SIAM J. OPTIM.
Vol. 33, No. 2, pp. 615–637

# MINI-BATCH RISK FORMS[*]

DARINKA DENTCHEVA[†] AND ANDRZEJ RUSZCZYŃSKI[‡]

**Abstract.** Risk forms are real functionals of two arguments: a bounded measurable function on a Polish space and a probability measure on that space. They are convenient mathematical structures adapting the coherent risk measures to the situation of a variable reference probability measure. We introduce a new class of risk forms called mini-batch forms. We construct them by using a random empirical probability measure as the second argument and by post-composition with the expected value operator. We prove that coherent and law invariant risk forms generate mini-batch risk forms which are well defined on the space of integrable random variables, and we derive their dual representation. We demonstrate how unbiased stochastic subgradients of such risk forms can be constructed. Then, we consider pre-compositions of mini-batch risk forms with nonsmooth and nonconvex functions, which are differentiable in a generalized way, and we derive generalized subgradients and unbiased stochastic subgradients of such compositions. Finally, we study the dependence of risk forms and mini-batch risk forms on perturbation of the probability measure and establish quantitative stability in terms of optimal transport metrics. We obtain finite-sample expected error estimates for mini-batch risk forms involving functions on a finite-dimensional space.

**Key words.** risk measures, empirical estimates, dual representation, stochastic subgradients, Wasserstein metric, error bounds

**MSC codes.** 90C15, 90C48, 49J52, 60F99

**DOI.** 10.1137/22M1503774

## 1. Introduction.

The theory of risk measures is one of the main directions of research in stochastic optimization, with many applications, beyond the original motivation in finance. The main setting is the following: a probability space $(\Omega, \mathscr{F}, P)$ is fixed and a risk measure is defined as a functional on a certain vector space of real-valued measurable functions on $\Omega$ (usually, $\mathscr{L}_p(\Omega, \mathscr{F}, P)$ with $p \in [1, \infty]$). The functional is required to satisfy several axioms, which we recall in the next section. The initial contributions were [21], [28], [1], [24], and [14]; we refer the reader to [15], [39], [32], and [40] for detailed presentation, applications, and further references. However, in many problems of risk-averse optimization and control, such as controlled Markov systems [36] or partially observable systems [13], we deal with variable and decision-dependent probability measures. This makes the extant risk measure theory insufficient.

In [9], we introduced *risk forms*: real-valued functionals $\varrho[Z, P]$ of two arguments, a bounded measurable function $Z$ on a Polish space $\mathfrak{D}$, and a probability measure $P$ on the Borel $\sigma$-field $\mathscr{B}(\mathfrak{D})$. Under less restrictive assumptions than in the fixed probability measure case, we proved a dual representation and a generalized Kusuoka representation of risk forms, which remain valid for all probability measures on $\mathscr{B}(\mathfrak{D})$.

[*] Received by the editors June 16, 2022; accepted for publication (in revised form) November 18, 2022; published electronically June 2, 2023.
https://doi.org/10.1137/22M1503774

**Funding:** This work was supported by the National Science Foundation award DMS-1907522 and by the Office of Naval Research award N00014-21-1-2161.

[†] Department of Mathematical Sciences, Stevens Institute of Technology, Hoboken, NJ 07030 USA (darinka.dentcheva@stevens.edu).

[‡] Department of Management Science and Information Systems, Rutgers University, Piscataway, NJ 08854 USA (rusz@rutgers.edu).

Our goal is to advance the theory of risk forms by considering a random empirical probability measure as their second argument. We interpret $\mathfrak{D}$ as the "data space" and evaluate risk on a small sample of the data, frequently referred to as a *mini-batch* in machine learning. The expected value of this risk evaluation is a new object, which we call the *mini-batch risk form*. We formally define these functionals in section 2. They inherit many properties of the "mother" risk forms, but they have the remarkable feature that they are well defined on *all* integrable random variables, not only bounded functions. In section 3 we develop an explicit dual representation of coherent and law invariant mini-batch risk forms. We use it in section 4 to show how unbiased stochastic subgradients of mini-batch risk forms can be constructed by simulation. This is in contrast to earlier works, such as [17, 18, 38], where major effort was needed to overcome the inherent bias in the estimation of stochastic subgradients of risk measures. In section 5, we analyze the pre-composition of coherent and law invariant mini-batch risk forms with possibly nonsmooth and nonconvex functions from a very broad class of functions that are differentiable in a generalized sense [27]. This class of functions contains all semismooth functions and covers virtually all structures arising in machine learning applications; we refer the reader to [18, 37] for an extensive discussion of this issue. We show, under quite general assumptions, that such a composition is differentiable in a generalized way itself, and we show a straightforward way to construct its stochastic subgradients. This opens the door for many applications of mini-batch risk forms in risk-averse machine learning [22]. Finally, in section 6, we study the continuity of risk forms and mini-batch risk forms with respect to the probability measure, by using transportation metrics. This allows us to develop finite-sample estimates of the difference between the mini-batch risk forms and their "mother" forms for the case of the Average Value at Risk, Kusuoka forms, and mean-semideviation forms of arbitrary orders. These results complement the asymptotic properties established in [6, 7, 35, 40].

## 2. Definition and elementary properties.

Consider a Polish space $\mathfrak{D}$ and its Borel $\sigma$-algebra $\mathscr{B}(\mathfrak{D})$. Let $\mathscr{P}(\mathfrak{D})$ be the set of probability measures on $\mathscr{B}(\mathfrak{D})$. The space of all real-valued bounded measurable functions on $\mathfrak{D}$ is denoted by $\mathbb{B}(\mathfrak{D})$. We use $D$ to denote an element of $\mathfrak{D}$ and $\delta_D$ to denote the Dirac measure concentrated at $D$. The symbol $\mathbb{1}$ stands for the function in $\mathbb{B}(\mathfrak{D})$ that is constantly equal to 1.

A *probabilistic model* is a pair $[Z,P]\in\mathbb{B}(\mathfrak{D})\times\mathscr{P}(\mathfrak{D})$. For two probabilistic models $[Z,P]$ and $[W,Q]$ the notation $[Z,P]\sim[W,Q]$ means that $P\{Z\le\eta\}=Q\{W\le\eta\}$ for all $\eta\in\mathbb{R}$ (both models have the same distribution function). The inequality $Z\le V$ between elements of $\mathbb{B}(\mathfrak{D})$ is always understood pointwise.

In [9], we proposed an approach to risk evaluation of a family of probabilistic models. In the definition below, the first four properties are the same as for a coherent measure of risk, with the second argument of the risk form fixed. The last two properties are specific for our model with two arguments.

Definition 2.1. *A measurable functional $\varrho:\mathbb{B}(\mathfrak{D})\times\mathscr{P}(\mathfrak{D})\to\mathbb{R}$ is called a risk form.*

(i) *It is* convex *if $\varrho[\lambda Z+(1-\lambda)W,P]\le\lambda\varrho[Z,P]+(1-\lambda)\varrho[W,P]$ for all $Z,W\in\mathbb{B}(\mathfrak{D})$, all $\lambda\in[0,1]$, and all $P\in\mathscr{P}(\mathfrak{D})$.*

(ii) *It is* monotonic *if $Z\le W$ implies $\varrho[Z,P]\le\varrho[W,P]$ for all $P\in\mathscr{P}(\mathfrak{D})$.*

(iii) *It is* translation equivariant *if for all $Z\in\mathbb{B}(\mathfrak{D})$, all $a\in\mathbb{R}$, and all $P\in\mathscr{P}(\mathfrak{D})$, $\varrho[a\mathbb{1}+Z,P]=a+\varrho[Z,P]$.*

(iv) *It is* positively homogeneous *if for all $Z\in\mathbb{B}(\mathfrak{D})$, all $\beta\in\mathbb{R}_+$, and all $P\in\mathscr{P}(\mathfrak{D})$, $\varrho[\beta Z,P]=\beta\varrho[Z,P]$.*

(v) *It is* law invariant *if* $[Z,P]\sim[W,Q]$ *implies that* $\varrho[Z,P]=\varrho[W,Q]$.

(vi) *It has the* support property *if* $\varrho[\mathbb{1}_{\text{supp}(P)}Z,P]=\varrho[Z,P]$ *for all* $(Z,P)\in\mathbb{B}(\mathfrak{D})\times\mathscr{P}(\mathfrak{D})$.

We say that a risk form is *coherent* if it satisfies the properties (i)–(iv) above.

A simple example of a risk form is the expected value, which is the well-understood bilinear form

$$\mathbb{E}[Z,P]=\int_{\mathfrak{D}} Z(v)\,P(\mathrm{d}v). \tag{2.1}$$

In our analysis, we are interested mainly in risk forms depending on each of the arguments in a nonlinear way. An example is the *mean-semideviation model* or order $p\in[1,\infty)$ (see [28, 29]):

$$\operatorname{msd}_p[Z,P]=\int_{\mathfrak{D}} Z(u)\,P(\mathrm{d}u)+\varkappa\left(\int_{\mathfrak{D}}\left[Z(u)-\int_{\mathfrak{D}} Z(v)\,P(\mathrm{d}v)\right]_+^p P(\mathrm{d}u)\right)^{1/p},\ \varkappa\in[0,1]. \tag{2.2}$$

Yet another example, rarely used in the risk measure theory, due to its conservative nature, but very relevant for us, is the *worst-case risk form*:

$$\operatorname{es}[Z,P]=\inf\left\{b\in\mathbb{R}:P[v:Z(v)\le b]=1\right\}. \tag{2.3}$$

All three examples above are coherent and law invariant risk forms having the support property.

Our concept of law invariance is broader than that for the measures of risk, because it allows the probability measure to vary. If the risk form is law invariant, then it has the support property, because $[Z,P]\sim[\mathbb{1}_{\text{supp}(P)}Z,P]$.

We now introduce the main object of our study: a *mini-batch risk form*. Suppose $\varrho:\mathbb{B}(\mathfrak{D})\times\mathscr{P}(\mathfrak{D})\to\mathbb{R}$ is a risk form and $P\in\mathscr{P}(\mathfrak{D})$. If we draw a sample $D^{1:N}=(D^1,\dots,D^N)$, with $N$ independent random elements distributed according to $P$ in $\mathfrak{D}$, we obtain a random empirical measure

$$P^{(N)}=\frac{1}{N}\sum_{i=1}^N \delta_{D^i}.$$

It is a $\mathscr{P}(\mathfrak{D})$-valued random variable on the product space $(\mathfrak{D}^N,\mathscr{B}(\mathfrak{D}^N),P^N)$. Using it as the second argument of the risk form $\varrho$, we obtain a random risk form $\varrho[Z,P^{(N)}]$. For fixed $Z$ and $P$, it is a random variable on $(\mathfrak{D}^N,\mathscr{B}(\mathfrak{D}^N),P^N)$. This leads to the following definition.

DEFINITION 2.2. *For a risk form* $\varrho:\mathbb{B}(\mathfrak{D})\times\mathscr{P}(\mathfrak{D})\to\mathbb{R}$*, the corresponding* mini-batch risk form $\varrho^{(N)}:\mathbb{B}(\mathfrak{D})\times\mathscr{P}(\mathfrak{D})\to\mathbb{R}$ *is defined as*

$$\varrho^{(N)}[Z,P]=\mathbb{E}_{D^{1:N}\sim P^N}\left\{\varrho[Z,P^{(N)}]\right\}. \tag{2.4}$$

The following lemma summarizes the basic properties of a mini-batch risk form.

LEMMA 2.3. *If the risk form* $\varrho[\,\cdot\,,\cdot\,]$ *is convex (monotonic, translation equivariant, positively homogeneous, or has the support property), then the mini-batch risk form* $\varrho^{(N)}[\,\cdot\,,\cdot\,]$ *has the corresponding properties as well. If the risk form* $\varrho[\,\cdot\,,\cdot\,]$ *is law invariant, then the mini-batch risk form* $\varrho^{(N)}[\,\cdot\,,\cdot\,]$ *is law invariant as well.*

*Proof.* The inheritance by $\varrho^{(N)}[\,\cdot\,,\,\cdot\,]$ of convexity, monotonicity, translation equivariance, positively homogeneity, and the support property from $\varrho[\,\cdot\,,\,\cdot\,]$ is evident. It remains to prove the preservation of law invariance. First we establish that $\varrho[Z,P^{(N)}]$ is a function of the (unordered) set of realizations of $Z$ under $P^{(N)}$. Consider two discrete measures:

$$P^{(N)} = \frac{1}{N}\sum_{i=1}^{N}\delta_{D^i} \quad \text{and} \quad Q^{(N)} = \frac{1}{N}\sum_{i=1}^{N}\delta_{w^i}.$$

If the sets of the realizations of two functions, $Z(\cdot)$ and $U(\cdot)$, under $P^{(N)}$ and $Q^{(N)}$, respectively, are identical,

$$\big\{Z(D^1),\dots,Z(D^N)\big\} = \big\{U(w^1),\dots,U(w^N)\big\},$$

then the distribution of $Z$ under $P^{(N)}$ is the same as the distribution of $U$ under $Q^{(N)}$. By the law invariance,

$$\varrho\left[Z,P^{(N)}\right] = \varrho\left[U,Q^{(N)}\right].$$

This means that $\varrho[Z,P^{(N)}]$ is only a function of the set $\{Z(D^1),\dots,Z(D^N)\}$. Thus, a measurable function $\varPsi:\mathbb{R}^N\to\mathbb{R}$ exists, such that

$$\varrho[Z,P^{(N)}] = \varPsi\left(Z(D^1),\dots,Z(D^N)\right), \tag{2.5}$$

and for every permutation $\pi$ of $\{1,\dots,N\}$,

$$\varPsi\left(Z(D^1),\dots,Z(D^N)\right) = \varPsi\left(Z(D^{\pi(1)}),\dots,Z(D^{\pi(N)})\right). \tag{2.6}$$

Now, if $[Z,P]\sim[U,Q]$, then for $D^{1:N}\sim P^N$ and $C^{1:N}\sim Q^N$, the vectors $(Z(D^1),\dots,Z(D^N))$ and $(U(C^1),\dots,U(C^N))$ have the same distribution. Therefore,

$$\begin{aligned}\varrho^{(N)}[Z,P] &= \mathbb{E}_{D^{1:N}\sim P^N}\left\{\varPsi\left(Z(D^1),\dots,Z(D^N)\right)\right\}\\ &= \mathbb{E}_{C^{1:N}\sim Q^N}\left\{\varPsi\left(U(C^1),\dots,U(C^N)\right)\right\} = \varrho^{(N)}[U,Q].\end{aligned}$$ □

Let us consider the mini-batch risk form associated with (2.3):

$$\mathrm{es}^{(N)}[Z,P] = \mathbb{E}_{D^{1:N}\sim P^N}\left[\max_{1\le i\le N} Z(D^i)\right]. \tag{2.7}$$

In the special case of $N=2$ it has the mean-risk structure,

$$\begin{aligned}\mathrm{es}^{(2)}[Z,P] &= \mathbb{E}_{(D^1,D^2)\sim P^2}\left[\max\left(Z(D^1),Z(D^2)\right)\right]\\ &= \mathbb{E}_{(D^1,D^2)\sim P^2}\left[Z(D^1)+\max\left(0,Z(D^2)-Z(D^1)\right)\right] = \mathbb{E}[Z]+\varGamma[Z],\end{aligned}$$

with the Gini index [30, 43]

$$\varGamma[Z] = \frac{1}{2}\mathbb{E}_{(D^1,D^2)\sim P^2}\left[\left|Z(D^1)-Z(D^2)\right|\right].$$

For $N\ge 2$, (2.7) is a generalized mean-Gini model, with increasing degree of risk aversion. Its explicit form allows for more detailed analysis.

Let us recall that the Average Value at Risk is defined as

$$\mathrm{AVaR}_\alpha[Z,P] = \frac{1}{\alpha}\int_{1-\alpha}^{1} F_Z^{-1}(t)\,\mathrm{d}t, \quad \alpha\in(0,1], \tag{2.8}$$

with $F_Z^{-1}(\cdot)$ denoting the quantile function of $Z$ under $P$: $F_Z^{-1}(t) = \inf\{\tau : P[v : Z(v)\le\tau]\ge t\}$. A *spectral risk form* has the representation

$$\varrho[Z,P] = \int_0^1 \mathrm{AVaR}_\alpha[Z,P]\,\lambda(\alpha)\,\mathrm{d}\alpha,$$

with the *spectral density* $\lambda\ge 0$ satisfying $\int_0^1\lambda(\alpha)\,\mathrm{d}\alpha = 1$.

THEOREM 2.4. *The mini-batch risk form* (2.7) *is a spectral measure of risk with the spectral density* $N(N-1)\alpha(1-\alpha)^{N-2}$, $\alpha\in[0,1]$*. Furthermore, for a fixed* $P$*, the risk form* (2.7) *is well defined and finite for all* $Z\in\mathscr{L}_1(\mathfrak{D},\mathscr{B}(\mathfrak{D}),P)$.

*Proof.* The distribution function of the sample maximum is

$$P^N\left\{\mathrm{es}^{(N)}[Z,P]\le v\right\} = \left(P[Z\le v]\right)^N, \quad v\in\mathbb{R}.$$

Therefore, the mini-batch risk form (2.7) can be rewritten by changing the variables:

$$\mathrm{es}^{(N)}[Z,P] = \int_{-\infty}^{\infty} v\,\mathrm{d}\left(\left(P[Z\le v]\right)^N\right) = \int_0^1 F_Z^{-1}(\alpha)\,\mathrm{d}\left(\alpha^N\right).$$

The last expression is a dual (rank-dependent) utility functional [8, 33, 42]

$$\mathrm{es}^{(N)}[Z,P] = \int_0^1 F_Z^{-1}(\alpha)\,\mathrm{d}w(\alpha), \tag{2.9}$$

with the rank-dependent utility function $w(\alpha)=\alpha^N$.

For a fixed $P$, the functional (2.7) as a function of $Z$ is a coherent measure of risk. The expression (2.9) allows for the derivation of its Kusuoka representation. Changing the order of integration, we obtain

$$\begin{aligned}\mathrm{es}^{(N)}[Z,P] &= N\int_0^1 F_Z^{-1}(\alpha)\alpha^{N-1}\,\mathrm{d}\alpha = N(N-1)\int_0^1 F_Z^{-1}(\alpha)\int_0^\alpha \beta^{N-2}\,\mathrm{d}\beta\,\mathrm{d}\alpha\\ &= N(N-1)\int_0^1 \beta^{N-2}\int_\beta^1 F_Z^{-1}(\alpha)\,\mathrm{d}\alpha\,\mathrm{d}\beta\\ &= N(N-1)\int_0^1 (1-\beta)\beta^{N-2}\mathrm{AVaR}_{1-\beta}[Z]\,\mathrm{d}\beta.\end{aligned} \tag{2.10}$$

It is straightforward to check that $N(N-1)\int_0^1(1-\beta)\beta^{N-2}\,\mathrm{d}\beta = 1$, as required from the spectral density. □

*Example* 2.5. Consider the mini-batch mean-semideviation risk form derived from (2.2) with $p=1$:

$$\mathrm{msd}_1^{(N)}[Z,P] = \mathbb{E}_{D^{1:N}\sim P^N}\left\{\frac{1}{N}\sum_{j=1}^N Z(D^j) + \frac{\varkappa}{N}\sum_{j=1}^N\left(Z(D^j) - \frac{1}{N}\sum_{k=1}^N Z(D^k)\right)_+\right\}. \tag{2.11}$$

In the special case of $N=2$ it has a mean-risk form similar to the previous example:

$$\operatorname{msd}_1^{(2)}[Z,P]=\mathbb{E}[Z]+\frac{\varkappa}{4}\mathbb{E}_{(D^1,D^2)\sim P^2}\left[\left|Z(D^1)-Z(D^2)\right|\right]=\mathbb{E}[Z]+\frac{\varkappa}{2}\varGamma[Z].$$

Again, for a fixed $P$, the functional (2.11) as a function of $Z$ is a coherent measure of risk which is well-defined and finite on $\mathscr{L}_1(\mathfrak{D},\mathscr{B}(\mathfrak{D}),P)$.

## 3. Dual representation and extension to integrable random variables.

Suppose the risk form $\varrho:\mathbb{B}(\mathfrak{D})\times\mathscr{P}(\mathfrak{D})\to\mathbb{R}$ is coherent and law invariant. Then, for a fixed probability measure $P$, the function $\varrho^{(N)}[\,\cdot\,,P]$ is a coherent measure of risk on the space of bounded functions $\mathbb{B}(\mathfrak{D})$. We plan to derive its dual representation. In general, such a representation would have to involve elements from the space of finitely additive measures $\operatorname{ba}(\mathscr{B}(\mathfrak{D}))$, which is the topological dual of $\mathbb{B}(\mathfrak{D})$. In [9, Thm. 1] we have shown, however, that the dual representation of coherent and law invariant risk forms involves only countably additive measures. Now, we advance the analysis for the mini-batch risk forms, and we derive a more explicit representation in terms of measures which are absolutely continuous with respect to $P$.

Owing to (2.5), the random risk form $\varrho[\,\cdot\,,P^{(N)}]$ is a function $\varPsi(\cdot)$ of the vector $(Z(D^1),\dots,Z(D^N))$. Since $\varrho[\,\cdot\,,P^{(N)}]$ is coherent and law invariant, the function $\varPsi:\mathbb{R}^N\to\mathbb{R}$ is convex, nondecreasing, and positively homogeneous and has the translation property: $\varPsi(z_1+a,\dots,z_N+a)=\varPsi(z_1,\dots,z_N)+a$. Due to (2.6), it is invariant with respect to permutations of the arguments. Applying Fenchel duality to $\varPsi(\cdot)$ and reasoning as for the dual representation of a coherent measure of risk [39, Thm. 2.2], we obtain that a closed convex set

$$\mathscr{A}^{(N)}\subset S^{(N)}=\left\{\xi\in\mathbb{R}^N_+:\sum_{j=1}^N\xi_j=1\right\} \tag{3.1}$$

exists, such that for all $Z\in\mathbb{B}(\mathfrak{D})$

$$\varrho[Z,P^{(N)}]=\max_{\xi\in\mathscr{A}^{(N)}}\sum_{j=1}^N\xi_jZ(D^j). \tag{3.2}$$

The set $\mathscr{A}^{(N)}$ depends on $N$, but not on the specific sample $D^{1:N}$, and has the property that for every $\xi\in\mathscr{A}^{(N)}$ and for every permutation $\pi$ of $\{1,\dots,N\}$, the vector$(\xi_{\pi(1)},\dots,\xi_{\pi(N)})$ is an element of $\mathscr{A}^{(N)}$ as well. Therefore,

$$\varrho^{(N)}[Z,P]=\mathbb{E}_{D^{1:N}\sim P^N}\left\{\max_{\xi\in\mathscr{A}^{(N)}}\sum_{j=1}^N\xi_jZ(D^j)\right\}. \tag{3.3}$$

We now transform formula (3.3) into the standard dual representation of a coherent measure of risk. Consider the cylindrical multifunction $H:\mathfrak{D}^N\rightrightarrows S^{(N)}$ defined as

$$H(D^{1:N})=\mathscr{A}^{(N)}\quad\text{for all}\quad D^{1:N}\in\mathfrak{D}^N.$$

For each $D^{1:N}$ the maximizer in the braces in (3.3) exists; with no loss of generality we may assume that it is a measurable function of the sample. This follows from Berge's theorem about the measurability of the optimal value (see, e.g., [2, Th. 1.4.16]), and from the existence of measurable selectors [23]. We can thus interchange the maximization and expectation operators and rewrite the formula (3.3) as follows:

$$\varrho^{(N)}[Z,P] = \max_{\xi(\cdot)\lessdot H} \mathbb{E}_{D^{1:N}\sim P^N} \left\{ \sum_{j=1}^{N} \xi_j(D^{1:N})\, Z(D^j) \right\}. \tag{3.4}$$

Here, the notation $\xi(\cdot)\lessdot H$ means that $\xi(\cdot)$ is a measurable selector of $H(\cdot)$. The said maximizer in the braces in (3.3), as a function of $D^{1:N}$, constitutes a maximizer in (3.4).

Define linear continuous operators $\mathit{\Pi}_j^{(N)} : \mathscr{L}_\infty(\mathfrak{D}^N, \mathscr{B}(\mathfrak{D}^N), P^N) \to \mathscr{L}_\infty(\mathfrak{D}, \mathscr{B}(\mathfrak{D}), P)$, $j=1,\ldots,N$, as follows:

$$\left[\mathit{\Pi}_j^{(N)}(\xi)\right](v) = \mathbb{E}_{D^{1:N}\sim P^N} \left[ \xi_j(D^{1:N}) \,\middle|\, D^j = v \right], \quad v \in \mathfrak{D}. \tag{3.5}$$

Using the fact that all $D^j$ are distributed according to $P$, we obtain

$$\begin{aligned} \varrho^{(N)}[Z,P] &= \max_{\xi(\cdot)\lessdot H} \mathbb{E}_{D^{1:N}\sim P^N} \left\{ \sum_{j=1}^{N} \left[\mathit{\Pi}_j^{(N)}(\xi)\right](D^j)\, Z(D^j) \right\} \\ &= \max_{\xi(\cdot)\lessdot H} \int_{\mathfrak{D}} \sum_{j=1}^{N} \left[\mathit{\Pi}_j^{(N)}(\xi)\right](v)\, Z(v)\, P(\mathrm{d}v). \end{aligned} \tag{3.6}$$

In this way, we established the following result.

Theorem 3.1. *If the risk form $\varrho : \mathbb{B}(\mathfrak{D}) \times \mathscr{P}(\mathfrak{D}) \to \mathbb{R}$ is coherent and law invariant, then, for every $N \geq 1$, the corresponding mini-batch risk form has the following dual representation:*

$$\varrho^{(N)}[Z,P] = \sup_{Q \in \partial\varrho^{(N)}[0,P]} \int_{\mathfrak{D}} Z(v)\, Q(\mathrm{d}v), \tag{3.7}$$

*where*

$$\partial\varrho^{(N)}[0,P] = \left\{ Q \in \mathscr{P}(\mathfrak{D}) : \exists\, (\xi(\cdot)\lessdot H) \;\; \frac{\mathrm{d}Q}{\mathrm{d}P} = \sum_{j=1}^{N} \left[\mathit{\Pi}_j^{(N)}(\xi)\right] \right\}, \tag{3.8}$$

*with the operators $\mathit{\Pi}_j^{(N)}$ defined as in* (3.5).

We may check the essential properties of $\partial\varrho^{(N)}[0,P]$. Since $\mathscr{A}^{(N)}$ is convex, then the set of measurable selectors $\xi(\cdot)\lessdot H$ is convex as well. As the mapping $\xi \mapsto \sum_{j=1}^{N} \mathit{\Pi}_j^{(N)}(\xi)$ is linear, the resulting set of densities is indeed convex. Furthermore, for every measure $Q$ in the set (3.8), using (3.5) and (3.1), we obtain

$$Q(\mathfrak{D}) = \sum_{j=1}^{N} \mathbb{E}_{D^{1:N}\sim P^N}[\mathit{\Pi}_j^{(N)}(\xi)] = \sum_{j=1}^{N} \mathbb{E}_{D^{1:N}\sim P^N}[\xi_j] = \mathbb{E}_{D^{1:N}\sim P^N}\left[ \sum_{j=1}^{N} \xi_j \right] = 1.$$

Nonnegativity is evident by construction, and thus indeed $Q \in \mathscr{P}(\mathfrak{D})$, $Q \ll P$.

We conclude that for a fixed probability measure $P$, our mini-batch risk form $\varrho^{(N)}[\,\cdot\,,P]$, although originally defined on the set of bounded functions, has a dual representation in terms of countably additive measures, without any singular components. This is due to the involvement of the expectation over a finite sample, which "averages out" the singularities (this can be seen best on the form (2.7)). All probability measures in the dual representation are absolutely continuous with respect to the base measure $P$. Their densities are bounded by $N$, because all $[\mathit{\Pi}_j^{(N)}(\xi)](\cdot) \in [0,1]$, due to (3.1). Therefore, for a fixed probability measure $P$, we can consider the following mini-batch risk measure $r^{(N)} : \mathscr{L}_1(\mathfrak{D}, \mathscr{B}(\mathfrak{D}), P) \to \mathbb{R}$:

$$r^{(N)}[Z] = \varrho^{(N)}[Z, P], \quad Z \in \mathscr{L}_1(\mathfrak{D}, \mathscr{B}(\mathfrak{D}), P). \tag{3.9}$$

Corollary 3.2. *If the risk form $\varrho : \mathbb{B}(\mathfrak{D}) \times \mathscr{P}(\mathfrak{D}) \to \mathbb{R}$ is coherent and law invariant, then for every $N \geq 1$ the functional* (3.9) *is a coherent measure of risk on the space $\mathscr{L}_1(\mathfrak{D}, \mathscr{B}(\mathfrak{D}), P)$, and its subdifferential is given by the equation*

$$\partial r^{(N)}[Z] = \left\{ \zeta \in \mathscr{L}_\infty(\mathfrak{D}, \mathscr{B}(\mathfrak{D}), P) : \ \exists (\hat{\xi}(\cdot) \lessdot \hat{H}) \ \ \zeta = \sum_{j=1}^{N} \mathit{\Pi}_j^{(N)}(\hat{\xi}) \right\}, \tag{3.10}$$

*where the multifunction $\hat{H} : \mathfrak{D}^N \rightrightarrows S^{(N)}$ is defined as*

$$\hat{H}(D^{1:N}) = \operatorname*{Arg\,max}_{\xi \in \mathscr{A}^{(N)}} \sum_{j=1}^{N} \xi_j Z(D^j), \quad D^{1:N} \in \mathfrak{D}^N.$$

*Example* 3.3. Consider the mini-batch max risk form (2.7) of Example 2.4:

$$r^{(N)}[Z] = \mathbb{E}_{D^{1:N} \sim P^N} \left[ \max_{1 \leq i \leq N} Z(D^i) \right]; \tag{3.11}$$

the measure $P$ is fixed here. We have $\mathscr{A}^{(N)} = \mathscr{S}^{(N)}$ in this case, and the formula (3.10) applies directly.

An alternative way to derive the dual representation in this case follows from the Kusuoka representation. Formula (2.10) and Strassen's theorem imply that every subgradient of $r^{(N)}[0]$ has density of the form

$$\zeta(v) = N(N-1) \int_0^1 \alpha(1-\alpha)^{N-2} \eta(\alpha, v) \, \mathrm{d}\alpha, \quad v \in \mathfrak{D}, \tag{3.12}$$

where the function $\eta : (0,1] \times \mathfrak{D} \to \mathbb{R}_+$ satisfies the conditions

$$0 \leq \eta(\alpha, \cdot) \leq \frac{1}{\alpha}, \quad \alpha \in (0,1],$$
$$\int_{\mathfrak{D}} \eta(\alpha, v) \, P(\mathrm{d}v) = 1, \quad \alpha \in (0,1],$$

which express the requirement that $\eta(\alpha, \cdot) \in \partial \mathrm{AVaR}_\alpha[0]$. The collection of functions (3.12) is identical to (3.10) at $Z = 0$ with $\mathscr{A}^{(N)} = \mathscr{S}^{(N)}$.

**4. Unbiased stochastic subgradients.** Suppose we have a coherent and law invariant risk form $\varrho : \mathbb{B}(\mathfrak{D}) \times \mathscr{P}(\mathfrak{D}) \to \mathbb{R}$. For bounded measurable functions $Z : \mathfrak{D} \to \mathbb{R}$ we can define the mini-batch risk form $\varrho^{(N)}[Z, P]$ as in (2.4). As discussed in section 3, when $P$ is fixed and we view $\varrho^{(N)}[\,\cdot\,, P]$ as a function of its first argument, it is well defined on the space of integrable random variables $\mathscr{L}_1(\mathfrak{D}, \mathscr{B}(\mathfrak{D}), P)$. Particularly relevant for our purposes is formula (3.3). The set $\mathscr{A}^{(N)}$ is a closed convex subset of the simplex (3.1). Explicit representations of the set $\mathscr{A}^{(N)}$ for some popular measures of risk are readily available.

For a sample $D^{1:N}=(D^1,\dots,D^N)$ and a bounded measurable function $Z(\cdot)$, we construct a random probability measure $\mu(Z;D^{1:N})$ in the following way. We solve the maximization problem inside (3.3), obtaining a vector $\hat{\xi}(Z;D^{1:N})$. Then we set

$$\mu(Z;D^{1:N})=\sum_{j=1}^N \hat{\xi}_j(Z;D^{1:N})\,\delta_{D^j}. \tag{4.1}$$

For any $W\in\mathbb{B}(\mathfrak{D})$, the following inequality holds by construction:

$$\begin{aligned}\max_{\xi\in\mathscr{A}^{(N)}}\sum_{j=1}^N \xi_j W(D^j) &\ge \sum_{j=1}^N \hat{\xi}_j(Z;D^{1:N})Z(D^j)+\sum_{j=1}^N \hat{\xi}_j(Z;D^{1:N})\left(W(D^j)-Z(D^j)\right)\\ &=\sum_{j=1}^N \hat{\xi}_j(Z;D^{1:N})Z(D^j)+\int_{\mathfrak{D}}(W(v)-Z(v))\,\left[\mu(Z;D^{1:N})\right](\mathrm{d}v).\end{aligned}$$

Taking the expected value of both sides with respect to the sample $D^{1:N}$, we conclude that

$$\varrho^{(N)}[W,P]\ge\varrho^{(N)}[Z,P]+\int_{\mathfrak{D}}(W(v)-Z(v))\,[\bar{\mu}(Z)]\,(\mathrm{d}v), \tag{4.2}$$

where

$$\bar{\mu}(Z)=\mathbb{E}_{D^{1:N}\sim P^N}\left[\mu(Z;D^{1:N})\right].$$

The expected value is understood in the weak* sense: for every bounded measurable function $f:\mathfrak{D}\to\mathbb{R}$,

$$\mathbb{E}_{D^{1:N}\sim P^N}\left[\int f(v)\,\left[\mu(Z;D^{1:N})\right](\mathrm{d}v)\right]=\int f(v)\,[\bar{\mu}(Z)]\,(\mathrm{d}v). \tag{4.3}$$

It follows from (4.2) that the probability measure $\bar{\mu}(Z)$ is a subgradient of $\varrho^{(N)}[\,\cdot\,,P]$ at $Z$, and the random measure $\mu(Z;D^{1:N})$ can be interpreted as an unbiased stochastic subgradient at $Z$.

Using (4.3), for any bounded measurable function $W(\cdot)$ we obtain

$$\begin{aligned}\int_{\mathfrak{D}}(W(v)-Z(v))\,[\bar{\mu}(Z)]\,(\mathrm{d}v)&=\mathbb{E}_{D^{1:N}\sim P^N}\left[\int_{\mathfrak{D}}(W(v)-Z(v))\,\left[\mu(Z;D^{1:N})\right](\mathrm{d}v)\right]\\ &=\mathbb{E}_{D^{1:N}\sim P^N}\left[\sum_{j=1}^N \hat{\xi}_j(Z;D^{1:N})\left(W(D^j)-Z(D^j)\right)\right].\end{aligned} \tag{4.4}$$

As in (3.5) we define

$$\left[\varPi_j^{(N)}(\hat{\xi})\right](v)=\mathbb{E}\left[\hat{\xi}_j(Z;D^{1:N})\,\middle|\,D^j=v\right],\quad j=1,\dots,N.$$

Using the fact that all $D^j$ are distributed according to $P$, we obtain from (4.2) and (4.4) the inequality

$$\begin{aligned}\varrho^{(N)}[W,P]-\varrho^{(N)}[Z,P]&\ge\mathbb{E}_{D^{1:N}\sim P^N}\left\{\sum_{j=1}^N\left[\varPi_j^{(N)}(\hat{\xi})\right](D^j)\left(W(D^j)-Z(D^j)\right)\right\}\\ &=\int_{\mathfrak{D}}\sum_{j=1}^N\left[\varPi_j^{(N)}(\hat{\xi})\right](v)\,(W(v)-Z(v))\;P(\mathrm{d}v).\end{aligned}$$

This means that the subgradient $\bar{\mu}(Z)$ is absolutely continuous with respect to $P$, with the density

$$\zeta = \sum_{j=1}^{N} \mathit{\Pi}_j^{(N)}(\hat{\xi}).$$

If we extend the domain of $\varrho^{(N)}[\,\cdot\,,P]$ to $\mathscr{L}_1(\mathfrak{D},\mathscr{B}(\mathfrak{D}),P)$, we get $\zeta \in \partial\varrho^{(N)}[Z,P]$.

We shall calculate unbiased stochastic subgradients of the mini-batch risk forms defined in Examples 2.4 and 2.5.

*Example* 4.1. Consider the mini-batch risk form (2.7). For a function $Z(\cdot)$ and a sample $D^{1:N}$, we find $j^*(Z;D^{1:N})$ such that

$$Z(D^{j^*(Z;D^{1:N})}) = \max_{1\le j\le N} Z(D^j).$$

Then it follows from (4.1) that the random measure $\mu(Z;D^{1:N}) = \delta_{D^{j^*(Z;D^{1:N})}}$ is a stochastic subgradient of $\mathrm{es}^{(N)}[\,\cdot\,,P]$ at $Z$. Its weak* expected value has the density in (3.10).

*Example* 4.2. For the mini-batch risk form (2.11), given a function $Z$ and a sample $D^{1:N}$, we can find the vector $\hat{\xi}(Z;D^{1:N})$ by calculating

$$\lambda_k = \begin{cases} \frac{\varkappa}{N} & \text{if } Z(D^k) \ge \frac{1}{N}\sum_{j=1}^{N} Z(D^j),\\ 0 & \text{otherwise,}\end{cases} \qquad k=1,\dots,N,$$

and setting $\hat{\xi}_k(Z;D^{1:N}) = \frac{1}{N} + \lambda_k - \frac{1}{N}\sum_{j=1}^{N}\lambda_j$, $k=1,\dots,N$. Then the formula (4.1) provides an unbiased stochastic subgradient of $\mathrm{msd}_1^{(N)}[\,\cdot\,,P]$ at $Z$.

Stochastic subgradients of mini-batch risk forms are random probability measures. However, they rarely occur in isolation; rather, they occur in compositions that we discuss in the next section.

## 5. Compositions with nonsmooth and nonconvex functions.

In applications, we usually deal with compositions of risk measures with some random functions of our decision variables. A typical situation is the following: we have a "loss function" $\ell:\mathbb{R}^n\times\mathfrak{D}\to\mathbb{R}$ and we consider the operator $\mathit{\Lambda}:\mathbb{R}^n\to\mathscr{L}_1(\mathfrak{D},\mathscr{B}(\mathfrak{D}),P)$, defined as

$$[\mathit{\Lambda}(x)]\,(v) = \ell(x,v), \quad v\in\mathfrak{D}.$$

Then, for a risk measure $r:\mathscr{L}_1(\mathfrak{D},\mathscr{B}(\mathfrak{D}),P)\to\mathbb{R}$, we formulate the optimization problem

$$\min_{x\in X}\; r[\mathit{\Lambda}(x)], \tag{5.1}$$

with some feasible set $X\subset\mathbb{R}^n$. Of course, additional conditions (to be discussed in due course) are needed to make the composition $F = r\circ\mathit{\Lambda}$ well defined. Furthermore, in many situations, such as adversarial machine learning models [18], or multistage stochastic programming [10], the loss function $\ell(x,D)$ is neither smooth nor convex with respect to $x$. This creates theoretical challenges in the analysis of the composition $F = r\circ\mathit{\Lambda}$ and in the construction of solution methods. The extant approaches are limited to specially structured risk measures, such as the semideviations [17, 18] or the Average Value at Risk [20]. Our intention is to show that when we use the risk measure $r^{(N)}[\,\cdot\,] = \varrho^{(N)}[\,\cdot\,, P]$ derived from a mini-batch risk form, we can analyze and solve problems of form (5.1) for a very general class of nonsmooth and nonconvex loss functions.

In our analysis, we focus on the broad subclass of locally Lipschitz functions introduced in [27] and called "differentiable in a generalized sense" there. Here, we call them "Norkin differentiable" for brevity.

DEFINITION 5.1. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is* Norkin differentiable at a point $x \in \mathbb{R}^n$ *if an open set $\mathscr{U} \subset \mathbb{R}^n$ containing $x$, and a nonempty, convex, compact valued, and upper semicontinuous multifunction $\hat{\partial} f : \mathscr{U} \rightrightarrows \mathbb{R}^n$ exist, such that for all $y \in \mathscr{U}$ and all $g \in \hat{\partial} f(y)$ the following equation is true*:

$$f(y) = f(x) + \langle g, y - x\rangle + o(x, y, g),$$

*with*

$$\lim_{y\to x} \sup_{g\in\hat{\partial} f(y)} \frac{o(x,y,g)}{\|y-x\|} = 0.$$

*The set $\hat{\partial} f(y)$ is the Norkin subdifferential of $f$ at $y$. If a function is Norkin differentiable at every $x \in \mathbb{R}^n$ with the same subdifferential mapping $\hat{\partial} f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, we call it Norkin differentiable. A vector function $f : \mathbb{R}^n \to \mathbb{R}^m$ is Norkin differentiable if each of its component functions, $f_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \dots, m$, has this property.*

The class of such functions is contained in the set of locally Lipschitz functions and contains all subdifferentially regular functions [5], Whitney stratifiable Lipschitz functions [12], semismooth functions [25], and their compositions. If a function is Norkin differentiable and has directional derivatives at $x$ in every direction, then it is semismooth at $x$. The Clarke subdifferential $\partial f(x)$ is an inclusion-minimal Norkin subdifferential, but the Norkin subdifferential mapping $\hat{\partial} f(\cdot)$ is not uniquely defined in Definition 5.1, which is important when considering compositions. For stochastic optimization, essential is the closure of the class of such functions with respect to expectation, which allows for easy generation of stochastic subgradients.

THEOREM 5.2. *Suppose the function $\ell(x, v)$ is Norkin differentiable with respect to $x$ for all $v \in \mathfrak{D}$, and $P$-integrable with respect to $v$ for all $x \in \mathbb{R}^n$. Suppose the multifunction $\hat{\partial}\ell : \mathbb{R}^n \times \mathfrak{D} \rightrightarrows \mathbb{R}^n$ is measurable with respect to $v$ for all $x \in \mathbb{R}^n$, and is a Norkin subdifferential mapping of $\ell(\cdot, v)$ for all $v \in \mathfrak{D}$. Furthermore, let for every compact set $K \subset \mathbb{R}^n$ a $P$-integrable function $L_K : \mathfrak{D} \to \mathbb{R}$ exist, such that $\sup_{x\in K} \sup_{g\in\hat{\partial}\ell(x,v)} \|g\| \le L_K(v)$, $v \in \mathfrak{D}$. If the risk form $\varrho : \mathbb{B}(\mathfrak{D}) \times \mathscr{P}(\mathfrak{D}) \to \mathbb{R}$ is coherent and law invariant, then for every $N \ge 1$ the function*

$$F(x) = \varrho^{(N)}[\Lambda(x), P], \quad x \in \mathbb{R}^n, \tag{5.2}$$

*is well defined and Norkin differentiable, and the multifunction $\Gamma : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ defined as*

$$\begin{aligned} \Gamma(x) = \Bigg\{ & \gamma \in \mathbb{R}^n : \exists\, g(\cdot) \in \hat{\partial}\ell(x,\cdot),\ \exists\, \zeta \in \partial r^{(N)}[Z] \text{ at } Z = \Lambda(x), \\ & \gamma = \int_{\mathfrak{D}} g(v)\, \zeta(v)\, P(\mathrm{d}v) \Bigg\}, \quad x \in \mathbb{R}^n, \end{aligned} \tag{5.3}$$

*is its Norkin subdifferential mapping.*

*Proof.* By formula (3.3),

$$F(x) = \mathbb{E}_{D^{1:N} \sim P^N} \left\{ \max_{\xi \in \mathscr{A}^{(N)}} \sum_{j=1}^{N} \xi_j \ell(x, D^j) \right\}. \tag{5.4}$$

Consider the function $f : \mathbb{R}^n \times \mathfrak{D}^N \to \mathbb{R}$ given by

$$f(x, D^{1:N}) = \max_{\xi \in \mathscr{A}^{(N)}} \sum_{j=1}^{N} \xi_j \ell(x, D^j). \tag{5.5}$$

It is a composition $f = m \circ \ell^{(N)}$ of the convex support function $m : \mathbb{R}^M \to \mathbb{R}$,

$$m(z_1, \dots, z_N) = \max_{\xi \in \mathscr{A}^{(N)}} \sum_{j=1}^{N} \xi_j z_j,$$

and the sample loss function

$$\ell^{(N)}(x, D^{1:N}) = \big(\ell(x, D^1), \dots, \ell(x, D^N)\big).$$

By virtue of [27, Prop. 7], $f(\cdot, D^{1:N})$ is Norkin differentiable, and

$$\hat{\partial} f(x, D^{1:N}) = \text{ conv} \left\{ s \in \mathbb{R}^n : s = \sum_{j=1}^{N} \hat{\xi}_j g_j, \text{ with} \right.$$
$$\left. \hat{\xi} \in \hat{\partial} m\big(\ell(x, D^1), \dots, \ell(x, D^N)\big) \text{ and } g_j \in \hat{\partial} \ell(x, D^j),\ j = 1, \dots, m \right\} \tag{5.6}$$

is its Norkin subdifferential mapping. Observe that $\hat{\xi} \in \hat{\partial} m(\ell(x, D^1), \dots, \ell(x, D^N))$ corresponds to the maximizers in (3.4) at $Z(D^j) = \ell(x, D^j)$, $j = 1, \dots, N$. As $\hat{\xi} \geq 0$, the convex hull operation in (5.6) is not needed, because the set in braces is convex already; for a similar argument, see [19, Thm. VI.4.3.1]. Furthermore, $\xi_j \in [0,1]$ implies that the norm of each subgradient of $f(x, D^{1:N})$ is bounded by the norm of a subgradient of $\ell(x, D^j)$ for some $j$. Hence, the assumptions of the theorem entail the following bound for every compact set $K$:

$$\sup_{x \in K} \sup_{\gamma \in \hat{\partial} f(x, D^{1:N})} \|g\| \leq \sum_{j=1}^{N} L_K(D^j).$$

The bound is integrable with respect to $D^{1:N} \sim P^N$. Observing that in (5.4) we have

$$F(x) = \mathbb{E}_{D^{1:N} \sim P^N} \left[ f(x, D^{1:N}) \right], \tag{5.7}$$

we obtain $\hat{\partial} F(x) = \mathbb{E}_{D^{1:N} \sim P^N} [\hat{\partial} f(x, D^{1:N})]$ by virtue of [26, Thm. 23.1]. This, together with (5.6), where we skip the unnecessary convex hull, results in formula (5.3). □

Due to (5.7), formula (5.6) can be directly used to obtain unbiased stochastic Norkin subgradients of the composition (5.2). First, we generate a sample $D^{1:N}$ from $P^N$. Then we calculate for each $D^j$ a subgradient $g_j$ of $\ell(x, D^j)$, $j = 1, \dots, N$. Further, we solve the maximization problem in (5.5) and obtain the vector $\hat{\xi}$; this is equivalent to the calculation of the random risk measure $\varrho[\Lambda(x), P^{(N)}]$. Finally, we construct the stochastic subgradient by setting $s = \sum_{j=1}^{N} \hat{\xi}_j g_j$; $\mathbb{E}_{D^{1:N} \sim P^N}[s] \in \hat{\partial} F(x)$ by Theorem 5.2. This allows for the application of the stochastic subgradient method of [37], which is capable of solving problems with nonconvex and nonsmooth functions that are differentiable in a generalized sense.

## 6. Dependence of risk forms on the probability measure.

We study the properties of risk forms when the measure $P$ changes and their implication for mini-batch risk forms, when the sample size changes. Throughout this section, we assume that the space $\mathfrak{D}$ is a finite-dimensional vector space with the norm $\|\cdot\|$ and that $Z$ is a fixed continuous function on $\mathfrak{D}$.

Recall the definition of the transportation distance between probability measures.[1]

For $\mu_1, \mu_2 \in \mathscr{P}(\mathfrak{D})$, we define the set of transportation plans

$$U(\mu_1, \mu_2) = \{\pi \in \mathscr{P}(\mathfrak{D} \times \mathfrak{D}) : \Pi_1 \pi = \mu_1,\ \Pi_2 \pi = \mu_2\},$$

where $\Pi_1 \pi$ and $\Pi_2 \pi$ denote the marginalizations of $\pi$ over the first and the second $\mathfrak{D}$-spaces. The *transportation distance* of order $p \in [1, \infty)$ between $\mu_1$ and $\mu_2$ is defined as

$$\mathscr{T}_p(\mu_1, \mu_2) = \inf_{\pi \in U(\mu_1, \mu_2)} \left( \int_{\mathfrak{D} \times \mathfrak{D}} \|v - w\|^p \, \pi(\mathrm{d}v\, \mathrm{d}w) \right)^{1/p}. \tag{6.1}$$

We restrict the space of probability measures on $\mathfrak{D}$ to measures which have finite moments or order $p$. We denote

$$M_p(\mu) = \int_{\mathfrak{D}} \|w\|^p \, \mu(\mathrm{d}w)$$

and consider the space

$$\mathscr{P}_p(\mathfrak{D}) = \{\mu \in \mathscr{P}(\mathfrak{D}) : M_p(\mu) < \infty\}.$$

The space $\mathscr{P}_p(\mathfrak{D})$ with the metric $\mathscr{T}_p(\cdot, \cdot)$ is a Polish space; see, e.g., [41].

This setting is very useful for our purposes, because it is known that for a finite-dimensional data space $\mathfrak{D}$ the empirical measures $P^{(N)}$ converge to $P$ in the expected distance $\mathscr{T}_p$ if $M_u(P) < \infty$ for some $u > p$. Furthermore, the expected distance can be bounded by an explicit expression involving the batch size, the dimension of the space, and the moment $M_u(P)$. The following inequality due to [11, 16] is true for all $N$:

$$\mathbb{E}\left[ \mathscr{T}_p\left(P^{(N)}, P\right) \right] \le C M_u^{p/u} \times \begin{cases} N^{-1/2} + N^{-(u-p)/u} & \text{if } p > d/2 \text{ and } u \neq 2p, \\ N^{-1/2} \ln(1+N) + N^{-(u-p)/u} & \text{if } p = d/2 \text{ and } u \neq 2p, \\ N^{-p/d} + N^{-(u-p)/u} & \text{if } p < d/2 \text{ and } u \neq \frac{d}{d-p}, \end{cases} \tag{6.2}$$

where $d = \dim(\mathfrak{D})$, $u$ is an arbitrary real number greater than $p$, and $C$ is a constant depending only on $p$, $u$, and $d$.

[1] It is called the Monge–Kantorovich, the Earth Mover's, or the Wasserstein distance; we refer the reader to the monographs [34] and [41] for an extensive exposition and historical account.

It follows that the continuity of the mapping $P \mapsto \varrho[Z,P]$ is germane to our study. We investigate this property and its implications for the analysis of the effect of the batch size $N$ in two important special cases.

### 6.1. Average Value at Risk and Kusuoka representations.

We first consider the special case of the Average Value at Risk (2.8) as the base risk form $\varrho[\,\cdot\,,\,\cdot\,]$. It has an equivalent extremal representation,

$$\operatorname{AVaR}_\alpha[Z,P] = \inf_{\eta\in\mathbb{R}} \left\{ \eta + \frac{1}{\alpha}\mathbb{E}_{D\sim P}\left[\max(0, Z(D)-\eta)\right]\right\}, \tag{6.3}$$

and the dual representation

$$\operatorname{AVaR}_\alpha[Z,P] = \sup_{Q\in\mathscr{A}(P)} \mathbb{E}_{D\sim Q}\left[Z(D)\right], \tag{6.4}$$

where

$$\mathscr{A}(P) = \left\{ Q \in \mathscr{P}(\mathfrak{D}),\ Q \ll P,\ \frac{dQ}{dP} \le \frac{1}{\alpha} \right\}. \tag{6.5}$$

If $P \in \mathscr{P}_1(\mathfrak{D})$, then $\mathscr{A}(P) \subset \mathscr{P}_1(\mathfrak{D})$, and thus $\mathscr{A}$ can be viewed as a multifunction from $\mathscr{P}_1(\mathfrak{D})$ to its subsets.

The Hausdorff distance between sets $A$ and $B$ in $\mathscr{P}_p(\mathfrak{D})$ is understood here as

$$\operatorname{dist}_p(A,B) = \max\left( \sup_{P\in A}\inf_{Q\,\in B} \mathscr{T}_p(P,Q)\,,\ \sup_{Q\in B}\inf_{P\,\in A} \mathscr{T}_p(P,Q)\right).$$

Lemma 6.1. *The multifunction* $\mathscr{A} : \mathscr{P}_1(\mathfrak{D}) \rightrightarrows \mathscr{P}_1(\mathfrak{D})$ *given by* (6.5) *is Lipschitz continuous in the Hausdorff distance* $\operatorname{dist}_1(\cdot,\cdot)$ *with the constant* $1/\alpha$:

$$\operatorname{dist}_1(\mathscr{A}(P_1), \mathscr{A}(P_2)) \le \frac{1}{\alpha}\mathscr{T}_1(P_1,P_2) \quad \forall\, P_1,P_2 \in \mathscr{P}_1(\mathfrak{D}).$$

*Proof.* Consider two probability measures $P_1, P_2 \in \mathscr{P}_1(\mathfrak{D})$, and let $\pi \in U(P_1,P_2)$ be the optimal transportation plan:[2]

$$\int_{\mathfrak{D}\times\mathfrak{D}} \|x-y\|\,\pi(\mathrm{d}x\,\mathrm{d}y) = \mathscr{T}_1(P_1,P_2). \tag{6.6}$$

It always exists; see, e.g., [41, Thm. 4.1]. We disintegrate the measure $\pi$ into the marginal $P_1$ on $\mathfrak{D}$ and a kernel $K : \mathfrak{D} \to \mathscr{P}_1(\mathfrak{D})$ (see, e.g., [4, Thm. IV.2.18]), so that

$$\pi(A\times B) = \int_A\int_B K(\mathrm{d}y|x)\,P_1(\mathrm{d}x) \quad \forall\, A,B \in \mathscr{B}(\mathfrak{D}).$$

For an arbitrary $Q_1 \in \mathscr{A}(P_1)$ we construct a measure $Q_2 \in \mathscr{P}_1(\mathfrak{D})$ by setting (see, e.g., [4, Thm. I.6.11])

$$Q_2(B) = \int_{\mathfrak{D}}\int_B K(\mathrm{d}y|x)\,Q_1(\mathrm{d}x) \quad \forall\, B \in \mathscr{B}(\mathfrak{D}). \tag{6.7}$$

[2] In some proofs in this section, to improve readability, we use $x$ and $y$ to denote elements of the space $\mathfrak{D}$.

Now we verify that $Q_2 \in \mathscr{A}(P_2)$. Let $\varphi = \frac{dQ_1}{dP_1}$ be the Radon–Nikodym derivative. From (6.7), after returning to the product measure, we obtain

$$\begin{aligned} Q_2(B) &= \int_{\mathfrak{D}}\int_B K(\mathrm{d}y|x)\,\varphi(x)\,P_1(\mathrm{d}x) = \int_{\mathfrak{D}}\int_B \varphi(x)\,\pi(\mathrm{d}x\,\mathrm{d}y)\\ &\le \frac{1}{\alpha}\int_{\mathfrak{D}}\int_B \pi(\mathrm{d}x\,\mathrm{d}y) = \frac{1}{\alpha}P_2(B) \quad \forall\, B \in \mathscr{B}(\mathfrak{D}). \end{aligned}$$

This means that $Q_2 \in \mathscr{A}(P_2)$.

Consider the transportation plan

$$\lambda(A\times B) = \int_A\int_B K(\mathrm{d}y|x)\,Q_1(\mathrm{d}x) \quad \forall\, A, B \in \mathscr{B}(\mathfrak{D}).$$

Equation (6.7) implies that $\lambda \in U(Q_1, Q_2)$. Therefore,

$$\begin{aligned} \mathscr{T}_1(Q_1,Q_2) &\le \int_{\mathfrak{D}\times\mathfrak{D}} \|x-y\|\,\lambda(\mathrm{d}x\,\mathrm{d}y) = \int_{\mathfrak{D}\times\mathfrak{D}} \|x-y\|\,\varphi(x)\,\pi(\mathrm{d}x\,\mathrm{d}y)\\ &\le \frac{1}{\alpha}\int_{\mathfrak{D}\times\mathfrak{D}} \|x-y\|\,\pi(\mathrm{d}x\,\mathrm{d}y) = \frac{1}{\alpha}\mathscr{T}_1(P_1,P_2), \end{aligned} \tag{6.8}$$

where we have used (6.6). Thus,

$$\sup_{Q_1\in\mathscr{A}(P_1)}\ \inf_{Q_2\in\mathscr{A}(P_2)} \mathscr{T}_1(Q_1,Q_2) \le \frac{1}{\alpha}\mathscr{T}_1(P_1,P_2).$$

Reversing the roles of $P_1$ and $P_2$, we obtain the assertion. ☐

Recall that a function $Z : \mathfrak{D} \to \mathbb{R}$ admits a modulus of continuity $\psi : [0,\infty) \to [0,\infty)$ if

$$|Z(v) - Z(w)| \le \psi\left(\|v-w\|\right) \quad \forall\, v, w \in \mathfrak{D},$$

where $\lim_{t\downarrow 0}\psi(t) = \psi(0) = 0$. It is well known that if $Z(\cdot)$ admits a modulus of continuity, then it also admits a nondecreasing modulus of continuity. If the modulus of continuity has an affine majorant or if the domain of $Z(\cdot)$ is convex, then a concave nondecreasing modulus of continuity exists. A good example is $\psi(t) = Lt^a$ with $a \in (0,1]$.

Theorem 6.2. (i) *If $Z(\cdot)$ is continuous and satisfies the inequality*

$$|Z(w)| \le C_Z\left(1 + \|w\|\right) \quad \forall\, w \in \mathfrak{D}, \tag{6.9}$$

*then the functional* $\mathrm{AVaR}_\alpha[\mathrm{Z},\cdot]$ *is continuous on the space* $\mathscr{P}_1(\mathfrak{D})$.
(ii) *If $Z(\cdot)$ admits a concave nondecreasing modulus of continuity $\psi(\cdot)$, then*

$$|\mathrm{AVaR}_\alpha[\mathrm{Z},\mathrm{P}_1] - \mathrm{AVaR}_\alpha[\mathrm{Z},\mathrm{P}_2]| \le \psi\left(\frac{1}{\alpha}\mathscr{T}_1(P_1,P_2)\right) \quad \forall\, P_1, P_2 \in \mathscr{P}_1(\mathfrak{D}). \tag{6.10}$$

*Proof.* Assume that $Z(\cdot)$ is continuous and satisfies the growth condition (6.9). In this case, $\mathrm{AVaR}_\alpha[Z,P] < \infty$ for all $P \in \mathscr{P}_1(\mathfrak{D})$. Let $\{P_n\}$ be a sequence of measures converging to $P$ in the space $\mathscr{P}_1(\mathfrak{D})$. Suppose $\alpha \in (0,1)$. It follows from (6.3) that

$$\mathrm{AVaR}_\alpha[Z,P] = \bar{\eta} + \frac{1}{\alpha}\mathbb{E}_{D\sim P}\left[\max(0, Z(D) - \bar{\eta})\right], \tag{6.11}$$

with $\bar{\eta}$ being an $(1-\alpha)$-quantile of $Z$ under $P$. This yields the following estimate:

$$\begin{aligned}
&\mathrm{AVaR}_\alpha[Z,P_n] - \mathrm{AVaR}_\alpha[Z,P] \\
&\quad = \min_\eta \left\{ \eta + \frac{1}{\alpha}\mathbb{E}_{D\sim P_n}\left[\max(0, Z(D)-\eta)\right] \right\} - \bar{\eta} - \frac{1}{\alpha}\mathbb{E}_{D\sim P}\left[\max(0, Z(D)-\bar{\eta})\right] \\
&\quad \le \frac{1}{\alpha}\left(\mathbb{E}_{D\sim P_n}[\max(0, Z(D)-\bar{\eta}) - \mathbb{E}_{D\sim P}[\max(0, Z(D)-\bar{\eta})\right).
\end{aligned} \tag{6.12}$$

Notice that $|\max(0, Z(D)-\bar{\eta})| \le |Z(D)-\bar{\eta}|$, and, thus, it satisfies the growth condition (6.9). Hence, the right-hand side of (6.12) converges to zero, by virtue of [41, Definition 6.4 (iv)]. We infer that

$$\limsup_{n\to\infty} \mathrm{AVaR}_\alpha[Z,P_n] \le \mathrm{AVaR}_\alpha[Z,P].$$

Let $\bar{Q} \in \mathscr{A}(P)$ be such that

$$\mathrm{AVaR}_\alpha[Z,P] = \sup_{Q\in\mathscr{A}(P)} \mathbb{E}_{D\sim Q}\left[Z(D)\right] = \mathbb{E}_{D\sim\bar{Q}}\left[Z(D)\right].$$

Using Lemma 6.1, we construct $Q_n \in \mathscr{A}(P_n)$ such that

$$\mathscr{T}_1(Q_n,\bar{Q}) \le \frac{1}{\alpha}\mathscr{T}_1(P_n,P) + \frac{1}{n}.$$

Hence, the sequence $\{Q_n\}$ converges to $\bar{Q}$ in the space $\mathscr{P}_1(\mathfrak{D})$. We obtain

$$\begin{aligned}
\mathrm{AVaR}_\alpha[Z,P_n] - \mathrm{AVaR}_\alpha[Z,P] &= \sup_{Q\in\mathscr{A}(P_n)} \mathbb{E}_{D\sim Q}\left[Z(D)\right] - \mathbb{E}_{D\sim\bar{Q}}\left[Z(D)\right] \\
&\ge \mathbb{E}_{D\sim Q_n}\left[Z(D)\right] - \mathbb{E}_{D\sim\bar{Q}}\left[Z(D)\right].
\end{aligned} \tag{6.13}$$

Letting $n\to\infty$, we obtain that the right-hand side of (6.13) converges to zero. Hence,

$$\liminf_{n\to\infty} \mathrm{AVaR}_\alpha[Z,P_n] \ge \mathrm{AVaR}_\alpha[Z,P].$$

This concludes the proof of the continuity of $\mathrm{AVaR}_\alpha[Z,\cdot\,]$ for $\alpha\in(0,1)$. If $\alpha=1$, then $\mathrm{AVaR}_1[Z,P] = \mathbb{E}_{D\sim P}[Z(D)]$, and the continuity follows directly from [41, Definition 6.4 (iv)].

We now pass to the assertion (6.10) involving a concave nondecreasing modulus of continuity. Observe that the growth condition (6.9) is satisfied in this case. Let $P_1, P_2 \in \mathscr{P}_1(\mathfrak{D})$. Let $Q_1 \in \mathscr{A}(P_1)$ be such that $\mathbb{E}_{D\sim Q_1}[Z(D)] = \mathrm{AVaR}_\alpha[Z,P_1]$.

As in Lemma 6.1, we can construct $Q_2 \in \mathscr{A}(P_2)$ such that (6.8) is satisfied. We have the estimate

$$\begin{aligned}
\mathrm{AVaR}_\alpha[Z,P_2] \ge \mathbb{E}_{D\sim Q_2}\left[Z(D)\right] &= \mathbb{E}_{D\sim Q_1}\left[Z(D)\right] + \left(\mathbb{E}_{D\sim Q_2}\left[Z(D)\right] - \mathbb{E}_{D\sim Q_1}\left[Z(D)\right]\right) \\
&= \mathrm{AVaR}_\alpha[Z,P_1] + \left(\mathbb{E}_{D\sim Q_2}\left[Z(D)\right] - \mathbb{E}_{D\sim Q_1}\left[Z(D)\right]\right).
\end{aligned} \tag{6.14}$$

Consider the case when $Z(\cdot)$ admits the modulus of continuity $\psi(\cdot)$. For any transportation plan $\lambda \in U(Q_1,Q_2)$, applying the modulus of continuity and Jensen's inequality, we obtain

$$\begin{aligned}
\left|\mathbb{E}_{D\sim Q_2}\left[Z(D)\right] - \mathbb{E}_{D\sim Q_1}\left[Z(D)\right]\right| &= \left|\int_{\mathfrak{D}\times\mathfrak{D}} \left[Z(v)-Z(w)\right]\,\lambda(\mathrm{d}v\,\mathrm{d}w)\right| \\
&\le \int_{\mathfrak{D}\times\mathfrak{D}} \psi(\|v-w\|)\,\lambda(\mathrm{d}v\,\mathrm{d}w) \le \psi\left(\int_{\mathfrak{D}\times\mathfrak{D}} \|v-w\|\,\lambda(\mathrm{d}v\,\mathrm{d}w)\right).
\end{aligned}$$

Taking the infimum over the feasible transportation plans on the right-hand side, and keeping in mind the monotonicity of $\psi(\cdot)$, we obtain

$$\left|\mathbb{E}_{D\sim Q_2}\left[Z(D)\right]-\mathbb{E}_{D\sim Q_1}\left[Z(D)\right]\right|\le\psi\left(\mathscr{T}_1(Q_1,Q_2)\right). \tag{6.15}$$

Integrating (6.14), (6.15), and (6.8), we conclude that

$$\mathrm{AVaR}_\alpha[Z,P_1]-\mathrm{AVaR}_\alpha[Z,P_2]\le\psi\left(\mathscr{T}_1(Q_1,Q_2)\right)\le\psi\left(\frac{1}{\alpha}\mathscr{T}_1(P_1,P_2)\right).$$

Exchanging the roles of $P_1$ and $P_2$, we obtain the estimate (6.10). □

Let us pass to the mini-batch risk form

$$\mathrm{AVaR}_\alpha^{(N)}[Z,P]=\mathbb{E}_{D^{1:N}\sim P^N}\left[\mathrm{AVaR}_\alpha\left[Z,P^{(N)}\right]\right].$$

It is well known that the infimum of a sample average approximation is a lower bound of the infimum in (6.3):

$$\mathrm{AVaR}_\alpha^{(N)}[Z,P]\le\mathrm{AVaR}_\alpha[Z,P].$$

Now, we can obtain an explicit bound on the bias.

COROLLARY 6.3. *If $Z(\cdot)$ admits a concave nondecreasing modulus of continuity $\psi(\cdot)$, then*

$$\mathrm{AVaR}_\alpha[Z,P]-\mathrm{AVaR}_\alpha^{(N)}[Z,P]\le\psi\left(\frac{1}{\alpha}\mathbb{E}\left[\mathscr{T}_1(P^{(N)},P)\right]\right).$$

*Furthermore, if $M_u(P)<\infty$ for some $u>1$, then for all $N\ge1$*

$$\mathrm{AVaR}_\alpha[Z,P]-\mathrm{AVaR}_\alpha^{(N)}[Z,P]\le\psi\left(\frac{\tau_1(N)}{\alpha}\right),$$

*where the constant $\tau_1(N)$ is given by the right-hand side of* (6.2).

*Proof.* According to Theorem 6.2,

$$\mathrm{AVaR}_\alpha[Z,P]-\mathrm{AVaR}_\alpha[Z,P^{(N)}]\le\psi\left(\frac{1}{\alpha}\mathscr{T}_1(P^{(N)},P)\right).$$

Taking the expected value of both sides, and using Jensen's inequality, we obtain the assertion. □

These estimates allow us to analyze law invariant risk forms that admit a Kusuoka representation

$$\varrho[Z,P]=\sup_{\lambda\in\Lambda_\varrho}\int_0^1\mathrm{AVaR}_\alpha[Z,P]\,\lambda(\mathrm{d}\alpha), \tag{6.16}$$

where $\Lambda_\varrho$ is a closed convex set of probability measures on $[0,1]$. In [9], sufficient conditions for such a representation are provided.

With every $\lambda\in\Lambda_\varrho$ in the Kusuoka representation (6.16) of $\varrho[Z,P^{(N)}]$, we associate the function

$$\varphi_\lambda(\alpha)=\int_\alpha^1\frac{\lambda(\mathrm{d}t)}{t}. \tag{6.17}$$

We observe that $\varphi_\lambda(\cdot)$ is nonnegative and nonincreasing and can be viewed as a density because

$$\int_0^1 \varphi_\lambda(\alpha)\,\mathrm{d}\alpha = \int_0^1 \int_\alpha^1 \frac{\lambda(\mathrm{d}t)}{t}\mathrm{d}\alpha = \int_0^1 \int_0^t \mathrm{d}\alpha \frac{\lambda(\mathrm{d}t)}{t} = \int_0^1 \lambda(\mathrm{d}t) = 1. \tag{6.18}$$

The second equation is obtained by changing the order of integration.

We show that the resulting mini-batch risk measure is strongly consistent for bounded continuous functions $Z(\cdot)$.

THEOREM 6.4. *If $Z(\cdot)$ is continuous and bounded, the measure $P$ has connected support, and the risk form $\varrho[\,\cdot\,,\,\cdot\,]$ admits the representation* (6.16)*, then the mini-batch risk measure $\varrho^{(N)}[Z,P]$ converges to $\varrho[Z,P]$, when $N \to \infty$.*

*Proof.* We denote the quantile function of $Z$ by $G_Z$ and the generalized inverse of $F_Z^{(N)}$ by $G_Z^{(N)}$.

Noting that $\lambda(\mathrm{d}\alpha)/\alpha = -\mathrm{d}\varphi_\lambda(\alpha)$ and integrating by parts, we transform the integral in the Kusuoka representation as follows:

$$\begin{aligned}\int_0^1 \mathrm{AVaR}_\alpha[Z,P^{(N)}]\,\lambda(\mathrm{d}\alpha) &= \int_0^1 \frac{1}{\alpha}\int_{1-\alpha}^1 G_Z(t)\,\mathrm{d}t\,\lambda(\mathrm{d}\alpha)\\ &= -\int_0^1 \int_{1-\alpha}^1 G_Z(t)\,\mathrm{d}t\,\mathrm{d}\varphi_\lambda(\alpha) = -\int_0^1 \varphi_\lambda(\alpha) G_Z(\alpha)\,\mathrm{d}\alpha.\end{aligned} \tag{6.19}$$

We compare $\varrho[Z,P^{(N)}]$ and $\varrho[Z,P]$ using (6.19) in their Kusuoka representations as follows:

$$\begin{aligned}\varrho[Z,P^{(N)}] - \varrho[Z,P] &= \sup_{\lambda\in\Lambda_\varrho}\int_0^1 \varphi_\lambda(\alpha)G_Z(\alpha)\,\mathrm{d}\alpha - \sup_{\lambda\in\Lambda_\varrho}\int_0^1 \varphi_\lambda(\alpha)G_Z^{(N)}(\alpha)\,\mathrm{d}\alpha\\ &\le \sup_{\lambda\in\Lambda_\varrho}\int_0^1 \varphi_\lambda(\alpha)\left(G_Z(\alpha) - G_Z^{(N)}(\alpha)\right)\mathrm{d}\alpha\\ &\le \left\|G_Z^{(N)}(\alpha) - G_Z(\alpha)\right\|_\infty.\end{aligned} \tag{6.20}$$

The last inequality is obtained by applying Hölder's inequality and using (6.18). The norm $\|G_Z^{(N)}(\alpha) - G_Z(\alpha)\|_\infty$ is finite for any $Z$ with bounded support. Hence,

$$\left|\varrho[Z,P^{(N)}] - \varrho[Z,P]\right| \le \left\|G_Z^{(N)}(\alpha) - G_Z(\alpha)\right\|_\infty.$$

Under the assumptions of the theorem, $\|G_Z^{(N)}(\alpha) - G_Z(\alpha)\|_\infty$ converges a.s. to zero, when $N \to \infty$ [3], which implies that $\varrho[Z,P^{(N)}]$ converges a.s. to $\varrho[Z,P]$. Additionally, our estimate implies

$$|\varrho[Z,P^{(N)}]| \le |\varrho[Z,P]| + \left\|G_Z^{(N)}(\alpha) - G_Z(\alpha)\right\|_\infty \le |\varrho[Z,P]| + 2\left\|G_Z(\alpha)\right\|_\infty.$$

Hence, $\varrho^{(N)}[Z,P]$ converges to $\varrho[Z,P]$ when $N \to \infty$ by virtue of the Lebesgue dominated convergence theorem. ∎

To obtain error estimates, stronger assumptions on the spectrum are needed. However, we may drop the boundedness condition on $Z(\cdot)$.

THEOREM 6.5. *Suppose $P \in \mathscr{P}_p(\mathfrak{D})$ with $p \in [1,\infty)$ and $Z(\cdot)$ is Lipschitz continuous with the constant $L_Z$. Furthermore, suppose the risk form $\varrho[\,\cdot\,,\cdot\,]$ admits the representation* (6.16) *such that, with some constant $C_\varrho$,*

$$\|\varphi_\lambda(\cdot)\|_q \le C_\varrho \quad \forall\, \lambda \in \Lambda_\varrho, \quad 1/p+1/q=1.$$

*If $M_u(P)<\infty$ for some $u>p$, then for all $N \ge 1$*

$$\left|\varrho^{(N)}[Z,P]-\varrho[Z,P]\right| \le C_\varrho L_Z \tau_p(N),$$

*where the constant $\tau_p(N)$ is given by the right-hand side of* (6.2).

*Proof.* We compare $\varrho[Z,P^{(N)}]$ and $\varrho[Z,P]$ using (6.20):

$$\begin{aligned}\varrho[Z,P^{(N)}]-\varrho[Z,P] &\le \sup_{\lambda\in\Lambda_\varrho}\int_0^1 \varphi_\lambda(\alpha)\left(G_Z^{(N)}(\alpha)-G_Z(\alpha)\right)\mathrm{d}\alpha\\ &\le \sup_{\lambda\in\Lambda_\varrho}\left\|G_Z^{(N)}(\cdot)-G_Z(\cdot)\right\|_p\|\varphi_\lambda(\cdot)\|_q \le C_\varrho \mathscr{T}_p\left(P_Z^{(N)},P_Z\right). \qquad (6.21)\end{aligned}$$

The penultimate inequality is obtained by applying Hölder's inequality, while the last inequality follows by the assumptions on $\Lambda_\varrho$ and the representation of the distance by quantile functions: $\mathscr{T}_p(P_Z^{(N)},P_Z)=\|G_Z^{(N)}(\alpha)-G_Z(\alpha)\|_p$; see, e.g., [31].

Suppose $\lambda$ is the optimal transportation plan between $P^{(N)}$ and $P$ on $\mathfrak{D}\times\mathfrak{D}$. Define $\pi \in \mathscr{P}(\mathbb{R}\times\mathbb{R})$ as

$$\pi(A\times B)=\lambda\left(Z^{-1}(A)\times Z^{-1}(B)\right) \quad \forall\, A,B\in\mathscr{B}(\mathbb{R}).$$

Its left marginal is

$$\pi(A\times\mathbb{R})=\lambda\left(Z^{-1}(A)\times\mathfrak{D}\right)=P^{(N)}\left(Z^{-1}(A)\right)=P_Z^{(N)}(A) \quad \forall\, A\in\mathscr{B}(\mathbb{R}).$$

The right marginal is $P_Z$, in a similar way, and thus $\pi\in U(P_Z^{(N)},P_Z)$. Consequently,

$$\begin{aligned}\mathscr{T}_p\left(P_Z^{(N)},P_Z\right)^p &\le \int_{\mathbb{R}\times\mathbb{R}}|a-b|^p\,\pi(\mathrm{d}a\,\mathrm{d}b)=\int_{\mathfrak{D}\times\mathfrak{D}}|Z(v)-Z(w)|^p\,\lambda(\mathrm{d}v\,\mathrm{d}w)\\ &\le L_Z^p\int_{\mathfrak{D}\times\mathfrak{D}}\|v-w\|^p\,\lambda(\mathrm{d}v\,\mathrm{d}w)=L_Z^p\mathscr{T}_p(P^{(N)},P)^p.\end{aligned}$$

The substitution into (6.21) yields

$$\varrho[Z,P^{(N)}]-\varrho[Z,P]\le C_\varrho L_Z\mathscr{T}_p(P^{(N)},P).$$

Reversing the roles of $P^{(N)}$ and $P$, and taking the expected value, we obtain the assertion. ∎

**6.2. Mean-semideviation risk forms.** The method applied in the previous subsection can be extended to other risk forms with the use of higher order transportation metrics.

Consider the mean-semideviation risk form of order $p\in[1,\infty)$ defined in (2.2):

$$\mathrm{msd}_p[Z,P]=\mathbb{E}_P[Z]+\varkappa\left\|\max\left(0,Z-\mathbb{E}_P[Z]\right)\right\|_p.$$

To simplify notation, we write $\mathbb{E}_P[Z]$ for $\mathbb{E}_{D\sim P}[Z(D)]$, and we use the norm $\|\cdot\|_p$ in the space $\mathscr{L}_p(\mathfrak{D},\mathscr{B}(\mathfrak{D}),P)$. The dual representation of $\mathrm{msd}_p[\,\cdot\,,P]$ has the following form, with $1/p+1/q=1$:

$$\mathscr{A}(P)=\left\{Q\in\mathscr{P}(\mathfrak{D}):Q\ll P,\ \frac{dQ}{dP}=1+\zeta-\mathbb{E}_P[\zeta],\ \|\zeta\|_q\le\varkappa,\ \zeta\ge 0\right\}. \tag{6.22}$$

Hölder's inequality implies that $\mathscr{A}(P)\subset\mathscr{P}_1(\mathfrak{D})$ whenever $P\in\mathscr{P}_p(\mathfrak{D})$, and thus we can view $\mathscr{A}$ as a multifunction from $\mathscr{P}_p(\mathfrak{D})$ to $\mathscr{P}_1(\mathfrak{D})$.

Lemma 6.6. *The multifunction $\mathscr{A}:\mathscr{P}_p(\mathfrak{D})\rightrightarrows\mathscr{P}_1(\mathfrak{D})$ given by* (6.22) *satisfies the inequality*

$$\mathrm{dist}_1\left(\mathscr{A}(P_1),\mathscr{A}(P_2)\right)\le(1+\varkappa)\mathscr{T}_p(P_1,P_2)\quad\forall\,P_1,P_2\in\mathscr{P}_p(\mathfrak{D}). \tag{6.23}$$

*Proof.* Consider two probability measures $P_1,P_2\in\mathscr{P}_p(\mathfrak{D})$, and let $\pi\in U(P_1,P_2)$ be the optimal transportation plan:

$$\left(\int_{\mathfrak{D}\times\mathfrak{D}}\|x-y\|^p\,\pi(\mathrm{d}x\,\mathrm{d}y)\right)^{1/p}=\mathscr{T}_p(P_1,P_2). \tag{6.24}$$

Again, we disintegrate the measure $\pi$ into the marginal $P_1$ on $\mathfrak{D}$ and a kernel $K:\mathfrak{D}\to\mathscr{P}_p(\mathfrak{D})$:

$$\pi(A\times B)=\int_A\int_B K(\mathrm{d}y|x)\,P_1(\mathrm{d}x)\quad\forall\,A,B\in\mathscr{B}(\mathfrak{D}).$$

For an arbitrary $Q_1\in\mathscr{A}(P_1)$, we construct a measure $Q_2\in\mathscr{P}(\mathfrak{D})$ by formula (6.7). It follows from (6.22) that $\frac{dQ_1}{dP_1}=1+\zeta-\mathbb{E}_{P_1}[\zeta]$ for some $\zeta\ge 0$ integrable in the $q$th power with respect to $P_1$. Substituting into (6.7), we obtain

$$\begin{aligned}Q_2(B)&=\int_{\mathfrak{D}}\int_B K(\mathrm{d}y|x)\ \left(1+\zeta(x)-\mathbb{E}_{P_1}[\zeta]\right)\ P_1(\mathrm{d}x)\\&=P_2(B)\left(1-\mathbb{E}_{P_1}[\zeta]\right)+\int_{\mathfrak{D}}\int_B K(\mathrm{d}y|x)\,\zeta(x)\,P_1(\mathrm{d}x)\quad\forall\,B\in\mathscr{B}(\mathfrak{D}).\end{aligned} \tag{6.25}$$

Consider the measure defined by the third term in (6.25):

$$\mathit{\Gamma}(B)=\int_{\mathfrak{D}}\int_B K(\mathrm{d}y|x)\,\zeta(x)\,P_1(\mathrm{d}x)=\int_{\mathfrak{D}\times B}\zeta(x)\,\pi(\mathrm{d}x\,\mathrm{d}y)\quad\forall\,B\in\mathscr{B}(\mathfrak{D}).$$

As it is absolutely continuous with respect to $P_2$, it has a density $\gamma=\frac{\mathrm{d}\mathit{\Gamma}}{\mathrm{d}P_2}$ by the Radon–Nikodym theorem. Hence, for any function $g\in\mathscr{L}_p(\mathfrak{D},\mathscr{B}(\mathfrak{D}),P_2)$, we can write

$$\int_{\mathfrak{D}}g(y)\,\gamma(y)\,P_2(\mathrm{d}y)=\int_{\mathfrak{D}\times\mathfrak{D}}g(y)\zeta(x)\,\pi(\mathrm{d}x\,\mathrm{d}y). \tag{6.26}$$

Observe that we may formally regard both $g$ and $\zeta$ as members of $\mathscr{L}_p(\mathfrak{D}\times\mathfrak{D},\mathscr{B}(\mathfrak{D}\times\mathfrak{D}),\pi)$ and $\mathscr{L}_q(\mathfrak{D}\times\mathfrak{D},\mathscr{B}(\mathfrak{D}\times\mathfrak{D}),\pi)$, respectively. We apply Hölder's inequality to the right-hand side of (6.26) to obtain

$$\int_{\mathfrak{D}}g(y)\,\gamma(y)\,P_2(\mathrm{d}y)\le\|g\|_p\,\|\zeta\|_q,$$

with the norms in the said spaces. However, since $g(\cdot)$ is a function of the second variable only and $P_2$ is the second marginal of $\pi$, $\|g\|_p$ is the same in the space $\mathscr{L}_p(\mathfrak{D},\mathscr{B}(\mathfrak{D}),P_2)$. In a similar way, $\|\zeta\|_q$ is the same in $\mathscr{L}_q(\mathfrak{D},\mathscr{B}(\mathfrak{D}),P_1)$. It follows from the last displayed inequality that

$$\|\gamma\|_q \le \|\zeta\|_q \le \varkappa, \tag{6.27}$$

with the norm of $\gamma$ in $\mathscr{L}_q(\mathfrak{D},\mathscr{B}(\mathfrak{D}),P_2)$ and the norm of $\zeta$ in $\mathscr{L}_q(\mathfrak{D},\mathscr{B}(\mathfrak{D}),P_1)$. Furthermore, (6.26) with $g \equiv 1$ implies that

$$\mathbb{E}_{P_2}[\gamma] = \int_{\mathfrak{D}} \gamma(y)\, P_2(\mathrm{d}y) = \int_{\mathfrak{D}} \zeta(x)\, P_1(\mathrm{d}x) = \mathbb{E}_{P_1}[\zeta]. \tag{6.28}$$

Combining (6.25), (6.27), and (6.28), we conclude that

$$\frac{\mathrm{d}Q_2}{\mathrm{d}P_2} = 1 - \mathbb{E}_{P_2}[\gamma] + \gamma, \quad \|\gamma\|_q \le \varkappa, \quad \gamma \ge 0.$$

This means that $Q_2 \in \mathscr{A}(P_2)$.

To estimate the distance $\mathscr{T}_1(Q_1,Q_2)$ we consider the transportation plan

$$\lambda(A\times B) = \int_A \int_B K(\mathrm{d}y|x)\, Q_1(\mathrm{d}x) \quad \forall\, A,B \in \mathscr{B}(\mathfrak{D}).$$

By (6.7), $\lambda \in U(Q_1,Q_2)$. Therefore, by Hölder's inequality in the spaces $\mathscr{L}_p(\mathfrak{D}\times\mathfrak{D},\mathscr{B}(\mathfrak{D},\times\mathfrak{D}),\pi)$ and $\mathscr{L}_q(\mathfrak{D}\times\mathfrak{D},\mathscr{B}(\mathfrak{D},\times\mathfrak{D}),\pi)$, and by (6.24), we obtain the chain of relations

$$\begin{aligned}\mathscr{T}_1(Q_1,Q_2) &\le \int_{\mathfrak{D}\times\mathfrak{D}} \|x-y\|\,\lambda(\mathrm{d}x\,\mathrm{d}y)\\ &= \int_{\mathfrak{D}\times\mathfrak{D}} \|x-y\|\,(1+\zeta(x)-\mathbb{E}_{P_1}[\zeta])\,\pi(\mathrm{d}x\,\mathrm{d}y)\\ &\le \mathscr{T}_p(P_1,P_2)\,\|1+\zeta-\mathbb{E}_{P_1}[\zeta]\|_q\\ &\le \mathscr{T}_p(P_1,P_2)(1+\varkappa).\end{aligned} \tag{6.29}$$

In the last inequality, we used the fact that $0 \le \mathbb{E}_{P_1}[\zeta] \le \|\zeta\|_q \le \varkappa$. Thus,

$$\sup_{Q_1\in\mathscr{A}(P_1)}\ \inf_{Q_2\in\mathscr{A}(P_2)} \mathscr{T}_1(Q_1,Q_2) \le (1+\varkappa)\mathscr{T}_p(P_1,P_2).$$

Reversing the roles of $P_1$ and $P_2$, we obtain (6.23). □

This allows us to obtain an estimate of the difference between the mini-batch risk measure and its "mother" measure for any batch size $N$.

Corollary 6.7. *If $P \in \mathscr{P}_p(\mathfrak{D})$ and $Z(\cdot)$ admits a concave nondecreasing modulus of continuity $\psi(\cdot)$, then*

$$\left|\mathrm{msd_p[Z,P]} - \mathrm{msd_p^{(N)}[Z,P]}\right| \le \psi\left((1+\varkappa)\,\mathbb{E}\left[\mathscr{T}_p(P^{(N)},P)\right]\right).$$

*Furthermore, if $M_u(P) < \infty$ for some $u > p$, then for all $N \ge 1$*

$$\left|\mathrm{msd_p[Z,P]} - \mathrm{msd_p^{(N)}[Z,P]}\right| \le \psi\left((1+\varkappa)\,\tau_p(N)\right),$$

*where the constant $\tau_p(N)$ is given by the right-hand side of* (6.2).

## 7. Conclusions.

While the base risk forms are defined for bounded functions on the data space $\mathfrak{D}$ and probability measures on this space, the mini-batch risk forms are well defined for all integrable functions once the probability measure is fixed. They have the potential for applications in risk-averse decision, control, and learning, due to the fact that they allow for the construction of unbiased estimates of their subgradients, and generalized gradients of their compositions with random loss functions. Furthermore, restricting the data space to a finite-dimensional space allows for the use of transportation metrics in the spaces of probability measures and leads to quantitative estimates of the difference between the mini-batch risk evaluation and the base risk measure.

## REFERENCES

[1] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, *Coherent measures of risk*, Math. Finance, 9 (1999), pp. 203–228.

[2] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser Boston, Boston, MA, 2009.

[3] J. M. Bogoya, A. Böttcher, and E. A. Maximenko, *From convergence in distribution to uniform convergence*, Bol. Soc. Mat. Mex., 22 (2016), pp. 695–710.

[4] E. Çinlar, *Probability and Stochastics*, Springer, New York, 2011.

[5] F. H. Clarke, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.

[6] D. Dentcheva, Y. Lin, and S. Penev, *Stability and sample-based approximations of composite stochastic optimization problems*, Oper. Res., (2022), https://doi.org/10.1287/opre.2022.2308.

[7] D. Dentcheva, S. Penev, and A. Ruszczyński, *Statistical estimation of composite risk functionals and risk optimization problems*, Ann. Inst. Stat. Math., 69 (2017), pp. 737–760.

[8] D. Dentcheva and A. Ruszczyński, *Common mathematical foundations of expected utility and dual utility theories*, SIAM J. Optim., 23 (2013), pp. 381–405, https://doi.org/10.1137/120868311.

[9] D. Dentcheva and A. Ruszczyński, *Risk forms: Representation, disintegration, and application to partially observable two-stage systems*, Math. Program., 181 (2020), pp. 297–317.

[10] D. Dentcheva and A. Ruszczyński, *Subregular recourse in nonlinear multistage stochastic optimization*, Math. Program., 189 (2021), pp. 249–270.

[11] S. Dereich, M. Scheutzow, and R. Schottstedt, *Constructive quantization: approximation by empirical measures*, Ann. Inst. Henri Poincaré'Probab. Stat., 49 (2013), pp. 1183–1203.

[12] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis, *Curves of descent*, SIAM J. Control Optim., 53 (2015), pp. 114–138, https://doi.org/10.1137/130920216.

[13] J. Fan and A. Ruszczyński, *Risk measurement and risk-averse control of partially observable discrete-time Markov systems*, Math. Methods Oper. Res., 88 (2018), pp. 161–184.

[14] H. Föllmer and A. Schied, *Convex measures of risk and trading constraints*, Finance Stoch., 6 (2002), pp. 429–447.

[15] H. Föllmer and A. Schied, *Stochastic Finance*, de Gruyter, Berlin, 2011.

[16] N. Fournier and A. Guillin, *On the rate of convergence in Wasserstein distance of the empirical measure*, Probab. Theory Relat. Fields, 162 (2015), pp. 707–738.

[17] S. Ghadimi, A. Ruszczyński, and M. Wang, *A single timescale stochastic approximation method for nested stochastic optimization*, SIAM J. Optim., 30 (2020), pp. 960–979, https://doi.org/10.1137/18M1230542.

[18] M. Gürbüzbalaban, A. Ruszczyński, and L. Zhu, *A stochastic subgradient method for distributionally robust non-convex and non-smooth learning*, J. Optim. Theory Appl., 194 (2022), pp. 1014–1041.

[19] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms I: Fundamentals*, Grundlehren Math. Wiss. 305, Springer-Verlag, Berlin, 1993.

[20] D. S. Kalogerias, *Fast and stable convergence of online SGD for CVaR-based risk-aware learning*, in Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Washington, DC, 2022, pp. 6007–6011.

[21] M. Kijima and M. Ohnishi, *Mean-risk analysis of risk aversion and wealth effects on optimal portfolios with multiple investment opportunities*, Ann. Oper. Res., 45 (1993), pp. 147–163.

[22] U. Köse and A. Ruszczyński, *Risk-averse learning by temporal difference methods with Markov risk measures*, J. Mach. Learn. Res., 22 (2021), pp. 1–34.

[23] K. Kuratowski and C. Ryll-Nardzewski, *A general theorem on selectors*, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 13 (1965), pp. 397–403.

[24] S. Kusuoka, *On law-invariant coherent risk measures*, in Advances in Mathematical Economics, Vol. 3, S. Kusuoka and T. Maruyama, eds., Springer, Tokyo, 2001, pp. 83–95.

[25] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972, https://doi.org/10.1137/0315061.

[26] V. S. Mikhalevich, A. M. Gupal, and V. I. Norkin, *Nonconvex Optimization Methods*, Nauka, Moscow, 1987.

[27] V. I. Norkin, *Generalized-differentiable functions*, Cybernet. Syst. Anal., 16 (1980), pp. 10–12.

[28] W. Ogryczak and A. Ruszczyński, *From stochastic dominance to mean-risk models: Semideviations as risk measures*, European J. Oper. Res., 116 (1999), pp. 33–50.

[29] W. Ogryczak and A. Ruszczyński, *On consistency of stochastic dominance and mean-semideviation models*, Math. Program., 89 (2001), pp. 217–232.

[30] W. Ogryczak and A. Ruszczyński, *Dual stochastic dominance and related mean-risk models*, SIAM J. Optim., 13 (2002), pp. 60–78, https://doi.org/10.1137/S1052623400375075.

[31] V. M. Panaretos and Y. Zemel, *Statistical aspects of Wasserstein distances*, Ann. Rev. Stat. Appl., 6 (2019), pp. 405–431.

[32] G. Ch Pflug and W. Römisch, *Modeling, Measuring and Managing Risk*, World Scientific, Singapore, 2007.

[33] J. Quiggin, *A theory of anticipated utility*, J. Econ. Behav. Organ., 3 (1982), pp. 323–343.

[34] S. T. Rachev and L. Rüschendorf, *Mass Transportation Problems: Volume I: Theory*, Springer, New York, 1998.

[35] W. Römisch, *Stability of stochastic programming problems*, in Stochastic Programming, Handbooks Oper. Res. Management Sci. 10, A. Ruszczyński and A. Shapiro, eds., Elsevier, Amsterdam, 2003, pp. 483–554.

[36] A. Ruszczyński, *Risk-averse dynamic programming for Markov decision processes*, Math. Program., 125 (2010), pp. 235–261.

[37] A. Ruszczyński, *Convergence of a stochastic subgradient method with averaging for nonsmooth nonconvex constrained optimization*, Optim. Lett., 14 (2020), pp. 1615–1625.

[38] A. Ruszczyński, *A stochastic subgradient method for nonsmooth nonconvex multilevel composition optimization*, SIAM J. Control Optim., 59 (2021), pp. 2301–2320, https://doi.org/10.1137/20M1312952.

[39] A. Ruszczyński and A. Shapiro, *Optimization of convex risk functions*, Math. Oper. Res., 31 (2006), pp. 433–542.

[40] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, 3rd ed., MOS-SIAM Ser. Optim. 28, SIAM, Philadelphia, 2021, https://doi.org/10.1137/1.9781611976595.

[41] C. Villani, *Optimal Transport: Old and New*, Springer, Berlin, 2009.

[42] M. E. Yaari, *The dual theory of choice under risk*, Econometrica, 55 (1987), pp. 95–115.

[43] S. Yitzhaki, *Stochastic dominance, mean variance, and Gini's mean difference*, Amer. Econ. Rev., 72 (1982), pp. 178–185.