Big Data Portfolio Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Drawing on statistical learning theory, we propose a robust portfolio optimization mechanism agnostic to market distribution. In particular, our algorithm returns a linear investment policy based on the risk preferences of the investor, with guaranteed out of sample performance. We also provide guidance for big-data scenarios, ie. when the number of features is of the order of the size of the sample, thus enriching the learning theory literature. **[We conclude by contrasting small- and big-data scenario in a real application.]**

1 Introduction

Ever since it was formally theorized by Markowitz [4], one-step theoretical portfolio management has mostly kept the same approach: maximize the returns while minimizing the variance using a trade-off parameter. However, such an approach suffers from a fatal flaw, as it needs to make asumptions on the underlying distribution of the returns. While Markowitz considered gaussian returns, others have investigated more sophisticated distributions, using eg. jump diffusion, gamma returns, etc. **[citations needed.]** On the other side, [2] exhibits a clever stock-picking algorithm with asymptotic performance guarantees **[retravailler, plus de détails]**. However, unlike classical portfolio theory, this method assumes a risk-neutral behaviour, ie. an investor wishing only to maximize profits, with no regards to the possible loss. **[Et ici aussi.]**

This work is an attempt at bridging these two concepts. Using a size n sample of the (unknown) market distribution consisting of market features and market returns, a regularization parameter λ and by specifying an arbitrary concave utility function, we can derive an in-sample optimal linear investment policy by optimizing the certainty equivalent on the sample. We first show that the out-sample performance of the policy is bounded by a $O(1/\sqrt{n})$ error term. Second, We also investigate how this this method scales when the number of market features p is of the order of n, ie. in a *big-data* regime, and show that the performance scales linearly in the number p of available features. As far as we are aware, this situation has not been studied by the learning theory, and consequently we hope to enrich the field. **[Remanier.]** Finally, we determine the conditions under which the true optimal solution in regard to the market distribution can be attained. **[We conclude by presenting numerical results from different degenerated distributions.]**

The *market* considered by this document could be any asset traded on the market.**[Incorporer quelque part.]**

At a higher level, this document should be mostly understood as providing guidance to portfolio managers who would wish to incorporate general statistical and machine learning strategies in order to uncover market returns indicators. In fact, as more and more features are poured into a model (for example by considering polynomial kernels **[reference needed]**), there is real possibility that the out-sample performance becomes degraded, and we wish to show how it can be prevented.

Most of this work derives from statistical learning theory, and in particular from stability theory, as exposed by Bousquet and Elisseef in their seminal paper [1]. The author showed, using powerful

Submitted to 29th Conference on Neural Information Processing Systems (NIPS 2016). Do not distribute.

concentration inequalities, how the empirical risk minimization of a Lipschitz loss function with additional convexity driven by a ℓ_2 regularization on the decision would converge in the size of the sample toward the out-sample performance. In particular, their results were a departure from classical learning theory as the tools they were using stems strictly from algorithmic and convexity analysis.

We also improve on results from [6] **[anonynimize?]** who study the application of learning theory to a feature based newsvendor problem. However, while they explicitly consider the big-data regime, we believe our model is more general in the sense that we directly show the effects of p on the performance of the algorithm.

[Padder davantage, plus de details sur 1. theorie moderne de portefeuille, 2. portefeuille universel, 3. Theorie de la stabilité, 4. Donner plus de références.]

2 Model and Main Results

2.1 Assumptions and definitions

[Nécesaire?] In the following, $\boldsymbol{A}$ (capital boldface) are assumed to represent a real subset of any dimension, A (capital case) represents random variables (or distributions) and a (lower case) represents deterministic variables or realizations. $\mathscr{R}$ represents the real set, $\underset{\sim}{\subset}$ the support of a random variable, and $\|\cdot\|$ is the euclidean norm.

Our model considers the market M as being a $p+1$-variate random distribution, with on its first margin a random (finite) return $R \underset{\sim}{\subset} \boldsymbol{R} = [r_{\min}, r_{\max}] \subset \mathscr{R}$ **[Il peut être plus simple d'avoir $|R| \leq \bar{r}$, notamment dans l'expression de Ω]** and on the other margin a random vector of features $(X_1, \ldots, X_p)$, which would typically represent financial or economic news, etc. We will assume that all features are pairwise independant.

We also suppose that the investor is endowed with a monotonically increasing concave utility function $\bar{u} : \boldsymbol{R} \to \boldsymbol{U}$, such that $\bar{u}$ can be rescaled to u with $\bar{u}(r) = ku(r) + l$, with the additionnal requirements that $u(0) = 0$, $\lim_{r\to 0^+} \nabla u(r) = 1$ and that u is γ-Lipschitz, ie. such that for any $r_1, r_2 \in \boldsymbol{R}$, $|u(r_1) - u(r_2)| \leq \gamma |r_1 - r_2|$. For example, any piece-wise linear utility would fit the Lipschitz requirements.

Our method studies optimal linear investment decisions $q \in \boldsymbol{Q} \underset{\sim}{\subseteq} \mathscr{R}^p$ over the random features so as to maximize the certainty equivalent $\mathrm{CE}(q)$ of the portfolio, where:

$$\mathrm{CE}(q) = u^{-1}(\Psi(q)),$$

with

$$\Psi(q) = \boldsymbol{E}_M[u(R\, q^T X)]$$

the out-sample utility of q. We would typically add a riskless return rate to the equation, however we set it to 0 for the sake of simplicity. Additionnally, given a sample $\{(x_i, r_i)\}_{i=1}^n$ drawn from M^n, we will also study the sample certainty equivalent $\hat{\mathrm{CE}}$:

$$\hat{\mathrm{CE}}(q) = u^{-1}(\hat{\Psi}(q)),$$

with

$$\hat{\Psi}(q) = n^{-1} \sum_{i=1}^{n} u(r_i\, q^T x_i)$$

the in-sample utility of q.

2.2 Out-sample complexity

Supposing we have a size n sample drawn *i.i.d.* from the market, then a natural choice for q would be the optimal solution of the regularized in-sample utility:

$$\hat{q} = \arg\max_q \{\hat{\Psi}(q) - \lambda \|q\|^2\} = \arg\max_q \left\{ \frac{1}{n} \sum_{i=1}^{n} u(r_i\, q^T x_i) - \lambda \|q\|^2 \right\},$$

where $\lambda\|q\|^2$ is here to avoid overfitting on the training sample. We will sometimes refer to $\hat{q}$ as the algorithm mapping from a market sample to the decision vector, rather than the decision vector itself.

Before going on, we will assume that the random features vector is bounded, ie. $\|X\| \leq \xi$, although we will relax this hypothesis in the next subsection.

We are now in a position to present a bound on the out-sample error:

Theorem 1. *With probability $1-\delta$, the error between the in- and out-sample certainty equivalent is bounded by the following relation:*

$$\mathrm{CE}(\hat{q}) \geq \hat{\mathrm{CE}}(\hat{q}) - \Omega \cdot \nabla u^{-1}(\hat{\mathrm{CE}}(\hat{q})),$$

where

$$\Omega = \frac{(\bar{r}\xi)^2}{2\lambda}\left(\frac{\gamma^2}{n} + \frac{\gamma(1+3\gamma)}{\sqrt{2n}}\sqrt{\log(1/\delta)}\right).$$

In particular, this implies that the error bound shrinks at a $O(1/\sqrt{n})$ rate.

Our proof of Theorem 1 proceeds as follow. First, borrowing from the terminology introduced by [1], we show that the algorithm leading to $\hat{q}$ is β-stable. We then show that for any $\hat{q}$ generated from a sample of M, the utility derived from applying this decision will be absolutely bounded, regardless of the outcome from M. These last two conditions can therefore lead to a direct application of Bousquet-Ellisseef out-sample error bound theorem on the $\boldsymbol{U}$ space. We finally show how this result can be inverted back to the $\boldsymbol{R}$ space.

Lemma 1. *Let $q_1, q_2 \in \boldsymbol{Q}$, $x \sim X$ and $r \sim R$. The algorithm generating $\hat{q}$ has σ-admissiblity of $\gamma\bar{r}$, ie.*

$$|u(r\, q_1^T x) - u(r\, q_2^T x)| \leq \gamma\bar{r}|q_1^T x - q_2^T x|.$$

Proof. This lemma follows trivially from the Lipschitz property of u. See Definition 19 from [1] for more details. □

Lemma 2. *The algorithm generating $\hat{q}$ has β-stability, ie. with $s_n \sim M^n$ and s_n' differing from s_n by a single resampling from M, then,*

$$|u(R\,\hat{q}(s_n)^T X) - u(R\,\hat{q}(s_n')^T X)| \leq \beta,$$

where

$$\beta \leq \frac{(\gamma\bar{r}\xi)^2}{2\lambda n}.$$

Proof. Using Lemma 1, this follows directly from Theorem 22 in [1]. □

Lemma 3. *The norm of the decision $\hat{q}$ is bounded:*

$$\|\hat{q}\| \leq \frac{\gamma\bar{r}\xi}{2\lambda}.$$

Proof. Let $\{(x,r)\}_n \sim M^n$ be an *i.i.d.* sample of the market. The empirical decision algorithm is equivalent to

$$\begin{array}{ll} \text{maximize} & n^{-1}\sum_{i=1}^{n} u(r_i\, q^T x_i) - \lambda s^2 \\ \text{subject to} & s \geq 0 \\ & \|q\| = 1, \end{array}$$

where the optimization variables are now the direction q and the scale s. Therefore, for any direction q, we can define a concave function $g(s)$ which becomes the objective:

$$\begin{array}{ll} \text{maximize} & g(s) = n^{-1}\sum_{i=1}^{n} u(r_i\, s q^T x_i) - \lambda s^2 \\ \text{subject to} & s \geq 0. \end{array}$$

Because $g : \mathscr{R}_+ \to \mathscr{R}$ is concave, we can consider two cases: either the maximum is realized at the boundary, ie. $s^\star = 0$, or there exists an optimal value $s^\star > 0$ such that $\nabla g(s^\star) = 0$. To derive a bound on $s^\star$, we can seek a value $\bar{s}$ such that for any q, $\nabla g(\bar{s}) < 0$ and therefore $s^\star < \bar{s}$.

To do so, we first note that

$$\nabla g(s) = n^{-1} \sum_{i=1}^{n} r_i \, q^T x_i \, u'(r_i \, s q^T x_i) - 2\lambda s,$$

bu because $\|q\| = 1$, we have $q^T x_i \leq \|x_i\| \leq \xi$. We also have $r_i \leq \bar{r}$ and $u' \leq \gamma$, so that

$$\nabla g(s) \leq \gamma \bar{r} \xi - 2\lambda s.$$

Therefore, with

$$\bar{s} = \frac{\gamma \bar{r} \xi}{2\lambda},$$

we have $\nabla g(\bar{s}) \leq 0$. □

Lemma 4. *For any $(x, r) \sim M$ and any $\hat{q}$,*

$$-\frac{(\gamma \xi \bar{r})^2}{2\lambda} \leq u(r \, \hat{q}^T x) \leq \frac{\gamma (\xi \bar{r})^2}{2\lambda}.$$

Proof. The maximum utility will be realized when $r = \bar{r}$, so that

$$u(r \, \hat{q}^T x) \leq r \hat{q}^T x \leq \frac{\gamma (\xi \bar{r})^2}{2\lambda},$$

since the identity function bounds u above. Likewise for negative returns, although this time γ applies. □

The following theorem was first proven in [1], although its statement is adapted from [5] and is presented in accordance to our particular setting.

Theorem (Bousquet-Ellisseef Outsample Error Theorem). *Let $s_n = \{(x_i, r_i)\}_{i=1}^n$ by a size n sample drawn i.i.d. from M. If $\hat{q}$ has β-stability and $\hat{u}_{\min} \leq u(R \, \hat{q}^T X) \leq \hat{u}_{\max}$, then, with probability $1 - \delta$,*

$$\Psi(\hat{q}) \geq \hat{\Psi}(\hat{q}) - \Omega_u,$$

where

$$\Omega_u = \beta + (2n\beta + (\hat{u}_{\max} - \hat{u}_{\min})) \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Using directly Lemma 2 and 4, we therefore find the following outsample error bound on the utility

$$\Omega_u = \frac{(\bar{r}\xi)^2}{2\lambda} \left(\frac{\gamma^2}{n} + \frac{\gamma(1 + 3\gamma)}{\sqrt{2n}} \sqrt{\log(1/\delta)} \right).$$

We now show how to transform this last result on a bound on the CE of the decision. Note that for any convex function f, $f(a+b) \geq f(a) + b \cdot \nabla f(a)$. Therefore, from the out-sample error bounding theorem, we have

$$u^{-1}(\Psi(\hat{q})) \geq u^{-1}(\hat{\Psi}(\hat{q}) - \Omega_u) \geq u^{-1}(\hat{\Psi}(\hat{q})) - \Omega_u \cdot \nabla(u^{-1})(\hat{\Psi}(\hat{q})),$$

since u^{-1} is also a monotonic function. This proves Theorem 1.

2.3 Big Data Phenomenon

We now take a closer look on the effect the dimension of the feature space can have on the bound Ω stated in Theorem 1, and in particular on the bound ξ^2. If we let $Z^2 = \sum_{i=1}^n X_i^2$ be the random squared norm of X, we can show that Z^2 is of the order $O(p)$ with high probabiltiy . This implies that the algorithm $\hat{q}$ has in fact a sample complexity $O(p/\sqrt{n})$.

We present three cases, each with additional generalization properties. In what follows, we will assume with no loss of generality (because it is an affine transformation) that $\boldsymbol{E}X_i = 0$ and $\operatorname{Var} X_i = 1$, which already implies that $\boldsymbol{E}X_i^2 = 1$, and therefore $\boldsymbol{E}Z^2 = p$.

[Ajouter de l'intuition pour le lecteur.]

Let us first consider the specific case where $X \sim \mathscr{N}(0, I)$, ie. X is a p-mutlinormal random vector. It then follows that $Z^2 \sim \chi^2(p)$. But we know from [3] that a chi-square distribution has the following property for all t:

$$\boldsymbol{P}\{Z^2 - p \geq 2\sqrt{pt} + 2t\} \leq e^{-t},$$

which is equivalent, with probability $1-\delta$ to:

$$Z^2 < p + 2\sqrt{p\log(1/\delta)} + 2\log(1/\delta).$$

As a somewhat more natural example, without making any asumption on the distribution of the features, we can consider the case where each of them is bounded, either by truncation in the pre-processing step or because their support is known to be finite. If $X_i^2 \leq \nu_i$, and we let $\nu_0^2 = \sum_{i=1}^p \nu_i^2$, then, by Hoeffding's theorem,

$$\boldsymbol{P}\{Z^2 - p \geq t\} \leq \exp\left(-\frac{t^2}{\nu_0^2}\right),$$

which, again, can be reexpressed as the following inequality with probability $1-\delta$:

$$Z^2 < p + \nu_0\sqrt{\log(1/\delta)}.$$

Finally, Markov's inequality provides the most general theorem for the situation, since it simply states that with probability $1-\delta$,

$$Z^2 < \delta^{-1}p.$$

Theorem 2. *With probability $1-(\delta_1+\delta_2)$, the error between the in- and out-sample certainty equivalent is bounded by the following relation:*

$$\mathrm{CE}(\hat{q}) \geq \hat{\mathrm{CE}}(\hat{q}) - \Omega \cdot \nabla u^{-1}(\hat{\mathrm{CE}}(\hat{q})),$$

where

$$\Omega = \frac{\bar{r}^2 p}{2\lambda\delta_1}\left(\frac{\gamma^2}{n} + \frac{\gamma(1+3\gamma)}{\sqrt{2n}}\sqrt{\log(1/\delta_2)}\right).$$

If the features are bounded, then

$$\Omega = (p + \nu_0\sqrt{\log(1/\delta_1)})\frac{\bar{r}^2}{2\lambda}\left(\frac{\gamma^2}{n} + \frac{\gamma(1+3\gamma)}{\sqrt{2n}}\sqrt{\log(1/\delta_2)}\right).$$

In particular, this implies that the error bound shrinks at a $O(p/\sqrt{n})$ rate.

2.4 Market efficiency and true optimal

The last theoretical topic we want to discuss is how our model relates to the theory of market efficiency.

Definition. Let $q^\star = \arg\max_q \Psi(q)$. Then M is said to be efficient with respect to u if $\|q^\star\|$ is bounded.

3 Old.

Assumption. The random return has a finite support, ie. $R \subsetneq \boldsymbol{R} \subseteq [r_{\min}, r_{\max}]$. Additionally, $|R| \leq \bar{r}$.

Assumption. The portfolio manager is endowed with an utility function $\bar{u} : \boldsymbol{R} \to \boldsymbol{U}$ with these properties:

- $\bar{u}$ can be reexpressed as $\bar{u}(r) = ku(r) + l$, $k > 0$, with $u(0) = 0$ and $\lim_{r\to 0^+} u'(r) = 1$.
- $u(r) = o(r)$, ie. the investor is risk-averse;

- $|u(r_1) - u(r_2)| \leq \gamma |r_1 - r_2|$, ie. u is γ-Lipschitz;
- u is monotonically increasing;
- With $u(r) = u_-(r)\mathbf{1}_{\{r<0\}} + u_+(r)\mathbf{1}_{\{r\geq 0\}}$, then $u_+(r) = o(u_-(r))$. In other words, u_- decreases faster than u_+ increases.

Definition. Let $\ell : \boldsymbol{M} \times \boldsymbol{Q} \to \boldsymbol{U}$ be a loss function defined by

$$\ell(m, q) = \ell(x, r, q) = -u(rq^T x)$$

where r_0 is the risk free return rate. We also define the cost function $c : \boldsymbol{I} \times \boldsymbol{R} \to \boldsymbol{U}$ as

$$c(p, r) = -u(rp),$$

so that $\ell(x, r, q) = c(q^T x, r)$.

Definition. The in-sample risk $\hat{R} : \boldsymbol{M}^n \times \boldsymbol{Q} \to \boldsymbol{U}$ associated with decision q and market sample μ_n is given by

$$\hat{R}_{\mu_n}(q) = n^{-1} \sum_{i=1}^{n} \ell(m_i, q).$$

Definition. The empirical decision algorithm $\hat{A}_n : \boldsymbol{M}^n \to \boldsymbol{Q}$ associated with market sample μ_n is the optimal value of the problem

$$\text{minimize} \quad \hat{R}_{\mu_n}(q) + \lambda \|q\|^2.$$

From now on, as a notation shortcut, let $\hat{q}_n := \hat{A}_n(\mu_n)$ the in-sample decision associated with random market sample μ_n and $\hat{R} := \hat{R}_{\mu_n}$ the in-sample risk function.

Definition. The in-sample certainty equivalent $\hat{\mathrm{CE}} : \boldsymbol{M}^n \times \boldsymbol{Q} \to \boldsymbol{R}$ associated with decision q and market sample μ_n is given by

$$\hat{\mathrm{CE}}(q) = ku^{-1}(-\hat{R}(q)) + l.$$

Definition. The true risk $R : \boldsymbol{Q} \to \boldsymbol{U}$ associated with decision q is given by

$$R(q) = \boldsymbol{E}\ell(M, q).$$

Definition. The true certainty equivalent CE associated with decision q is given by

$$\mathrm{CE}(q) = ku^{-1}(-R(q)) + l.$$

3.1 Performance Bounds

We are concerned about how the in sample performance can deviate from the expected out sample performance, that is we want to identify f_1 such that

$$\mathrm{CE}(\hat{q}) \geq \hat{\mathrm{CE}}(\hat{q}) - f_1(n, p, \lambda)$$

with high probability. We are also interested in the suboptimality of the problem, namely the function f_2 such that

$$\mathrm{CE}(q^\star) \geq \mathrm{CE}(\hat{q}) - f_2(n, p, \lambda),$$

also with high probability.

The following theorem is adapted from [1], and is the starting point of our analysis.

Theorem 3. *The in-sample and out-sample performance of the algorithm given by $\hat{q}$ is bounded by the following expression with probability $1 - \delta$:*

$$\begin{aligned} R(\hat{q}) &\leq \hat{R}(\hat{q}) + \frac{(\gamma \bar{r} \xi)^2}{2\lambda n} + \left(\frac{(\gamma \bar{r} \xi)^2}{\lambda} + \frac{\gamma(\gamma+1)\xi^2 \, r_{\max} \bar{r}}{2\lambda} \right) \sqrt{\frac{\log 1/\delta}{2n}} \\ &:= \hat{R}(\hat{q}) + \Omega. \end{aligned}$$

Proof. **[See claim ????? for further details.]** $\square$

Theorem 4. *The following inequality holds with probability* $1-\delta$*:*

$$\mathrm{CE}(\hat{q}) \geq \hat{\mathrm{CE}}(\hat{q}) + ku^{-1}(\Omega).$$

Proof. The following steps follow directly from the monotonicity, convexity, and superadditivity of u^{-1} :

$$\begin{aligned}
& R(\hat{q}) \leq \hat{R}(\hat{q}) + \Omega \\
\iff & -R(\hat{q}) \geq -\hat{R}(\hat{q}) - \Omega \\
\iff & u^{-1}(-R(\hat{q})) \geq u^{-1}(-\hat{R}(\hat{q}) - \Omega) \\
\iff & u^{-1}(-R(\hat{q})) \geq u^{-1}(-\hat{R}(\hat{q})) + u^{-1}(-\Omega) \\
\iff & ku^{-1}(-R(\hat{q})) + l \geq ku^{-1}(-\hat{R}(\hat{q})) + l + ku^{-1}(-\Omega) \\
\iff & \mathrm{CE}(\hat{q}) \geq \hat{\mathrm{CE}}(\hat{q}) + ku^{-1}(-\Omega).
\end{aligned}$$

□

Theorem 5. *Likewise, the following inequality holds:*

$$\boldsymbol{E}_{\mu_n}[\mathrm{CE}(\hat{q})] \geq$$

Since $\Omega > 0$, it follows that $u^{-1}(-\Omega) > -\Omega$, therefore we have the following relation:

$$\mathrm{CE}(\hat{q}) \geq \hat{\mathrm{CE}}(\hat{q}) - O\left(\frac{\xi^2}{\sqrt{n}\lambda}\right).$$

3.2 Big Data Situation

The literature revolving around Theorem 3 and its applications ([Shai-Shalev, Rudin]) generally leaves the ξ as an afterthought, but in real big-data contexts, if n is insufficiently large compared to p, than out-of-sample convergence might not be certain. Actually, with $n = o(p^2)$, divergence is almost certain.

Let's assume that $\boldsymbol{E}X_i = 0$ for all features, and let $Z^2 = \sum_{i=1}^p X_i^2$. In a general setting, if $\boldsymbol{E}X_i^2 \leq M$, then $\boldsymbol{E}Z^2 \leq Mp$, and Markov's inequality applies:

$$\boldsymbol{P}\{Z^2 \geq t\} \leq \frac{\boldsymbol{E}Z^2}{t} \leq \frac{Mp}{t}.$$

Equivalently, with probability $1-\delta$,

$$Z^2 \leq \delta^{-1}Mp = O(p).$$

If we further assume pairwise independance of features and that each feature is supported by a closed interval, wether because the support of the feature is known to be bounded or because it's been saturated in pre-processing, then each feature can be rescaled by a_i so that so that $a_iX_i = \tilde{X}_i \subsetneq [-1,1]$, or $\tilde{X}_i^2 \subsetneq [0,1]$. Then, using Hoeffding's theorem,

$$\boldsymbol{P}\{\tilde{Z}^2 \geq \tilde{M}p + t\} \leq \exp\left(-\frac{2t^2}{p}\right),$$

equivalently with probability $1-\delta$,

$$\tilde{Z}^2 \leq \tilde{M}p + \sqrt{\frac{p\log(1/\delta)}{2}} = O(p)$$

Of course, it is easy to see that such a transformation is reversible.

The point is that in general cases, we expect to have $\xi^2 = O(p)$, so that the convergence of our algorithm will be given by

$$\mathrm{CE}(\hat{q}) \geq \hat{\mathrm{CE}}(\hat{q}) - O\left(\frac{p}{\sqrt{n}\lambda}\right).$$

4 Empirical Results

5 Conclusion

6 Appendix

[Move this elsewhere and put in context.]

Claim 1. *The uniform stability β of our algorithm is given by:*

$$|\ell(\hat{q}_n, m) - \ell(\hat{q}_{n-1}, m)| \leq \frac{(\gamma \bar{r} \xi)^2}{2\lambda n}$$

Claim 2. *The following inequality holds:*

$$|c(p_1, r) - c(p_2, r)| \leq \gamma\, \bar{r}\, |p_1 - p_2|.$$

Proof. Using the Lipschitz property of u, the claim follows trivially. $\square$

Claim 3. *The following inequality holds:*

$$\|\hat{q}\| \leq \frac{\gamma \xi r_{\max}}{2\lambda} = O\left(\frac{\xi}{\lambda}\right).$$

Proof. Let μ_n be a sample of the market. The empirical decision algorithm

$$\text{minimize} \quad n^{-1} \sum_{i=1}^{n} \ell(m_i, q) + \lambda \|q\|^2$$

is equivalent to

$$\begin{array}{ll} \text{minimize} & n^{-1} \sum_{i=1}^{n} \ell(m_i, sq) + \lambda s^2 \\ \text{subject to} & s \geq 0 \\ & \|q\|_2 = 1, \end{array}$$

where the optimization variables are now on the direction (q) and the scale (s). Therefore, for any direction q, we can define a convex function $g(s)$ which becomes the objective:

$$\begin{array}{ll} \text{minimize} & g(s) \\ \text{subject to} & s \geq 0, \end{array}$$

where

$$g(s) = n^{-1} \sum_{i=1}^{n} \ell(m_i, sq) + \lambda s^2.$$

Because $g : (0, +\infty) \to \mathscr{R}$ is convex, we can consider two cases: either the minimum is realized at the boundary, ie. $s^\star = 0$, or there exists an optimal value $s^\star > 0$ such that $g'(s^\star) = 0$. To derive a bound on $s^\star$, we can seek a value $\bar{s}$ such that for any q, $g'(\bar{s}) > 0$ and therefore $\bar{s} > s^\star$.

To do so, we first note that

$$\begin{aligned} g'(s) &= \nabla_s \left[n^{-1} \sum_{i=1}^{n} \ell(m_i, sq) + \lambda s^2 \right] \\ &= 2\lambda s - n^{-1} \sum_{i=1}^{n} \nabla_s u(r_i\, s q^T x_i) \\ &= 2\lambda s - n^{-1} \sum_{i=1}^{n} r_i\, q^T x_i\, u'(r_i\, s q^T x_i). \end{aligned}$$

Now, because $\|q\| = 1$, we have $q^T x_i \leq \|x_i\| \leq \xi$. We also have $r_i \leq r_{\max}$ and $u' \leq \gamma$, so that

$$n^{-1} \sum_{i=1}^{n} r_i \, q^T x_i \, u'(r_i \, s q^T x_i) \leq \gamma \xi r_{\max}.$$

Therefore, with

$$\bar{s} := \frac{\gamma \xi r_{\max}}{2\lambda},$$

for any $s > \bar{s}$,

$$g'(s) \geq 0. \qquad \square$$

Claim 4. *The following inequality holds:*

$$|\hat{p}| = |\hat{q}^T x| \leq \bar{p} = \frac{\gamma \xi^2 r_{\max}}{2\lambda} = O\left(\frac{\xi^2}{\lambda}\right).$$

Proof. The claim follows directly from Hölder's inequality and Claim 3. $\square$

Claim 5. *The following inequalities hold:*

$$-\frac{\gamma \xi^2 r_{\max} \bar{r}}{2\lambda} \leq \ell(M, \hat{q}) \leq \frac{\gamma^2 \xi^2 r_{\max} \bar{r}}{2\lambda}.$$

First, the maximum loss (ie. worst realized utility) is reached when $\hat{p} = \bar{p}$ as expressed by Claim 4, and $r = -\bar{r}$, ie. $c(\bar{p}, -\bar{r}) \leq \gamma^2 \xi^2 r_{\max} \bar{r} / 2\lambda$. Likewise, the minimum loss (best utility) occurs with $\hat{p} = \bar{p}$ and $r = \bar{r}$, so that $c(\bar{p}, \bar{r}) \geq -\gamma \xi^2 r_{\max} \bar{r} / 2\lambda$.

Claim 6. *The out of sample average returns are at least $u^{-1}(-\hat{R}(\hat{q}_n) - \Omega_n)$.*

Proof. Using the convexity of u^{-1} and Jensen's inequality, we first have

$$\begin{aligned} \boldsymbol{E}[u^{-1}(-\ell(M, \hat{q}_n))] &\geq u^{-1}(\boldsymbol{E}[-\ell(M, \hat{q}_n)]) \\ &= u^{-1}(-R(\hat{q}_n)). \end{aligned}$$

From Theorem 4, we also have

$$-R(\hat{q}_n) \geq -\Omega_n - \hat{R}(\hat{q}_n).$$

Since u^{-1} is monotonically increasing, we finally obtain

$$u^{-1}(-R(\hat{q}_n)) \geq u^{-1}(-\Omega_n - \hat{R}(\hat{q}_n)). \qquad \square$$

Claim 7. *The following inequality holds:*

$$|R(q_1) - R(q_2)| \leq \gamma \bar{r} \xi \|q_1 - q_2\|.$$

Proof. We have the following chain of inequality:

$$\begin{aligned} |R(q_1) - R(q_2)| &= |\boldsymbol{E}[\ell(q_1, M)] - \boldsymbol{E}[\ell(q_2, M)]| \\ &= |\boldsymbol{E}[\ell(q_1, M) - \ell(q_2, M)]| \\ &\leq \boldsymbol{E}[|\ell(q_1, M) - \ell(q_2, M)|] \\ &= \boldsymbol{E}[|u(R q_1^T X) - u(R q_2^T X)|] \\ &\leq \gamma \boldsymbol{E}[|R(q_1 - q_2)^T X|] \\ &\leq \gamma \bar{r} \xi \|q_1 - q_2\|. \end{aligned} \qquad \square$$

Claim 8. *The following inequality holds:*

$$|R(q^\star) - R_\lambda(q_\lambda^\star)| \leq \lambda \|q^\star\|_2^2.$$

Proof. We first have that

$$R(q^\star) = \min_q R(q) \le \min_q R(q) + \lambda\|q\|_2^2 = R_\lambda(q_\lambda^\star).$$

We also have that

$$R_\lambda(q_\lambda^\star) = \min_q R(q) + \lambda\|q\|_2^2 \le R(q^\star) + \lambda\|q^\star\|_2^2,$$

and therefore

$$R_\lambda(q_\lambda^\star) - \lambda\|q^\star\|_2^2 \le R(q^\star) \le R_\lambda(q_\lambda^\star),$$

leading to

$$0 \le R_\lambda(q_\lambda^\star) - R(q^\star) \le \lambda\|q^\star\|_2^2. \qquad \square$$

Claim 9. *If $u : \mathscr{R} \to \mathscr{R}$ monotonically increasing is such that $u(0) = 0$, $u(x) = u_-(x)\mathbf{1}_{\{x<0\}} + u_+(x)\mathbf{1}_{\{x\ge 0\}}$ and $u_+(x) = o(u_-(x))$, then, for any real random variable Z with $|Z| \le M$ and $\lim_{x\to 0^-} F_Z(x) > 0$,*

$$\arg\max_{k>0} \boldsymbol{E}[u(kZ)]$$

is finite.

Proof. We first note that $\lim_{k\to\infty} \boldsymbol{E}[u(kZ)] = -\infty$ is a sufficient condition. Next, by hypothesis, there exists a $\delta < 0$ such that $\boldsymbol{P}\{Z < \delta\} = p > 0$. Let B a discrete random variable such that $\boldsymbol{P}\{B = \delta\} = 1 - \boldsymbol{P}\{B = M\} = p$. Then $\boldsymbol{P}\{B \ge z\} \ge \boldsymbol{P}\{Z \ge z\}$ for any z. This in turn implies that $\boldsymbol{E}[u(kB)] \ge \boldsymbol{E}[u(kZ)]$ **[Too fast?]**. But

$$\lim_{k\to\infty} \boldsymbol{E}[u(kB)] = \lim_{k\to\infty} p\, u(k\delta) + (1-p)u(kM) = -\infty. \qquad \square$$

References

[1] Olivier Bousquet and André Elisseeff. “Stability and generalization”. In: *The Journal of Machine Learning Research* 2 (2002), pp. 499–526.

[2] Thomas M Cover. “Universal portfolios”. In: *Mathematical finance* 1.1 (1991), pp. 1–29.

[3] Beatrice Laurent and Pascal Massart. “Adaptive estimation of a quadratic functional by model selection”. In: *Annals of Statistics* (2000), pp. 1302–1338.

[4] Harry Markowitz. “Portfolio selection”. In: *The journal of finance* 7.1 (1952), pp. 77–91.

[5] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[6] Cynthia Rudin and Gah-Yi Vahn. “The big data newsvendor: Practical insights from machine learning”. In: *Available at SSRN 2559116* (2015).