which generates the given Polish topology. For $\mu, \nu \in \mathcal{P}(E)$ and $q \geq 1$ define the q -Wasserstein distance $\mathcal{W}_q(\mu, \nu)$ by

$$\mathcal{W}_q^q(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{E \times E} d^q(x, y) \pi(dx, dy), \tag{1.6}$$

where $\Pi(\mu, \nu)$ is the set of probability measures on $E \times E$ with first marginal μ and second marginal ν . In Section 4.2 we will prove the following.

Corollary 1.1. (Wasserstein convergence rate.) *Let d be any compatible metric on E . Let $q > r \geq 1$, and let $\mu \in \mathcal{P}(E)$ satisfy $\int_E d^q(x, x_0) \mu(dx) < \infty$ for some (equivalently, for any) $x_0 \in E$. Then, for each $a > 0$,*

$$\limsup_{n \rightarrow \infty} n^{q/r-1} \mu^n(\mathcal{W}_r(L_n, \mu) \geq a) < \infty.$$

In particular,

$$\limsup_{n \rightarrow \infty} n^{q-1} \mu^n(\mathcal{W}_1(L_n, \mu) \geq a) < \infty.$$

In other words, $\mu^n(\mathcal{W}_r(L_n, \mu) \geq a) = O(n^{1-q/r})$. In the $r = 1$ case, a comparison with a more classical setting reveals that this rate is the right one, in a sense. Suppose X_i are i.i.d. real-valued random variables with law μ , mean zero, and $\mathbb{E}|X_1|^q < \infty$. Then the a.s. inequality

$$\frac{1}{n} \sum_{i=1}^n X_i \leq \mathcal{W}_1 \left(\frac{1}{n} \sum_{i=1}^n \delta_{X_i}, \mu \right)$$

and Corollary 1.1 give $\mathbb{P}(X_1 + \dots + X_n \geq an) = O(n^{1-q})$ for each $a > 0$. It is known in this context that $\mathbb{P}(X_1 + \dots + X_n > an) = o(n^{1-q})$, and this exponent cannot be improved under the sole assumption of a finite q th moment [48, Chapter IX, Theorems 27–28]. A similar argument in the case $r > 1$ indicates that the exponent $q/r - 1$ is sharp in the first claim of Corollary 1.3.

There is now a substantial literature on rates of convergence of empirical measures of i.i.d. sequences in Wasserstein distance, and we refer to the recent paper [30] for the state of the

art and an overview of the many applications in quantization, interacting particle systems, *etc.* Yet, our result seems quite new in several respects. First, while the $n \rightarrow \infty$ convergence rate of the *expected distance*

$$M_n^{(r)} := \int_{E^n} \mathcal{W}_r^r(L_n, \mu) d\mu^n$$

is well understood, the (asymptotic) rate of convergence in n of the deviation probabilities given in Corollary 1.1 appears to be new. Case (3) in Theorem 2 of [30] gives some non-asymptotic bounds on these probabilities which are worse than ours in the $n \rightarrow \infty$ regime; the closest counterpart among their results is a bound of $O(n^{1-q+\epsilon})$ for any $\epsilon > 0$, but it is given only for $a > 1$ and $r < q/2$.

A second novelty of Corollary 1.1 is that it is valid in arbitrary Polish spaces, whereas most of the prior literature deals with Euclidean spaces. In the setting of tail probability bounds, a notable exception is the work of Boissard [12], which shows exponential decay in n of the probabilities $\mu^n(\mathcal{W}_1(L_n, \mu) \geq a)$ but under assumptions that the measure μ has finite exponential moments or satisfies a transport inequality. In the study of the expected distances $M_n^{(r)}$, it is well known that the dimension of the underlying space (or more generally a notion of metric entropy as in [19] and [54]) must absolutely come into play. For example, in Euclidean space $E \subset \mathbb{R}^d$ with $d > 2$, $M_n^{(1)}$ is known to be asymptotic to $n^{-1/d}$, at least when E is compact and μ is absolutely continuous with respect to Lebesgue measure [19]. Corollary 1.1 shows that this dimension dependence disappears from the probabilistic rate of convergence. Note that this leads to no contradiction: writing

$$M_n^{(1)} = \int_0^\infty \mu^n(\mathcal{W}_1(L_n, \mu) \geq a) da$$

and applying Corollary 1.1 does not imply $n^{q-1}M_n^{(1)} \rightarrow 0$, as the dominated convergence theorem does not apply here.

1.1.2. *Cramér's upper bound.* While Cramér's theorem in full generality, like Sanov's, does not require any finite moments, the upper bound is often vacuous when the underlying random variables have heavy tails. This simple observation has driven a large and growing literature on large deviation asymptotics for sums of i.i.d. random variables, to be reviewed shortly. This literature is full of *precise asymptotics*, mostly out of reach of our abstract framework. However, from Theorem 1.2 we can derive a modest alternative to Cramér's upper bound which is notable in its wide applicability. See Section 4.2 for a proof of the following.

Corollary 1.2. (Cramér upper bound.) *Let $q \in (1, \infty)$, and let E be a separable Banach space. Let $(X_i)_{i=1}^\infty$ be i.i.d. E -valued random variables with $\mathbb{E}\|X_1\|^q < \infty$. Define $\Lambda : E^* \rightarrow \mathbb{R} \cup \{\infty\}$ by*

$$\Lambda(x^*) = \inf\{m \in \mathbb{R} : \mathbb{E}[(1 + \langle x^*, X_1 \rangle - m)^+]^q \leq 1\},$$

and define $\Lambda^(x) = \sup_{x^* \in E^*} (\langle x^*, x \rangle - \Lambda(x^*))$ for $x \in E$. Then, for every closed set $A \subset E$,*

$$\limsup_{n \rightarrow \infty} n^{q-1} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i \in A\right) \leq \left(\inf_{x \in A} \Lambda^*(x)\right)^{-q}.$$

Here, E^ denotes the continuous dual of E .*

In analogy with the classical Cramér's theorem, the function Λ in Corollary 1.2 plays the role of the cumulant generating function. In both Theorem 1.2 and Corollary 1.2, notice that as

soon as the constant on the right-hand side is finite we may conclude that the probabilities in question are $O(n^{1-q})$, consistent with some now-standard results on one-dimensional heavy-tailed sums for events of the form $A = [r, \infty)$, for $r > 0$. For instance, as we mentioned in the previous subsection, if $(X_i)_{i=1}^\infty$ are i.i.d. real-valued random variables with mean zero and $\mathbb{E}|X_1|^q < \infty$, then the sharpest result possible under these assumptions is $\mathbb{P}(X_1 + \dots + X_n > nr) = o(n^{1-q})$. For $q > 2$, the Fuk–Nagaev inequality gives a related non-asymptotic bound; see [43, Corollary 1.8], or [25] for a Banach space version.

In general, we cannot expect a matching lower bound in Corollary 1.2, and thus we cannot expect one in Theorem 1.2. If stronger assumptions are made on X_i , such as regular variation, then corresponding lower bounds are known for certain sets A , but it remains unclear whether or not our abstract approach can recover such lower bounds. Refer to [13], [29], and [41] for detailed overviews of such results, as well as the more recent [17], [49], and references therein. Indeed, *precise asymptotics* require detailed assumptions on the shape of the tails of X_i , and this is especially true in multivariate and infinite-dimensional contexts. An interesting recent line of work extends the theory of regular variation to metric spaces [15, 34, 35, 40], but again the assumptions on the underlying μ are much stronger than mere existence of a finite moment.

The only real strengths of our Corollary 1.2, compared to the deep literature on sums of heavy-tailed random variables, is its broad applicability. It requires only finite moments, applies in general (separable) Banach spaces, and allows for arbitrary closed sets A , the latter point being useful in that it enables contraction principle arguments.

Before turning to the next application, it is worth mentioning a few more loosely related papers. In connection with concentration of measure, the papers of Bobkov and Ding [11, 18] studied transport inequalities involving functionals like (1.5), resulting in characterizations of certain non-exponential tail bounds. Less closely related, Atar *et al.* [4] exploited a variational representation for exponential integrals involving the functional (1.5) and showed how to use it to bound, for example, a large deviation probability for one model in terms of an alternative more tractable model; their work does not, however, appear to be applicable to situations with heavy tails.

1.1.3. Stochastic optimization. Let \mathcal{X} be another Polish space. Consider a continuous function $h: \mathcal{X} \times E \rightarrow \mathbb{R}$ bounded from below, and define $V: \mathcal{P}(E) \rightarrow \mathbb{R}$ by

$$V(\nu) = \inf_{x \in \mathcal{X}} \int_E h(x, w) \nu(dw).$$

Fix $\mu \in \mathcal{P}(E)$ again as a reference measure. The most common and natural approach to solving the optimization problem $V(\mu)$ numerically is to construct i.i.d. samples X_1, X_2, \dots with law μ and instead study $V(L_n(X_1, \dots, X_n))$, where as usual

$$L_n(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}.$$

The two obvious questions are then as follows.

- (A) Does $V(L_n(X_1, \dots, X_n))$ converge to $V(\mu)$?
- (B) Do the minimizers of $V(L_n(X_1, \dots, X_n))$ converge to those of $V(\mu)$ in some sense?

The answers to these questions are known to be affirmative in very general settings, using a form of set-convergence for question (B); see [21], [36], and [38]. Given this, we then hope

to quantify the rate of convergence for both of these questions. This is done in the language of large deviations in a paper of Kaniovski *et al.* [37], under a strong exponential integrability assumption derived from Cramér’s condition. In this section we complement their results by showing that under weaker integrability assumptions we can still obtain polynomial rates of convergence.

Theorem 1.3. *Suppose \mathcal{X} is compact. Suppose the function h is jointly continuous, and its sub-level sets are compact. Let $q \in (1, \infty)$ and $\mu \in \mathcal{P}(E)$ be such that, if*

$$\psi(w) := \left(\sup_{x \in \mathcal{X}} h(x, w) \right)^+,$$

then $\int_E \psi^q d\mu < \infty$. Then, for each $\epsilon > 0$,

$$\limsup_{n \rightarrow \infty} n^{q-1} \mu^n (|V(L_n) - V(\mu)| \geq \epsilon) < \infty.$$

The proof is given in Section 4.3, where we also present a related result on the rate of convergence of the optimizers themselves, addressing question (B) above.

1.2. Uniform upper bounds and martingales

Certain classes of dependent sequences admit uniform upper bounds, which we derive from Theorem 1.1 by working with

$$\alpha(v) = \inf_{\mu \in M} H(v | \mu),$$

for a given convex weakly compact set $M \subset \mathcal{P}(E)$. The conjugate ρ , unsurprisingly, is $\rho(f) = \sup_{\mu \in M} \log \int e^f d\mu$, and ρ_n turns out to be tractable as well, that is,

$$\rho_n(f) = \sup_{\mu \in M_n} \log \int_{E^n} e^f d\mu,$$

where M_n is defined as the set of laws $\mu \in \mathcal{P}(E^n)$ with $\mu_{k-1,k} \in M$ for each $k = 1, \dots, n$, μ -almost surely; in other words, M_n is the set of laws of E^n -valued random variables (X_1, \dots, X_n) , when the law of X_1 belongs to M and so does the conditional law of X_k given (X_1, \dots, X_{k-1}) , almost surely, for each $k = 2, \dots, n$. Theorem 1.1 becomes

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \sup_{\mu \in M_n} \int_{E^n} e^{nF \circ L_n} d\mu = \sup_{\mu \in M, v \in \mathcal{P}(E)} (F(v) - H(v | \mu)) \quad \text{for } F \in C_b(\mathcal{P}(E)).$$

From this we derive a uniform large deviation upper bound, for closed sets $A \subset \mathcal{P}(E)$:

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log \sup_{\mu \in M_n} \mu(L_n \in A) \leq - \inf_{\mu \in M, v \in A} H(v | \mu). \tag{1.7}$$

With a prudent choice of M , this specializes to an asymptotic relative of the Azuma–Hoeffding inequality. The novel feature here is that we can work with arbitrary closed sets and in multiple dimensions.

Theorem 1.4. *Let $\varphi: \mathbb{R}^d \rightarrow \mathbb{R}$, and define $S_{d,\varphi}$ to be the set of \mathbb{R}^d -valued martingales $(S_k)_{k=0}^n$, defined on a common but arbitrary probability space, satisfying $S_0 = 0$ and*

$$\mathbb{E}[\exp(\langle y, S_k - S_{k-1} \rangle) | S_0, \dots, S_{k-1}] \leq e^{\varphi(y)} \quad \text{a.s. for } k = 1, \dots, n, y \in \mathbb{R}^d.$$

Then, for closed sets $A \subset \mathbb{R}^d$, we have

$$\limsup_{n \rightarrow \infty} \sup_{(S_k)_{k=0}^n \in \mathcal{S}_{d,\varphi}} \frac{1}{n} \log \mathbb{P}(S_n/n \in A) \leq - \inf_{x \in A} \varphi^*(x),$$

where $\varphi^*(x) = \sup_{y \in \mathbb{R}^d} (\langle x, y \rangle - \varphi(y))$.

By taking $(S_k)_{k=0}^n$ to be a random walk (i.e. the increments are i.i.d.) such that $\varphi(y) = \log \mathbb{E}[e^{\langle y, S_1 - S_0 \rangle}] < \infty$, it is readily checked that the bound of Theorem 1.4 coincides with the upper bound from Cramér’s theorem and is thus sharp. Föllmer and Knispel [26] found some results which loosely resemble (1.7) (see Corollary 5.3 therein), based on an analysis of the same risk measure ρ . See also [33] and [31] for somewhat related results on large deviations for capacities.

1.3. Laws of large numbers

Some specializations of Theorem 1.1 appear to have nothing to do with large deviations. For example, suppose $M \subset \mathcal{P}(E)$ is convex and compact, and let

$$\alpha(v) = \begin{cases} 0 & \text{if } v \in M, \\ \infty & \text{otherwise.} \end{cases}$$

It can be shown that $\rho_n(f) = \sup_{\mu \in M_n} \int_{E^n} f \, d\mu$, where M_n is defined as in Section 1.2, for instance by a direct computation using (1.4). Theorem 1.1 then becomes

$$\lim_{n \rightarrow \infty} \sup_{\mu \in M_n} \int_{E^n} F \circ L_n \, d\mu = \sup_{\mu \in M} F(\mu) \quad \text{for each } F \in C_b(\mathcal{P}(E)). \tag{1.8}$$

When $M = \{\mu\}$ is a singleton, so is $M_n = \{\mu^n\}$, and this simply expresses the weak convergence $\mu^n \circ L_n^{-1} \rightarrow \delta_\mu$. The general case can be interpreted as a robust law of large numbers, where ‘robust’ refers to perturbations of the joint law of an i.i.d. sequence. More precisely, noting that $\sup_{\mu \in M} F(\mu) = \sup_{Q \in \mathcal{P}(M)} \int F \, dQ$, one can derive from (1.8) certain forms of set-convergence (e.g. Painlevé–Kuratowski) of the sequence $\{\mu \circ L_n^{-1} : \mu \in M_n\}$ toward $\mathcal{P}(M) := \{Q \in \mathcal{P}(E) : Q(M) = 1\}$, though we refrain from lengthening the paper with further details. In another direction, (1.8) is closely related to laws of large numbers under nonlinear expectations [46].

1.4. Optimal transport costs

Another interesting consequence of Theorem 1.1 comes from choosing α as an optimal transport cost. Fix $\mu \in \mathcal{P}(E)$ and a lower semicontinuous function $c : E^2 \rightarrow [0, \infty]$, and define

$$\alpha(v) = \inf_{\pi \in \Pi(\mu, v)} \int c \, d\pi,$$

where $\Pi(\mu, v)$ was defined immediately after (1.6). Under a modest additional assumption on c (stated shortly in Corollary 1.3, proved later in Lemma 6.2), α satisfies our standing assumptions.

The dual ρ can be identified using Kantorovich duality, and ρ_n turns out to be the value of a stochastic optimal control problem. To illustrate this, it is convenient to work with probabilistic

notation. Suppose $(X_i)_{i=1}^\infty$ is a sequence of i.i.d. E -valued random variables with common law μ , defined on some fixed probability space. For each n , let \mathcal{Y}_n denote the set of E^n -valued random variables (Y_1, \dots, Y_n) where Y_k is (X_1, \dots, X_k) -measurable for each $k = 1, \dots, n$. We think of elements of \mathcal{Y}_n as adapted control processes. For each $n \geq 1$ and each $f \in B(E^n)$, we show in Proposition 6.1 that

$$\rho_n(f) = \sup_{(Y_1, \dots, Y_n) \in \mathcal{Y}_n} \mathbb{E} \left[f(Y_1, \dots, Y_n) - \sum_{i=1}^n c(X_i, Y_i) \right]. \tag{1.9}$$

The expression (1.9) yields the following corollary of Theorem 1.1.

Corollary 1.3. *Suppose that for each compact set $K \subset E$, the function $h_K(y) := \inf_{x \in K} c(x, y)$ has pre-compact sub-level sets. That is, the closure of $\{y \in E : h_K(y) \leq m\}$ is compact for each $m \geq 0$. This assumption holds, for example, if E is a subset of Euclidean space and there exists $y_0 \in E$ such that $c(x, y) \rightarrow \infty$ as $d(y, y_0) \rightarrow \infty$, uniformly for x in compacts. For each $F \in C_b(\mathcal{P}(E))$, we have*

$$\begin{aligned} & \lim_{n \rightarrow \infty} \sup_{(Y_k)_{k=1}^n \in \mathcal{Y}_n} \mathbb{E} \left[F(L_n(Y_1, \dots, Y_n)) - \frac{1}{n} \sum_{i=1}^n c(X_i, Y_i) \right] \\ &= \sup_{\nu \in \mathcal{P}(E)} (F(\nu) - \alpha(\nu)) \\ &= \sup_{\pi \in \Pi(\mu)} \left(F(\pi(E \times \cdot)) - \int_{E \times E} c \, d\pi \right), \end{aligned} \tag{1.10}$$

where $\Pi(\mu) = \cup_{\nu \in \mathcal{P}(E)} \Pi(\mu, \nu)$.

This can be seen as a long-time limit of the optimal value of the control problems. However, the renormalization in n is a bit peculiar in that it enters inside the terminal cost F , and there does not seem to be a direct connection with ergodic control. A direct proof of (1.10) is possible but seems to be no simpler and potentially narrower in scope.

While the pre-limit expression in (1.10) may look peculiar, we include this example in part because it is remarkably tractable and in part because the limiting object is quite ubiquitous, encompassing a wide variety of variational problems involving optimal transport costs. Two notable recent examples can be found in the study of Cournot–Nash equilibria in large-population games [10] and in the theory of Wasserstein barycenters [2].

1.5. Alternative choices of α

There are many other natural choices of α for which the implications of Theorem 1.1 remain unclear. For example, consider the φ -divergence

$$\alpha(\nu) = \int_E \varphi(d\nu/d\mu) \, d\mu \quad \text{for } \nu \ll \mu, \quad \alpha(\nu) = \infty \quad \text{otherwise,}$$

where $\mu \in \mathcal{P}(E)$ and $\varphi: \mathbb{R}_+ \rightarrow \mathbb{R}$ is convex and satisfies $\varphi(x)/x \rightarrow \infty$ as $x \rightarrow \infty$. This α has weakly compact sub-level sets, according to [16, Lemma 6.2.16], and it is clearly convex. The dual, known in the risk literature as the *optimized certainty equivalent*, was computed by Ben-Tal and Teboulle [7, 8] to be

$$\rho(f) = \inf_{m \in \mathbb{R}} \left(\int_E \varphi^*(f(x) - m) \mu(dx) + m \right),$$

where $\varphi^*(x) = \sup_{y \in \mathbb{R}} (xy - \varphi(y))$ is the convex conjugate. We did not find any good expressions or estimates for ρ_n or α_n , so the interpretation of the main Theorem 1.1 eludes us in this case.

A related choice is the *shortfall risk measure* introduced by Föllmer and Schied [27]:

$$\rho(f) = \inf \left\{ m \in \mathbb{R} : \int_E \ell(f(x) - m) \mu(dx) \leq 1 \right\}.$$

This choice of ρ and the corresponding (tractable!) α are discussed briefly in Section 4.1. The choice of $\ell(x) = [(1+x)^+]^q$ corresponds to (1.5), and we make extensive use of this in Section 4, as was discussed in Section 1.1. The choice of $\ell(x) = e^x$ recovers the classical case $\rho(f) = \log \int_E e^f d\mu$. Aside from these two examples, for general ℓ , we found no useful expressions or estimates for ρ_n or α_n . In connection with tails of random variables, shortfall risk measures have an intuitive appeal stemming from the following simple analog of Chernoff’s bound, observed in [39, Proposition 3.3]. If $\gamma(\lambda) = \rho(\lambda f)$ for all $\lambda \geq 0$, where f is some given measurable function, then $\mu(f > t) \leq 1/\ell(\gamma^*(t))$ for all $t \geq 0$, where $\gamma^*(t) = \sup_{\lambda \geq 0} (\lambda t - \gamma(\lambda))$.

It is worth pointing out the natural but ultimately fruitless idea of working with

$$\rho(f) = \varphi^{-1} \left(\int_E \varphi(f) d\mu \right),$$

where φ is increasing. Such functionals were studied first – it seems – by Hardy, Littlewood, and Pólya [32, Chapter 3], providing necessary and sufficient conditions for ρ to be convex (rediscovered in [7]). Using the formula (1.4) to compute ρ_n , this choice would lead to the exceptionally pleasant formula

$$\rho_n(f) = \varphi^{-1} \left(\int_{E^n} \varphi(f) d\mu^n \right),$$

which we observed already in the classical case $\varphi(x) = e^x$. Unfortunately, however, such a ρ cannot come from a functional α on $\mathcal{P}(E)$, in the sense that (1.1) cannot hold unless φ is affine or exponential. The problem, as is known in the risk measure literature, is that the additivity property $\rho(f+c) = \rho(f) + c$ for all $c \in \mathbb{R}$ and $f \in B(E)$ fails unless φ is affine or exponential (cf. [28, Proposition 2.46]).

1.6. Interpreting Theorem 1.1 in terms of risk measures

It is straightforward to rewrite Theorem 1.1 in a language more in line with the literature on convex risk measures, for which we again defer to [28] for background. Let (Ω, \mathcal{F}) be a measurable space, and suppose φ is a convex risk measure on the set $B(\Omega, \mathcal{F})$ of bounded measurable functions. That is, $\varphi: B(\Omega, \mathcal{F}) \rightarrow \mathbb{R}$ is convex, $\varphi(f+c) = \varphi(f) + c$ for all $f \in B(\Omega, \mathcal{F})$ and $c \in \mathbb{R}$, and $\varphi(f) \geq \varphi(g)$ whenever $f \geq g$ pointwise. Suppose we are given a sequence of E -valued random variables $(X_i)_{i=1}^\infty$, i.e. measurable maps $X_i: \Omega \rightarrow E$. Assume X_i have the following independence property, identical to Peng’s notion of independence under nonlinear expectations [47]: for $n \geq 1$ and $f \in B(E^n)$,

$$\varphi(f(X_1, \dots, X_n)) = \varphi[\varphi(f(X_1, \dots, X_{n-1}, x))|_{x=X_n}].$$

In particular, $\varphi(f(X_i)) = \varphi(f(X_1))$ for all i . Define $\alpha: \mathcal{P}(E) \rightarrow (-\infty, \infty]$ by

$$\alpha(v) = \sup_{f \in B(E)} \left(\int_E f dv - \varphi(f(X_1)) \right).$$

Additional assumptions on φ (see e.g. Theorem 2.2 below) can ensure that α has weakly compact sub-level sets, so that Theorem 1.1 applies. Then, for $F \in C_b(\mathcal{P}(E))$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \varphi(nF(L_n(X_1, \dots, X_n))) = \sup_{\nu \in \mathcal{P}(E)} (F(\nu) - \alpha(\nu)). \tag{1.11}$$

Indeed, in our previous notation, $\rho_n(f) = \varphi(f(X_1, \dots, X_n))$ for $f \in B(E^n)$.

In the risk measure literature, one thinks of $\varphi(f)$ as the risk associated with an uncertain financial loss $f \in B(\Omega, \mathcal{F})$. With this in mind, and with $Z_n = F(L_n(X_1, \dots, X_n))$, the quantity $\varphi(nZ_n)$ appearing in (1.11) is the risk-per-unit of an investment in n units of Z_n . One might interpret Z_n as capturing the *composition* of the investment, while the multiplicative factor n represents the *size* of the investment. As n increases, say to $n + 1$, the investment is ‘rebalanced’ in the sense that one additional independent component, X_{n+1} , is incorporated and the size of the total investment is increased by one unit. The limit in (1.11) is then an asymptotic evaluation of the risk-per-unit of this rebalancing scheme.

1.7. Extensions

Broadly speaking, the book of Dupuis and Ellis [22] and numerous subsequent works illustrate how the classical convex duality between relative entropy and cumulant generating functions can serve as a foundation from which to derive an impressive range of large deviation principles. Similarly, each alternative dual pair (α, ρ) should provide an alternative foundation for a potentially equally wide range of limit theorems. From this perspective, our work raises more questions than it answers by restricting attention to analogs of the two large deviation principles of Sanov and Cramér. It is possible, for instance, that an analog of Mogulskii’s theorem (see [42] or [22, Section 3]) holds in our context, though one must not expect any such analog to look too much like a heavy-tailed large deviation principle, in light of the negative result of [49, Section 4.4]. These speculations are pursued no further but are meant to convey the versatility of our framework. In fact, extensions and applications of our framework have appeared since the first version of this paper. First, [23] extended the ideas beyond the i.i.d. setting, to the study of occupation measures of Markov chains. More recently, [5] applied Theorem 1.1 to obtain new limit theorems for Brownian motion, with connections to Schilder’s theorem, vanishing noise limits of BSDEs and PDEs, and Schrödinger problems.

1.8. Outline of the paper

The remainder of the paper is organized as follows. Section 2 begins by clarifying the (α, ρ) duality, explaining some useful properties of ρ and ρ_n and extending their definitions to unbounded functions before giving. In Section 2.2 we give the proof of an extension of Theorem 1.1, which contains Theorem 1.1 as a special case but is extended to stronger topologies and unbounded functions F . See also Section 2.3 for abstract analogs of the contraction principle and Cramér’s theorem. Section 3 elaborates on the additional topological assumptions needed for the extension of the main theorem. Then, Section 4 focuses on the particular choice of α in (1.5), providing proofs of the claims of Section 1.1. Sections 5 and 6 respectively elaborate on the examples of 1.2 and 1.4. Appendix A proves a different representations of ρ_n , namely those of (1.4). Finally, two minor technical results are relegated to Appendix B.

2. Convex duality and an extension of Theorem 1.1

We begin by outlining the key features of the (α, ρ) duality, as a first step toward stating and proving an extension of the main theorem as well as an abstract analog of Cramér’s theorem.

The first two theorems below are borrowed from the literature on convex risk measures, for which an excellent reference is the book of Föllmer and Schied [28]. While we will make use of some of the properties listed in Theorem 2.1, the goal of the first two theorems is more to illustrate how one can make ρ the starting point rather than α . In particular, Theorem 2.2 will not be needed below. For proofs of Theorems 2.1 and 2.2, refer to Bartl [6, Theorem 2.6].

Theorem 2.1. *Suppose $\alpha : \mathcal{P}(E) \rightarrow (-\infty, \infty]$ is convex and has weakly compact sub-level sets. Define $\rho : B(E) \rightarrow \mathbb{R}$ as in (1.1). Then the following hold.*

- (R1) *If $f \geq g$ pointwise then $\rho(f) \geq \rho(g)$.*
- (R2) *If $f \in B(E)$ and $c \in \mathbb{R}$, then $\rho(f + c) = \rho(f) + c$.*
- (R3) *If $f, f_n \in B(E)$ with $f_n \uparrow f$ pointwise, then $\rho(f_n) \uparrow \rho(f)$.*
- (R4) *If $f_n \in C_b(E)$ and $f \in B(E)$ with $f_n \downarrow f$ pointwise, then $\rho(f_n) \downarrow \rho(f)$.*

Moreover, for $v \in \mathcal{P}(E)$ we have

$$\alpha(v) = \sup_{f \in C_b(E)} \left(\int_E f \, d\nu - \rho(f) \right). \tag{2.1}$$

Theorem 2.2. *Suppose $\rho : B(E) \rightarrow \mathbb{R}$ is convex and satisfies properties (R1–R4) of Theorem 2.1. Define $\alpha : \mathcal{P}(E) \rightarrow (-\infty, \infty]$ by (2.1). Then α is convex and has weakly compact sub-level sets. Moreover, the identity (1.1) holds.*

For the rest of the paper, unless stated otherwise, we work at all times with the standing assumptions on α described in the Introduction.

Standing assumptions. The function $\alpha : \mathcal{P}(E) \rightarrow (-\infty, \infty]$ is convex, has weakly compact sub-level sets, and is not identically equal to ∞ . Lastly, ρ is defined as in (1.1).

We next extend the domain of ρ and ρ_n to unbounded functions. Let $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. We adopt the convention that $\infty - \infty := -\infty$, although this will have few consequences aside from streamlined definitions. In particular, if $\nu \in \mathcal{P}(E^n)$ and a measurable function $f : E^n \rightarrow \overline{\mathbb{R}}$ and satisfies $\int f^- \, d\nu = \int f^+ \, d\nu = \infty$, we define $\int f \, d\nu = -\infty$.

Definition 2.1. For $n \geq 1$ and measurable $f : E^n \rightarrow \overline{\mathbb{R}}$, define

$$\rho_n(f) = \sup_{\nu \in \mathcal{P}(E^n)} \left(\int_{E^n} f \, d\nu - \alpha_n(\nu) \right).$$

As usual, abbreviate $\rho \equiv \rho_1$. It is worth emphasizing that while $\rho(f)$ is finite for bounded f , it can be either $+\infty$ or $-\infty$ when f is unbounded.

2.1. Stronger topologies on $\mathcal{P}(E)$

As a last preparation, we discuss a well-known class of topologies on subsets of $\mathcal{P}(E)$ with which we will work frequently. Given a continuous function $\psi : E \rightarrow \mathbb{R}_+ := [0, \infty)$, define

$$\mathcal{P}_\psi(E) = \left\{ \mu \in \mathcal{P}(E) : \int_E \psi \, d\mu < \infty \right\}.$$

Endow $\mathcal{P}_\psi(E)$ with the (Polish) topology generated by the maps $\nu \mapsto \int_E f \, d\nu$, where $f: E \rightarrow \mathbb{R}$ is continuous and $|f| \leq 1 + \psi$; we call this the ψ -weak topology. A useful fact about this topology is that a set $M \subset \mathcal{P}_\psi(E)$ is pre-compact if and only if for every $\epsilon > 0$ there exists a compact set $K \subset E$ such that

$$\sup_{\mu \in M} \int_{K^c} \psi \, d\mu \leq \epsilon.$$

This is easily proved directly using Prokhorov’s theorem, or refer to [28, Corollary A.47]. It is worth noting that if d is a compatible metric on E and $\psi(x) = d^p(x, x_0)$ for some fixed $x_0 \in E$ and $p \geq 1$, then the ψ -weak topology is simply the p -Wasserstein topology associated with the metric d [52, Theorem 7.12].

2.2. An extension of Theorem 1.1

In this section we state and prove a useful generalization of Theorem 1.1 for stronger topologies and unbounded functions, taking advantage of the preparations of the previous sections. At all times in this section, the standing assumptions on (α, ρ) (stated early in Section 2) are in force. Part of Theorem 2.3 below requires the assumption that the sub-level sets of α are pre-compact in $\mathcal{P}_\psi(E)$, and this rather opaque assumption will be explored in more detail in Section 3.1.

Theorem 2.3. *Let $\psi: E \rightarrow \mathbb{R}_+$ be continuous. If $F: \mathcal{P}_\psi(E) \rightarrow \mathbb{R} \cup \{\infty\}$ is lower semicontinuous (with respect to the ψ -weak topology) and bounded from below, then*

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \rho_n(nF \circ L_n) \geq \sup_{\nu \in \mathcal{P}_\psi(E)} (F(\nu) - \alpha(\nu)).$$

Suppose also that the sub-level sets of α are pre-compact subsets of $\mathcal{P}_\psi(E)$. If $F: \mathcal{P}_\psi(E) \rightarrow \mathbb{R} \cup \{-\infty\}$ is upper semicontinuous and bounded from above, then

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \rho_n(nF \circ L_n) \leq \sup_{\nu \in \mathcal{P}_\psi(E)} (F(\nu) - \alpha(\nu)).$$

Proof of lower bound. Let us first prove the lower bound. It is immediate from the definition that $n^{-1}\alpha_n(\nu^n) = \alpha(\nu)$ for each $\nu \in \mathcal{P}(E)$, recalling that ν^n denotes the n -fold product measure. Thus

$$\begin{aligned} \frac{1}{n} \rho_n(nF(L_n)) &= \sup_{\nu \in \mathcal{P}(E^n)} \left\{ \int_{E^n} F \circ L_n \, d\nu - \frac{1}{n} \alpha_n(\nu) \right\} \\ &\geq \sup_{\nu \in \mathcal{P}(E)} \left\{ \int_{E^n} F \circ L_n \, d\nu^n - \frac{1}{n} \alpha_n(\nu^n) \right\} \\ &= \sup_{\nu \in \mathcal{P}(E)} \left\{ \int_{E^n} F \circ L_n \, d\nu^n - \alpha(\nu) \right\}. \end{aligned} \tag{2.2}$$

For $\nu \in \mathcal{P}(E)$, the law of large numbers (see [20, Theorem 11.4.1]) implies $\nu^n \circ L_n^{-1} \rightarrow \delta_\nu$ weakly, i.e. in $\mathcal{P}(\mathcal{P}(E))$. For $\nu \in \mathcal{P}_\psi(E)$, the convergence takes place in $\mathcal{P}(\mathcal{P}_\psi(E))$. Lower semicontinuity of F on $\mathcal{P}_\psi(E)$ then implies (e.g. by [22, Theorem A.3.12]), for each $\nu \in \mathcal{P}_\psi(E)$,

$$\begin{aligned} \liminf_{n \rightarrow \infty} \frac{1}{n} \rho_n(nF(L_n)) &\geq \liminf_{n \rightarrow \infty} \int_{E^n} F \circ L_n \, d\nu^n - \alpha(\nu) \\ &\geq F(\nu) - \alpha(\nu). \end{aligned}$$

Take the supremum over ν to complete the proof of the lower bound. □

Proof of upper bound, F bounded. The upper bound is more involved. First we prove it in four steps under the assumption that F is bounded.

Step 1. First we simplify the expression somewhat. For each $\nu \in \mathcal{P}(E^n)$ the definition of α_n and convexity of α imply

$$\begin{aligned} \frac{1}{n}\alpha_n(\nu) &= \frac{1}{n} \sum_{k=1}^n \int_{E^n} \alpha(\nu_{k-1,k}(x_1, \dots, x_{k-1})) \nu(dx_1, \dots, dx_n) \\ &\geq \int_{E^n} \alpha \left(\frac{1}{n} \sum_{k=1}^n \nu_{k-1,k}(x_1, \dots, x_{k-1}) \right) \nu(dx_1, \dots, dx_n). \end{aligned}$$

Combine this with (2.2) to get

$$\frac{1}{n}\rho_n(nF(L_n)) \leq \sup_{\nu \in \mathcal{P}(E^n)} \int_{E^n} \left[F(L_n) - \alpha \left(\frac{1}{n} \sum_{k=1}^n \nu_{k-1,k} \right) \right] d\nu. \tag{2.3}$$

Now choose arbitrarily some μ_f such that $\alpha(\mu_f) < \infty$. The choice $\nu = \mu_f^n$ and boundedness of F show that the supremum in (2.3) is bounded below by $-\|F\|_\infty - \alpha(\mu_f)$, where $\|F\|_\infty := \sup_{\nu \in \mathcal{P}_\psi(E)} |F(\nu)|$. For each n , choose $\nu^{(n)} \in \mathcal{P}(E^n)$ attaining the supremum in (2.3) to within $1/n$. Then

$$\int_{E^n} \alpha \left(\frac{1}{n} \sum_{k=1}^n \nu_{k-1,k}^{(n)} \right) d\nu^{(n)} \leq 2\|F\|_\infty + \alpha(\mu_f) + \frac{1}{n}. \tag{2.4}$$

It is convenient to switch now to a probabilistic notation. On some sufficiently rich probability space, find an E^n -valued random variable (Y_1^n, \dots, Y_n^n) with law $\nu^{(n)}$. Define the random measures

$$S_n := \frac{1}{n} \sum_{k=1}^n \nu_{k-1,k}^{(n)}(Y_1^n, \dots, Y_{k-1}^n), \quad \tilde{S}_n := \frac{1}{n} \sum_{k=1}^n \delta_{Y_k^n}.$$

Use (2.3) and unwrap the definitions to find

$$\frac{1}{n}\rho_n(nF(L_n)) \leq \mathbb{E}[F(\tilde{S}_n) - \alpha(S_n)] + 1/n. \tag{2.5}$$

Moreover, (2.4) implies

$$\sup_n \mathbb{E}[\alpha(S_n)] \leq 2\|F\|_\infty + \alpha(\mu_f) + 1 < \infty. \tag{2.6}$$

Step 2. We next show that the sequence (S_n, \tilde{S}_n) is tight, viewed as $\mathcal{P}_\psi(E) \times \mathcal{P}_\psi(E)$ -valued random variables. Here we use the assumption that the sub-level sets of α are ψ -weakly compact subsets of $\mathcal{P}_\psi(E)$. It then follows from (2.6) that (S_n) is tight (see e.g. [22, Theorem A.3.17]).

To see that the pair (S_n, \tilde{S}_n) is tight, it remains to check that $(\tilde{S}_n)_n$ is tight. To this end, we first notice that S_n and \tilde{S}_n have the same mean measure for each n , in the sense that for every

$f \in B(E)$ we have

$$\begin{aligned} \mathbb{E} \left[\int_E f \, dS_n \right] &= \mathbb{E} \left[\frac{1}{n} \sum_{k=1}^n \mathbb{E}[f(Y_k^n) \mid Y_1^n, \dots, Y_{k-1}^n] \right] \\ &= \mathbb{E} \left[\frac{1}{n} \sum_{k=1}^n f(Y_k^n) \right] \\ &= \mathbb{E} \left[\int_E f \, d\tilde{S}_n \right]. \end{aligned} \tag{2.7}$$

To prove (\tilde{S}_n) is tight, it suffices (by Prokhorov's theorem) to show that for all $\epsilon > 0$ there exists a ψ -weakly compact set $K \subset \mathcal{P}_\psi(E)$ such that $P(\tilde{S}_n \notin K) \leq \epsilon$. We will look for K of the form

$$K = \bigcap_{k=1}^\infty \left\{ \nu : \int_{C_k^c} \psi \, d\nu \leq 1/k \right\},$$

where $(C_k)_{k=1}^\infty$ is a sequence of compact subsets of E to be specified later; indeed, sets K of this form are pre-compact in $\mathcal{P}_\psi(E)$ according to a form of Prokhorov's theorem discussed in Section 2.1 (see also [28, Corollary A.47]). For such a set K , use Markov's inequality and (2.7) to compute

$$\begin{aligned} P(\tilde{S}_n \notin K) &\leq \sum_{k=1}^\infty P \left(\int_{C_k^c} \psi \, d\tilde{S}_n > 1/k \right) \\ &\leq \sum_{k=1}^\infty k \mathbb{E} \int_{C_k^c} \psi \, d\tilde{S}_n \\ &= \sum_{k=1}^\infty k \mathbb{E} \int_{C_k^c} \psi \, dS_n. \end{aligned} \tag{2.8}$$

By a form of Jensen's inequality (see Lemma B.2),

$$\sup_n \alpha(\mathbb{E}S_n) \leq \sup_n \mathbb{E}[\alpha(S_n)] < \infty,$$

where $\mathbb{E}S_n$ is the probability measure on E defined by $(\mathbb{E}S_n)(A) = \mathbb{E}[S_n(A)]$. Hence, the sequence $(\mathbb{E}S_n)$ is pre-compact in $\mathcal{P}_\psi(E)$, thanks to the assumption that sub-level sets of α are pre-compact subsets of $\mathcal{P}_\psi(E)$. It follows that for every $\epsilon > 0$ there exists a compact set $C \subset E$ such that $\sup_n \mathbb{E} \int_{C^c} \psi \, dS_n \leq \epsilon$. With this in mind, we may choose C_k to make (2.8) arbitrarily small, uniformly in n . This shows that (\tilde{S}_n) is tight, completing Step 2.

Step 3. We next show that every limit in distribution of (S_n, \tilde{S}_n) is concentrated on the diagonal $\{(\nu, \nu) : \nu \in \mathcal{P}_\psi(E)\}$. By definition of $\nu_{k-1,k}^{(n)}$, we have

$$\mathbb{E} \left[f(Y_k^n) - \int_E f \, d\nu_{k-1,k}^{(n)}(Y_1^n, \dots, Y_{k-1}^n) \mid Y_1^n, \dots, Y_{k-1}^n \right] = 0 \quad \text{for } k = 1, \dots, n$$

for every $f \in B(E)$. That is, the terms inside the expectation form a martingale difference sequence. Thus, for $f \in B(E)$, we have

$$\begin{aligned} \mathbb{E} \left[\left(\int_E f \, dS_n - \int_E f \, d\tilde{S}_n \right)^2 \right] &= \mathbb{E} \left[\left(\frac{1}{n} \sum_{k=1}^n \left(f(Y_k^n) - \int_E f \, d\nu_{k-1,k}^{(n)}(Y_1^n, \dots, Y_{k-1}^n) \right) \right)^2 \right] \\ &= \frac{1}{n^2} \sum_{k=1}^n \mathbb{E} \left[\left(f(Y_k^n) - \int_E f \, d\nu_{k-1,k}^{(n)}(Y_1^n, \dots, Y_{k-1}^n) \right)^2 \right] \\ &\leq 2\|f\|_\infty^2/n, \end{aligned} \tag{2.9}$$

where $\|f\|_\infty := \sup_{x \in E} |f(x)|$. It is straightforward to check that (2.9) implies that every weak limit of (S_n, \tilde{S}_n) is concentrated on (i.e. almost surely belongs to) the diagonal $\{(\nu, \tilde{\nu}) : \nu \in \mathcal{P}(E)\}$ (cf. [22, Lemma 2.5.1(b)]). Indeed, if (S, \tilde{S}) is some $\mathcal{P}_\psi(E) \times \mathcal{P}_\psi(E)$ -valued random variable such that (S_n, \tilde{S}_n) converges in law to (S, \tilde{S}) , then (2.9) implies

$$\mathbb{E} \left[\left(\int_E f \, dS - \int_E f \, d\tilde{S} \right)^2 \right] = 0,$$

for each $f \in C_b(E)$, by continuity of the map

$$\mathcal{P}_\psi(E) \times \mathcal{P}_\psi(E) \ni (\nu, \tilde{\nu}) \mapsto \left(\int_E f \, d\nu - \int_E f \, d\tilde{\nu} \right)^2.$$

Hence, $\int_E f \, dS = \int_E f \, d\tilde{S}$ a.s. for each $f \in C_b(E)$, and arguing with a countable separating family from $C_b(E)$ (see e.g. [45, Theorem 6.6]) allows us to deduce that $S = \tilde{S}$ a.s.

Step 4. We can now complete the proof of the upper bound. With Step 3 in mind, fix a subsequence and a $\mathcal{P}_\psi(E)$ -valued random variable η such that $(S_n, \tilde{S}_n) \rightarrow (\eta, \eta)$ in distribution (where we relabeled the subsequence). Recall that α is bounded from below and ψ -weakly lower semicontinuous, whereas F is upper semicontinuous and bounded. Returning to (2.5), we conclude now that

$$\begin{aligned} \limsup_{n \rightarrow \infty} \frac{1}{n} \rho_n(nF(L_n)) &\leq \limsup_{n \rightarrow \infty} \mathbb{E}[F(\tilde{S}_n) - \alpha(S_n)] \\ &\leq \mathbb{E}[F(\eta) - \alpha(\eta)] \\ &\leq \sup_{\nu \in \mathcal{P}_\psi(E)} \{F(\nu) - \alpha(\nu)\}. \end{aligned}$$

Of course, we abused notation by relabeling the subsequences, but we have argued that for every subsequence there exists a further subsequence for which this bound holds, which proves the upper bound for F bounded. □

Proof of upper bound, unbounded F. With the proof complete for bounded F , we now remove the boundedness assumption using a natural truncation procedure. Let $F : \mathcal{P}(E) \rightarrow E \cup \{-\infty\}$ be upper semicontinuous and bounded from above. For $m > 0$ let $F_m := F \vee (-m)$. Since F_m is bounded and upper semicontinuous, the previous step yields

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \rho_n(nF_m(L_n)) \leq \sup_{\nu \in \mathcal{P}_\psi(E)} \{F_m(\nu) - \alpha(\nu)\} =: S_m,$$

for each $m > 0$. Since $F_m \geq F$, we have

$$\rho_n(nF_m(L_n)) \geq \rho_n(nF(L_n))$$

for each m , and it remains only to show that

$$\lim_{m \rightarrow \infty} S_m = \sup_{\nu \in \mathcal{P}_\psi(E)} \{F(\nu) - \alpha(\nu)\} =: S. \tag{2.10}$$

Clearly $S_m \geq S$, since $F_m \geq F$. Note that $S < \infty$, as F and α are bounded from above and from below, respectively. If $S = -\infty$, then $F(\nu) = -\infty$ whenever $\alpha(\nu) < \infty$, and we conclude that, as $m \rightarrow \infty$,

$$S_m \leq -m - \inf_{\nu \in \mathcal{P}(E)} \alpha(\nu) \rightarrow -\infty = S.$$

Now suppose instead that S is finite. Fix $\epsilon > 0$. For each $m > 0$, find $\nu_m \in \mathcal{P}(E)$ such that

$$F_m(\nu_m) - \alpha(\nu_m) + \epsilon \geq S_m \geq S. \tag{2.11}$$

Since F is bounded from above and $S > -\infty$, it follows that $\sup_m \alpha(\nu_m) < \infty$. The sub-level sets of α are ψ -weakly compact, and thus the sequence (ν_m) has a limit point (in $\mathcal{P}_\psi(E)$). Let ν_∞ denote any limit point, and suppose $\nu_{m_k} \rightarrow \nu_\infty$. Note that $\inf_m F_m(\nu_m) > -\infty$ in light of (2.11), because α is bounded from below. Hence, for all sufficiently large m , we have $F_m(\nu_m) = F(\nu_m)$. Thus

$$\limsup_{k \rightarrow \infty} \{F_{m_k}(\nu_{m_k}) - \alpha(\nu_{m_k})\} \leq F(\nu_\infty) - \alpha(\nu_\infty) \leq S,$$

where the second inequality follows from upper semicontinuity of F and lower semicontinuity of α . This holds for any limit point of the pre-compact sequence (ν_m) , and it follows from (2.11) that

$$S \leq \limsup_{m \rightarrow \infty} S_m \leq \limsup_{m \rightarrow \infty} \{F_m(\nu_m) - \alpha(\nu_m)\} + \epsilon \leq S + \epsilon.$$

Since $\epsilon > 0$ was arbitrary, this proves (2.10). □

Remark 2.1. Several natural choices of α in fact have sub-level sets which are compact in topology induced by bounded measurable test functions, i.e. the topology on $\mathcal{P}(E)$ generated by the maps $\nu \mapsto \int_E f \, d\nu$, where $f \in B(E)$. While this topology is stronger than the usual weak convergence topology, the conclusion of Theorem 1.1 will likely still hold for bounded functions F which are continuous in this stronger (non-metrizable) topology. This is known to be true in the classical case $\alpha(\cdot) = H(\cdot \mid \mu)$ (see e.g. [16, Section 6.2]), where we recall the definition of relative entropy H from (1.2). For the sake of brevity, we do not pursue this generalization.

2.3. Contraction principles and an abstract form of Cramér’s theorem

Viewing Theorem 2.3 as an abstract form of Sanov’s theorem, we may derive from it a form of Cramér’s theorem. The key tool is an analog of the contraction principle from classical large deviations (cf. [16, Theorem 4.2.1]). In its simplest form, if $\varphi: \mathcal{P}(E) \rightarrow E'$ is continuous for some topological space E' , then for $F \in C_b(E')$ we have from Theorem 1.1

$$\lim_{n \rightarrow \infty} \frac{1}{n} \rho_n(nF \circ \varphi \circ L_n) = \sup_{\nu \in \mathcal{P}(E)} (F(\varphi(\nu)) - \alpha(\nu)) = \sup_{x \in E'} (F(x) - \alpha_\varphi(x)),$$

where we define $\alpha_\varphi: E' \rightarrow (-\infty, \infty]$ by

$$\alpha_\varphi(x) := \inf\{\alpha(\nu): \nu \in \mathcal{P}(E), \varphi(\nu) = x\}.$$

This line of reasoning leads to the following extension of Cramér’s theorem.

Theorem 2.4. *Let $(E, \|\cdot\|)$ be a separable Banach space with continuous dual E^* . Define $\Lambda^*: E \rightarrow \mathbb{R} \cup \{\infty\}$ by*

$$\Lambda^*(x) = \sup_{x^* \in E^*} (\langle x^*, x \rangle - \rho(x^*)).$$

Define $S_n: E^n \rightarrow E$ by

$$S_n(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i.$$

If $F: E \rightarrow \mathbb{R} \cup \{\infty\}$ is lower semicontinuous and bounded from below, then

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \rho_n(nF \circ S_n) \geq \sup_{x \in E} (F(x) - \Lambda^*(x)).$$

Suppose also that the sub-level sets of α are pre-compact subsets of $\mathcal{P}_\psi(E)$, for $\psi(x) := \|x\|$. If $F: E \rightarrow \mathbb{R} \cup \{-\infty\}$ is upper semicontinuous and bounded from above, then

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \rho_n(nF \circ S_n) \leq \sup_{x \in E} (F(x) - \Lambda^*(x)).$$

The proof makes use of a proposition, interesting in its own right, which generalizes the well-known result that the functions

$$t \mapsto \log \int_{\mathbb{R}} e^{tx} \mu(dx) \quad \text{and} \quad t \mapsto \inf \left\{ H(v \mid \mu) : v \in \mathcal{P}(\mathbb{R}), \int_{\mathbb{R}} x v(dx) = t \right\}$$

are convex conjugates of each other (see e.g. [22, Lemma 3.3.3]).

Proposition 2.1. *Let $(E, \|\cdot\|)$ be a separable Banach space, and let $\psi(x) = \|x\|$. Suppose the sub-level sets of α are pre-compact subsets of $\mathcal{P}_\psi(E)$. Define $\Psi: E \rightarrow \mathbb{R} \cup \{\infty\}$ by*

$$\Psi(x) = \inf \left\{ \alpha(v) : v \in \mathcal{P}_\psi(E), \int_E z v(dz) = x \right\},$$

where the integral is in the sense of Bochner. Define Ψ^* on the continuous dual E^* by

$$\Psi^*(x^*) = \sup_{x \in E} (\langle x^*, x \rangle - \Psi(x)).$$

Then Ψ is convex and lower semicontinuous, and $\Psi^*(x^*) = \rho(x^*)$ for every $x^* \in E^*$. In particular,

$$\Psi(x) = \sup_{x^* \in E^*} (\langle x^*, x \rangle - \rho(x^*)). \tag{2.12}$$

Proof. We first show that Ψ is convex. Let $t \in (0, 1)$ and $x_1, x_2 \in E$. Fix $\epsilon > 0$, and find $v_1, v_2 \in \mathcal{P}_\psi(E)$ such that $\int_E z v_i(dz) = x_i$ and $\alpha(v_i) \leq \Psi(x_i) + \epsilon$. Convexity of α yields

$$\begin{aligned} \Psi(tx_1 + (1-t)x_2) &\leq \alpha(tv_1 + (1-t)v_2) \\ &\leq t\alpha(v_1) + (1-t)\alpha(v_2) \\ &\leq t\Psi(x_1) + (1-t)\Psi(x_2) + \epsilon. \end{aligned}$$

To prove that Ψ is lower semicontinuous, first note that Ψ is bounded from below since α is. Let $x_n \rightarrow x$ in E , and find $v_n \in \mathcal{P}_\psi(E)$ such that $\alpha(v_n) \leq \Psi(x_n) + 1/n$ and $\int_E z v_n(dz) = x_n$ for

each n . Fix a subsequence $\{x_{n_k}\}$ such that $\Psi(x_{n_k}) < \infty$ for all k and $\Psi(x_{n_k})$ converges to a finite value (if no such subsequence exists, then there is nothing to prove, as $\Psi(x_n) \rightarrow \infty$). Then $\sup_k \alpha(v_{n_k}) < \infty$, and because α has ψ -weakly compact sub-level sets there exists a further subsequence (again denoted n_k) and some $v_\infty \in \mathcal{P}_\psi(E)$ such that $v_{n_k} \rightarrow v_\infty$. The convergence $v_{n_k} \rightarrow v_\infty$ in the ψ -weak topology implies

$$x = \lim_{k \rightarrow \infty} x_{n_k} = \lim_{k \rightarrow \infty} \int_E z v_{n_k}(dz) = \int_E z v_\infty(dz).$$

Using lower semicontinuity of α we conclude

$$\Psi(x) \leq \alpha(v_\infty) \leq \liminf_{k \rightarrow \infty} \alpha(v_{n_k}) \leq \liminf_{k \rightarrow \infty} \Psi(x_{n_k}). \tag{2.13}$$

For every sequence (x_n) in E and any subsequence thereof, this argument shows that there exists a further subsequence for which (2.13) holds, and this proves that Ψ is lower semicontinuous. Next, compute Ψ^* as follows:

$$\begin{aligned} \Psi^*(x^*) &= \sup_{x \in E} (\langle x^*, x \rangle - \Psi(x)) \\ &= \sup_{x \in E} \sup \left\{ \langle x^*, x \rangle - \alpha(v) : v \in \mathcal{P}_\psi(E), \int_E z v(dz) = x \right\} \\ &= \sup_{v \in \mathcal{P}_\psi(E)} \left(\left\langle x^*, \int_E z v(dz) \right\rangle - \alpha(v) \right) \\ &= \sup_{v \in \mathcal{P}_\psi(E)} \left(\int_E \langle x^*, z \rangle v(dz) - \alpha(v) \right) \\ &= \rho(x^*). \end{aligned}$$

Indeed, we can take the supremum equivalently over $\mathcal{P}_\psi(E)$ or over $\mathcal{P}(E)$ in the last two steps, thanks to the assumption that $\alpha = \infty$ away from $\mathcal{P}_\psi(E)$ and our convention $\infty - \infty = -\infty$. Because Ψ is lower semicontinuous and convex, we conclude from the Fenchel–Moreau theorem [55, Theorem 2.3.3] that it is equal to its biconjugate, which is precisely what (2.12) says. □

Proof of Theorem 2.4. The map

$$\mathcal{P}_\psi(E) \ni \mu \mapsto F \left(\int_E z \mu(dz) \right)$$

is upper (resp. lower) semicontinuous as soon as F is upper (resp. lower) semicontinuous. The claims then follow from Theorem 2.3 and Proposition 2.1. □

3. Compactness of sub-level sets of α in $\mathcal{P}_\psi(E)$

Several results of the previous section, such as the upper bound of Theorem 2.3, operate under the assumption that the sub-level sets of α are pre-compact subsets of $\mathcal{P}_\psi(E)$. This section compiles some related properties of (ρ, α) which will be useful when we encounter specific examples later in the paper.

3.1. Cramér’s condition

A first useful result is a condition under which the effective domain of α is contained in $\mathcal{P}_\psi(E)$.

Proposition 3.1. *Fix a measurable function $\psi : E \rightarrow \mathbb{R}_+$. Suppose $\rho(\lambda\psi) < \infty$ for some $\lambda > 0$. Then, for each $\nu \in \mathcal{P}(E)$ satisfying $\alpha(\nu) < \infty$, we have $\int \psi \, d\nu < \infty$.*

Proof. By definition, for each $\nu \in \mathcal{P}(E)$,

$$\infty > \rho(\lambda\psi) \geq \lambda \int \psi \, d\nu - \alpha(\nu).$$

If $\alpha(\nu) < \infty$ then certainly $\int \psi \, d\nu < \infty$. □

The next and more important proposition identifies a condition under which the sub-level sets of α are not only weakly compact (which we always assume) but also ψ -weakly compact.

Proposition 3.2. *Fix a continuous function $\psi : E \rightarrow \mathbb{R}_+$. Suppose*

$$\lim_{m \rightarrow \infty} \rho(\lambda\psi 1_{\{\psi \geq m\}}) = \rho(0) \quad \text{for all } \lambda > 0. \tag{3.1}$$

Then, for each $c \in \mathbb{R}$, the weak and ψ -weak topologies coincide on $\{\nu \in \mathcal{P}(E) : \alpha(\nu) \leq c\} \subset \mathcal{P}_\psi(E)$; in particular, the sub-level sets of α are ψ -weakly compact.

A first step in the proof comes from the following simple lemma, worth stating separately for emphasis.

Lemma 3.1. *Fix a continuous function $\psi : E \rightarrow \mathbb{R}_+$. Suppose (3.1) holds. Then $\rho(\lambda\psi) < \infty$ for every $\lambda \geq 0$. In particular, for each $\nu \in \mathcal{P}(E)$ satisfying $\alpha(\nu) < \infty$, we have $\int \psi \, d\nu < \infty$.*

Proof. The second claim is just Proposition 3.1. For $m, \lambda > 0$ we have $\lambda\psi \leq \lambda m + \lambda\psi 1_{\{\psi \geq m\}}$, and thus properties (R1) and (R2) of Theorem 2.1 imply

$$\rho(\lambda\psi) \leq \lambda m + \rho(\lambda\psi 1_{\{\psi \geq m\}}).$$

By (3.1), for m sufficiently large the right-hand side is finite. □

Proof of Proposition 3.2. Fix $c \in \mathbb{R}$, and abbreviate $S = \{\nu \in \mathcal{P}(E) : \alpha(\nu) \leq c\}$. Assume $S \neq \emptyset$. Note that Lemma 3.1 implies $S \subset \mathcal{P}_\psi(E)$. It suffices to prove that the map $\nu \mapsto \int_E f \, d\nu$ is weakly continuous on S for every continuous $f : E \rightarrow \mathbb{R}$ with $|f| \leq 1 + \psi$. Note that for $\eta_n, \eta \in \mathcal{P}(\mathbb{R})$ with $\eta_n \rightarrow \eta$ weakly, we have $\int g \, d\eta_n \rightarrow \int g \, d\eta$ for each continuous function g which is uniformly integrable, in the sense that

$$\lim_{m \rightarrow \infty} \sup_n \int_{\{|g| \geq m\}} |g| \, d\eta_n = 0.$$

(See [22, Theorem A.3.19].) Applying this to the image measures $\{\nu \circ f^{-1} : \nu \in S\}$ for f as above, we find that it suffices to prove the uniform integrability condition

$$\lim_{m \rightarrow \infty} \sup_{\nu \in S} \int_{\{\psi \geq m\}} \psi \, d\nu = 0.$$

By definition of ρ , for $m > 0$ and $\nu \in S$,

$$\lambda \int_{\{\psi \geq m\}} \psi \, d\nu \leq \rho(\lambda\psi 1_{\{\psi \geq m\}}) + \alpha(\nu) \leq \rho(\lambda\psi 1_{\{\psi \geq m\}}) + c. \tag{3.2}$$

Given $\epsilon > 0$, choose $\lambda > 0$ large enough that $(\epsilon + \rho(0) + c)/\lambda \leq \epsilon$. Then choose m large enough that $\rho(\lambda\psi 1_{\{\psi \geq m\}}) \leq \epsilon + \rho(0)$, which is possible because of assumption (3.1). It then follows from (3.2) that $\int_{\{\psi \geq m\}} \psi \, d\nu \leq \epsilon$, and the proof is complete. \square

We refer to (3.1) as the *strong Cramér condition*. Several extensions of the classical form of Sanov’s theorem to stronger topologies rely on what might be called a ‘strong Cramér condition’. For instance, if $\psi : E \rightarrow \mathbb{R}_+$ is continuous, the results of Schied [50] indicate that Sanov’s theorem can be extended to the ψ -weak topology if (and essentially only if) $\log \int_E e^{\lambda\psi} \, d\mu < \infty$ for every $\lambda \geq 0$; see also [53] and [24].

The form of our strong Cramér condition (3.1) was heavily inspired by the work of Owari [44] on continuous extensions of monotone convex functionals. In several cases of interest (namely, Propositions 4.1 and 5.3 below), it turns out that a converse to Lemma 3.1 is true, that is, the strong Cramér condition (3.1) is equivalent to the statement that $\rho(\lambda\psi) < \infty$ for all $\lambda > 0$. In general, however, the strong Cramér condition is the strictly stronger statement. Consider the following simple example, borrowed from [44, Example 3.7]. Let $E = \{0, 1, \dots\}$ be the natural numbers, and define $\mu_n \in \mathcal{P}(E)$ by $\mu_1\{0\} = 1$, $\mu_n\{0\} = 1 - 1/n$, and $\mu_n\{n\} = 1/n$. Let M denote the closed convex hull of (μ_n) . Then M is convex and weakly compact. Define $\alpha(\mu) = 0$ for $\mu \in M$ and $\alpha(\mu) = \infty$ otherwise. Then α satisfies our standing assumptions, and $\rho(f) = \sup_{\mu \in M} \int f \, d\mu = \sup_n \int f \, d\mu_n$. Finally, let $\psi(x) = x$ for $x \in E$. Then $\rho(\lambda\psi) = \lambda < \infty$ because $\int \psi \, d\mu_n = 1$ for all n , and similarly $\rho(\lambda\psi 1_{\{\psi \geq m\}}) = \lambda$ because $\int \psi 1_{\{\psi \geq m\}} \, d\mu_n = 1_{\{n \geq m\}}$. In particular, $\rho(\lambda\psi) < \infty$ for all $\lambda > 0$, but the strong Cramér condition fails.

Finally, we remark that it is conceivable that a converse to Proposition 3.1 might hold, that is, the strong Cramér condition (3.1) may be *equivalent* to the pre-compactness of the sub-level sets of α in $\mathcal{P}_\psi(E)$. Indeed, the results of Schied [50, Theorem 2] and Owari [44, Theorem 3.8] suggest that this may be the case. But this remains an open problem.

3.2. Implications of ψ -weakly compact sub-level sets

This section contains two results to be used occasionally below. First is a useful lemma that aids in the computation of $\rho(f)$ for certain unbounded f in Section 4.

Lemma 3.2. *If $f : E \rightarrow \mathbb{R}$ is upper semicontinuous and bounded from above, then*

$$\rho(f) = \lim_{m \rightarrow \infty} \rho(f \vee (-m)) = \inf_{m \geq 0} \rho(f \vee (-m)).$$

If $f : E \rightarrow \mathbb{R}$ is measurable and bounded from below, then

$$\rho(f) = \lim_{m \rightarrow \infty} \rho(f \wedge m) = \sup_{m \geq 0} \rho(f \wedge m).$$

Lastly, let $\psi : E \rightarrow \mathbb{R}_+$ be continuous. If the sub-level sets of α are pre-compact subsets of $\mathcal{P}_\psi(E)$, and if $f : E \rightarrow \mathbb{R}$ is measurable with $f \geq -c(1 + \psi)$ pointwise for some $c \geq 0$, then

$$\rho(f) = \lim_{m \rightarrow \infty} \rho(f \wedge m) = \sup_{m \geq 0} \rho(f \wedge m).$$

Proof. The second claim is a special case of the final claim with $\psi \equiv 0$. To prove the final claim, note first that $\rho(f \wedge m)$ is non-decreasing in m (see (R1) of Theorem 2.1). We find

$$\rho(f) = \sup_{\nu \in \mathcal{P}_\psi(E)} \left(\int_E f \, d\nu - \alpha(\nu) \right) = \sup_{m \geq 0} \sup_{\nu \in \mathcal{P}_\psi(E)} \left(\int_E f \wedge m \, d\nu - \alpha(\nu) \right) = \sup_{m \geq 0} \rho(f \wedge m).$$

Indeed, for each $\nu \in \mathcal{P}_\psi(E)$, the monotone convergence theorem applies because $f \wedge m$ for $m \geq 0$ are bounded from below by the ν -integrable function $-c(1 + \psi)$. To prove the first claim, abbreviate $f_m = f \vee (-m)$ for $m \geq 0$. Monotonicity of ρ implies $\inf_{m \geq 0} \rho(f_m) \geq \rho(f)$, so we need only prove the reverse inequality. Assume without loss of generality that $\inf_{m \geq 0} \rho(f_m) > -\infty$. For each n , we may find for each n some $\nu_n \in \mathcal{P}_\psi(E)$ such that

$$-\infty < \inf_{m \geq 0} \rho(f_m) \leq \rho(f_n) \leq \int_E f_n \, d\nu_n - \alpha(\nu_n) + 1/n. \tag{3.3}$$

This implies $\sup_n \alpha(\nu_n) < \infty$, because f is bounded from above. Pre-compactness of the sub-level sets of α allows us to extract a subsequence n_k and $\nu_\infty \in \mathcal{P}(E)$ such that $\nu_{n_k} \rightarrow \nu_\infty$ weakly. By Skorokhod’s representation, we may construct random variables X_k and X_∞ with respective laws ν_{n_k} and ν_∞ such that $X_k \rightarrow X_\infty$ a.s. The upper semicontinuity assumption implies $\limsup_{k \rightarrow \infty} \int_E f_{n_k}(X_k) \leq \int_E f(X_\infty)$ almost surely. We then conclude from Fatou’s lemma that

$$\limsup_{k \rightarrow \infty} \int_E f_{n_k} \, d\nu_{n_k} = \limsup_{k \rightarrow \infty} \mathbb{E}[f_{n_k}(X_k)] \leq \mathbb{E}[f(X_\infty)] = \int_E f \, d\nu_\infty.$$

Since α is weakly lower semicontinuous, we conclude from (3.3) that

$$\inf_{m \geq 0} \rho(f_m) \leq \int_E f \, d\nu_\infty - \alpha(\nu_\infty) \leq \sup_{\nu \in \mathcal{P}(E)} \left(\int_E f \, d\nu - \alpha(\nu) \right) = \rho(f). \quad \square$$

4. Non-exponential large deviations

The goal of this section is to prove Theorem 1.2 and its consequences detailed in Section 1.1, but along the way we will explore a particularly interesting class of (α, ρ) pairs.

4.1. Shortfall risk measures

Fix $\mu \in \mathcal{P}(E)$ and a non-decreasing, non-constant, convex function $\ell: \mathbb{R} \rightarrow \mathbb{R}_+$ satisfying $\ell(x) < 1$ for all $x < 0$. Let $\ell^*(y) = \sup_{x \in \mathbb{R}} (xy - \ell(x))$ denote the convex conjugate, and define $\alpha: \mathcal{P}(E) \rightarrow [0, \infty]$ by

$$\alpha(\nu) = \begin{cases} \inf_{t > 0} \frac{1}{t} \left(1 + \int_E \ell^* \left(t \frac{d\nu}{d\mu} \right) d\mu \right) & \text{if } \nu \ll \mu, \\ \infty & \text{otherwise.} \end{cases} \tag{4.1}$$

Note that $\ell^*(x) \geq -\ell(0) \geq -1$, by assumption and by continuity of ℓ , so that $\alpha \geq 0$. Define ρ as usual by (1.1). It is known [28, Proposition 4.115] that, for $f \in B(E)$,

$$\rho(f) = \inf \left\{ m \in \mathbb{R} : \int_E \ell(f(x) - m) \mu(dx) \leq 1 \right\}. \tag{4.2}$$

Refer to the book of Föllmer and Schied [28, Section 4.9] for a thorough study of the properties of ρ . Notably, they show that ρ satisfies all of properties (R1–R4) of Theorem 2.1, and that both dual formulas hold:

$$\rho(f) = \sup_{\nu \in \mathcal{P}(E)} \left(\int_E f \, d\nu - \alpha(\nu) \right), \quad \alpha(\nu) = \sup_{f \in B(E)} \left(\int_E f \, d\nu - \rho(f) \right).$$

If $\ell(x) = e^x$ we recover $\rho(f) = \log \int_E e^f \, d\mu$ and $\alpha(\nu) = H(\nu \mid \mu)$. If $\ell(x) = [(1+x)^+]^q$ for some $q \geq 1$, then

$$\alpha(\nu) = \|\mathrm{d}\nu/\mathrm{d}\mu\|_{L^p(\mu)} - 1 \quad \text{for } \nu \ll \mu, \quad \alpha(\nu) = \infty \quad \text{otherwise,} \tag{4.3}$$

where $p = q/(q - 1)$, and where of course

$$\|f\|_{L^p(\mu)} = \left(\int |f|^p \, d\mu \right)^{1/p}.$$

See [28, Example 4.118] or [39, Section 3.1] for this computation. The -1 is a convenient normalization, ensuring that $\alpha(\nu) = 0$ if and only if $\nu = \mu$.

In the rest of this subsection we work with α and ρ given as in (4.1) and (4.2). The following result shows how the strong Cramér condition (3.1) simplifies in the present context. It is essentially contained in [44, Proposition 7.3], but we include the short proof.

Proposition 4.1. *Let $\psi : E \rightarrow \mathbb{R}_+$ be measurable. Suppose*

$$\int_E \ell(\lambda \psi(x)) \mu(\mathrm{d}x) < \infty \quad \text{for all } \lambda > 0.$$

Then the strong Cramér condition holds,

$$\lim_{m \rightarrow \infty} \rho(\lambda \psi 1_{\{\psi \geq m\}}) \rightarrow 0 \quad \text{for each } \lambda > 0.$$

In particular, the sub-level sets of α are compact subsets of $\mathcal{P}_\psi(E)$.

Proof. The final claim is simply an application of Proposition 3.2. Fix $\epsilon > 0$ and $\lambda > 0$. Since ℓ is non-decreasing, the following two limits hold:

$$\lim_{m \rightarrow \infty} \mu(\psi < m) = 1, \quad \lim_{m \rightarrow \infty} \int_{\{\psi \geq m\}} \ell(\lambda \psi(x) - \epsilon) \mu(\mathrm{d}x) = 0.$$

Since $\ell(-\epsilon) < 1$, it follows that, for sufficiently large m ,

$$\begin{aligned} 1 &\geq \ell(-\epsilon) \mu(\psi < m) + \int_{\{\psi \geq m\}} \ell(\lambda \psi(x) - \epsilon) \mu(\mathrm{d}x) \\ &= \int_E \ell(\lambda \psi(x) 1_{\{\psi \geq m\}}(x) - \epsilon) \mu(\mathrm{d}x). \end{aligned}$$

Next, the second assertion of Lemma 3.2 implies $\rho(\lambda \psi 1_{\{\psi \geq m\}}) = \sup_{n \geq 0} \rho(n \wedge (\lambda \psi 1_{\{\psi \geq m\}}))$. For each n , we use the identity (4.2), which is valid for bounded f , to get, for sufficiently large m ,

$$\begin{aligned} \rho(\lambda \psi 1_{\{\psi \geq m\}}) &= \sup_{n \geq 0} \rho(n \wedge (\lambda \psi 1_{\{\psi \geq m\}})) \\ &= \sup_{n \geq 0} \inf \left\{ c \in \mathbb{R} : \int_E \ell(n \wedge (\lambda \psi 1_{\{\psi \geq m\}}) - c) \, d\mu \leq 1 \right\} \end{aligned}$$

$$\begin{aligned} &\leq \inf \left\{ c \in \mathbb{R} : \int_E \ell(\lambda \psi 1_{\{\psi \geq m\}} - c) \, d\mu \leq 1 \right\} \\ &\leq \epsilon. \end{aligned} \quad \square$$

Note that (4.2) is only valid, *a priori*, for bounded f , although the expression on the right-hand side certainly makes sense for unbounded f . The next results provide some useful cases for which the identity (4.2) carries over to unbounded functions, and these will be needed in the proof of Corollary 1.2. In the following, define $\ell(\pm\infty) = \lim_{x \rightarrow \pm\infty} \ell(x)$.

Proposition 4.2. *Let $\psi : E \rightarrow \mathbb{R}_+$ be continuous, and suppose $\int_E \ell(\lambda \psi(x)) \mu(dx) < \infty$ for all $\lambda > 0$. Suppose $f : E \rightarrow \mathbb{R}$ is continuous with $|f| \leq c(1 + \psi)$ pointwise for some $c \geq 0$. Then the identity (4.2) holds.*

Proof. Let $H(f)$ denote the right-hand side of (4.2), well-defined for any measurable function $f : E \rightarrow \mathbb{R}$. We must show $\rho(f) = H(f)$ for f as in the statement of the proposition. As was mentioned above, it is known from [28, Proposition 4.115] that $\rho(f) = H(f)$ for bounded f .

Step 1. Assume first that f is continuous and bounded from above, with $|f| \leq c(1 + \psi)$. Let $f_n = f \vee (-n)$ for $n \geq 0$. Since f_n is bounded for each n , we have $\rho(f_n) = H(f_n)$. The first assertion of Lemma 3.2 then implies $\rho(f) = \lim_{n \rightarrow \infty} \rho(f_n) = \lim_{n \rightarrow \infty} H(f_n)$. It remains to show $H(f_n) \rightarrow H(f)$. Clearly $H(f_n) \geq H(f_{n+1}) \geq H(f)$ for each n since $f_n \geq f_{n+1}$ pointwise and ℓ is non-decreasing, so the sequence $H(f_n)$ has a limit. As ℓ is continuous and strictly increasing in a neighborhood of the origin, note that $H(f)$ is the unique solution $c \in \mathbb{R}$ of the equation

$$\int_E \ell(f(x) - c) \mu(dx) = 1. \tag{4.4}$$

Similarly, $H(f_n)$ uniquely solves

$$\int_E \ell(f_n(x) - H(f_n)) \mu(dx) = 1.$$

Let $c = \lim_{n \rightarrow \infty} H(f_n)$, and note that the integrands $\ell(f_n(x) - H(f_n))$ are uniformly bounded and converge pointwise to $\ell(f(x) - c)$. Passing to the limit using dominated convergence shows that c solves the equation (4.4), which implies $c = H(f)$.

Step 2. We now turn to general continuous f satisfying $|f| \leq c(1 + \psi)$. Define $f_n = f \wedge n$ for $n \geq 0$, so that f_n is bounded from above. By Step 2, $\rho(f_n) = H(f_n)$ for each n . By Proposition 4.1, the sub-level sets of α are ψ -weakly compact, and the third assertion of Lemma 3.2 yields $\rho(f_n) \rightarrow \rho(f)$. It remains to show $H(f_n) \rightarrow H(f)$. To see this, note first that $H(f_n) \leq H(f_{n+1}) \leq H(f)$ for each n since $f_n \leq f_{n+1}$ pointwise. Let $\epsilon > 0$ and $c = H(f) - \epsilon$, and note that the definition of H and monotonicity of ℓ imply $\int_E \ell(f(x) - c) \mu(dx) > 1$. By monotone convergence, there exists n such that $\int_E \ell(f_n(x) - c) \mu(dx) > 1$. The definition of H now implies $H(f_n) > c = H(f) - \epsilon$. As $\epsilon > 0$ was arbitrary, we conclude $H(f_n) \rightarrow H(f)$. \square

We record here for later use a simple but useful lemma.

Lemma 4.1. *Define α as in (4.3). Let $\psi : E \rightarrow \mathbb{R}_+$ be continuous, and suppose $\int_E \psi^q \, d\mu < \infty$. Suppose $A \subset \mathcal{P}_\psi(E)$ is closed (in the ψ -weak topology), and $\mu \notin A$. Then $\inf_{v \in A} \alpha(v) > 0$.*

Proof. Recall that α as in (4.3) is the special case of (4.1) corresponding to $\ell(x) = [(1 + x)^+]^q$. Thus Proposition 4.1 and the assumption $\int_E \psi^q \, d\mu < \infty$ ensure that the sub-level sets

of α are ψ -weakly compact. If $\inf_{v \in A} \alpha(v) = 0$, we may find $v_n \in A$ such that $\alpha(v_n) \rightarrow 0$. The sequence (v_n) admits a ψ -weak limit point v^* , which must of course belong to the ψ -weakly closed set A . Lower semicontinuity and non-negativity of α imply $\alpha(v^*) = 0$. This implies $v^* = \mu$, as $t \mapsto t^p$ is strictly convex, and this contradicts the assumption that $\mu \notin A$. \square

4.2. Proofs of Theorem 1.2, Corollary 1.1, and Corollary 1.2

With these generalities in hand, we now turn toward the proof of Theorem 1.2. The idea is to apply Theorem 2.3 with α defined as in (4.3). The following estimate is crucial.

Lemma 4.2. *Let $q \in (1, \infty]$, and let $p = q/(q - 1)$ denote the conjugate exponent. Let α be as in (4.3). Then, for each $n \geq 1$ and $\nu \in \mathcal{P}(E^n)$ with $\nu \ll \mu^n$,*

$$\alpha_n(\nu) \leq n^{1/q} \|\mathrm{d}\nu/\mathrm{d}\mu^n\|_{L^p(\mu^n)}. \tag{4.5}$$

Proof. The case $p = \infty$ and $q = 1$ follows by sending $p \rightarrow \infty$ in (4.5), so we prove only the case $p < \infty$. As we will be working with conditional expectations, it is convenient to work with a more probabilistic notation. Fix n , and endow $\Omega = E^n$ with its Borel σ -field as well as the probability $P = \mu^n$. Let $X_i: E^n \rightarrow E$ denote the natural projections, and let $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$ denote the natural filtration, for $k = 1, \dots, n$, with $\mathcal{F}_0 := \{\emptyset, \Omega\}$. For $\nu \in \mathcal{P}(E^n)$ and $k = 1, \dots, n$, let ν_k denote a version of the regular conditional law of X_k given \mathcal{F}_{k-1} under ν , or symbolically $\nu_k := \nu(X_k \in \cdot \mid \mathcal{F}_{k-1})$. Let \mathbb{E}^ν denote integration with respect to ν . Since $P(X_k \in \cdot \mid \mathcal{F}_{k-1}) = \mu$ a.s., if $\nu \ll P$ then

$$\frac{\mathrm{d}\nu_k}{\mathrm{d}\mu} = \frac{\mathbb{E}^P[\mathrm{d}\nu/\mathrm{d}P \mid \mathcal{F}_k]}{\mathbb{E}^P[\mathrm{d}\nu/\mathrm{d}P \mid \mathcal{F}_{k-1}]} =: \frac{M_k}{M_{k-1}} \text{ a.s., where } \frac{0}{0} := 0.$$

Therefore

$$\alpha(\nu_k) = \mathbb{E}^P \left[\left(\frac{M_k}{M_{k-1}} \right)^p \mid \mathcal{F}_{k-1} \right]^{1/p} - 1.$$

Note that $(M_k)_{k=0}^n$ is a non-negative martingale, with $M_0 = 1$ and $M_n = \mathrm{d}\nu/\mathrm{d}P$. Then

$$\begin{aligned} \alpha_n(\nu) &= \mathbb{E}^\nu \left[\sum_{k=1}^n \alpha(\nu_k) \right] \\ &= \mathbb{E}^P \left[M_n \sum_{k=1}^n \left(\mathbb{E}^P \left[\left(\frac{M_k}{M_{k-1}} \right)^p \mid \mathcal{F}_{k-1} \right]^{1/p} - 1 \right) \right] \\ &= \mathbb{E}^P \left[\sum_{k=1}^n (\mathbb{E}^P [M_k^p \mid \mathcal{F}_{k-1}]^{1/p} - M_{k-1}) \right]. \end{aligned}$$

Subadditivity of $x \mapsto x^{1/p}$ implies

$$(\mathbb{E}^P [M_k^p \mid \mathcal{F}_{k-1}])^{1/p} \leq (\mathbb{E}^P [M_k^p - M_{k-1}^p \mid \mathcal{F}_{k-1}])^{1/p} + M_{k-1},$$

where the right-hand side is well-defined because

$$\mathbb{E}^P [M_k^p \mid \mathcal{F}_{k-1}] \geq \mathbb{E}^P [M_k \mid \mathcal{F}_{k-1}]^p = M_{k-1}^p.$$

Concavity of $x \mapsto x^{1/p}$ and Jensen’s inequality yield

$$\begin{aligned} \alpha_n(v) &\leq \mathbb{E}^P \left[\sum_{k=1}^n (\mathbb{E}^P [M_k^p - M_{k-1}^p \mid \mathcal{F}_{k-1}])^{1/p} \right] \\ &\leq n^{1-1/p} \left(\mathbb{E}^P \left[\sum_{k=1}^n \mathbb{E}^P [M_k^p - M_{k-1}^p \mid \mathcal{F}_{k-1}] \right] \right)^{1/p} \\ &= n^{1/q} (\mathbb{E}^P [M_n^p - M_0^p])^{1/p} \\ &\leq n^{1/q} (\mathbb{E}^P [M_n^p])^{1/p}. \end{aligned} \quad \square$$

Proof of Theorem 1.2. Again, let $q \in (1, \infty)$ and $p = q/(q - 1)$, and let α be as in (4.3), noting that it corresponds to (4.1) with $\ell(x) = [(1 + x)^+]^q$. Then Proposition 4.1 and the assumption that $\int \psi^q d\mu < \infty$ imply that the sub-level sets of α are pre-compact subsets of $\mathcal{P}_\psi(E)$. Hence, Theorem 2.3 applies to the ψ -weakly upper semicontinuous function $F: \mathcal{P}_\psi(E) \rightarrow [-\infty, 0]$ defined by $F(v) = 0$ if $v \in A$ and $F(v) = -\infty$ otherwise. This yields

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \rho_n(nF \circ L_n) \leq - \inf_{v \in A} \alpha(v). \tag{4.6}$$

Now use Lemma 4.2, noting that $(1/n)n^{1/q} = n^{-1/p}$, to get

$$\begin{aligned} \frac{1}{n} \rho_n(nF \circ L_n) &= \sup_{v \in \mathcal{P}(E^n)} \left(\int_{E^n} F \circ L_n dv - \frac{1}{n} \alpha_n(v) \right) \\ &= - \inf \left\{ \frac{1}{n} \alpha_n(v) : v \in \mathcal{P}(E^n), v(L_n \in A) = 1 \right\} \\ &\geq - \inf \{ n^{-1/p} \|dv/d\mu^n\|_{L^p(\mu^n)} : v \in \mathcal{P}(E^n), v \ll \mu^n, v(L_n \in A) = 1 \}. \end{aligned}$$

Set $B_n = \{x \in E^n : L_n(x) \in A\}$, and define $v \ll \mu^n$ by $dv/d\mu^n = 1_{B_n}/\mu^n(B_n)$. A quick computation yields

$$\|dv/d\mu^n\|_{L^p(\mu^n)} = \mu^n(B_n)^{(1-p)/p} = \mu^n(B_n)^{-1/q}.$$

Thus

$$\frac{1}{n} \rho_n(nF \circ L_n) \geq -(n^{1/p} \mu^n(B_n)^{1/q})^{-1}.$$

Combine this with (4.6) to get

$$\limsup_{n \rightarrow \infty} -(n^{1/p} \mu^n(L_n \in A)^{1/q})^{-1} \leq - \inf_{v \in A} \alpha(v).$$

Recalling the definition of α from (4.3) and noting that $q/p = q - 1$, this inequality can be rewritten as the desired result. □

Proof of Corollary 1.1. Define a continuous function $\psi: E \rightarrow \mathbb{R}_+$ by $\psi(x) = d^r(x, x_0)$. Note that \mathcal{W}_r then metrizes $\mathcal{P}_\psi(E)$ (see [52, Theorem 7.12]). Hence, the set

$$A = \{v \in \mathcal{P}_\psi(E) : \mathcal{W}_r(v, \mu) \geq a\}$$

is closed in $\mathcal{P}_\psi(E)$. Because $\int \psi^{q/r} d\mu = \int d^q(x, x_0) \mu(dx) < \infty$ by assumption, we may apply Theorem 1.2 with q/r in place of q to get

$$\limsup_{n \rightarrow \infty} n^{q/r-1} \mu^n(\mathcal{W}_r(L_n, \mu) \geq a) \leq \left(\inf_{v \in A} \alpha(v) \right)^{-q/r},$$

where α is defined as in (4.3) with $p = (q/r)/(q/r - 1)$. It remains to show that $\inf_{\nu \in A} \alpha(\nu) > 0$. But this follows from Lemma 4.1, since A is closed, $\mu \notin A$, and $\int \psi^{q/r} d\mu < \infty$. \square

Proof of Corollary 1.2. Again, let α be as in (4.3), and note that it corresponds to the shortfall risk measure (4.2) with $\ell(x) = [(1 + x)^+]^q$. Let $\psi(x) = \|x\|$, and consider the $\mathcal{P}_\psi(E)$ -closed set

$$B = \left\{ \mu \in \mathcal{P}_\psi(E) : \int_E z \mu(dz) \in A \right\},$$

where the integral is defined in the Bochner sense. Proposition 4.1 and the assumption that $\int \psi^q d\mu = \mathbb{E}[\|X_1\|^q] < \infty$ imply that the sub-level sets of α are pre-compact subsets of $\mathcal{P}_\psi(E)$. We may then apply Theorem 1.2 to get

$$\limsup_{n \rightarrow \infty} n^{q-1} \mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n X_i \in A \right) \leq \left(\inf_{\nu \in B} \alpha(\nu) \right)^{-q},$$

where again α is as in (4.3). It remains to simplify the right-hand side. Proposition 2.1 yields

$$\sup_{x^* \in E^*} (\langle x^*, x \rangle - \rho(x^*)) = \inf \left\{ \alpha(\nu) : \nu \in \mathcal{P}_\psi(E), \int_E z \nu(dz) = x \right\} \quad \text{for } x \in E.$$

Infimize over $x \in A$ on both sides to get

$$\inf_{\nu \in B} \alpha(\nu) = \inf_{x \in A} \sup_{x^* \in E^*} (\langle x^*, x \rangle - \rho(x^*)). \tag{4.7}$$

According to Proposition 4.2, for $x^* \in E^*$ we have

$$\rho(x^*) = \inf \left\{ m \in \mathbb{R} : \int_E [(1 + x^*(x) - m)^+]^q \mu(dx) \leq 1 \right\} = \Lambda(x^*),$$

where the latter equality is simply the definition of Λ given in the statement of Corollary 1.2. Hence, the identity (4.7) becomes $\inf_{\nu \in B} \alpha(\nu) = \inf_{x \in A} \Lambda^*(x)$, and the proof is complete. \square

4.3. Stochastic optimization with heavy tails

This section elaborates on the application discussed in Section 1.1.3, concerning the convergence of Monte Carlo estimates for stochastic optimization problems. We use the notation of Section 1.1.3, and we begin with the proof of Theorem 1.3.

Proof of Theorem 1.3. Let $A = \{ \nu \in \mathcal{P}_\psi(E) : |V(\nu) - V(\mu)| \geq \epsilon \}$. The map

$$\mathcal{X} \times \mathcal{P}_\psi(E) \ni (x, \nu) \mapsto \int_E h(x, w) \nu(dw)$$

is jointly continuous. By Berge’s theorem [3, Theorem 17.31], V is continuous on $\mathcal{P}_\psi(E)$, so A is closed. Theorem 1.2 implies

$$\limsup_{n \rightarrow \infty} n^{q-1} \mu^n(|V(L_n) - V(\mu)| \geq \epsilon) = \limsup_{n \rightarrow \infty} n^{q-1} \mu^n(L_n \in A) \leq \left(\inf_{\nu \in A} \alpha(\nu) \right)^{-q}.$$

Note that $q/p = q - 1$, and finally use Lemma 4.1 to conclude $\inf_{\nu \in A} \alpha(\nu) > 0$. \square

Remark 4.1. The joint continuity and compactness assumptions in Theorem 1.3 could likely be weakened, but we focus on the more novel integrability issues to ease the exposition.

Now we have shown that the optimal value itself converges, we turn to the convergence of optimizers themselves.

Theorem 4.1. *Grant the assumptions of Theorem 1.3. Let $\hat{x}: \mathcal{P}_\psi(E) \rightarrow \mathcal{X}$ be any measurable function satisfying*

$$\hat{x}(\nu) \in \arg \min_{x \in \mathcal{X}} \int_E h(x, w) \nu(dw) \quad \text{for each } \nu.$$

Suppose there exist a measurable function $\varphi: \mathbb{R} \rightarrow \mathbb{R}$ and a compatible metric d on \mathcal{X} such that

$$\varphi(d(\hat{x}(\mu), x)) \leq \int_E h(x, w) \mu(dw) - \int_E h(\hat{x}(\mu), w) \mu(dw).$$

Then, for any $\epsilon > 0$,

$$\limsup_{n \rightarrow \infty} n^{q-1} \mu^n(\varphi(d(\hat{x}(\mu), \hat{x}(L_n))) \geq \epsilon) < \infty.$$

In particular, if φ is strictly increasing with $\varphi(0) = 0$, then for any $\epsilon > 0$,

$$\limsup_{n \rightarrow \infty} n^{q-1} \mu^n(d(\hat{x}(\mu), \hat{x}(L_n)) \geq \epsilon) < \infty.$$

Such a function \hat{x} exists because $(x, \nu) \mapsto \int_E h(x, w) \nu(dw)$ is measurable in ν and continuous in x ; see e.g. [3, Theorem 18.19].

Proof. Note that for $\epsilon > 0$, on the event $\{\varphi(d(\hat{x}(\mu), \hat{x}(L_n))) \geq \epsilon\}$ we have

$$\begin{aligned} \epsilon &\leq \varphi(d(\hat{x}(\mu), \hat{x}(L_n))) \\ &\leq \int_E h(\hat{x}(L_n), w) \mu(dw) - \int_E h(\hat{x}(\mu), w) \mu(dw) \\ &\leq |V(L_n) - V(\mu)| + \sup_{x \in \mathcal{X}} \int_E h(x, w) [\mu - L_n](dw). \end{aligned}$$

The first term converges at the right rate, thanks to Theorem 1.3, and it remains to check that

$$\limsup_{n \rightarrow \infty} n^{q-1} \mu^n \left(\sup_{x \in \mathcal{X}} \int_E h(x, w) [\mu - L_n](dw) \geq \epsilon \right) < \infty.$$

The map $(x, \nu) \mapsto \int_E h(x, w) \nu(dw)$ is continuous on $\mathcal{X} \times \mathcal{P}_\psi(E)$, so the map

$$\mathcal{P}_\psi(E) \ni \nu \mapsto \sup_{x \in \mathcal{X}} \int_E h(x, w) [\mu - \nu](dw)$$

is continuous by Berge’s theorem [3, Theorem 17.31]. Hence, the set

$$B := \left\{ \nu \in \mathcal{P}_\psi(E) : \sup_{x \in \mathcal{X}} \int_E h(x, w) [\mu - \nu](dw) \geq \epsilon \right\}$$

is closed in $\mathcal{P}_\psi(E)$. Theorem 1.2 then implies

$$\limsup_{n \rightarrow \infty} n^{q-1} \mu^n \left(\sup_{x \in \mathcal{X}} \int_E h(x, w) [\mu - L_n](dw) \geq \epsilon \right) \leq \left(\inf_{\nu \in B} \alpha(\nu) \right)^{-q},$$

where α is defined as in (4.3). Finally, Lemma 4.1 implies that $\inf_{\nu \in B} \alpha(\nu) > 0$. □

Under the assumption $\int_E \psi^q d\mu < \infty$, we see that the value $V(L_n)$ always converges to $V(\mu)$ with the polynomial rate n^{1-q} . To see when Theorem 4.1 applies, notice that in many situations, \mathcal{X} is a convex subset of a normed vector space, and we have uniform convexity in the following form. There exists a strictly increasing function φ such that $\varphi(0) = 0$ and, for all $t \in (0, 1)$ and $x, y \in \mathcal{X}$,

$$\begin{aligned} & \int_E h(tx + (1-t)y, w)\mu(dw) \\ & \leq t \int_E h(x, w)\mu(dw) + (1-t) \int_E h(y, w)\mu(dw) - t(1-t)\varphi(\|x - y\|). \end{aligned}$$

See [37, pages 202–203] for more on this.

5. Uniform large deviations and martingales

This section returns to the example of Section 1.2. We first record a useful abstract theorem of Föllmer and Schied [28], which will allow us to verify tightness of the sub-level sets of α before knowing it is convex by checking a property of ρ .

Proposition 5.1. (Proposition 4.30 of [28].) *Suppose a functional $\rho: B(E) \rightarrow \mathbb{R}$ admits the representation*

$$\rho(f) = \sup_{\nu \in \mathcal{P}(E)} \left(\int_E f d\nu - \alpha(\nu) \right) \quad \text{for } f \in C_b(E),$$

for some functional $\alpha: \mathcal{P}(E) \rightarrow (-\infty, \infty]$. Suppose also that there is a sequence (K_n) of compact subsets of E such that

$$\lim_{n \rightarrow \infty} \rho(\lambda 1_{K_n}) = \rho(\lambda) \quad \text{for all } \lambda \geq 1.$$

Then α has tight sub-level sets.

Fix a convex weakly compact family of probability measures $M \subset \mathcal{P}(E)$. Define

$$\alpha(\nu) = \inf_{\mu \in M} H(\nu | \mu), \tag{5.1}$$

where the relative entropy was defined in (1.2). In light of the classical formula [22, Proposition 1.4.2]

$$\sup_{\nu \in B(E)} \left(\int_E f d\nu - H(\nu | \mu) \right) = \log \int_E e^f d\mu,$$

the ρ corresponding to the functional α given by (5.1) is then

$$\begin{aligned} \rho(f) & := \sup_{\nu \in B(E)} \left(\int_E f d\nu - \alpha(\nu) \right) \\ & = \sup_{\nu \in B(E)} \sup_{\mu \in M} \left(\int_E f d\nu - H(\nu | \mu) \right) \\ & = \sup_{\mu \in M} \log \int_E e^f d\mu. \end{aligned} \tag{5.2}$$

Let us also take note of the famous Donsker–Varadhan formula [22, Lemma 1.4.3]

$$H(\nu \mid \mu) = \sup_{f \in C_b(E)} \left(\int_E f \, d\nu - \log \int_E e^f \, d\mu \right). \tag{5.3}$$

Lemma 5.1. *The functional α defined in (5.1) satisfies the standing assumptions. That is, it is convex and bounded from below, and its sub-level sets are weakly compact.*

Proof. Note first that $-\log \int_E e^f \, d\mu$ is convex and weakly continuous in μ as well as concave and sup-norm continuous in f . Thus, using (5.3) and Sion’s minimax theorem [51], we find

$$\begin{aligned} \alpha(\nu) &= \inf_{\mu \in M} \sup_{f \in C_b(E)} \left(\int_E f \, d\nu - \log \int_E e^f \, d\mu \right) \\ &= \sup_{f \in C_b(E)} \inf_{\mu \in M} \left(\int_E f \, d\nu - \log \int_E e^f \, d\mu \right) \\ &= \sup_{f \in C_b(E)} \left(\int_E f \, d\nu - \rho(f) \right). \end{aligned}$$

This shows that α is convex and lower semicontinuous. It remains to prove that α has tight sub-level sets, which will follow from Proposition 5.1 once we check the second assumption therein. By Prokhorov’s theorem, there exist compact sets $K_1 \subset K_2 \subset \dots$ such that $\sup_{\mu \in M} \mu(K_n^c) \leq 1/n$. Then, for $\lambda \geq 0$, using the formula for ρ of (5.2),

$$\begin{aligned} \lambda \geq \rho(\lambda 1_{K_n}) &= \sup_{\mu \in M} \log \int_E \exp(\lambda 1_{K_n}) \, d\mu \\ &= \sup_{\mu \in M} \log [(e^\lambda - 1)\mu(K_n) + 1] \\ &\geq \log [(e^\lambda - 1)(1 - 1/n) + 1]. \end{aligned}$$

As $n \rightarrow \infty$, the right-hand side converges to λ , which shows $\rho(\lambda 1_{K_n}) \rightarrow \lambda = \rho(\lambda)$. □

To compute ρ_n , recall that for $M \subset \mathcal{P}(E)$ we define M_n as the set of $\mu \in \mathcal{P}(E^n)$ satisfying $\mu_{0,1} \in M$ and $\mu_{k-1,k}(x_1, \dots, x_{k-1}) \in M$ for all $k = 2, \dots, n$ and $x_1, \dots, x_{n-1} \in E$. (Recall that the conditional measures $\mu_{k-1,k}$ were defined in the Introduction.) Notice that $M_1 = M$.

Proposition 5.2. *For each $n \geq 1$, $\alpha_n(\nu) = \inf_{\mu \in M_n} H(\nu \mid \mu)$. Moreover, for each measurable $f: E^n \rightarrow \mathbb{R} \cup \{-\infty\}$ satisfying $\int_{E^n} e^f \, d\mu < \infty$ for every $\mu \in M_n$,*

$$\rho_n(f) = \sup_{\mu \in M_n} \log \int_{E^n} e^f \, d\mu.$$

Proof. Given the first claim, the second follows from the well-known duality

$$\sup_{\nu \in \mathcal{P}(E^n)} \left(\int_{E^n} f \, d\nu - H(\nu \mid \mu) \right) = \log \int_{E^n} e^f \, d\mu,$$

which holds for $\mu \in \mathcal{P}(E^n)$ as long as e^f is μ -integrable (see e.g. the proof of [22, 1.4.2]). Indeed, this implies

$$\begin{aligned} \rho_n(f) &= \sup_{\nu \in \mathcal{P}(E^n)} \left(\int_{E^n} f \, d\nu - \alpha_n(\nu) \right) \\ &= \sup_{\mu \in M_n} \sup_{\nu \in \mathcal{P}(E^n)} \left(\int_{E^n} f \, d\nu - H(\nu \mid \mu) \right) \\ &= \sup_{\mu \in M_n} \log \int_{E^n} e^f \, d\mu. \end{aligned}$$

To prove the first claim, note that by definition

$$\alpha_n(\nu) = \sum_{k=1}^n \int_{E^n} \inf_{\mu \in M} H(\nu_{k-1,k}(x_1, \dots, x_{k-1}) \mid \mu) \nu(dx_1, \dots, dx_n).$$

For $k = 2, \dots, n$, let \mathcal{Y}_k denote the set of measurable maps from E^{k-1} to M , and let $\mathcal{Y}_1 = M$. Then the usual measurable selection argument [9, Proposition 7.50] yields

$$\alpha_n(\nu) = \sum_{k=1}^n \inf_{\eta_k \in \mathcal{Y}_k} \int_{E^n} H(\nu_{k-1,k}(x_1, \dots, x_{k-1}) \mid \eta_k(x_1, \dots, x_{k-1})) \nu(dx_1, \dots, dx_n).$$

Now, if $(\eta_1, \dots, \eta_n) \in \prod_{k=1}^n \mathcal{Y}_k$, then the measure

$$\mu(dx_1, \dots, dx_n) = \eta_1(dx_1) \prod_{k=2}^n \eta_k(x_1, \dots, x_{k-1})(dx_k)$$

is in M , and $\mu_{k-1,k} = \eta_k$ is a version of the conditional law. Thus

$$\alpha_n(\nu) \geq \inf_{\mu \in M} \sum_{k=1}^n \int_{E^n} H(\nu_{k-1,k}(x_1, \dots, x_{k-1}) \mid \mu_{k-1,k}(x_1, \dots, x_{k-1})) \nu(dx_1, \dots, dx_n).$$

On the other hand, for every $\mu \in M_n$, the vector $(\mu_{0,1}, \mu_{1,2}, \dots, \mu_{n-1,n})$ belongs to $\prod_{k=1}^n \mathcal{Y}_k$, and we deduce the opposite inequality. Hence

$$\begin{aligned} \alpha_n(\nu) &\geq \inf_{\mu \in M} \sum_{k=1}^n \int_{E^n} H(\nu_{k-1,k}(x_1, \dots, x_{k-1}) \mid \mu_{k-1,k}(x_1, \dots, x_{k-1})) \nu(dx_1, \dots, dx_n) \\ &= \inf_{\mu \in M} H(\nu \mid \mu), \end{aligned}$$

where the last equality follows from the chain rule for relative entropy [22, Theorem B.2.1]. □

Theorem 2.3 now leads to the following uniform large deviation bound.

Corollary 5.1. *For $F \in C_b(\mathcal{P}(E))$, we have*

$$\lim_{n \rightarrow \infty} \sup_{\mu \in M_n} \frac{1}{n} \log \int_{E^n} e^{nF \circ L_n} \, d\mu = \sup_{\nu \in \mathcal{P}(E), \mu \in M} (F(\nu) - H(\nu \mid \mu)).$$

For closed sets $A \subset \mathcal{P}(E)$, we have

$$\lim_{n \rightarrow \infty} \sup_{\mu \in M_n} \frac{1}{n} \log \mu(L_n \in A) \leq - \inf\{H(v \mid \mu) : v \in A, \mu \in M\}.$$

Proof. The first claim is an immediate consequence of Theorem 2.3 and the calculation of ρ_n in Proposition 5.2. To prove the second claim, define F on $\mathcal{P}(E)$ by

$$F(v) = \begin{cases} 0 & \text{if } v \in A, \\ -\infty & \text{otherwise.} \end{cases}$$

Then F is upper semicontinuous and bounded from above. Use Proposition 5.2 to compute

$$\rho_n(nF \circ L_n) = \sup_{\mu \in M_n} \log \int_{E^n} \exp(nF \circ L_n) d\mu = \sup_{\mu \in M_n} \log \mu(L_n \in A).$$

The proof is completed by applying Theorem 2.3 with this function F . □

The following proposition simplifies the strong Cramér condition (3.1) in the present context.

Proposition 5.3. *Let $\psi : E \rightarrow \mathbb{R}_+$ be measurable. Suppose that for every $\lambda > 0$ we have*

$$\sup_{\mu \in M} \int_E e^{\lambda \psi} d\mu < \infty. \tag{5.4}$$

Then the strong Cramér condition holds, i.e. $\lim_{m \rightarrow \infty} \rho(\lambda \psi 1_{\{\psi \geq m\}}) \rightarrow 0$ for all $\lambda > 0$. In particular, the sub-level sets of α are pre-compact subsets of $\mathcal{P}_\psi(E)$.

Proof. Because $e^{\lambda \psi}$ is μ -integrable for each $\mu \in M$ and $\lambda > 0$, Proposition 5.2 implies

$$\begin{aligned} \rho(\lambda \psi 1_{\{\psi \geq m\}}) &= \sup_{\mu \in M} \log \int_E \exp(\lambda \psi 1_{\{\psi \geq m\}}) d\mu \\ &\leq \sup_{\mu \in M} \log \left(1 + \int_{\{\psi \geq m\}} e^{\lambda \psi} d\mu \right). \end{aligned}$$

Now note that

$$1_{\{\psi \geq m\}} \leq \frac{\psi}{m} \leq \frac{1}{m} e^\psi$$

pointwise, and thus the assumption (5.4) yields

$$\lim_{m \rightarrow \infty} \sup_{\mu \in M} \int_{\{\psi \geq m\}} e^{\lambda \psi} d\mu \leq \lim_{m \rightarrow \infty} \sup_{\mu \in M} \frac{1}{m} \int_E e^{(1+\lambda)\psi} d\mu = 0. \tag{□}$$

We are finally ready to specialize Corollary 5.1 to prove Theorem 1.4, similarly to how we specialized Theorem 1.2 to prove Corollary 1.2 in Section 4.

Proof of Theorem 1.4. Define

$$M = \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) : \log \int_{\mathbb{R}^d} e^{(y,x)} \mu(dx) \leq \varphi(y) \text{ for all } y \in \mathbb{R}^d \right\}.$$

We claim that M is weakly compact. Indeed, it is clearly convex, and closedness follows from Fatou's lemma (cf. [22, Theorem A.3.12]). To prove tightness, let e_1, \dots, e_d denote the standard basis vectors in \mathbb{R}^d . Write $x = (x_1, \dots, x_d)$ for a generic element of \mathbb{R}^d . For each $\mu \in M$ and $t > 0$, Markov's inequality yields

$$\begin{aligned} \mu\{x \in \mathbb{R}^d : \max_{i=1, \dots, d} |x_i| > t\} &\leq \sum_{k=1}^d (\mu\{x \in \mathbb{R}^d : x_i > t/2\} + \mu\{x \in \mathbb{R}^d : -x_i > t/2\}) \\ &\leq \sum_{k=1}^d e^{-t/2} \int_{\mathbb{R}^d} (e^{x_i} + e^{-x_i}) \mu(dx) \\ &\leq e^{-t/2} \sum_{k=1}^d e^{\varphi(e_i) + \varphi(-e_i)}, \end{aligned}$$

and we deduce that M is tight. Now define $\psi(x) = \sum_{i=1}^d |x_i|$ and note that

$$\sup_{\mu \in M} \int_{\mathbb{R}^d} \exp(\lambda \psi) d\mu < \infty \quad \text{for all } \lambda \geq 0.$$

Proposition 5.3 then shows that the strong Cramér condition holds. Define a closed set $B \subset \mathcal{P}_\psi(\mathbb{R}^d)$ by

$$B = \left\{ \nu \in \mathcal{P}_\psi(\mathbb{R}^d) : \int_{\mathbb{R}^d} z \nu(dz) \in A \right\},$$

where A was the given closed subset of \mathbb{R}^d . Corollary 5.1 yields

$$\limsup_{n \rightarrow \infty} \sup_{\mu \in M_n} \frac{1}{n} \log \mu(L_n \in B) \leq - \inf \left\{ \alpha(\nu) : \nu \in \mathcal{P}_\psi(\mathbb{R}^d), \int x \nu(dx) \in A \right\},$$

Now let $(S_0, \dots, S_n) \in \mathcal{S}_{d,\varphi}$. The law of S_1 belongs to M , and the conditional law of $S_k - S_{k-1}$ given S_1, \dots, S_{k-1} belongs almost surely to M , for each k , and so the law of $(S_1, S_2 - S_1, \dots, S_n - S_{n-1})$ belongs to M_n . Thus

$$\mathbb{P}(S_n/n \in A) \leq \sup_{\mu \in M_n} \mu(L_n \in B),$$

and all that remains is to prove that

$$\inf \left\{ \alpha(\nu) : \nu \in \mathcal{P}_\psi(\mathbb{R}^d), \int z \nu(dz) \in A \right\} \geq \inf_{x \in A} \varphi^*(x).$$

To prove this, it suffices to show $\Psi(x) \geq \varphi^*(x)$ for every $x \in \mathbb{R}^d$, where

$$\Psi(x) := \inf \left\{ \alpha(\nu) : \nu \in \mathcal{P}_\psi(\mathbb{R}^d), \int z \nu(dz) = x \right\}.$$

To this end, note that for all $y \in \mathbb{R}^d$

$$\rho(\cdot, y) = \sup_{\mu \in M} \log \int_E e^{(z,y)} \mu(dz) \leq \varphi(y),$$

and then use the representation of 2.1 to get

$$\Psi(x) = \sup_{y \in \mathbb{R}^d} ((x, y) - \rho(\cdot, y)) \geq \sup_{y \in \mathbb{R}^d} ((x, y) - \varphi(y)) = \varphi^*(x). \quad \square$$

6. Optimal transport and control

This section discusses the example of Section 1.4 in more detail. Again let E be a Polish space, and fix a lower semicontinuous function $c: E^2 \rightarrow [0, \infty]$ which is not identically equal to ∞ . Fix $\mu \in \mathcal{P}(E)$, and define

$$\alpha(\nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int c \, d\pi,$$

where $\Pi(\mu, \nu)$ is the set of probability measures on $E \times E$ with first marginal μ and second marginal ν . Assume that $\int_E c(x, x)\mu(dx) < \infty$; in many practical cases, $c(x, x) = 0$ for all x , so this is not a restrictive assumption and merely ensures that $\alpha(\mu) < \infty$. Kantorovich duality [52, Theorem 1.3] shows that

$$\alpha(\nu) = \sup \left(\int_E f \, d\nu - \int_E g \, d\mu : f, g \in C_b(E), f(y) - g(x) \leq c(x, y) \text{ for all } x, y \right).$$

This immediately shows that α is convex and weakly lower semicontinuous. The next two lemmas identify, respectively, the dual ρ and the modest conditions that ensure that α has compact sub-level sets.

Lemma 6.1. *Given α as above, and defining ρ as usual by (1.1), we have*

$$\rho(f) = \int_E R_c f \, d\mu \quad \text{for all } f \in B(E), \tag{6.1}$$

where $R_c f: E \rightarrow \mathbb{R}$ is defined by

$$R_c f(x) = \sup_{y \in E} (f(y) - c(x, y)).$$

Proof. Note that $R_c f$ is universally measurable (e.g. by [9, Proposition 7.50]), so the integral in (6.1) makes sense. Now compute

$$\begin{aligned} \rho(f) &= \sup_{\nu \in \mathcal{P}(E)} \left(\int_E f \, d\nu - \alpha(\nu) \right) \\ &= \sup_{\nu \in \mathcal{P}(E)} \sup_{\pi \in \Pi(\mu, \nu)} \left(\int_E f \, d\nu - \int_{E^2} c \, d\pi \right) \\ &= \sup_{\pi \in \Pi(\mu)} \int_{E^2} (f(y) - c(x, y))\pi(dx, dy), \end{aligned}$$

where $\Pi(\mu)$ is the set of $\pi \in \mathcal{P}(E \times E)$ with first marginal μ . Use the standard measurable selection theorem [9, Proposition 7.50] to find a measurable map $Y: E \rightarrow E$ such that $R_c f(x) = f(Y(x)) - c(x, Y(x))$ for μ -a.e. x . Then, choosing $\pi(dx, dy) = \mu(dx)\delta_{Y(x)}(dy)$ shows

$$\rho(f) \geq \int_E (f(Y(x)) - c(x, Y(x)))\mu(dx) = \int_E R_c f \, d\mu.$$

On the other hand, it is clear that for every $\pi \in \Pi(\mu)$ we have

$$\int_{E^2} (f(y) - c(x, y))\pi(dx, dy) \leq \int_E \sup_{y \in E} (f(y) - c(x, y))\mu(dx) = \int_E R_c f \, d\mu. \quad \square$$

Lemma 6.2. *Suppose that for each compact set $K \subset E$, the function $h_K(y) := \inf_{x \in K} c(x, y)$ has pre-compact sub-level sets. Then α has compact sub-level sets. In fact, since c is lower semi-continuous, so is h_K (see [3, Lemma 17.30]). Thus, our assumption is equivalent to requiring $\{y \in E: h_K(y) \leq m\}$ to be compact for each $m \geq 0$.*

Proof. We already know that α has closed sub-level sets, so we must show only that they are tight. Fix $\nu \in \mathcal{P}(E)$ such that $\alpha(\nu) < \infty$ (noting that such ν certainly exist, as μ is one example). Fix $\epsilon > 0$, and find $\pi \in \Pi(\mu, \nu)$ such that

$$\int c \, d\pi \leq \alpha(\nu) + \epsilon < \infty. \tag{6.2}$$

As finite measures on Polish spaces are tight, we may find a compact set $K \subset E$ such that $\mu(K^c) \leq \epsilon$. Set $K_n := \{y \in E: h_K(y) < n\}$ for each n , and note that this set is pre-compact by assumption. Disintegrate π by finding a measurable map $E \ni x \mapsto \pi_x \in \mathcal{P}(E)$ such that $\pi(dx, dy) = \mu(dx)\pi_x(dy)$. By Markov's inequality, for each $n > 0$ and each $x \in K$ we have

$$\pi_x(K_n^c) \leq \pi_x\{y \in E: c(x, y) \geq n\} \leq \frac{1}{n} \int_E c(x, y)\pi_x(dy).$$

Using this and inequality (6.2) along with the assumption that c is non-negative,

$$\begin{aligned} \nu(K_n^c) &= \int_E \mu(dx)\pi_x(K_n^c) \\ &\leq \mu(K^c) + \int_K \mu(dx)\pi_x(K_n^c) \\ &\leq \epsilon + \frac{1}{n} \int_K \mu(dx) \int_E \pi_x(dy)c(x, y) \\ &\leq \epsilon + \frac{1}{n} \int_{E \times E} c \, d\pi \\ &\leq \left(1 + \frac{1}{n}\right)\epsilon + \frac{1}{n}\alpha(\nu). \end{aligned}$$

As ϵ was arbitrary, we have $\nu(K_n^c) \leq \alpha(\nu)/n$. Thus, for each $m > 0$, the sub-level set $\{\nu \in \mathcal{P}(E): \alpha(\nu) \leq m\}$ is contained in the tight set

$$\bigcap_{n=1}^{\infty} \{\nu \in \mathcal{P}(E): \nu(K_n^c) \leq m/n\}. \quad \square$$

Let us now compute ρ_n . It is convenient to work with more probabilistic notation, so let us suppose $(X_i)_{i=1}^{\infty}$ is a sequence of i.i.d. E -valued random variables with common law μ , defined on some fixed probability space. For each n , let \mathcal{Y}_n denote the set of equivalence classes of a.s. equal E^n -valued random variables (Y_1, \dots, Y_n) where Y_k is (X_1, \dots, X_k) -measurable for each $k = 1, \dots, n$.

Proposition 6.1. *For each $n \geq 1$ and each $f \in B(E)$,*

$$\rho_n(f) = \sup_{(Y_1, \dots, Y_n) \in \mathcal{Y}_n} \mathbb{E} \left[f(Y_1, \dots, Y_n) - \sum_{i=1}^n c(X_i, Y_i) \right].$$

Proof. The proof is by induction. Let us first rewrite ρ in our probabilistic notation:

$$\rho(f) = \mathbb{E} \left[\sup_{y \in E} [f(y) - c(X_1, y)] \right].$$

Using a standard measurable selection argument [9, Proposition 7.50], we deduce that

$$\rho(f) = \sup_{Y_1 \in \mathcal{Y}_1} \mathbb{E}[f(Y_1) - c(X_1, Y_1)].$$

The inductive step proceeds as follows. Suppose we have proved the claim for a given n . Fix $f \in B(E^{n+1})$ and define $g \in B(E^n)$ by

$$g(x_1, \dots, x_n) := \rho(f(x_1, \dots, x_n, \cdot)),$$

so that by Proposition A.1 we have $\rho_{n+1}(f) = \rho_n(g)$. Since X_1 and X_{n+1} have the same distribution, we may relabel to find

$$\begin{aligned} g(x_1, \dots, x_n) &= \sup_{Y_1 \in \mathcal{Y}_1} \mathbb{E}[f(x_1, \dots, x_n, Y_1) - c(X_1, Y_1)] \\ &= \sup_{Y_{n+1} \in \mathcal{Y}_{n+1}^1} \mathbb{E}[f(x_1, \dots, x_n, Y_{n+1}) - c(X_{n+1}, Y_{n+1})], \end{aligned}$$

where we define \mathcal{Y}_{n+1}^1 to be the set of X_{n+1} -measurable E -valued random variables. Now note that any (Y_1, \dots, Y_n) in \mathcal{Y}_n is (X_1, \dots, X_n) -measurable, and independence of $(X_i)_{i=1}^\infty$ implies

$$g(Y_1, \dots, Y_n) = \sup_{Y_{n+1} \in \mathcal{Y}_{n+1}^1} \mathbb{E}[f(Y_1, \dots, Y_n, Y_{n+1}) - c(X_{n+1}, Y_{n+1}) \mid Y_1, \dots, Y_n].$$

We claim that

$$\mathbb{E}[g(Y_1, \dots, Y_n)] = \sup_{Y_{n+1}} \mathbb{E}[f(Y_1, \dots, Y_n, Y_{n+1}) - c(X_{n+1}, Y_{n+1})], \tag{6.3}$$

where the supremum is over (X_1, \dots, X_{n+1}) -measurable E -valued random variables Y_{n+1} . Indeed, once this is established, we conclude as desired (using Proposition A.1) that

$$\begin{aligned} \rho_{n+1}(f) &= \rho_n(g) \\ &= \sup_{(Y_1, \dots, Y_n) \in \mathcal{Y}_n} \mathbb{E} \left[g(Y_1, \dots, Y_n) - \sum_{i=1}^n c(X_i, Y_i) \right] \\ &= \sup_{(Y_1, \dots, Y_n) \in \mathcal{Y}_n} \sup_{Y_{n+1}} \mathbb{E} \left[f(Y_1, \dots, Y_n, Y_{n+1}) - \sum_{i=1}^{n+1} c(X_i, Y_i) \right]. \end{aligned}$$

Hence, the rest of the proof is devoted to justifying (6.3), which is really an interchange of supremum and expectation.

Note that \mathcal{Y}_{n+1}^1 is a Polish space when topologized by convergence in measure. The function $h: E^n \times \mathcal{Y}_{n+1}^1 \rightarrow \mathbb{R}$ given by

$$h(x_1, \dots, x_n; Y_{n+1}) := \mathbb{E}[f(x_1, \dots, x_n, Y_{n+1}) - c(X_{n+1}, Y_{n+1})]$$

is jointly measurable. Note as before that independence implies that for every $(Y_1, \dots, Y_n) \in \mathcal{Y}_n$ and $Y_{n+1} \in \mathcal{Y}_{n+1}^1$ we have, for a.e. ω ,

$$h(Y_1(\omega), \dots, Y_n(\omega); Y_{n+1}) = \mathbb{E}[f(Y_1, \dots, Y_n, Y_{n+1}) - c(X_{n+1}, Y_{n+1}) \mid Y_1, \dots, Y_n](\omega). \tag{6.4}$$

Using the usual measurable selection theorem [9, Proposition 7.50] we get

$$\begin{aligned} \mathbb{E}[g(Y_1, \dots, Y_n)] &= \mathbb{E} \left[\sup_{Y_{n+1} \in \mathcal{Y}_{n+1}^1} h(Y_1(\cdot), \dots, Y_n(\cdot); Y_{n+1}) \right] \\ &= \sup_{H \in \tilde{\mathcal{Y}}_{n+1}^1} \mathbb{E}[h(Y_1(\cdot), \dots, Y_n(\cdot); H(Y_1, \dots, Y_n))], \end{aligned}$$

where $\tilde{\mathcal{Y}}_{n+1}^1$ denotes the set of measurable maps $H: E^n \rightarrow \mathcal{Y}_{n+1}^1$. But a measurable map $H: E^n \rightarrow \mathcal{Y}_{n+1}^1$ can be identified almost everywhere with an (X_1, \dots, X_{n+1}) -measurable random variable Y_{n+1} . Precisely, by Lemma B.1 (in the appendix) there exists a jointly measurable map $\varphi: E^{n+1} \rightarrow E$ such that, for μ^n -a.e. $(x_1, \dots, x_n) \in E^n$, we have

$$\varphi(x_1, \dots, x_{n+1}) = H(x_1, \dots, x_n)(x_{n+1}) \quad \text{for } \mu\text{-a.e. } x_{n+1} \in E.$$

Define $Y_{n+1} = \varphi(X_1, \dots, X_{n+1})$, and note that (6.4) implies, for a.e. ω ,

$$\begin{aligned} &h(Y_1(\omega), \dots, Y_n(\omega); H(Y_1, \dots, Y_n)) \\ &= \mathbb{E}[f(Y_1, \dots, Y_n, Y_{n+1}) - c(X_{n+1}, Y_{n+1}) \mid Y_1, \dots, Y_n](\omega). \end{aligned}$$

This identification of $\tilde{\mathcal{Y}}_{n+1}^1$ and the tower property of conditional expectations leads to (6.3). □

Appendix A. A recursive formula for ρ_n

In this brief section we make rigorous the claim in (1.4). To do so requires a brief review of analytic sets, needed only for this section. A subset of a Polish space is *analytic* if it is the image of a Borel subset of another Polish space through a Borel-measurable function. A real-valued function f on a Polish space is *upper semianalytic* if $\{f \geq c\}$ is an analytic set for each $c \in \mathbb{R}$. It is well known that every analytic set is universally measurable [9, Corollary 7.42.1], and thus every upper semianalytic function is universally measurable. The defining formula for ρ_n given in (1.1) makes sense even when $f: E^n \rightarrow \mathbb{R}$ is bounded and universally measurable, or in particular when f is upper semianalytic.

Proposition A.1 *Let $n > 1$. Suppose $f: E^n \rightarrow \mathbb{R}$ is upper semianalytic. Define $g: E^{n-1} \rightarrow \mathbb{R}$ by*

$$g(x_1, \dots, x_{n-1}) = \rho(f(x_1, \dots, x_{n-1}, \cdot)).$$

Then g is upper semianalytic, and $\rho_n(f) = \rho_{n-1}(g)$.

Proof. To show that g is upper semianalytic, note that

$$\begin{aligned} g(x_1, \dots, x_{n-1}) &= \rho(f(x_1, \dots, x_{n-1}, \cdot)) \\ &= \sup_{\nu \in \mathcal{P}(E)} \left(\int_E f(x_1, \dots, x_{n-1}, \cdot) \, d\nu - \alpha(\nu) \right). \end{aligned}$$

Clearly α is Borel-measurable, as its sub-level sets are compact. It follows from [9, Proposition 7.48] that the term in parentheses is upper semianalytic as a function of $(x_1, \dots, x_{n-1}, \nu)$. Hence, g is itself upper semianalytic, by [9, Proposition 7.47].

We now turn toward the proof of the recursive formula for ρ_n . Note first that the definition of α_n can be written recursively by setting $\alpha_1 = \alpha$ and, for $\nu \in \mathcal{P}(E^n)$ and a kernel K from E^n to E (i.e. a Borel-measurable map $x \mapsto K_x$ from E^n to $\mathcal{P}(E)$), setting

$$\alpha_{n+1}(\nu(dx)K_x(dx_{n+1})) = \int_{E^n} \alpha(K_x)\nu(dx) + \alpha_n(\nu). \tag{A.1}$$

Fix $f \in B(E^{n+1})$, and note that $g(x_1, \dots, x_n) := \rho(f(x_1, \dots, x_n, \cdot))$ is upper semianalytic by the above argument. By definition,

$$\rho_n(g) = \sup_{\nu \in \mathcal{P}(E^n)} \left\{ \int_{E^n} g \, d\nu - \alpha_n(\nu) \right\}. \tag{A.2}$$

By a well-known measurable selection argument [9, Proposition 7.50], for each $\nu \in \mathcal{P}(E^n)$ it holds that

$$\begin{aligned} \int_{E^n} g \, d\nu &= \int_{E^n} \sup_{\eta \in \mathcal{P}(E)} \left(\int_E f(x_1, \dots, x_n, x_{n+1})\eta(dx_{n+1}) - \alpha(\eta) \right) \nu(dx) \\ &= \sup_K \left(\int_{E^n} \int_E f(x_1, \dots, x_{n+1})K_x(dx_{n+1})\nu(dx) - \int_{E^n} \alpha(K_x)\nu(dx) \right), \end{aligned}$$

where we have abbreviated $x = (x_1, \dots, x_n)$, and where the supremum is over all kernels from E^n to E , i.e. all Borel-measurable maps from E^n to $\mathcal{P}(E)$. *A priori*, the supremum should be taken over maps K from E^n to $\mathcal{P}(E)$ which are measurable with respect to the smallest σ -field containing the analytic sets. But any such map is universally measurable and thus agrees ν -a.e. with a Borel-measurable map. Every probability measure on E^{n+1} can be written as $\nu(dx)K_x(dx_{n+1})$ for some $\nu \in \mathcal{P}(E^n)$ and some kernel K from E^n to E . Thus, in light of (A.1) and (A.2),

$$\begin{aligned} \rho_n(g) &= \sup_{\nu \in \mathcal{P}(E^n)} \sup_K \left[\int_{E^n} \int_E f(x_1, \dots, x_{n+1})K_x(dx_{n+1})\nu(dx) - \int_{E^n} \alpha(K_x)\nu(dx) - \alpha_n(\nu) \right] \\ &= \sup_{\nu \in \mathcal{P}(E^{n+1})} \left(\int_{E^n} f \, d\nu - \alpha_{n+1}(\nu) \right) \\ &= \rho_n(f). \end{aligned} \tag{□}$$

In general, the function g in Proposition A.1 can fail to be Borel-measurable. For instance, if E is compact and $\alpha \equiv 0$, then our standing assumptions hold. In this case $\rho(f) = \sup_{x \in E} f(x)$ for $f \in B(E)$. For $f \in B(E^2)$ we have $\rho(f(x, \cdot)) = \sup_{y \in E} f(x, y)$. If $f(x, y) = 1_A(x, y)$ for a Borel set $A \subset E^2$ whose projections are not Borel, then $\rho(f(x, \cdot))$ is not Borel. Credit is due to Daniel Bartl for pointing out this simple counterexample to an inaccurate claim in an earlier version of the paper; his paper [6] shows why semianalytic functions are essential in this context.

Appendix B. Two technical lemmas

Here we state and prove a technical lemma that was used in the proof of Proposition 6.1 as well as a simple extension of Jensen’s inequality to convex functions of random measures.

The first lemma essentially says that if $f = f(x, y)$ is a function of two variables such that the map $x \mapsto f(x, \cdot)$ is measurable, from E into an appropriate function space, then f is essentially jointly measurable.

Lemma B.1. *Let (Ω, \mathcal{F}, P) be a standard Borel probability space, let E be a Polish space, and let $\mu \in \mathcal{P}(E)$. Let L^0 denote the set of equivalence classes of μ -a.e. equal measurable functions from E to \mathbb{R} , and endow L^0 with the topology of convergence in measure. If $H: \Omega \rightarrow L^0$ is measurable, then there exists a jointly measurable function $h: \Omega \times E \rightarrow \mathbb{R}$ such that, for P -a.e. ω , we have $H(\omega)(x) = h(\omega, x)$ for μ -a.e. $x \in E$.*

Proof. By Borel isomorphism, we may assume without loss of generality that $\Omega = E = [0, 1]$. In particular, $H(\omega)(x) \in [0, 1]$ for all $\omega, x \in [0, 1]$. Let L^1 denote the set of $P \times \mu$ -integrable (equivalence classes of a.s. equal) measurable functions from $[0, 1]^2$ to \mathbb{R} . Define a linear functional $T: L^1 \rightarrow \mathbb{R}$ by

$$T(\varphi) = \int P(d\omega) \int \mu(dx) H(\omega)(x) \varphi(\omega, x).$$

This is well-defined because the function

$$\omega \mapsto \int \mu(dx) H(\omega)(x) \varphi(\omega, x)$$

is measurable; indeed, this is easily checked for φ of the form $\varphi(\omega, x) = f(\omega)g(x)$, for f and g bounded and measurable, and the general case follows from a monotone class argument. Because $|H(\omega)(x)| \leq 1$, it is readily checked that T is continuous. Thus T belongs to the continuous dual of L^1 , and there exists a bounded measurable function $h: [0, 1]^2 \rightarrow \mathbb{R}$ such that

$$T(\varphi) = \int P(d\omega) \int \mu(dx) h(\omega, x) \varphi(\omega, x),$$

for all $\varphi \in L^1$. It is straightforward to check that this h has the desired property. □

Our final lemma, an infinite-dimensional form of Jensen's inequality, is surely known, but we were unable to locate a precise reference, and the proof is quite short.

Lemma B.2. *Fix $P \in \mathcal{P}(\mathcal{P}(E))$, and define the mean measure $\bar{P} \in \mathcal{P}(E)$ by*

$$\bar{P}(A) = \int_{\mathcal{P}(E)} m(A) P(dm).$$

Then, for any function $G: \mathcal{P}(E) \rightarrow (-\infty, \infty]$ which is convex, bounded from below, and weakly lower semicontinuous, we have

$$G(\bar{P}) \leq \int_{\mathcal{P}(E)} G dP.$$

Proof. Define (on some probability space) i.i.d. $\mathcal{P}(E)$ -valued random variables $(\mu_i)_{i=1}^\infty$ with common law P . Define the partial averages

$$S_n = \frac{1}{n} \sum_{i=1}^n \mu_i.$$

For any $f \in C_b(E)$, the law of large numbers implies

$$\lim_{n \rightarrow \infty} \int_E f \, dS_n = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \int_E f \, d\mu_i = \mathbb{E} \int_E f \, d\mu_1 = \int_E f \, d\bar{P} \text{ a.s.}$$

This easily shows that $S_n \rightarrow \bar{P}$ weakly a.s. Use Fatou's lemma and the assumptions on G to get

$$G(\bar{P}) \leq \liminf_{n \rightarrow \infty} \mathbb{E}[G(S_n)] \leq \liminf_{n \rightarrow \infty} \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n G(\mu_i) \right] = \mathbb{E}[G(\mu_1)] = \int_{\mathcal{P}(E)} G \, dP. \quad \square$$

Acknowledgements

The author is indebted to Stephan Eckstein and Daniel Bartl as well as two anonymous referees for their careful feedback, which greatly improved the exposition and accuracy of the paper.

References

- [1] ACCIAIO, B. AND PENNER, I. (2011). *Dynamic risk measures*. In *Advanced Mathematical Methods for Finance*, eds G. Di Nunno AND B. Øksendal, pp. 1–34. Springer.
- [2] AGUEH, M. AND CARLIER, G. (2011). Barycenters in the Wasserstein space. *SIAM J. Math. Anal.* **43** (2), 904–924.
- [3] ALIPRANTIS, C. AND BORDER, K. (2007). *Infinite Dimensional Analysis: A Hitchhiker's Guide*, 3rd edn. Springer.
- [4] ATAR, R., CHOWDHARY, K. AND DUPUIS, P. (2015). Robust bounds on risk-sensitive functionals via Rényi divergence. *SIAM/ASA J. Uncertain. Quantif.* **3** (1), 18–33.
- [5] BACKHOFF-VERAGUAS, J., LACKER, D. AND TANGPI, L. (2018). Non-exponential Sanov and Schilder theorems on Wiener space: BSDEs, Schrödinger problems and control. Available at [arXiv:1810.01980](https://arxiv.org/abs/1810.01980).
- [6] BARTL, D. (2019). Conditional nonlinear expectations. *Stoch. Process Appl.* doi.org/10.1016/j.spa.2019.03.014.
- [7] BEN-TAL, A. AND TEOULLE, M. (1986). Expected utility, penalty functions, and duality in stochastic nonlinear programming. *Manag. Sci.* **32** (11), 1445–1466.
- [8] BEN-TAL, A. AND TEOULLE, M. (2007). An old–new concept of convex risk measures: the optimized certainty equivalent. *Math. Finance* **17** (3), 449–476.
- [9] BERTSEKAS, D. AND SHREVE, S. (1996). *Stochastic Optimal Control: The Discrete Time Case*. Athena Scientific.
- [10] BLANCHET, A. AND CARLIER, G. (2015). Optimal transport and Cournot–Nash equilibria. *Math. Operat. Res.* **41** (1), 125–145.
- [11] BOBKOV, S. AND DING, Y. (2014). Optimal transport and Rényi informational divergence. Preprint.
- [12] BOISSARD, E. (2011). Simple bounds for the convergence of empirical and occupation measures in 1-Wasserstein distance. *Electron. J. Probab.* **16**, 2296–2333.
- [13] BOROVKOV, A. A. AND BOROVKOV, K. A. (2008). *Asymptotic Analysis of Random Walks: Heavy-Tailed Distributions* (Encyclopedia Math. Appl. **118**). Cambridge University Press.
- [14] CHERIDITO, P. AND KUPPER, M. (2011). Composition of time-consistent dynamic monetary risk measures in discrete time. *Internat. J. Theoret. Appl. Finance* **14** (1), 137–162.
- [15] DE HAAN, L. AND LIN, T. (2001). On convergence toward an extreme value distribution in $C[0, 1]$. *Ann. Probab.* **29**, 467–483.
- [16] DEMBO, A. AND ZEITOUNI, O. (2009). *Large Deviations Techniques and Applications* (Stoch. Model. Appl. Prob. **38**). Springer Science & Business Media.
- [17] DENISOV, D., DIEKER, A. B. AND SHNEER, V. (2008). Large deviations for random walks under subexponentiality: the big-jump domain. *Ann. Probab.* **36** (5), 1946–1991.
- [18] DING, Y. (2014). Wasserstein-Divergence transportation inequalities and polynomial concentration inequalities. *Statist. Prob. Lett.* **94**, 77–85.
- [19] DUDLEY, R. M. (1969). The speed of mean Glivenko–Cantelli convergence. *Ann. Math. Statist.* **40** (1), 40–50.
- [20] DUDLEY, R. M. (2018). *Real Analysis and Probability*. CRC Press.
- [21] DUPACOVÁ, J. AND WETS, R. J. B. (1988). Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *Ann. Statist.* **16**, 1517–1549.
- [22] DUPUIS, P. AND ELLIS, R. S. (2011). *A Weak Convergence Approach to the Theory of Large Deviations* (Wiley Series Prob. Statist. **902**). John Wiley.

- [23] ECKSTEIN, S. (2019). Extended Laplace principle for empirical measures of a Markov chain. *Adv. Appl. Prob.* **51** (1), 136–167.
- [24] EICHELSBACHER, P. AND SCHMOCK, U. (1996). Large deviations of products of empirical measures and U -empirical measures in strong topologies. Sonderforschungsbereich 343, Diskrete Strukturen in der Math., Universität Bielefeld.
- [25] EINMAHL, U. AND LI, D. (2008). Characterization of LIL behavior in Banach space. *Trans. Amer. Math. Soc.* **360** (12), 6677–6693.
- [26] FÖLLMER, H. AND KNISPEL, T. (2011). Entropic risk measures: coherence vs. convexity, model ambiguity and robust large deviations. *Stoch. Dynamics* **11** (02n03), 333–351.
- [27] FÖLLMER, H. AND SCHIED, A. (2002). Convex measures of risk and trading constraints. *Finance Stochast.* **6** (4), 429–447.
- [28] FÖLLMER, H. AND SCHIED, A. (2011). *Stochastic Finance: An Introduction in Discrete Time*. Walter de Gruyter.
- [29] FOSS, S., KORSHUNOV, D. AND ZACHARY, S. (2011). *An Introduction to Heavy-Tailed and Subexponential Distributions* (Springer Series Operat. Research Financial Eng. 6). Springer.
- [30] FOURNIER, N. AND GUILLIN, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Prob. Theory Relat. Fields* **162** (3–4), 707–738.
- [31] FUQING, G. AND MINGZHOU, X. (2012). Relative entropy and large deviations under sublinear expectations. *Acta Math. Sci.* **32** (5), 1826–1834.
- [32] HARDY, G. H., LITTLEWOOD, J. E. AND PÓLYA, G. (1952). *Inequalities*. Cambridge University Press.
- [33] HU, F. (2010). On Cramér's theorem for capacities. *Comptes Rendus Mathématique* **348** (17), 1009–1013.
- [34] HULT, H. AND LINDSKOG, F. (2006). Regular variation for measures on metric spaces. *Publ. Inst. Math. (Beograd)* (NS) **80** (94), 121–140.
- [35] HULT, H., LINDSKOG, F., MIKOSCH, T. AND SAMORODNITSKY, G. (2005). Functional large deviations for multivariate regularly varying random walks. *Ann. Appl. Prob.* **15** (4), 2651–2680.
- [36] KALL, P. (1987). *On Approximations and Stability in Stochastic Programming*. In *Parametric Optimization and Related Topics*, eds J. Guddat et al. pp. 387–407. Akademie, Berlin.
- [37] KANIOVSKI, Y. M., KING, A. J. AND WETS, R. J. B. (1995). Probabilistic bounds (via large deviations) for the solutions of stochastic programming problems. *Ann. Operat. Res.* **56** (1), 189–208.
- [38] KING, A. J. AND WETS, R. J. B. (1991). Epi-consistency of convex stochastic programs. *Stoch. Stoch. Reports* **34** (1–2), 83–92.
- [39] LACKER, D. (2018). Liquidity, risk measures, and concentration of measure. *Math. Operat. Res.* **34** (3), 693–1050, C2.
- [40] LINDSKOG, F., RESNICK, S. I. AND ROY, J. (2014). Regularly varying measures on metric spaces: hidden regular variation and hidden jumps. *Prob. Surveys* **11**, 270–314.
- [41] MIKOSCH, T. AND NAGAEV, A. V. (1998). Large deviations of heavy-tailed sums with applications in insurance. *Extremes* **1** (1), 81–110.
- [42] MOGUL'SKII, A. A. (1977). Large deviations for trajectories of multi-dimensional random walks. *Theory Prob. Appl.* **21** (2), 300–315.
- [43] NAGAEV, S. V. (1979). Large deviations of sums of independent random variables. *Ann. Prob.* **7** (5), 745–789.
- [44] OWARI, K. (2014). Maximum Lebesgue extension of monotone convex functions. *J. Funct. Anal.* **266** (6), 3572–3611.
- [45] PARTHASARATHY, K. R. (2005). *Probability Measures on Metric Spaces* (Prob. Math. Statist. 352). American Mathematical Society.
- [46] PENG, S. (2007). Law of large numbers and central limit theorem under nonlinear expectations. Available at [arXiv:math/0702358](https://arxiv.org/abs/math/0702358).
- [47] PENG, S. (2010). Nonlinear expectations and stochastic calculus under uncertainty. Available at [arXiv:1002.4546](https://arxiv.org/abs/1002.4546).
- [48] PETROV, V. (2012). *Sums of Independent Random Variables*. Springer Science & Business Media.
- [49] RHEE, C.-H., BLANCHET, J. AND ZWART, B. (2016). Sample path large deviations for heavy-tailed Lévy processes and random walks. Available at [arXiv:1606.02795](https://arxiv.org/abs/1606.02795).
- [50] SCHIED, A. (1998). Cramer's condition and Sanov's theorem. *Statist. Prob. Lett.* **39** (1), 55–60.
- [51] SION, M. (1958). On general minimax theorems. *Pacific J. Math.* **8** (1), 171–176.
- [52] VILLANI, C. (2003). *Topics in Optimal Transportation* (Graduate Studies Math. 58). American Mathematical Society.
- [53] WANG, R., WANG, X. AND WU, L. (2010). Sanov's theorem in the Wasserstein distance: a necessary and sufficient condition. *Statist. Prob. Lett.* **80** (5), 505–512.
- [54] WEED, J. AND BACH, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli* **25** (4A), 2620–2648.
- [55] ZALINESCU, C. (2002). *Convex Analysis in General Vector Spaces*. World Scientific.