Distributionally Robust Optimisation (DRO), and Risk Estimation with Wasserstein distances

Jan Obłój
Mathematical Institute
University of Oxford

joint works with
Daniel Bartl, Samuel Drapeau, Johannes Wiesel

CREST Doctoral Course
ENSAE Paris, May 2025

Oxford Mathematics

St John's College
Oxford

University of Oxford
Mathematical Institute

European Research Council
Established by the European Commission

Model's neighbourhoods & Wasserstein distances

Model neighbourhood

Measure μ (or $\mathbb{P}$) will denote a model, such as

- $\mu = \hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{x^i}$ is the empirical measure of the observations/test set.
- μ comes from a mathematical modelling effort, e.g., an SDE;

Model neighbourhood

Measure μ (or $\mathbb{P}$) will denote a model, such as

- $\mu = \hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{x^i}$ is the empirical measure of the observations/test set.
- μ comes from a mathematical modelling effort, e.g., an SDE;

There are MANY ways to build a neighbourhood $B_\delta(\mu)$ of μ:

- data perturbation
- support estimates
- moments contraints
- density constraints
- Prokhorov distance
- Hellinger distance
- Kullback–Leibler divergence/entropy bounds
- and more...

Wasserstein distance

For $p \geq 1$, $\mu, \nu \in \mathcal{P}(\mathcal{S})$ with p^{th} moments, set

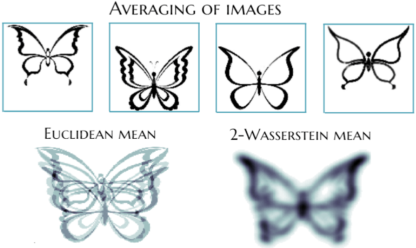
$$W_p(\mu, \nu) = \inf \left\{ \int_{\mathcal{S}\times\mathcal{S}} d(x, y)^p \, \pi(dx, dy) \colon \pi \in \mathrm{Cpl}(\mu, \nu) \right\}^{1/p},$$

where $\mathrm{Cpl}(\mu, \nu) = \{\pi : \pi(\cdot \times \mathcal{S}) = \mu \text{ and } \pi(\mathcal{S} \times \cdot) = \nu\}$.

$$\text{metric } d \text{ on } \mathcal{S} \quad \Longrightarrow \quad \text{metric } W \text{ on } \mathcal{P}(\mathcal{S})$$

Observe historical returns $r^1, \ldots, r^N$ assumed to follow a time-homogeneous ergodic Markov chain on $\mathbb{R}^d$ with an invariant distribution μ. Should we work with

the data points $(r^i)_{i=1}^N$ or the empirical measure $\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^N \delta_{r^i}$?



Source: J. Ebert, V. Spokoiny, A. Suvorikova arXiv:1703.03658

Wasserstein vs Euclidean mean (MNIST data)

Wasserstein vs Euclidean mean (MNIST data)

Small uncertainty limit

UNIVERSITY OF OXFORD
Mathematical Institute

Key property: $\hat{\mu}_N \xrightarrow{W_p} \mu$ + cnv rates, see Fournier & Guillin '14

Esfahani & Kuhn '18 argue that using Wasserstein balls gives

- finite sample guarantees,
- asymptotic consistency,
- tractability (see also Eckstein & Kupper '19)

Large uncertainty limit

Pflug, Pichler & Wozabal '12 use Wasserstein balls for robust portfolio selection:

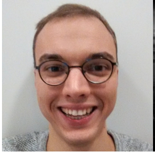
$$\inf_{a:\langle a,\mathbf{1}\rangle=1}\ \sup_{\nu\in B_\delta(\mu)}\Big(\mathbb{E}_\nu[\langle a,R\rangle]+\gamma\mathrm{Var}_\nu[\langle a,R\rangle]\Big)$$

and show that

$$a^*(\delta)\xrightarrow{\delta\to\infty}\left(\frac{1}{N},\dots,\frac{1}{N}\right)$$

which may not be true for weaker or stronger metrics.

OT & Distributionally Robust Optimization



based on Bartl, Drapeau, O. and Wiesel, *Proc. R. Soc. A* 477: 20210176, 2021
O. and Wiesel, *Math. Finance* 31(4): 1454–1493, 2021.

Problem Setting

Consider the following optimisation problem

$$V = \inf_{a \in \mathcal{A}} \int_{\mathcal{S}} f(a, x)\mu(dx),$$

where $\mathcal{A}$ is the set of controls, $\mathcal{S}$ is the state space and μ is the model.

Consider the following optimisation problem

$$V = \inf_{a \in \mathcal{A}} \int_{\mathcal{S}} f(a,x)\mu(dx),$$

where $\mathcal{A}$ is the set of controls, $\mathcal{S}$ is the state space and μ is the model.
Examples:

- risk neutral pricing: $\mathbb{E}_{\mathbb{Q}}[f(S_T)]$,
- optimal investment: $\inf_{a \in \mathcal{A}} \mathbb{E}_{\mathbb{P}}[-U(x + \langle a, S_T - S_0 \rangle)]$,
- optimised certainty equivalents: $\inf_{a \in \mathbb{R}} \mathbb{E}_{\mathbb{P}}[a - U(X + a)]$
- marginal utility pricing (Davis' price)...

Consider the following optimisation problem

$$V = \inf_{a \in \mathcal{A}} \int_{\mathcal{S}} f(a, x) \mu(dx),$$

where $\mathcal{A}$ is the set of controls, $\mathcal{S}$ is the state space and μ is the model.

Examples:

- risk neutral pricing: $\mathbb{E}_{\mathbb{Q}}[f(S_T)]$,
- optimal investment: $\inf_{a \in \mathcal{A}} \mathbb{E}_{\mathbb{P}}[-U(x + \langle a, S_T - S_0 \rangle)]$,
- optimised certainty equivalents: $\inf_{a \in \mathbb{R}} \mathbb{E}_{\mathbb{P}}[a - U(X + a)]$
- marginal utility pricing (Davis' price)...
- OLS regression: $\inf_{a \in \mathbb{R}^d} \frac{1}{N} \sum_{i=1}^N (y^i - \langle a, x^i \rangle)^2$,
- ML/NN: $\inf \frac{1}{N} \sum_{i=1}^N |y^i - ((A_2(\cdot) + b_2) \circ \sigma \circ (A_1(\cdot) + b_1))(x^i)|^p$ over $a = (A_1, A_2, b_1, b_2) \in \mathcal{A} = \mathbb{R}^{k \times d} \times \mathbb{R}^{d \times k} \times \mathbb{R}^k \times \mathbb{R}^d$, where $(x^i, y^i)_{i=1}^N$ is the training set.
-

Given our optimisation problem

$$V = \inf_{a \in \mathcal{A}} \int_{\mathcal{S}} f(a,x)\mu(dx),$$

we want to understand its dependence on the "model" μ.

We are interested in computing

$$\frac{\partial V}{\partial \mu} \quad \text{– the uncertainty sensitivity of the problem}$$

- parametric programming and statistical inference
  see Armacost & Fiacco '76 ... Bonnans & Shapiro '13;
- qualitative/quantitative stability in μ
  see Dupačová '90, Römisch '03
- robust optimisation
  see Bertsimas, Gupta & Kallus '18

Distributionally Robust Optimisation (DRO) considers

$$V(\delta) = \inf_{a \in \mathcal{A}} \sup_{\nu \in B_\delta(\mu)} \int_{\mathcal{S}} f(a, x)\nu(dx),$$

see Scarf '58, ... ,Rahimian & Mehrotra '19, where

$B_\delta(\mu)$ is a δ–neighbourhood of the model μ.

Distributionally Robust Optimisation (DRO) considers

$$V(\delta) = \inf_{a \in \mathcal{A}} \sup_{\nu \in B_\delta(\mu)} \int_{\mathcal{S}} f(a, x)\nu(dx),$$

see SCARF '58, ... ,RAHIMIAN & MEHROTRA '19, where

$B_\delta(\mu)$ is a δ–neighbourhood of the model μ.

We propose to compute

$$\Upsilon := V'(0) = \lim_{\delta \searrow 0} \frac{V(\delta) - V(0)}{\delta} \quad \text{and} \quad \beth := \lim_{\delta \searrow 0} \frac{a^*(\delta) - a^*(0)}{\delta},$$

with $B_\delta(\mu)$ being Wasserstein balls around μ.

Υ *the sensitivity of the value w.r.t.* $\Upsilon\pi o\delta\varepsilon\gamma\mu\alpha$, *the Model.*

$\beth$ *the sensitivity of* בקרה, the control, w.r.t. the Model.

Aside on convex duality

UNIVERSITY OF OXFORD
Mathematical Institute

Let E be a normed vector space. For $\Theta : E \to \mathbb{R} \cup \{+\infty\}$ convex, we consider

$$\Theta^*(p) := \sup_{x \in E} \left[\langle p, z \rangle - \Theta(z) \right], \quad p \in E^*.$$

Theorem (F-R duality)

Let Θ, Ξ be two convex functions on E, s.t. $\exists x_0 \in E$, $\Theta(x_0) < \infty$, $\Xi(x_0) < \infty$ and Θ continuous at x_0. Then

$$\inf_{x \in E} \left(\Theta(x) + \Xi(x) \right) = \max_{p \in E^*} \left(-\Theta^*(-p) - \Xi^*(p) \right).$$

$$I = \sup_{\pi \in \Phi_{\mu,\delta}} \int f(y)\pi(dx,dy), \quad \Phi_{\mu,\delta} = \left\{ \pi \in \bigcup_{\nu \in \mathcal{P}(S)} \Pi(\mu,\nu) : \int c d\pi \leq \delta \right\}$$

$$J = \inf \left\{ \underbrace{\lambda\delta + \int \phi d\mu}_{J(\lambda,\phi)} : \underbrace{\lambda \geq 0, \phi(x) + \lambda c(x,y) \geq f(y)}_{\Lambda_{c,f}} \right\}$$

Theorem (Blanchet & Murthy '19)

Let S be Polish, $c \geq$ LSC and $c(x,y) = 0$ iff $x = y$, $\mu \in \mathcal{P}(S)$, $f \in L^1(\mu)$ and USC. Then

$$I = J = \inf_{\lambda \geq 0} \left\{ \lambda\delta + \int \phi_\lambda d\mu \right\},$$

with J attained and where

$$\phi_\lambda(x) := \sup_{y \in S} \{ f(y) - \lambda c(x,y) \}.$$

Regularized optimization

University of Oxford
Mathematical Institute

- Square-root LASSO: Take $c = \|\cdot\|_q^2$.

$$\inf_{\beta \in \mathbb{R}^d} \sup_{d(\mu,\nu) \le \delta} \mathbb{E}_\nu[(y - \beta^\intercal x)^2] = \inf_{\beta \in \mathbb{R}^d} \left\{ \sqrt{\mathbb{E}_\mu[(y - \beta^\intercal x)^2]} + \sqrt{\delta}\|\beta\|_p \right\}^2.$$

- Regularised logistic regression: Take $c = \|\cdot\|_q$.

$$\inf_{\beta \in \mathbb{R}^d} \sup_{d(\mu,\nu) \le \delta} \mathbb{E}_\nu[\log(1 + e^{-Y\beta^\intercal X})] = \inf_{\beta \in \mathbb{R}^d} \{\mathbb{E}_\mu[\log(1 + e^{-Y\beta^\intercal X})] + \delta\|\beta\|_p\}.$$

Regularized optimization

- Square-root LASSO: Take $c = \|\cdot\|_q^2$.

$$\inf_{\beta\in\mathbb{R}^d} \sup_{d(\mu,\nu)\le\delta} \mathbb{E}_\nu[(y-\beta^\intercal x)^2] = \inf_{\beta\in\mathbb{R}^d}\left\{\sqrt{\mathbb{E}_\mu[(y-\beta^\intercal x)^2]} + \sqrt{\delta}\|\beta\|_p\right\}^2.$$

- Regularised logistic regression: Take $c = \|\cdot\|_q$.

$$\inf_{\beta\in\mathbb{R}^d} \sup_{d(\mu,\nu)\le\delta} \mathbb{E}_\nu[\log(1+e^{-Y\beta^\intercal X})] = \inf_{\beta\in\mathbb{R}^d}\{\mathbb{E}_\mu[\log(1+e^{-Y\beta^\intercal X})] + \delta\|\beta\|_p\}.$$

- Distributionally robust average value-at-risk: Take $c = |\cdot|$.

$$\mathrm{AVaR}_\alpha = \sup_{\frac{d\eta}{d\mu}\le\alpha^{-1}} \mathbb{E}_\eta[X], \qquad \textbf{AVaR}_\alpha = \sup_{\frac{d\eta}{d\nu}\le\alpha^{-1},\, d(\mu,\nu)\le\delta} \mathbb{E}_\eta[X].$$

$$\textbf{AVaR}_\alpha = \mathrm{AVaR}_\alpha + \delta\alpha^{-1}.$$

Main results

Part I: Sensitivity of the value function

Uncertainty Sensitivity of DRO problems

Mathematical Institute

Recall our DRO problem (for simplicity $\mathcal{A} = \mathbb{R}^k$, $\mathcal{S} = \mathbb{R}^d$)

$$V(\delta) = \inf_{a \in \mathbb{R}^k} \sup_{\nu \in B_\delta(\mu)} \int_{\mathbb{R}^d} f(x, a)\, \nu(dx).$$

Uncertainty Sensitivity of DRO problems

Recall our DRO problem (for simplicity $\mathcal{A} = \mathbb{R}^k$, $\mathcal{S} = \mathbb{R}^d$)

$$V(\delta) = \inf_{a \in \mathbb{R}^k} \sup_{\nu \in B_\delta(\mu)} \int_{\mathbb{R}^d} f(x, a)\, \nu(dx).$$

Theorem

For $p > 1$, $\frac{1}{p} + \frac{1}{q} = 1$, and under suitable assumptions, we have

$$\Upsilon := V'(0) = \lim_{\delta \to 0} \frac{V(\delta) - V(0)}{\delta} = \inf_{a^* \in A^{\mathrm{opt}}(0)} \left(\int_{\mathbb{R}^d} |\nabla_x f(x, a^*)|^q \, \mu(dx) \right)^{1/q},$$

where $A^{\mathrm{opt}}(\delta)$ denotes the set of optimisers for $V(\delta)$.

Υ: uncertainty sensitivity of the value function

We can restate the result as

$$\inf_{a\in\mathbb{R}^k} \sup_{\nu\in B_\delta(\mu)} \int_{\mathbb{R}^d} f(x,a)\,\nu(dx) \approx \inf_{a\in\mathbb{R}^k} \int_{\mathbb{R}^d} f(x,a)\,\mu(dx) + \Upsilon\delta + o(\delta)$$

where

$$\Upsilon = \inf_{a^*\in A^{\mathrm{opt}}(0)} \left(\int_{\mathbb{R}^d} |\nabla_x f(x,a^*)|^q\,\mu(dx) \right)^{1/q}.$$

Υ: uncertainty sensitivity of the value function

We can restate the result as

$$\inf_{a\in\mathbb{R}^k} \sup_{\nu\in B_\delta(\mu)} \int_{\mathbb{R}^d} f(x,a)\,\nu(dx) \approx \inf_{a\in\mathbb{R}^k} \int_{\mathbb{R}^d} f(x,a)\,\mu(dx) + \Upsilon\delta + o(\delta)$$

where

$$\Upsilon = \inf_{a^*\in A^{\mathrm{opt}}(0)} \left(\int_{\mathbb{R}^d} |\nabla_x f(x,a^*)|^q\,\mu(dx) \right)^{1/q}.$$

- extends to DRO problems with linear constraints, e.g., martingale;
- extends to general semi-norms;
- extends to sensitivity at a fixed $\delta > 0$: $V'(\delta+)$;
- no first order loss from using $a^*(0)$ instead of $a^*(\delta)$.

Sketch of the proof (1)

Mathematical Institute

Sensitivity of the value function: "$\leq$"

$$V(\delta) - V(0) \leq \sup_{\pi \in C_\delta(\mu)} \int f(y, a^*) - f(x, a^*)\, \pi(dx, dy)$$

$$= \sup_{\pi \in C_\delta(\mu)} \int \int_0^1 \langle \nabla_x f(x + t(y-x), a^*), (y-x) \rangle \, dt \, \pi(dx, dy)$$

$$\leq \delta \sup_{\pi \in C_\delta(\mu)} \int_0^1 \Big(\int |\nabla_x f(x + t(y-x), a^*)|^q \pi(dx, dy) \Big)^{1/q} dt.$$

+ growth conditions + DCT.

Sketch of the proof (2)

Sensitivity of the value function: "$\geq$"

$$T(x) := \frac{\nabla_x f(x,a^*)}{|\nabla_x f(x,a^*)|^{2-q}} \left(\int |\nabla_x f(z,a^*)|^q \, \mu(dz) \right)^{1/q-1}$$

$$\pi^\delta := [x \mapsto (x, x + \delta T(x))]_{\#}\mu \in C_\delta(\mu)$$

We can use π^δ to get a lower bound:

$$\begin{aligned}
\frac{V(\delta) - V(0)}{\delta} &\geq \frac{1}{\delta} \int f(x + \delta T(x), a^\delta) - f(x, a^\delta) \, \mu(dx) \\
&= \int \int_0^1 \langle \nabla_x f(x + t\delta T(x), a^\delta), T(x) \rangle \, dt \, \mu(dx) \\
&\xrightarrow{\delta \to 0} \int \langle \nabla_x f(x, a^*), T(x) \rangle \, \mu(dx) = \left(\int |\nabla_x f(x, a^*)|^q \, \mu(dx) \right)^{1/q}.
\end{aligned}$$

Sketch of the proof (2)

Sensitivity of the value function: "$\geq$"

$$T(x) := \frac{\nabla_x f(x, a^*)}{|\nabla_x f(x, a^*)|^{2-q}} \left(\int |\nabla_x f(z, a^*)|^q \, \mu(dz) \right)^{1/q-1}$$

$$\pi^\delta := [x \mapsto (x, x + \delta T(x))]_{\#}\mu \in C_\delta(\mu)$$

We can use π^δ to get a lower bound:

$$\begin{aligned}
\frac{V(\delta) - V(0)}{\delta} &\geq \frac{1}{\delta} \int f(x + \delta T(x), a^\delta) - f(x, a^\delta) \, \mu(dx) \\
&= \int \int_0^1 \langle \nabla_x f(x + t\delta T(x), a^\delta), T(x) \rangle \, dt \, \mu(dx) \\
&\xrightarrow{\delta \to 0} \int \langle \nabla_x f(x, a^*), T(x) \rangle \, \mu(dx) = \left(\int |\nabla_x f(x, a^*)|^q \, \mu(dx) \right)^{1/q}.
\end{aligned}$$

Sensitivity of the optimisers: similar but more involved + Langrange multipliers + min-max

Example 1: AV@R minimisation

Consider $X \sim \mu$ vector of returns in $\mathbb{R}^d$ and $a \in \mathcal{A} \subset \mathbb{R}^d$ portfolio

$$V(0) = \inf_{a \in \mathcal{A}} \mathrm{AV@R}_\alpha (a \cdot X) = \inf_{a \in \mathcal{A}, m \in \mathbb{R}} \left\{ m + \frac{1}{\alpha} \int (a \cdot x - m)^+ \mu(dx) \right\}$$

And its robust version reads

$$V(\delta) = \inf_{a \in \mathcal{A}} \mathcal{R}\mathrm{AV@R}_\alpha (a \cdot X) = \inf_{a \in \mathcal{A}, m \in \mathbb{R}} \sup_{\nu \in B_\delta(\mu)} \left\{ m + \frac{1}{\alpha} \int (a \cdot x - m)^+ \nu(dx) \right\},$$

Example 1: AV@R minimisation

Consider $X \sim \mu$ vector of returns in $\mathbb{R}^d$ and $a \in \mathcal{A} \subset \mathbb{R}^d$ portfolio

$$V(0) = \inf_{a \in \mathcal{A}} \mathrm{AV@R}_\alpha (a \cdot X) = \inf_{a \in \mathcal{A}, m \in \mathbb{R}} \left\{ m + \frac{1}{\alpha} \int (a \cdot x - m)^+ \mu(dx) \right\}$$

And its robust version reads

$$V(\delta) = \inf_{a \in \mathcal{A}} \mathcal{R}\mathrm{AV@R}_\alpha (a \cdot X) = \inf_{a \in \mathcal{A}, m \in \mathbb{R}} \sup_{\nu \in B_\delta(\mu)} \left\{ m + \frac{1}{\alpha} \int (a \cdot x - m)^+ \nu(dx) \right\},$$

where $B_\delta(\mu) = \{\nu \in \mathcal{P}(\mathcal{S}) : W_p(\mu, \nu) \leq \delta\}$. A direct computation gives

$$\Upsilon = |a^*| \left(\frac{1}{\alpha^q} \int 1_{\{a^* \cdot x \geq V@R_\alpha(a^* \cdot L)\}} \right)^{\frac{1}{q}} \mu(dx) = \frac{|a^*|}{\alpha^{1/p}} \quad \text{, or}$$

$$\inf_{a \in \mathcal{A}} \mathcal{R}\mathrm{AV@R}_\alpha (a \cdot X) = \mathrm{AV@R}_\alpha (a^* \cdot X) + \frac{|a^*|}{\alpha^{1/p}} \delta + o(\delta).$$

Example 2: Mean-variance optimal investment

Consider $X \sim \mu$ vector of returns in $\mathbb{R}^d$ and $\mathcal{A} = \{a : \langle a, \mathbf{1} \rangle = 1\}$.

$$V(0) = \inf_{a \in \mathcal{A}} \mathbb{E}[\langle a, X \rangle] + \gamma \mathrm{VAR}_\mu(\langle a, X \rangle) = \inf_{a \in \mathcal{A}} \sup_{Z : \mathbb{E}[Z]=1, \mathbb{E}[Z^2]=1+\gamma^2} \mathbb{E}\Big[\langle a, X \rangle Z\Big]$$

And its robust version, for $p = q = 2$, reads

$$V(\delta) = \inf_{a \in \mathcal{A}} \sup_{(\xi, Z) : \mathbb{E}[\langle \xi, \xi \rangle] \le \delta^2, \mathbb{E}[Z]=1, \mathbb{E}[Z^2]=1+\gamma^2} \mathbb{E}\Big[\langle a, X + \xi \rangle Z\Big]$$

A two-step computation recovers the result in Pflug et al. '12:

$$\Upsilon = |a^*| \sqrt{1 + \gamma^2}.$$

Ex 1: Decision making: prefs representation

Let X be agent's wealth/consumption. Savage '51, von Neuman & Morgenstern '53 give

$$\mathbb{P} \succeq \check{\mathbb{P}} \quad \Leftrightarrow \quad \mathbb{E}_{\mathbb{P}}[u(X)] \geq \mathbb{E}_{\check{\mathbb{P}}}[u(X)].$$

Ex 1: Decision making: prefs representation

Let X be agent's wealth/consumption. Savage '51, von Neuman & Morgenstern '53 give

$$\mathbb{P} \succeq \check{\mathbb{P}} \quad \Leftrightarrow \quad \mathbb{E}_{\mathbb{P}}[u(X)] \geq \mathbb{E}_{\check{\mathbb{P}}}[u(X)].$$

An ambiguity averse agent of Gilboa & Schmeidler '89, might instead consider

$$\mathbb{P} \succeq_{\rho} \check{\mathbb{P}} \quad \Leftrightarrow \quad \min_{\tilde{\mathbb{P}} \in B_\delta(\mathbb{P})} \mathbb{E}_{\tilde{\mathbb{P}}}[u(X)] \geq \min_{\tilde{\mathbb{P}} \in B_\delta(\check{\mathbb{P}})} \mathbb{E}_{\tilde{\mathbb{P}}}[u(X)].$$

for $B_\delta(\mathbb{P})$ a δ-ball around $\mathbb{P}$ in some metric ρ,

(also called *constraint preferences* by Hansen & Sargent '01).

Variational prefs: relative entropy vs Wasserstein

The variational/constraint preferences with ρ-ball $B_\delta(\mathbb{P})$

$$\mathcal{U}(X) := \min_{\tilde{\mathbb{P}} \in B_\delta(\mathbb{P})} \mathbb{E}_{\tilde{\mathbb{P}}}[u(X)]$$

up to $o(\delta)$ are equivalent to:

$\rho =$ Rel. entropy

$$\mathcal{U}(X) \approx \mathbb{E}_{\mathbb{P}}[u(X))] - \delta\sqrt{2\mathrm{Var}_{\mathbb{P}}(u(X))}$$

(cf. Lam '16)

$\rho = W_2$ Wasserstein

$$\mathcal{U}(X) \approx \mathbb{E}_{\mathbb{P}}[u(X))] - \delta\sqrt{\mathbb{E}_{\mathbb{P}}\left[|u'(X)|^2\right]}$$

(cf. our Υ-sensitivity)

Example 2: EUM & Optimal investment

$X = S_T - S_0 \sim \mu$ vector of returns in $\mathcal{S} \subset \mathbb{R}^d$ and $\mathcal{A} \subseteq \mathbb{R}^d$ admissible strategies; wlog $r = 0$, initial capital $x = 0$.
$u : \mathbb{R} \to \mathbb{R}$ strictly concave, continuously differentiable, bounded from above. Consider the expected utility maximisation problem:

$$V(0) = \sup_{a \in \mathcal{A}} \mathbb{E}_\mu \left[u \left(\langle X, a \rangle \right) \right]$$

The optimal $a^\star \in \mathcal{A}$ is characterised through the FOC

$$\mathbb{E}_\mu \left[X \cdot u' \left(\langle X, a^\star \rangle \right) \right] = 0$$

Example 2: EUM & Optimal investment

$X = S_T - S_0 \sim \mu$ vector of returns in $\mathcal{S} \subset \mathbb{R}^d$ and $\mathcal{A} \subseteq \mathbb{R}^d$ admissible strategies; wlog $r = 0$, initial capital $x = 0$.
$u : \mathbb{R} \to \mathbb{R}$ strictly concave, continuously differentiable, bounded from above. Consider the expected utility maximisation problem:

$$V(\delta) = \sup_{a \in \mathcal{A}} \inf_{\nu \in B_\delta(\mu)} \mathbb{E}_\nu \left[u \left(\langle X, a \rangle \right) \right]$$

The optimal $a^\star = a^\star(0) \in \mathcal{A}$ is characterised through the FOC

$$\mathbb{E}_\mu \left[X \cdot u' \left(\langle X, a^\star \rangle \right) \right] = 0$$

and

$$V'(0) = - \left(\mathbb{E}_\mu \left[|u'(\langle X, a^\star \rangle)|^q \right] \right)^{1/q} |a^\star|$$

is the sensitivity to ambiguity aversion.
Note that $V'(0) < 0$ and is increasing in p.

Binomial model with an exponential utility

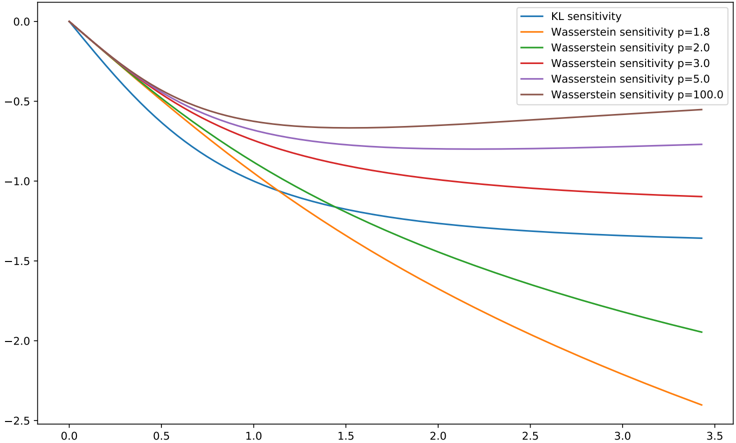


Figure: Sensitivities in function of the market's Sharpe ratio

$\mathcal{N}(m, \sigma^2)$ model with an exponential utility

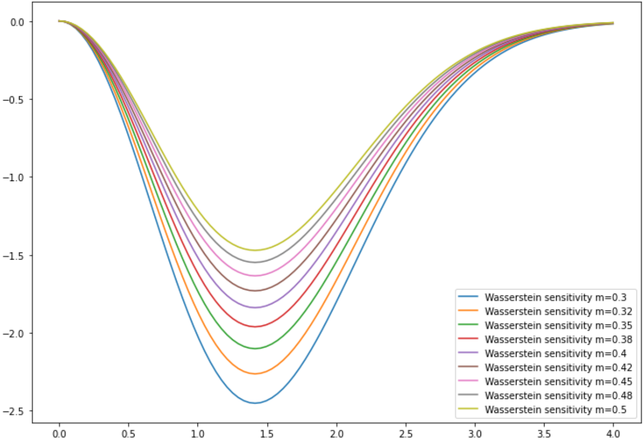


Figure: Sensitivities for $p = \infty$ in function of the market's Sharpe ratio $\frac{m}{\sigma}$

Ex 3: Robust call pricing (martingale constraint)

We optimise over measures $\nu \in B_\delta(\mu)$ satisfying $\int x\,\nu(dx) = S_0$.
A constrained version of our main results gives, for $p = 2$,

$$\Upsilon = \inf_{a^* \in \mathrm{A}^{\mathrm{opt}}(0)} \left(\int \left(\nabla_x f(x, a^*) - \int \nabla_x f(y, a^*)\,\mu(dy) \right)^2 \mu(dx) \right)^{1/2},$$

i.e., Υ is the standard deviation of $\nabla_x f(\cdot, a^*)$ under μ.

Ex 3: Robust call pricing (martingale constraint)

We optimise over measures $\nu \in B_\delta(\mu)$ satisfying $\int x\,\nu(dx) = S_0$.
A constrained version of our main results gives, for $p = 2$,

$$\Upsilon = \inf_{a^* \in \mathrm{A}^{\mathrm{opt}}(0)} \left(\int \left(\nabla_x f(x, a^*) - \int \nabla_x f(y, a^*)\,\mu(dy) \right)^2 \mu(dx) \right)^{1/2},$$

i.e., Υ is the standard deviation of $\nabla_x f(\cdot, a^*)$ under μ.

Let $\mu \sim S_T/S_0$ with (S_t) from the BS(σ) model and

$$\mathcal{R}BS(\delta) = \sup_{\nu \in B_\delta(\mu)} \left\{ \int (S_0 x - K)^+ \nu(dx) \colon \int x\nu(dx) = 1 \right\}$$

so that $\mathcal{R}BS(0) = BSCall(S_0, K, \sigma)$. For $p = 2$ we find

$$\Upsilon(K) = S_0 \sqrt{\Phi(d_-)(1 - \Phi(d_-))}.$$

Robust call: numerics

Exact value $\mathcal{R}BS(\delta)$, first-order (FO) approximation and the model (BS) price.

BS model with $S_0 = T = 1$, $K = 1.2$, $r = q = 0$, $\sigma = 0.2$. $\delta = 0.05$

Robust call: classical vs robust

Take $r = q = 0$, $T = 1$, $S_0 = 1$ and $\mu =$BS(σ) log-normal.

$$\mathcal{R}BS(\delta) = \sup_{\nu \in B_\delta(\mu)} \int_{\mathcal{S}} (s - K)^+ \nu(ds).$$

Parametric Approach

$$B_\delta(\mu) = \{\mathrm{BS}(\tilde{\sigma}) : |\tilde{\sigma} - \sigma| \leq \delta\}$$

Then

$$\mathcal{R}BS'(0) = \mathcal{V} = S_0 \phi(d_+).$$

Non-parametric Approach

$$B_\delta(\mu) = \{\nu : W_2(\mu, \nu) \leq \delta\}$$

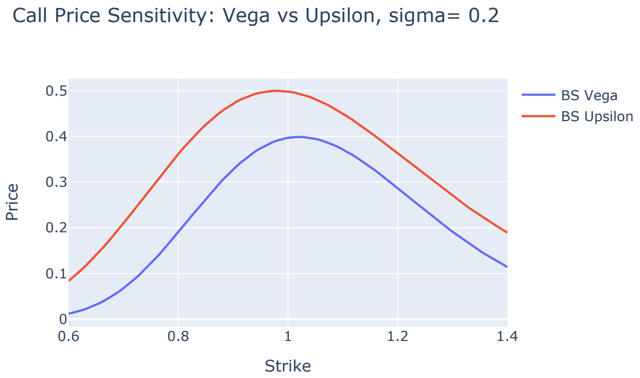
Then

$$\mathcal{R}BS'(0) = \Upsilon = S_0 \sqrt{\Phi(d_-)(1 - \Phi(d_-))}$$

BS Call: Vega($\mathcal{V}$) vs Upsilon(Υ)

Consider the simple example of a call option pricing.

Take $r = q = 0,\ T = 1,\ S_0 = 1$ and $\mu =$BS(σ) model.



Hedging: Δ–Vega vs Δ–ϒ (with S. Moliner '22)

Observe that $\Upsilon[aS_t + b] = 0$, *i.e., cash and stock carry no uncertainty.*

Comparison of two hedging approaches:

- Δ–Vega: at rebalancing buy/sell stock + ATM Call so that $\Delta = 0 = \mathcal{V}$
- Δ–ϒ: at rebalancing buy/sell stock + ATM Call so that $\Delta = 0$ and Υ is minimized

	Δ	$\Delta + \mathcal{V}$	$\Delta + \Upsilon$
Mean	-0.001	0.0	-0.0
Std	0.043	0.007	0.011
$\text{V@R}_{0.95}$	-0.086	-0.009	-0.018
$\text{ES}_{0.95}$	-0.110	-0.016	-0.024

Table 1: Risk measures with Heston Model $S_0 = T = 1$, $K = 1.05$, $v_0 = 0.04$, $\kappa = 1$, $\theta = 0.09$, $\sigma = 0.6$, $\rho = 0.5$

Hedging: Δ–Vega vs Δ–Υ (WITH S. MOLINER '22)

Observe that $\Upsilon[aS_t + b] = 0$, *i.e., cash and stock carry no uncertainty*

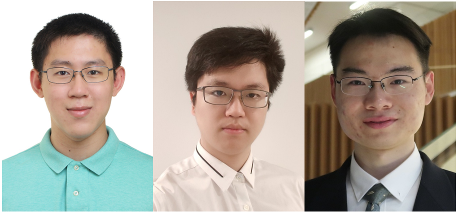
Comparison of two hedging approaches:

- Δ–Vega: at rebalancing buy/sell stock + ATM Call so that $\Delta = 0 = \mathcal{V}$
- Δ–Υ: at rebalancing buy/sell stock + ATM Call so that $\Delta = 0$ and Υ is minimized

	Δ	$\Delta + \mathcal{V}$	$\Delta + \Upsilon$
Mean	-0.015	-0.001	-0.002
Std	0.095	0.01	0.014
$V@R_{0.95}$	-0.190	-0.016	-0.028
$ES_{0.95}$	-0.296	-0.032	-0.045

Table 2: Risk measures with Bates Model $S_0 = T = 1$, $K = 1.05$, $v_0 = 0.04$, $\kappa = 1$, $\theta = 0.09$, $\sigma = 0.6$, $\rho = 0.5$, $\lambda = 15$, $\mu_J = 0$, $\sigma_J = 0.1$

W-Distributional Robustness of NNs



with X. Bai, G. He, Y. Jiang

NeurIPS 23

GitHub: JanObloj/W-DRO-Adversarial-Methods

Image classification setup

- An image is interpreted as a tuple $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where x denotes the feature vector and y denotes the class.
- W.l.o.g, we take $\mathcal{X} = [0,1]^n$ and $\mathcal{Y} = \{1, \ldots, m\}$.
- $\mathbb{P}$ is a given data distribution on $\mathcal{X} \times \mathcal{Y}$.
- A neural network is a map $f_\theta : \mathcal{X} \to \mathbb{R}^m$

$$f_\theta(x) = f^l \circ \cdots \circ f^1(x), \qquad \text{where } f^i(x) = \sigma(w^i x + b^i).$$

- Prediction of x under f_θ is given by $\arg\max_{1 \le i \le m}\{f_\theta(x)_i\}$.

Image classification setup

- An image is interpreted as a tuple $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where x denotes the feature vector and y denotes the class.
- W.l.o.g, we take $\mathcal{X} = [0,1]^n$ and $\mathcal{Y} = \{1, \ldots, m\}$.
- $\mathbb{P}$ is a given data distribution on $\mathcal{X} \times \mathcal{Y}$.
- A neural network is a map $f_\theta : \mathcal{X} \to \mathbb{R}^m$

$$f_\theta(x) = f^l \circ \cdots \circ f^1(x), \qquad \text{where } f^i(x) = \sigma(w^i x + b^i).$$

- Prediction of x under f_θ is given by $\arg\max_{1 \le i \le m}\{f_\theta(x)_i\}$.

The aim of image classification is to find a model with high accuracy

$$A := \mathbb{P}(\arg\max_{1 \le i \le m}\{f_\theta(x)_i\} = y) = \mathbb{P}(S).$$

This is achieved by training the network f_θ according to:

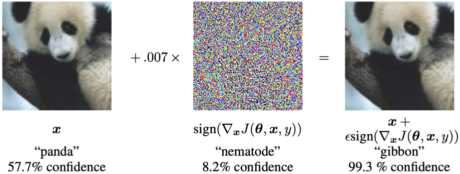
$$\inf_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}}[J(\theta, x, y)] \quad \text{where } J(\theta, x, y) = L(f_\theta(x), y).$$

NN & adversarial attacks

Mathematical
Institute

Consider data (x, y) from $\mathbb{P}$ and a NN trained according to:

$$\inf_{\theta} \int |J(\theta, x, y)| \, \mathbb{P}(dx, dy).$$



Source: Goodfellow, Shlens & Szegedy ICLR 2015

Background on adv attacks/training

Adversarial attack:

- Fast Gradient Sign Method (FGSM), see Goodfellow, Shlens & Szegedy '14
- Projected Gradient Descent (PGD), see Madry et al. '18
- Black-box attacks: Zeroth order optimization (Chen et al. '17), query-limited attack (Ilyas et al. '18) ...
- Autoattack, see Croce & Hein '20

Adversarial training:

- Random data generation by GAN/ diffusion models, see Gowal et al. '21 and Wang et al. '23
- Robustness–accuracy tradeoff, see TRADES Zhang et al. '19, MART Wang et al. '20, SCORE Pang et al. '22
- W-DRO based methods: Sinha, Namkoong & Duchi '18, Trillos & Trillos '22, Bui et al. '22 ...

Adversarial robustness dataset and benchmarks

Mathematical Institute

- Adversarial attacks and defence is a large field in ML
- RobustBench tracks over 3000 papers and maintains a leaderboard for CIFAR datasets

RobustBench Leaderboards Paper FAQ Contribute Model Zoo 🚀

RobustBench

A standardized benchmark for adversarial robustness

The goal of **RobustBench** is to systematically track the *real* progress in adversarial robustness. There are already more than 3'000 papers on this topic, but it is still unclear which approaches really work and which only lead to overestimated robustness. We start from benchmarking common corruptions, ℓ_∞- and ℓ_2-robustness since these are the most studied settings in the literature. We use AutoAttack, an ensemble of white-box and black-box attacks, to standardize the evaluation (for details see our paper) of the ℓ_p robustness and CIFAR-10-C for the evaluation of robustness to common corruptions. Additionally, we open source the RobustBench library that contains models used for the leaderboard to facilitate their usage for downstream applications.

To prevent potential overadaptation of new defenses to AutoAttack, we also welcome external evaluations based on *adaptive attacks*, especially where AutoAttack flags a potential overestimation of robustness. For each model, we are interested in the best known robust accuracy and see AutoAttack and adaptive attacks as complementary.

News:

- **May 2022:** We have extended the common corruptions leaderboard on ImageNet with 3D Common Corruptions (ImageNet-3DCC). ImageNet-3DCC evaluation is interesting since (1) it includes more realistic corruptions and (2) it can be used to assess generalization of the existing models which may have overfitted to ImageNet-C. For a quickstart, click here. See the new leaderboard with ImageNet-C and ImageNet-3DCC here (also mCE metrics can be found here).
- **May 2022:** We fixed the preprocessing issue for ImageNet corruption evaluations: previously we used resize to 256x256 and central crop to 224x224 which wasn't necessary since the ImageNet-C images are already 224x224. Note that this changed the ranking between the top-1 and top-2 entries.

Up-to-date leaderboard based on 120+ models

Unified access to 80+ state-of-the-art robust models via Model Zoo

W-DRO formulation

Clean training:

$$\inf_{\theta\in\Theta} \mathbb{E}_{\mathbb{P}}[L(f_\theta(x), y)].$$

Adversarial training (Madry et al. '18):

$$\inf_{\theta\in\Theta} \mathbb{E}_{\mathbb{P}}\left[\max_{\|x-x'\|_r\leq\delta} L(f_\theta(x'), y)\right].$$

W-DRO formulation

Clean training:

$$\inf_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}}[L(f_\theta(x), y)].$$

Adversarial training (Madry et al. '18):

$$\inf_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}} \left[\max_{\|x - x'\|_r \leq \delta} L(f_\theta(x'), y) \right].$$

W-DRO adversarial training:

$$\inf_{\theta \in \Theta} \sup_{\mathbb{Q} \in B_\delta(\mathbb{P})} \mathbb{E}_{\mathbb{Q}}[L(f_\theta(x), y)],$$

where $B_\delta(\mathbb{P})$ is the **p**-Wasserstein ball induced by a 'distance' d on $\mathcal{X} \times \mathcal{Y}$ defined by, $\mathbf{r} > 1$,

$$d((x, y), (x', y')) = \|x - x'\|_r + \infty \mathbf{1}_{\{y \neq y'\}}.$$

Taking the ∞-Wasserstein ball reduces W-DRO to Madry et al..

Remark that in reality training is done using

$$\hat{\mathbb{P}} = \frac{1}{M}\sum_{i=1}^{M}\delta_{(x_i,y_i)},$$

where $\{(x_i, y_i) : i = 1, \ldots, M\}$ is the training set.

Remark that in reality training is done using

$$\hat{\mathbb{P}} = \frac{1}{M}\sum_{i=1}^{M} \delta_{(x_i, y_i)},$$

where $\{(x_i, y_i) : i = 1, \dots, M\}$ is the training set.
A $(\mathcal{W}_\infty, l_\infty)$ δ-ball around $\hat{\mathbb{P}}$ contains all the measures

$$\frac{1}{M}\sum_{i=1}^{M} \delta_{(x_i', y_i)}, \quad \|x_i - x_i'\|_\infty \leq \delta \text{ for } i = 1, \dots, M,$$

i.e., it recovers pointwise perturbations of the pixels.

Remark that in reality training is done using

$$\hat{\mathbb{P}} = \frac{1}{M} \sum_{i=1}^{M} \delta_{(x_i, y_i)},$$

where $\{(x_i, y_i) : i = 1, \dots, M\}$ is the training set.
A $(\mathcal{W}_\infty, l_\infty)$ δ-ball around $\hat{\mathbb{P}}$ contains all the measures

$$\frac{1}{M} \sum_{i=1}^{M} \delta_{(x_i', y_i)}, \quad \|x_i - x_i'\|_\infty \leq \delta \text{ for } i = 1, \dots, M,$$

i.e., it recovers pointwise perturbations of the pixels.
However, a $(\mathcal{W}_2, l_2)$ δ-ball around $\hat{\mathbb{P}}$ contains many more measures, discrete and continuous, e.g., uniform measure over

$$\mathcal{X} \cap \bigcup_{i=1}^{M} \{(x, y_i) : |x_i^k - x^k| \leq \varepsilon \text{ for } k = 1, \dots n\}$$

for ε small enough ($\varepsilon^3 < 3\delta^2/2n$).

First order approximation

Let $J_\theta(x,y) = L(f_\theta(x),y)$ and $V(\delta) = \sup_{\mathbb{Q}\in B_\delta(\mathbb{P})} \mathbb{E}_{\mathbb{Q}}[L(f_\theta(x),y)]$.

Theorem

Assuming J_θ is Lipschitz, the following first order approximations hold:

(i) $V(\delta) = V(0) + \delta\Upsilon + o(\delta)$, *where*

$$\Upsilon = \Big(\mathbb{E}_{\mathbb{P}}\|\nabla_x J_\theta(x,y)\|_s^q\Big)^{1/q}.$$

(ii) $V(\delta) = \mathbb{E}_{\mathbb{Q}_\delta}[J_\theta(x,y)] + o(\delta)$, *where*

$$\mathbb{Q}_\delta = \Big[(x,y) \mapsto \big(x + \delta h(\nabla_x J_\theta(x,y))\|\Upsilon^{-1}\nabla_x J_\theta(x,y)\|_s^{q-1}, y\big)\Big]_\# \mathbb{P},$$

and h is uniquely determined by $\langle h(x), x\rangle = \|x\|_s$.

Wassserstein distributionally adversarial attacks

Based on the first order approximation, we propose W-FGSM attack given by

$$x' = x + \delta h(\nabla_x J_\theta(x^t, y)) \|\Upsilon^{-1} \nabla_x J_\theta(x, y)\|_s^{q-1}, \tag{1}$$

Similarly, we propose W-PGD attack as

$$x^{t+1} = \mathrm{proj}_\delta \big(x^t + \alpha h(\nabla_x J_\theta(x^t, y)) \|\Upsilon^{-1} \nabla_x J_\theta(x^t, y)\|_s^{q-1} \big), \tag{2}$$

where α is the stepsize, proj_δ is the projection onto Wasserstein ball $B_\delta(\mathbb{P})$ and $t = 1, \ldots, t_{max}$.

In particular, under the case $(\mathcal{W}_\infty, \ell_\infty)$ we retrieve FGSM attack given by

$$x' = x + \delta \, \mathrm{sgn}(\nabla_x J_\theta(x, y)).$$

Loss functions

Mathematical Institute

For pointwise attacks a combination of cross-entropy (CE) and Difference of Logits Ratio (DLR) losses works well. For $z = (z_1, \ldots, z_m) = f_\theta(x)$ and $z_{(1)} \geq \cdots \geq z_{(m)}$ the order statistics of z,

$$\mathrm{DLR}(z, y) = \begin{cases} -\dfrac{z_y - z_{(2)}}{z_{(1)} - z_{(3)}}, & \text{if } z_y = z_{(1)}, \\ -\dfrac{z_y - z_{(1)}}{z_{(1)} - z_{(3)}}, & \text{else.} \end{cases}$$

Under *distributional* threat models, we propose ReDLR (Rectified DLR) loss:

$$\mathrm{ReDLR}(z, y) = -(\mathrm{DLR})^{-}(z, y) = \begin{cases} -\dfrac{z_y - z_{(2)}}{z_{(1)} - z_{(3)}}, & \text{if } z_y = z_{(1)}, \\ 0, & \text{else.} \end{cases} \tag{3}$$

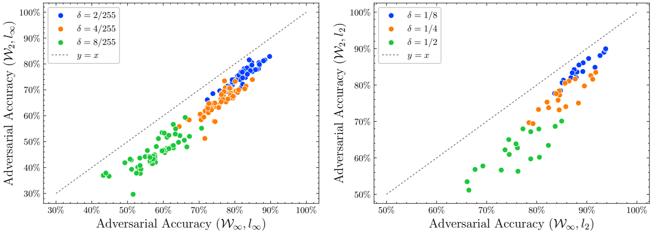
Comparison of adversarial attacks

CIFAR-10 dataset: 60k (50k+10k) color (3 channels) images across 10 classes.
We normalize the input feature as $x \in [0,1]^{3\times 32\times 32}$.

Recall $S = \{(x,y) \in \mathcal{X} \times \mathcal{Y} : y = \arg\max_{1\le i\le m}\{f_\theta(x)_i\}\}$.
Define the adversarial accuracy A_δ as

$$A_\delta := \inf_{\mathbb{Q}\in B_\delta(\mathbb{P})} \mathbb{Q}(S)$$

and compare it under classical $\mathcal{W}_\infty$ and $\mathcal{W}_2$ threat models:



Comparison of adversarial attacks

CIFAR-10 dataset: 60k (50k+10k) color (3 channels) images across 10 classes. We normalize the input feature as $x \in [0,1]^{3\times 32\times 32}$.

Recall $S = \{(x,y) \in \mathcal{X} \times \mathcal{Y} : y = \arg\max_{1\le i\le m}\{f_\theta(x)_i\}\}$.
Define the adversarial accuracy A_δ as

$$A_\delta := \inf_{\mathbb{Q}\in B_\delta(\mathbb{P})} \mathbb{Q}(S)$$

and compare it under classical $\mathcal{W}_\infty$ and $\mathcal{W}_2$ threat models:

	$\mathcal{W}_\infty$	$\mathcal{W}_2$		
Methods	AutoAttack	W-PGD-CE	W-PGD-DLR	W-PGD-ReDLR
l_∞	57.66%	61.32%	79.00%	**45.46%**
l_2	75.78%	74.62%	78.69%	**61.69%**

Bounds on adversarial accuracy

University of Oxford
Mathematical Institute

We write $\mathcal{R}_\delta := A_\delta / A$ as a metric of robustness for neural networks. Any admissible attack gives an upper bound on adversarial accuracy:

$$\mathcal{R}_\delta \leq \mathcal{R}_\delta^u := \mathbb{Q}_\delta(S)/A.$$

Bounds on adversarial accuracy

Mathematical Institute

We write $\mathcal{R}_\delta := A_\delta / A$ as a metric of robustness for neural networks. Any admissible attack gives an upper bound on adversarial accuracy:

$$\mathcal{R}_\delta \leq \mathcal{R}_\delta^u := \mathbb{Q}_\delta(S)/A.$$

To obtain a lower bound we impose:

- $0 < \mathbb{Q}(S) < 1$,
- $\mathcal{W}_p(\mathbb{P}(\cdot \mid S), \mathbb{Q}(\cdot \mid S)) + \mathcal{W}_p(\mathbb{P}(\cdot \mid S^c), \mathbb{Q}(\cdot \mid S^c)) = o(\delta)$,

for any $\mathbb{Q} \in B_\delta(\mathbb{P})$.

Bounds on adversarial accuracy

We write $\mathcal{R}_\delta := A_\delta / A$ as a metric of robustness for neural networks. Any admissible attack gives an upper bound on adversarial accuracy:

$$\mathcal{R}_\delta \leq \mathcal{R}^u_\delta := \mathbb{Q}_\delta(S)/A.$$

Theorem (lower bound)

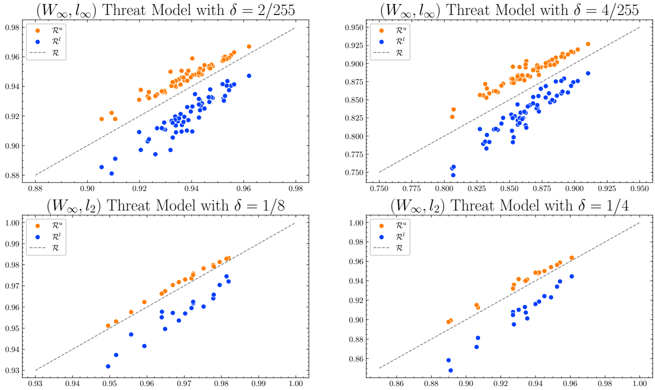
We write $W(0) = \mathbb{E}_{\mathbb{P}}[J_\theta(x,y)|S^c]$. Under suitable assumptions, we have an asymptotic lower bound as $\delta \to 0$

$$\mathcal{R}_\delta \geq \frac{W(0) - V(\delta)}{W(0) - V(0)} + o(\delta) = \mathcal{R}^l_\delta + o(\delta) \tag{4}$$

where $\mathcal{R}^l_\delta = \min\{\widetilde{\mathcal{R}}^l_\delta, \overline{\mathcal{R}}^l_\delta\}$ and the first order approximations are given by

$$\widetilde{\mathcal{R}}^l_\delta = \frac{W(0) - \mathbb{E}_{\mathbb{Q}_\delta}[J_\theta(x,y)]}{W(0) - V(0)} \quad \textit{and} \quad \overline{\mathcal{R}}^l_\delta = \frac{W(0) - V(0) - \delta\Upsilon}{W(0) - V(0)}. \tag{5}$$

Bounds on $\mathcal{W}_\infty$-adversarial accuracy



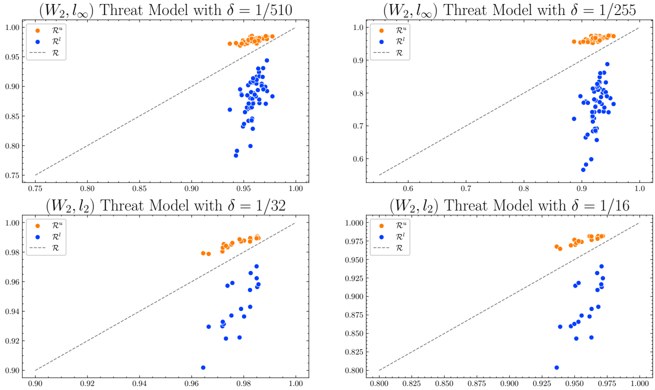
$\mathcal{R}^l$ computed using CE loss. Blue dot takes around $1-2\%$ of computational time compared to the diagonal.

Comparison of $(\mathcal{W}_\infty, l_\infty)$ computational times

	PreAct ResNet-18	ResNet -18	ResNet -50	WRN -28-10	WRN -34-10	WRN -70-16
$\mathcal{R}$	197	175	271	401	456	2369
$\mathcal{R}^l \& \mathcal{R}^u$	0.52	0.49	0.17	0.55	0.53	1.46

Computation times of $(\mathcal{W}_\infty, l_\infty)$, $\delta = 8/255$ attack for one mini-batch of size 100, in seconds. We compute $\mathcal{R}$ by AutoAttack and average the computation time over models on RobustBench grouped by their architecture.

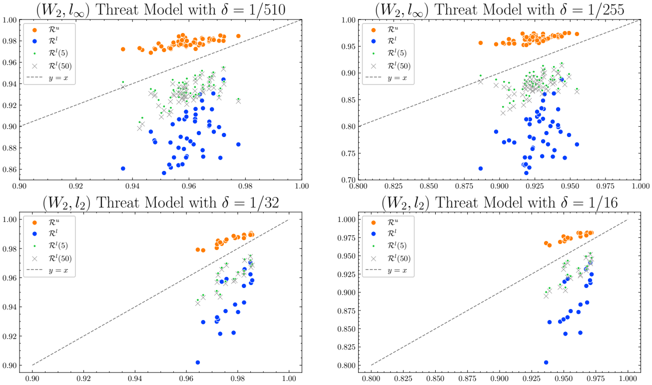
Bounds on $\mathcal{W}_2$-adversarial accuracy



$\mathcal{R}^l$ computed using Rectified DLR loss. Blue dot takes 2% of computational time compared to the diagonal.

Improved bounds on $\mathcal{W}_2$-adversarial accuracy

UNIVERSITY OF OXFORD
Mathematical Institute



$\mathcal{R}^l$ computed using Rectified DLR loss and $1, 5$ *and* 50 *iterations.*

Out of sample performance

Theorem

Let $\varepsilon > 0$. *With probability at least* $1 - K\exp(-KN\varepsilon^n)$ *we have*

$$V(\delta) \leq \widehat{V}(\delta) + \varepsilon \sup_{Q \in B^*_\delta(\widehat{\mathbb{P}})} \left(\mathbb{E}_Q \|\nabla_x J_\theta(x,y)\|_s^q \right)^{1/q} + o(\varepsilon) \leq \widehat{V}(\delta) + L\varepsilon$$

where $B^*_\delta(\widehat{\mathbb{P}}) = \arg\max_{Q \in B_\delta(\widehat{\mathbb{P}})} \mathbb{E}_Q[J_\theta(x,y)]$ *and* K *only depends on* p *and* n.

Corollary

With probability at least $1 - K\exp(-KN\delta^n)$ *we have*

$$A(\mathbb{P}) - A_\delta(\mathbb{P}) \leq \frac{\widehat{V}(\delta) - \widehat{V}(0)}{\widehat{W}(0) - \widehat{C}(0)} + \frac{2L\delta}{\widehat{W}(0) - \widehat{C}(0)} + o(\delta),$$

where $\widehat{\cdot}$ *are computed using the training set.*

Limitations and constraints

A (W_∞, l_∞) threat model with budget $\delta = 8/255$ can make significant changes to the image.

WideResNet-28-10 (Gowal et al., 2020), the confidence goes $73\% \rightsquigarrow 61\% \rightsquigarrow 60\%$.

Limitations and constraints

A (W_∞, l_∞) threat model with budget $\delta = 8/255$ can make significant changes to the image.

WideResNet-28-10 (Gowal et al., 2020), the confidence goes
73% $\rightsquigarrow$ 61% $\rightsquigarrow$ 60%.

Limitations and constraints

A (W_∞, l_∞) threat model with budget $\delta = 8/255$ can make significant changes to the image.

WideResNet-28-10 (Gowal et al., 2020), the confidence goes $73\% \rightsquigarrow 61\% \rightsquigarrow 60\%$.

Limitations and constraints

Our theoretical bounds for (W_∞, l_∞) threat model with budget $\delta = 8/255$. Bounds fail in some cases as we are outside of the first order approximation regime.

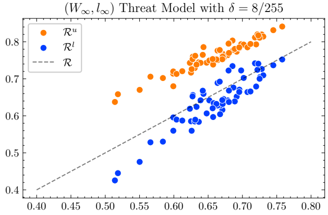


Figure: $\mathcal{R}^u \& \mathcal{R}^l$ versus $\mathcal{R}$.

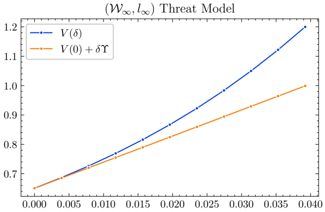


Figure: First order approximation.

LHS uses 60 models on RobustBench. RHS uses WideResNet-28-10 (Gowal et al., 2020).

W-DRO Training as fine-tuning

Networks	Clean Acc	$\mathcal{W}_\infty$ Adversarial Acc	$\mathcal{W}_2$ Adversarial Acc
Zhang et al. '19	83.71	59.99 (+2.95)	50.53 (+7.54)
Chen et al. '24	85.44	62.12 (+1.98)	53.42 (+9.66)
Gowal et al. '20	85.93	63.39 (-3.05)	52.14 (+1.15)
Cui et al. '23	88.88	68.71 (-2.21)	58.02 (+4.86)
Wang et al. '23	91.45	69.19 (-1.43)	55.93 (+3.79)

Main results

Part II: Sensitivity of the optimisers

Sensitivity of optimisers

Mathematical Institute

Theorem

For $p = q = 2$, under suitable regularity and growth assumptions,

$$\lim_{\delta \to 0} \frac{a^*(\delta) - a^*}{\delta} = -\frac{1}{\Upsilon} (\nabla_a^2 V(0, a^*))^{-1} \int \nabla_x \nabla_a f(x, a^*) \nabla_x f(x, a^*)\, \mu(dx),$$

where $a^* := a^*(0)$.

The results extends to general $p > 1$ and semi-norms.

Example 1: Square-root LASSO

Consider $\|(x,y)\|_* = |x|_r \mathbf{1}_{\{y=0\}} + \infty \mathbf{1}_{\{y\neq 0\}}$, $r>1$, $(x,y)\in\mathbb{R}^k\times\mathbb{R}$.
Then (see Blanchet, Kang & Murthy '19)

$$\inf_{a\in\mathbb{R}^k} \sup_{\nu\in B_\delta(\hat{\mu}_N)} \int (y-\langle x,a\rangle)^2\,d\nu = \inf_{a\in\mathbb{R}^k}\left(\sqrt{\int (y-\langle a,x\rangle)^2\,d\mu} + \delta|a|_s\right)^2,$$

where $1/r+1/s=1$. $\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^N \delta_{(x^i,y^i)}$ encodes the observations.
System is overdetermined so that $D=\int xx^T\,\mu(dx)$ is invertible.
$\delta=0$ case is the ordinary least squares regression: $a^* = \frac{1}{N}D^{-1}\int yx d\mu$.

Example 1: Square-root LASSO

Consider $\|(x,y)\|_* = |x|_r \mathbf{1}_{\{y=0\}} + \infty \mathbf{1}_{\{y\neq 0\}}$, $r > 1$, $(x,y) \in \mathbb{R}^k \times \mathbb{R}$.
Then (see Blanchet, Kang & Murthy '19)

$$\inf_{a\in\mathbb{R}^k} \sup_{\nu \in B_\delta(\hat{\mu}_N)} \int (y - \langle x, a\rangle)^2 \, d\nu = \inf_{a\in\mathbb{R}^k} \left(\sqrt{\int (y - \langle a, x\rangle)^2 \, d\mu} + \delta |a|_s \right)^2,$$

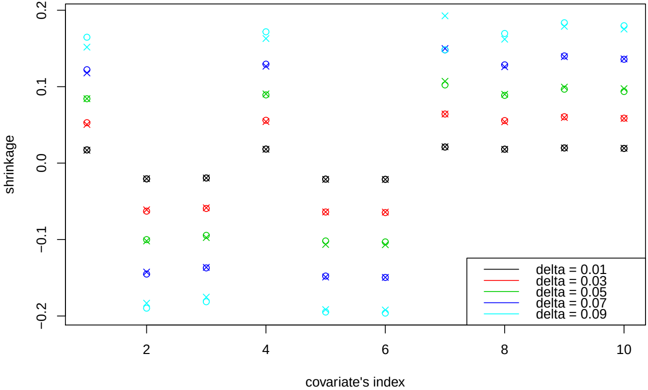
where $1/r + 1/s = 1$. $\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^N \delta_{(x^i, y^i)}$ encodes the observations.
System is overdetermined so that $D = \int xx^T \mu(dx)$ is invertible.
$\delta = 0$ case is the ordinary least squares regression: $a^* = \frac{1}{N} D^{-1} \int yx d\mu$.
$\delta > 0$, $s = 1 \rightsquigarrow$ RHS = square-root LASSO regression Belloni et al. '11
$\delta > 0$, $s = 2 \rightsquigarrow$ RHS $\approx$ Ridge regression
Then $a^*(\delta)$ is approximately, for $s = 1$ and $s = 2$ (cf. Tibshirani '96):

$$a^* - \sqrt{V(0)} D^{-1} \mathrm{sgn}(a^*) \delta \quad \text{and} \quad a^* \left(1 - \frac{\sqrt{V(0)}}{|a^*|_2} D^{-1} \delta\right)$$

Square-root LASSO: numerics

Comparison of exact (o) and first-order (x) approximation of square-root LASSO coefficients for 2000 data generated from: (with all X_i, ε i.i.d. $\mathcal{N}(0,1)$)

$$Y = 1.5X_1 - 3X_2 - 2X_3 + 0.3X_4 - 0.5X_5 - 0.7X_6 + 0.2X_7 + 0.5X_8 + 1.2X_9 + 0.8X_{10} + \varepsilon.$$



Ex 2: Marginal utility (Davis') price

Recall the EUM setup. For a continuous payoff $g \geq 0$ consider

$$V(\varepsilon, p_d) := \sup_{a \in \mathcal{A}} \mathbb{E}_\mu \left[u \left(-\varepsilon + \langle X, a \rangle + \frac{\varepsilon}{p_d} g(X) \right) \right],$$

Definition

Suppose that for each $p_d > 0$, the function $\varepsilon \mapsto V(\varepsilon, p_d)$ is differentiable at $\varepsilon = 0$ and $\hat{p}_d$ is a solution to

$$\partial_\varepsilon V(0, p_d) = 0.$$

Then $\hat{p}_d$ is called a marginal utility price of the option g.

Characterisation of the marginal utility price

UNIVERSITY OF OXFORD
Mathematical Institute

Theorem (Davis (1997))

Under mild technical assumptions $\hat{p}_d$ is unique and satisfies

$$\hat{p}_d = \frac{\mathbb{E}_\mu\left[u'(\langle X, a^\star\rangle) g(X)\right]}{\mathbb{E}_\mu\left[u'(\langle X, a^\star\rangle)\right]}.$$

In this way $\hat{p}_d$ is the price under a subjective martingale measure:

$$X = S_T - S_0 \quad \text{and} \quad \mathbb{E}_\mu\left[u'(\langle X, a^\star\rangle) X\right] = 0.$$

Robust marginal utility price

Mathematical Institute

Definition

Let us define

$$V(\delta, \varepsilon, p_d) = \sup_{a \in \mathcal{A}} \inf_{\nu \in B_\delta(\mu)} \mathbb{E}_\nu \left[u \left(-\varepsilon + \langle X, a \rangle + \frac{\varepsilon}{p_d} g(X) \right) \right].$$

Suppose that for each $p_d > 0$ the function $\varepsilon \mapsto V(\delta, \varepsilon, p_d)$ is differentiable. A number $\hat{p}_d(\delta)$, which satisfies

$$\partial_\varepsilon V(\delta, 0, \hat{p}_d(\delta)) = 0.$$

is called a robust marginal utility price of g at the uncertainty level δ.

Characterisation of DR marginal utility price

University of Oxford
Mathematical Institute

Theorem

Fix $\delta \geq 0, p_d > 0$. Under mild technical assumptions the robust marginal utility price $\hat{p}_d(\delta)$ is given by

$$\hat{p}_d(\delta) = \frac{\mathbb{E}_{\mu^\star}\left[u'(\langle X - X_0, a^\star_\delta\rangle)\, g(X)\right]}{\mathbb{E}_{\mu^\star}\left[u'(\langle X - X_0, a^\star_\delta\rangle)\right]}$$

for any pair of optimisers $a^\star_\delta \in \mathcal{A}$ and $\mu^\star \in B_\delta(\mu)$.

As before, $\hat{p}_d(\delta)$ is the price under a subjective martingale measure but which also depends on δ.

Characterisation of DR marginal utility price

UNIVERSITY OF OXFORD
Mathematical Institute

Theorem

Fix $\delta \geq 0, p_d > 0$. *Under mild technical assumptions the robust marginal utility price* $\hat{p}_d(\delta)$ *is given by*

$$\hat{p}_d(\delta) = \frac{\mathbb{E}_{\mu^\star}\left[u'(\langle X - X_0, a^\star_\delta\rangle)\, g(X)\right]}{\mathbb{E}_{\mu^\star}\left[u'(\langle X - X_0, a^\star_\delta\rangle)\right]}$$

for any pair of optimisers $a^\star_\delta \in \mathcal{A}$ *and* $\mu^\star \in B_\delta(\mu)$.

As before, $\hat{p}_d(\delta)$ is the price under a subjective martingale measure but which also depends on δ.

Special cases: $\hat{p}_d = \hat{p}_d(\delta)$ for all $\delta > 0$, e.g., for $\mu = \mathcal{N}(m, \sigma^2)$, $p = \infty$ and an agent with an exponential utility.

Sensitivity of the marginal utility price

University of Oxford
Mathematical Institute

Theorem

Under mild technical assumptions the following holds:

(i) *If $a^\star = 0$, then the Davis price $\hat{p}_d(\delta)$ satisfies*

$$\hat{p}_d'(0) = -\left(\mathbb{E}_\mu\left[|\nabla g(x)|^q\right]\right)^{1/q}.$$

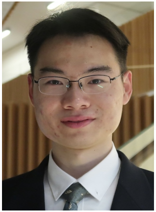
(ii) *If $a^\star \neq 0$ then*

$$\hat{p}_d'(0) = \frac{1}{\mathbb{E}_\mu\left[u'(\langle X, a^\star\rangle)\right]}\Bigg(\mathbb{E}_\mu\Big[u''(\langle X, a^\star\rangle)\cdot\Big(\langle T(X), a^\star\rangle - \langle X, a'(0)\rangle\Big)$$

$$\cdot\left(\mathbb{E}_{\hat{\mu}}\left[g(X)\right] - g(X)\right)\Big]\Bigg) - \mathbb{E}_{\hat{\mu}}\left[\langle \nabla g(X), T(X)\rangle\right],$$

where $\frac{d\hat{\mu}}{d\mu} \propto u'(\langle X, a^\star\rangle)$ *and* $T(x) \propto \frac{a^\star}{|a^\star|}|u'(\langle x, a^\star\rangle)|^{q-1}$.

Dynamic setting: Causal Wasserstein DRO



based on Bartl and Wiesel *SIFIN '23*, Jiang *arXiv:2401.16556*
and Jiang and O. *arXiv:2408.17109*

Sensitivity of causal DRO

Mathematical Institute

Let $p > 1$ and $1/p + 1/q = 1$. Take $c(x, y) = \|\Delta x - \Delta y\|^p$ for $p > 1$, where

$$\Delta(x_1, x_2, \ldots, x_N) = (x_1, x_2 - x_1, \ldots, x_N - x_{N-1}).$$

Write $\mathbb{D} = (\mathbb{D}_1, \ldots, \mathbb{D}_N)$ as the pullback of ∇ under Δ, i.e., $\mathbb{D}_n = \sum_{l \geq n} \partial_l$.

Sensitivity of causal DRO

University of Oxford
Mathematical Institute

Let $p > 1$ and $1/p + 1/q = 1$. Take $c(x, y) = \|\Delta x - \Delta y\|^p$ for $p > 1$, where

$$\Delta(x_1, x_2, \dots, x_N) = (x_1, x_2 - x_1, \dots, x_N - x_{N-1}).$$

Write $\mathbb{D} = (\mathbb{D}_1, \dots, \mathbb{D}_N)$ as the pullback of ∇ under Δ, i.e., $\mathbb{D}_n = \sum_{l \geq n} \partial_l$.
Under suitable assumptions, we have

$$\Upsilon := \lim_{\delta \to 0} \frac{v(\delta) - v(0)}{\delta} = L^* \left(\mathbb{E}_\mu \left[\sum_{n=1}^{N} |\mathbb{E}_\mu [\mathbb{D}_n f(X) | \mathcal{F}_n]|^q \right]^{1/q} \right) = L^* (\| {}^{\mathrm{o}}\mathbb{D} f \|_q).$$

Extensions

- Martingale constraint on the model.

$$\Upsilon_{\mathrm{Mart}} = L^*(\| {}^{\mathrm{o}}\mathbb{D}f - {}^{\mathrm{p}}\mathbb{D}f \|_2).$$

Extensions

- Martingale constraint on the model.

$$\Upsilon_{\mathrm{Mart}} = L^*(\| {}^{\mathrm{o}}\mathbb{D}f - {}^{\mathrm{p}}\mathbb{D}f\|_2).$$

- Pass limit to the continuous time!
  - Hyperbolic scaling — drift uncertainty.

$$c(x, y) = \lim_{N\to\infty} N^{p-1} \sum_{n=1}^{N} |\Delta x_n - \Delta y_n|^p = \|\partial_t(x - y)\|^p.$$

A *pathwise* Malliavin derivative leads to $\Upsilon = L^*(\| {}^{\mathrm{o}}\mathbb{D}f\|_q)$.

Extensions

- Martingale constraint on the model.

$$\Upsilon_{\mathrm{Mart}} = L^*(\| {}^{\mathrm{o}}\mathbb{D}f - {}^{\mathrm{p}}\mathbb{D}f\|_2).$$

- Pass limit to the continuous time!
  - Hyperbolic scaling — drift uncertainty.

$$c(x,y) = \lim_{N\to\infty} N^{p-1} \sum_{n=1}^{N} |\Delta x_n - \Delta y_n|^p = \|\partial_t(x-y)\|^p.$$

    A *pathwise* Malliavin derivative leads to $\Upsilon = L^*(\| {}^{\mathrm{o}}\mathbb{D}f\|_q)$.
  - Parabolic scaling — volatility uncertainty. Focus on $p=2$ and $\mu = \gamma$.

$$c(x,y) = \lim_{N\to\infty} \sum_{n=1}^{N} |\Delta x_n - \Delta y_n|^2 = [x-y]_T.$$

    An *extended* Skorokhod integral gives Υ_{Mart}.

AVaR of an exotic option

We consider $\mathbf{AVaR}_\alpha$ of an exotic option.

- $X = (X_1, X_2)$ – underlying asset.
- K – shifted strike price.
- $f(x) = (x_2 - x_1 + 1 - K)^+$ – payoff of the option.
- $(X_1, X_2) \sim (S_{0.5}, S_1)$. X follows the marginal distribution of a geometric Brownian motion S

$$\mathrm{d}S_t = \sigma S_t \,\mathrm{d}W_t, \quad S_0 = 1.$$

We take $\alpha = 0.95$, $\sigma = 0.2$, $c(x, y) = \|x - y\|^2$, and $L = +\infty \mathbf{1}_{(0.3^2, +\infty]}$.

AVaR of an exotic option

UNIVERSITY OF OXFORD
Mathematical Institute

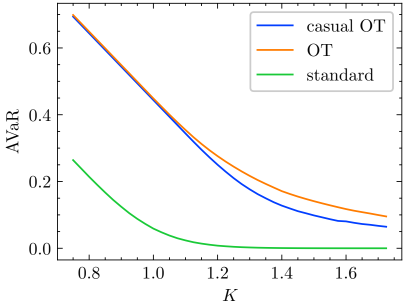


Figure: Comparison of AVaR for the option under a causal transport-type ambiguity (in blue), a classical transport-type ambiguity (in orange), and no ambiguity (in green). Take $\alpha = 0.95$, $\sigma = 0.2$, $c(x, y) = \|x - y\|^2$, and $L = +\infty \mathbf{1}_{(0.3^2, +\infty]}$.

Note that in some cases, there is no reduction in risk when restricting to a non-anticipative perturbation, e.g., $f(x) = (x_2 - K)^+ - (x_1 - K)^+$.

Asian option (disc. time sensitivity)

Mathematical
Institute

- A discrete-monitored Asian option with payoff

$$f(X) = \max\left\{0, \bar{X} - K\right\} \quad \text{with} \quad \bar{X} = \frac{1}{N}\sum_{n=1}^{N} X_n.$$

- Let μ be the reference risk-neutral measure.

Asian option (disc. time sensitivity)

UNIVERSITY OF OXFORD
Mathematical Institute

- A discrete-monitored Asian option with payoff

$$f(X) = \max\left\{0, \bar{X} - K\right\} \quad \text{with} \quad \bar{X} = \frac{1}{N}\sum_{n=1}^{N} X_n.$$

- Let μ be the reference risk-neutral measure.
- Notice that

$$\mathbb{D}_n f(X) = (N + 1 - n)\mathbf{1}_{\{\bar{X}>K\}}.$$

- The nonparametric 'Greek' of the Asian option is given by

$$\begin{aligned}\Upsilon_{\mathrm{Mart}} &= \left(\mathbb{E}_\mu\left[\sum_{n=1}^{N} \left|\mathbb{E}_\mu[\mathbb{D}_n f(X)|\mathcal{F}_n] - \mathbb{E}_\mu[\mathbb{D}_n f(X)|\mathcal{F}_{n-1}]\right|^2\right]\right)^{1/2} \\ &= \left(\mathbb{E}_\mu\left[\sum_{n=1}^{N}(N+1-n)^2\left|\mu(\bar{X} > K|\mathcal{F}_n) - \mu(\bar{X} > K|\mathcal{F}_{n-1})\right|^2\right]\right)^{1/2}.\end{aligned}$$

Merton's problem (cont. time sensitivity)

UNIVERSITY OF OXFORD
Mathematical Institute

- The stock follows the standard BS model, with S solving

$$dS_t = \zeta S_t dt + \sigma S_t dX_t.$$

- Agent's wealth process

$$dK_t^\theta = (r + \lambda\theta_t\sigma)K_t^\theta dt + \sigma\theta_t K_t^\theta dX_t,$$

  where $\lambda = (\zeta - r)/\sigma$, known as the market price of risk.
- Merton's problem of maximizing $\mathbb{E}[\log(K_T^\theta)]$ over θ is solved taking $\theta_t = \lambda/\sigma$.
- This gives $K_T^* = \kappa \exp((r + \lambda^2/2)T + \lambda X_T)$.

Merton's problem (cont., time sensitivity)

Mathematical Institute

- The stock follows the standard BS model, with S solving

$$dS_t = \zeta S_t dt + \sigma S_t dX_t.$$

- Agent's wealth process

$$dK_t^\theta = (r + \lambda\theta_t\sigma)K_t^\theta dt + \sigma\theta_t K_t^\theta dX_t,$$

where $\lambda = (\zeta - r)/\sigma$, known as the market price of risk.

- Merton's problem of maximizing $\mathbb{E}[\log(K_T^\theta)]$ over θ is solved taking $\theta_t = \lambda/\sigma$.
- This gives $K_T^* = \kappa \exp((r + \lambda^2/2)T + \lambda X_T)$.
- The general sensitivity to model uncertainty, around μ the Wiener measure, can be computed for

$$f(X) = \log(K_T^*) = \log(\kappa) + (r + \lambda^2/2)T + \lambda X_T.$$

Taking $p = 2$ and $L = +\infty\mathbf{1}_{(\sqrt{T},\infty)}$, we obtain $\Upsilon = \lambda\sqrt{T}$.

- The parametric sensitivity gives $\frac{\partial}{\partial\lambda}\mathbb{E}[\log(K_T^*)] = \Upsilon$.

OT & Data-Driven Approach: risk estimation example

$$(r_1, \ldots, r_N) \in \mathbb{R}^{dN} \quad \text{v.s.} \quad \hat{\mathbb{P}}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{r_i} \in \mathcal{P}(\mathbb{R}^d)$$

based on O. and Wiesel, *Ann. Stat.* **49**(1): 508–530, 2021.

Data set: historical returns

UNIVERSITY OF OXFORD
Mathematical Institute

Public information also includes historical stock returns. How can we use this information in a coherent and consistent way?

Data set: historical returns

Public information also includes historical stock returns. How can we use this information in a coherent and consistent way?

- Model specific: typically ignored. This is "physical measure" information hard to combine with "risk neutral measure"

Data set: historical returns

UNIVERSITY OF OXFORD
Mathematical Institute

Public information also includes historical stock returns. How can we use this information in a coherent and consistent way?

- Model specific: typically ignored. This is "physical measure" information hard to combine with "risk neutral measure"
- Robust approach: no $\mathbb{P}$ vs $\mathbb{Q}$ conflict.
  - indirect - agents can use to form beliefs/private information.
  - direct - non-parametric statistical estimation of superhedging prices (w/ Johannes Wiesel)

Take I: Plugin estimator

A simple setting: d assets, one-period, no other traded options.
Information: historical returns $r_1, \dots, r_N$ assumed i.i.d. from $\mathbb{P}$.

Aim: Build an estimator for

$$\pi^{\mathbb{P}}(\xi) = \inf \left\{ x \in \mathbb{R} \mid \exists H \in \mathbb{R}^d \text{ s.t. } \; x + H(r-1) \geq \xi(r) \; \mathbb{P}\text{-a.s.} \right\}$$

Take I: Plugin estimator

Mathematical Institute

A simple setting: d assets, one-period, no other traded options.
Information: historical returns $r_1, \dots, r_N$ assumed i.i.d. from $\mathbb{P}$.

Aim: Build an estimator for

$$\pi^{\mathbb{P}}(\xi) = \inf\left\{x \in \mathbb{R} \mid \exists H \in \mathbb{R}^d \text{ s.t. } \; x + H(r-1) \geq \xi(r) \; \mathbb{P}\text{-a.s.}\right\}$$

Theorem

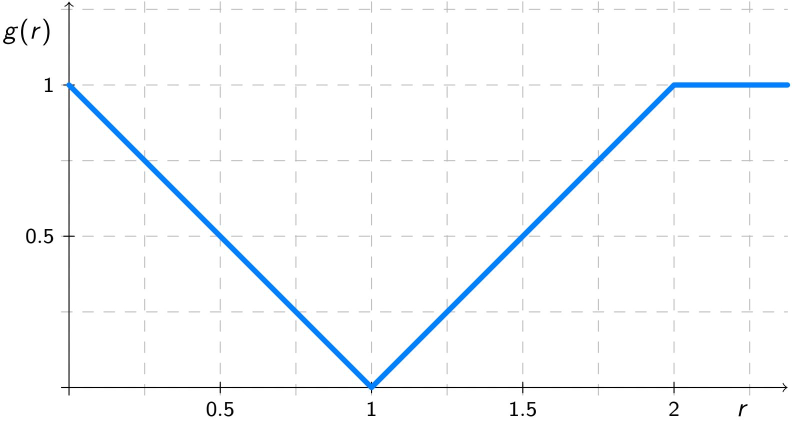
Let $\xi : \mathbb{R}_+^d \to \mathbb{R}$ be Borel-measurable. Define the empirical measure $\hat{\mathbb{P}}_N = \frac{1}{N}\sum_{i=1}^N \delta_{r_i}$. Then

$$\lim_{N\to\infty} \pi^{\hat{\mathbb{P}}_N}(\xi) = \pi^{\mathbb{P}}(\xi) \qquad \mathbb{P}^\infty\text{-a.s.},$$

where $\mathbb{P}^\infty$ denotes the product measure on $\Pi_{i=1}^\infty \mathbb{R}_+^d$.

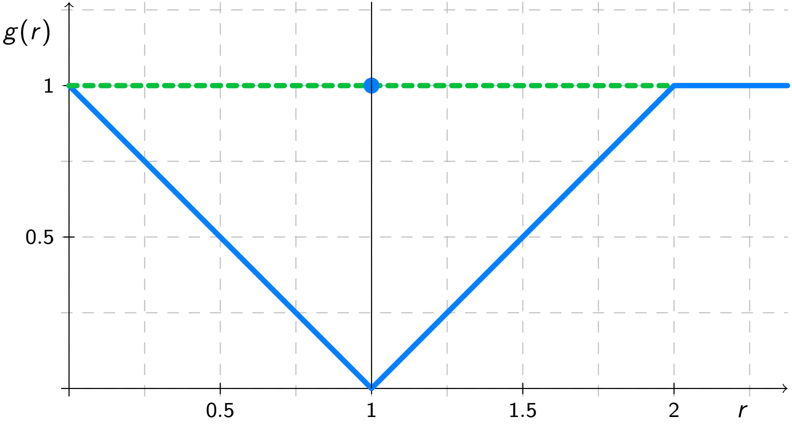
Example for consistency (1)

Let's take $\xi(r) = |r-1| \wedge 1$ and $\mathbb{P} = \frac{\lambda_{[0,2]}}{2}$.



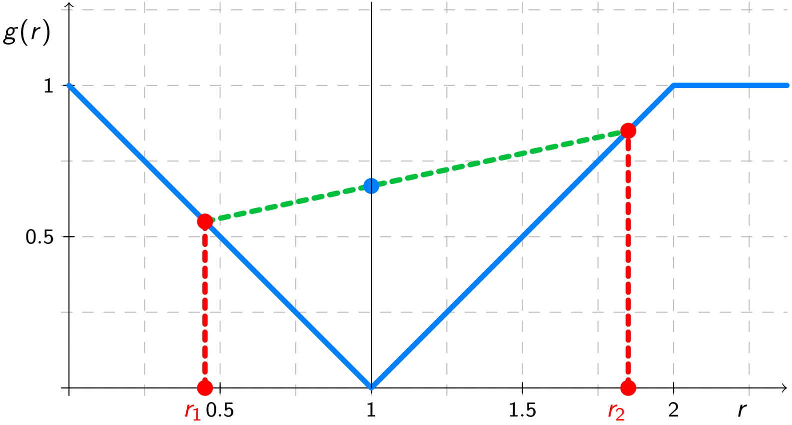
Example for consistency (2)

Let's take $\xi(r) = |r-1| \wedge 1$ and $\mathbb{P} = \frac{\lambda_{[0,2]}}{2}$.



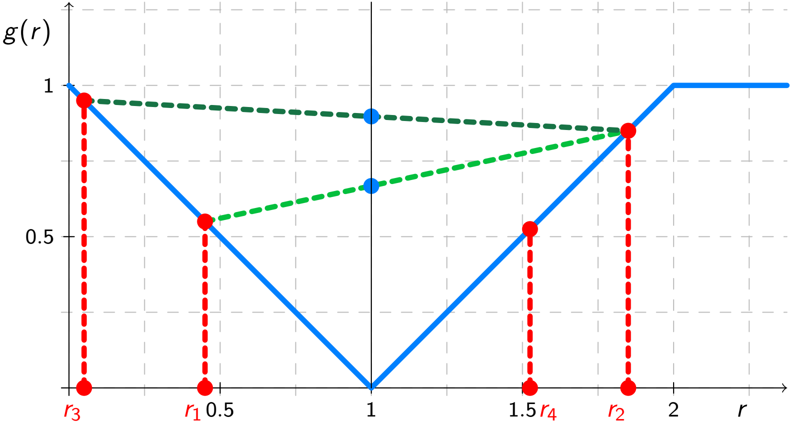
Example for consistency (3)

Let's take $\xi(r) = |r-1| \wedge 1$ and $\mathbb{P} = \frac{\lambda_{[0,2]}}{2}$.



Example for consistency (4)

Let's take $\xi(r) = |r-1| \wedge 1$ and $\mathbb{P} = \frac{\lambda_{[0,2]}}{2}$.



Concave envelope in two dimensions

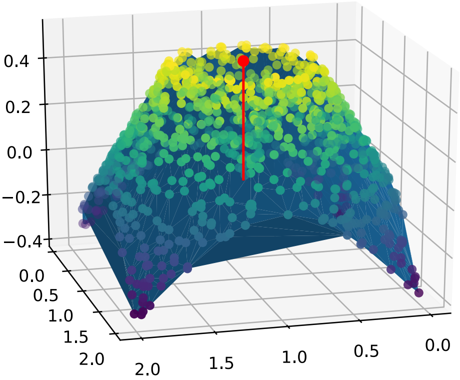


Figure: Concave envelope in 2 dimensions with $\mathbb{P} = \lambda|_{[0,2]^2}/4$, $\xi(r) = |r-1|\mathbb{1}_{\{|r-1|<1/2\}} + (1-|r-1|)\mathbb{1}_{\{|r-1|\geq 1/2\}}$

Problems with the plugin estimator

UNIVERSITY OF OXFORD
Mathematical Institute

The plugin estimator $\pi^{\hat{\mathbb{P}}_N}(\xi)$ is not robust!

- Not Financially: it underestimates the superhedging price $\pi^{\hat{\mathbb{P}}_N} \leq \pi^{\mathbb{P}}$.
- Not Statistically: (in the sense of Hampel). This applies to any estimator in fact:

Lemma

Let $\xi : \mathbb{R}_+^d \to \mathbb{R}$ be continuous and fix $\mathbb{P}$ on $\mathbb{R}_+^d$. Any consistent estimator T_N of $\pi^{\mathbb{P}}(\xi)$ is robust at $\mathbb{P}$ only if

$$\pi^{\mathbb{P}}(\xi) = \sup_{\mathbb{Q}\in\mathcal{M}} \mathbb{E}_{\mathbb{Q}}[\xi].$$

$\Longrightarrow$ need to control the support $\Longrightarrow$ robustness w.r.t. $\mathcal{W}^\infty$.

Positive results

- $\mathcal{W}^p$-approach.
- $\mathcal{W}^\infty$-robustness, estimating quantiles.
- Penalisation approach akin to risk measures.
- Convergence of superhedging strategies.
- Extension to law-invariant convex risk measures.
- Extension to multi-period models.

$\mathcal{W}^p$-approach

Fix $p \geq 1$. Assume we can find confidence bounds for the Glivenko-Cantelli theorem (see Dereich, Scheutzow, Schottstedt, 2011, Fournier, Guillin, 2013):

$$\mathbb{P}^N(\mathcal{W}^p(\mathbb{P}, \hat{\mathbb{P}}_N) \geq \varepsilon_N(\beta_N)) \leq \beta_N.$$

Definition

For a sequence $(k_N)_{N\in\mathbb{N}}$ such that $k_N \to \infty$ and $k_N = o(1/\varepsilon_N(\beta_N))$ we define

$$\hat{\mathcal{Q}}_N = \left\{ \mathbb{Q} \in \mathcal{M} \;\middle|\; \exists \nu \in B^p_{\varepsilon_N(\beta_N)}(\hat{\mathbb{P}}_N),\ \left\| \frac{d\mathbb{Q}}{d\nu} \right\|_\infty \leq k_N \right\}.$$

$\mathcal{W}^p$-approach: Consistency

Mathematical Institute

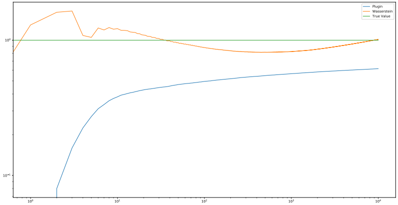
Theorem

Let g be Lipschitz continuous and bounded from below or continuous and bounded and $p \geq 1$. Pick a sequence $k_N = o(1/\varepsilon_N(\beta_N))$. Then

$$\lim_{N\to\infty} \sup_{\mathbb{Q}\in\hat{\mathcal{Q}}_N} \mathbb{E}_{\mathbb{Q}}[\xi] = \pi^{\mathbb{P}}(\xi) \qquad \mathbb{P}^\infty - a.s.,$$

if NA($\mathbb{P}$) holds.

Convergence of Wasserstein estimators



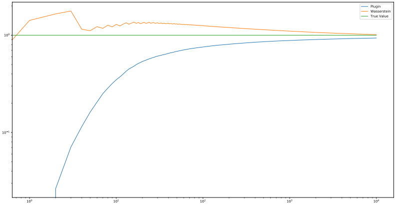


Figure: Wasserstein estimators with $g(r) = (1-r)\mathbb{1}_{\{r\leq 1\}} - \sqrt{r-1}\mathbb{1}_{\{r>1\}}$, $\mathbb{P} = \mathrm{Exp}(1)$ (left) and $g(r) = (r-2)^+$, $\mathbb{P} = \exp(\mathcal{N}(0,1))$ (right).

Robust Superhedging Price estimator

Take $k_N \to \infty$ and $k_N \varepsilon_N(\beta_N) \to 0$. Let

$$\pi_{\hat{\mathcal{Q}}_N}(\xi) = \sup_{\mathbb{P} \in B^p_{\varepsilon_N}(\hat{\mathbb{P}}_N)} \sup_{\mathbb{Q} \in \mathcal{M}: \|d\mathbb{Q}/d\mathbb{P}\|_\infty \le k_N} \mathbb{E}_{\mathbb{Q}}[\xi]$$

Robust Superhedging Price estimator

Take $k_N \to \infty$ and $k_N \varepsilon_N(\beta_N) \to 0$. Let

$$\pi_{\hat{\mathbb{Q}}_N}(\xi) = \sup_{\mathbb{P}\in B^p_{\varepsilon_N}(\hat{\mathbb{P}}_N)} \sup_{\mathbb{Q}\in\mathcal{M}:\|d\mathbb{Q}/d\mathbb{P}\|_\infty \le k_N} \mathbb{E}_{\mathbb{Q}}[\xi]$$

$$= \sup_{\mathbb{P}\in B^p_{\varepsilon_N}(\hat{\mathbb{P}}_N)} \sup_{\|d\mathbb{Q}/d\mathbb{P}\|_\infty \le k_N} \inf_{H\in\mathbb{R}^d} \mathbb{E}_{\mathbb{Q}}[\xi - H(r-1)]$$

$$= \inf_{H\in\mathbb{R}^d} \sup_{\mathbb{P}\in B^p_{\varepsilon_N}(\hat{\mathbb{P}}_N)} \sup_{\|d\mathbb{Q}/d\mathbb{P}\|_\infty \le k_N} \mathbb{E}_{\mathbb{Q}}[\xi - H(r-1)]$$

$$= \inf_{H\in\mathbb{R}^d} \sup_{\mathbb{P}\in B^p_{\varepsilon_N}(\hat{\mathbb{P}}_N)} AV@R^{\mathbb{P}}_{\frac{k_N-1}{k_N}}(\xi - H(r-1))$$

$$= \inf\left\{x\in\mathbb{R} \mid \exists H\in\mathbb{R}^d \text{ s.t. } \sup_{\mathbb{P}\in B^p_{\varepsilon_N}(\hat{\mathbb{P}}_N)} AV@R^{\mathbb{P}}_{\frac{k_N-1}{k_N}}(\xi - H(r-1) - x) \le 0\right\}$$

$\mathcal{W}^p$-approach: Robustness

Definition

Let $\mathfrak{P}, \tilde{\mathfrak{P}} \subseteq \mathcal{P}(\mathbb{R}^d_+)$. We define p-Wasserstein-Hausdorff metric

$$\mathcal{W}^p(\mathfrak{P}, \tilde{\mathfrak{P}}) = \max\left(\sup_{\mathbb{P}\in\mathfrak{P}} \inf_{\tilde{\mathbb{P}}\in\tilde{\mathfrak{P}}} \mathcal{W}^p(\mathbb{P}, \tilde{\mathbb{P}}), \sup_{\tilde{\mathbb{P}}\in\tilde{\mathfrak{P}}} \inf_{\mathbb{P}\in\mathfrak{P}} \mathcal{W}^p(\mathbb{P}, \tilde{\mathbb{P}})\right).$$

Theorem

The estimator $\sup_{\mathbb{Q}\in\hat{\mathcal{Q}}_N} \mathbb{E}_{\mathbb{Q}}[g]$ *is robust with respect to the* $\mathcal{W}^p$ *in the sense that*

$$\sup_{g\in\mathcal{L}_1} \left| \sup_{\mathbb{Q}\in\hat{\mathcal{Q}}^1_N} \mathbb{E}_{\mathbb{Q}}[g] - \sup_{\mathbb{Q}\in\hat{\mathcal{Q}}^2_N} \mathbb{E}_{\mathbb{Q}}[g] \right| \leq \mathcal{W}^p(\hat{\mathcal{Q}}^1_N, \hat{\mathcal{Q}}^2_N),$$

where $\hat{\mathcal{Q}}^i_N$ *are defined corresponding to* $\mathbb{P}^i \in \mathcal{P}(\mathbb{R}^d_+)$, $i = 1, 2$.

Superhedging with respect to risk measures (1)

Consider $\rho_{\mathbb{P}}$ with Kusuoka representation:

$$\rho_{\mathbb{P}}(\xi) = \sup_{\mu \in \mathfrak{P}} \int_0^1 \mathsf{AV@R}_\alpha^{\mathbb{P}}(\xi) d\mu(\alpha)$$

for a set $\mathfrak{P}$ of probability measures on $[0,1]$ ($\Rightarrow$ law-invariant coherent risk measures). Introduce

$$\pi^{\rho}_{B^p_{\varepsilon_N(\beta_N)}(\hat{\mathbb{P}}_N)}(\xi)$$

$$:= \inf \left\{ x \in \mathbb{R}^d \;\middle|\; \exists H \in \mathbb{R}^d \text{ s.t. } \sup_{\nu \in B^p_{\varepsilon_N(\beta_N)}(\hat{\mathbb{P}}_N)} \rho_\nu(\xi - x - H(r-1)) \leq 0 \right\}.$$

Superhedging with respect to risk measures (2)

Consistency

UNIVERSITY OF OXFORD
Mathematical Institute

Theorem

Assume g satisfies $|\xi(r) - \xi(\tilde{r})| \leq L_\gamma |r - \tilde{r}|^\gamma$ *for some* $\gamma \leq 1$ *and* $L_\gamma \in \mathbb{R}$. *Then*

$$\lim_{n\to\infty} \pi^{\rho}_{B^{p}_{\varepsilon_N(\beta_N)}(\hat{\mathbb{P}}_N)}(\xi) = \pi^{\rho_{\mathbb{P}}}(\xi) \qquad \mathbb{P}^{\infty}\text{-}a.s.$$

Plugin estimator and option prices

Mathematical Institute

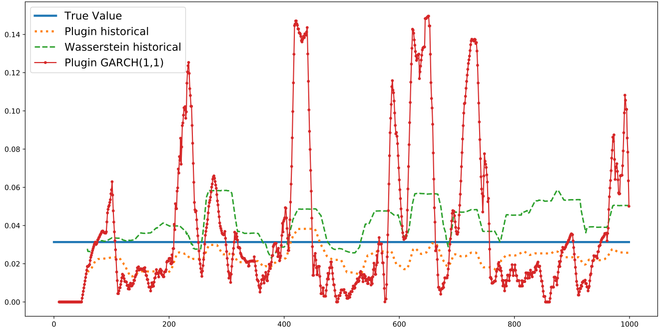
Corollary

Let $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d_+)$ and $\xi : \mathbb{R}^d_+ \to \mathbb{R}$ be Borel-measurable. In addition to the assets S, assume that there are $\tilde{d}$ traded options with continuous payoffs $f_1(r)$ and prices f_0 in the market. Then, if the observations $r_1, r_2, \dots$ are i.i.d. samples from $\mathbb{P}$, and under NA, we have

$$\lim_{N\to\infty} \inf\{x \in \mathbb{R} \mid \exists H, \tilde{H} \text{ s.t. } x + H(r_i - 1) + \tilde{H}(f_1 - f_0) \geq \xi(r_i)\ \forall i = 1, \dots, N\}$$

$$= \sup_{\mathbb{Q}\sim\mathbb{P},\ \mathbb{Q}\in\mathcal{M},\ \mathbb{E}_{\mathbb{Q}}(f_1)=f_0} \mathbb{E}_{\mathbb{Q}}[\xi].$$

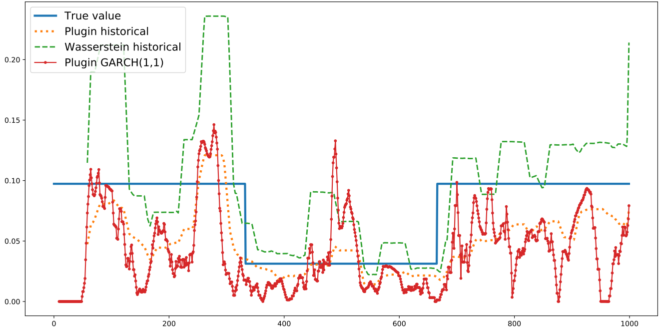
Estimates for $\pi^{\mathrm{AV@R}_{0.95}^{\tilde{\mathbb{P}}}}((r-1)^+)$



Rolling window of 50 data points, average of the last 10 estimates.
The data is from $\mathbb{P} \sim \mathrm{GARCH}(1,1)$.

Estimates for $\pi^{\mathrm{AV@R}^{\tilde{\mathbb{P}}}_{0.95}}((r-1)^+)$

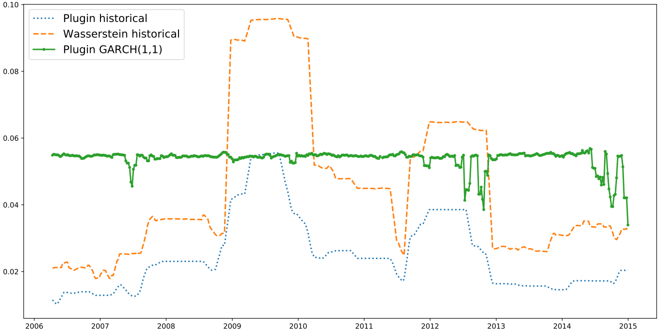
University of Oxford
Mathematical Institute



Rolling window of 50 data points, average of the last 10 estimates.
The data is from $\mathbb{P} \sim \mathrm{GARCH}(1,1)$.

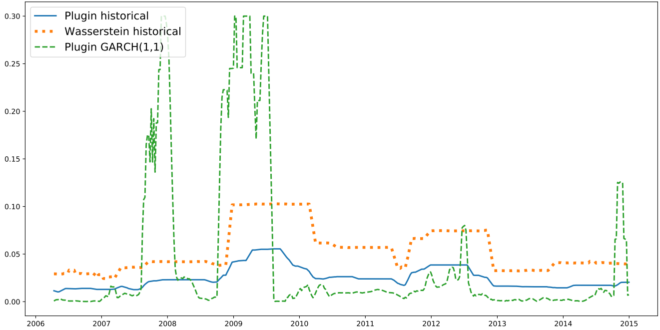
Estimates for $\pi^{\mathrm{AV@R}^{\tilde{\mathbb{P}}}_{0.95}}((r-1)^+)$

Mathematical Institute



Rolling window of 50 data points, average of the last 5 estimates.
Weekly S&P500 returns.

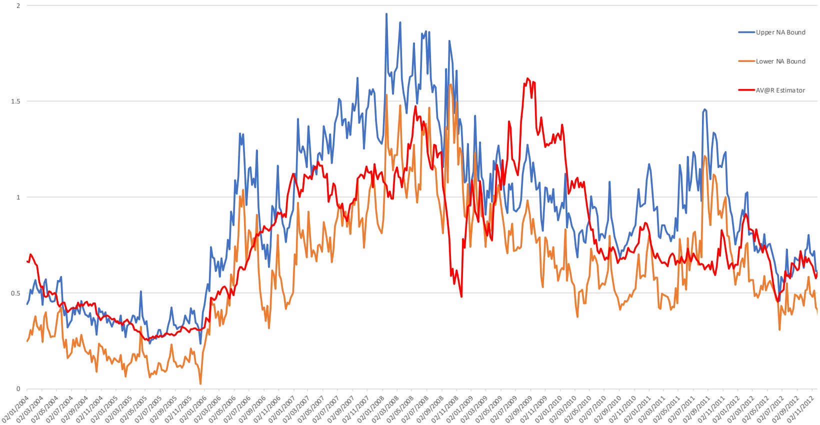
Estimates for $\pi^{\mathrm{AV@R}^{\tilde{\mathbb{P}}}_{0.95}}((r-1)^+)$



Rolling window of 50 data points, average of the last 5 estimates.
Weekly S&P500 log-returns.

Estimation divergence as an information signal

Mathematical
Institute



Tyssen ATM 1W Call: AV@R Estimator vs Bloomberg's IVol Synthetic bounds.

Conclusions

UNIVERSITY OF OXFORD
Mathematical Institute

- Robust approach builds risk estimates from market data without any modelling assumptions.
- OT allows to conceptualise and quantify the impact of model uncertainty
- Data/Information is used to endogenously specify models.
- The case of information on traded options' prices leads to an Optimal Transport problem with a martingale constraint. We develop numerical methods to solve it.
- DRO conceptually appealing. Applications in finance, statistics, UQ, ML and more!
- Wasserstein balls lead to statistical estimators for robust outputs directly from historical returns

University of Oxford
Mathematical Institute

Thank You

list of references to follow
some papers available at `www.maths.ox.ac.uk/people/jan.obloj`