http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 65, No. 10, October 2019, pp. 4656–4675
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Submodular Risk Allocation

Samim Ghamami,[a] Paul Glasserman[b]

[a] Center for Risk Management Research, Department of Economics, University of California, Berkeley, California 94720; [b] Graduate School of Business, Columbia Business School, New York, New York 10027

Contact: samim_ghamami@berkeley.edu (SG); pg20@columbia.edu, http://orcid.org/0000-0002-9577-0205 (PG)

Received: August 1, 2017
Revised: April 25, 2018
Accepted: June 12, 2018
Published Online in Articles in Advance: May 15, 2019

https://doi.org/10.1287/mnsc.2018.3156

Abstract. We analyze the optimal allocation of trades to portfolios when the cost associated with an allocation is proportional to each portfolio's risk. Our investigation is motivated by changes in the over-the-counter derivatives markets, under which some contracts may be traded bilaterally or through central counterparties, splitting a set of trades into two or more portfolios. A derivatives dealer faces risk-based collateral and capital costs for each portfolio, and it seeks to minimize total margin requirements through its allocation of trades to portfolios. When margin requirements are submodular, the problem becomes a submodular intersection problem. Its dual provides per-trade margin attributions, and assigning trades to portfolios based on the lowest attributed costs yields an optimal allocation. As part of this investigation, we derive conditions under which standard deviation and other risk measures are submodular functions of sets of trades. We compare systemwide optimality with individually optimal allocations in a market with multiple dealers.

History: Accepted by Yinyu Ye, optimization.

Keywords: financial institutions • markets • risk • correlation

1. Introduction

This paper studies the problem of allocating transactions or other individual sources of risk to portfolios, to minimize a sum of risk-based costs for the portfolios.

Our investigation is motivated by changes in the over-the-counter (OTC) derivatives market. Prior to the financial crisis, the market for swaps and other OTC derivatives was largely unregulated, and it operated as a diffuse network of bilateral contracts between market participants. In 2009, regulatory authorities from the G-20 countries agreed to reforms that have reshaped the market.

The reforms have two key elements. First, they require that all sufficiently standardized contracts be cleared through central counterparties (CCPs). A CCP (or clearinghouse) interposes itself between the two original parties to a contract (two dealers, for example) by entering into two back-to-back contracts with the original parties. To protect itself against losses from the default of a counterparty, the CCP collects collateral (also referred to as margin) from both parties.

The second key element of the OTC derivatives reforms addresses the part of the market that continues to trade bilaterally, rather than through CCPs. For these contracts, the reforms require that the parties exchange collateral (margin) to protect each party against the possible failure of the other party.

In the centrally cleared market, the margin collected by a CCP from a market participant depends on the riskiness of the participant's portfolio of trades at that CCP. In the bilateral market, the total amount of margin exchanged between two parties similarly depends on the riskiness of their portfolio of trades, though risk may be measured differently in the bilateral and centrally cleared settings. The total collateral cost faced by a derivatives dealer is the sum of its collateral costs across multiple portfolios—one portfolio for each CCP, and one or more portfolios for each bilateral counterparty.

The collateral cost associated with each portfolio increases with the risk of the portfolio. A dealer will often have some flexibility in choosing the CCP through which to clear a contract or deciding to trade the contract bilaterally. To minimize its total collateral costs, the dealer needs to allocate trades to portfolios in a way that minimizes the sum of risk-based costs over the portfolios. This is the problem we investigate.

To make the problem more explicit, consider a dealer with a set of trades S with a counterparty. Suppose the dealer is limited to two trading channels—two CCPs, for example, or one CCP and the bilateral market. For any subset of trades $A \subseteq S$, let $F(A)$ and $G(S\backslash A)$ denote the collateral costs associated with assigning portfolio A to the first channel and assigning the remaining trades $S\backslash A$ to the other channel. The dealer's optimization problem is then

$$\min_{A\subseteq S}\{F(A) + G(S\backslash A)\}. \tag{1}$$

Regulators have sought to increase bilateral margin requirements to incentivize greater use of central clearing, which, in the setting of (1), means increasing costs under G to encourage allocation to F. Regulators and some market participants have also expressed concern about whether the global supply of high-quality collateral is sufficient to meet increased collateral requirements, which entails a concern for an efficient systemwide allocation. See Heller and Vause (2012), Sidanius and Zikes (2012), Duffie et al. (2015), and Ghamami and Glasserman (2017) for details on these regulatory changes and empirical examinations of their impact. Another instance of (1) arises in the setting of bank capital requirements. A bank faces separate regulatory capital charges associated with securities held in the "banking book" and the "trading book." In some cases, a bank has flexibility to classify a security either way, for example by stating that the security will be held to maturity or potentially sold. If we take S to be the set of securities for which the bank has this discretion, and if we write F and G for the mappings from portfolios to capital charges in the banking and trading books, then (1) becomes the problem of allocating securities to minimize capital charges. Along the same lines, an international banking institution faces choices in selecting a subsidiary through which to execute a transaction—a unit based in London or New York, for example. The parent may face separate capital requirements for each unit, in which case the allocation problem takes the form in (1).

A further instance of (1) arises in the setting of portfolio liquidation. A portfolio manager seeking to liquidate a set S of securities would like to minimize total transaction costs or market impact costs in doing so. Because selling one security may move the price of other closely related securities, the total execution cost is best viewed at the portfolio level, rather than at the level of individual securities; see Schneider and Lillo (2019) and Tsoukalas et al. (2017) for models of portfolio trading costs that incorporate cross-asset impact. To minimize total costs, the portfolio manager may split the total portfolio across different trading venues or across different trading periods. These allocation problems are instances of (1), if we take $F(A)$ and $G(S\backslash A)$ as reduced-form representations of the execution costs resulting from trading the two subportfolios separately. For a different problem of allocating transaction costs to portfolios, see Iancu and Trichakis (2014).

We study the application of (1) when the cost functions F and G are submodular, a setting that has received extensive study since the work of Edmonds (1970). In our motivating application, the cost associated with a portfolio is proportional to its risk, so F and G represent measures of portfolio risk. We interpret submodularity as a strong version of the notion that diversification reduces risk. Under submodularity, the marginal change in risk from adding an asset to a portfolio decreases with the addition of another asset.

As a first step, we therefore investigate the submodularity of portfolio risk measures, with particular emphasis on standard deviation, viewed as set functions defined over a finite set of assets. Despite the vast literature on properties of risk measures growing out of Artzner et al. (1999), including convexity and subadditivity, the submodular case has received little prior attention. Conditions for submodularity are therefore of independent interest.

Once we have submodularity, problem (1) leads to several interesting and important properties. If F and G are monotone as well as submodular, (1) becomes the polymatroid intersection problem studied by Edmonds (1970). Extensions to the submodular case without monotonicity are treated in the books by Fujishige (2005) and Schrijver (2003). These results provide a dual characterization of (1) through which an individual cost may be attributed to each trade for each of the two portfolios. Allocating each trade to the portfolio for which it has a lower attributed cost yields an optimal allocation. This representation provides the dealer with a margin attribution for each trade under an optimal allocation. Some of these properties can also be interpreted through the framework of convex games, in the sense of Shapley (1971) and Topkis (1998).

When two or more dealers seek to allocate overlapping sets of trades, their optimal allocations may conflict. Differences in their cost attributions characterize payments between dealers that would reconcile their conflicting allocations. In some cases, these payments may be described as "valuation adjustments" of the type widely used in industry practice (see Gregory (2015), Andersen et al. (2019), and the references therein). Our framework provides a rigorous mechanism for the ad hoc practice of decomposing portfolio-level valuation adjustments into trade-level adjustments.

We also compare systemwide costs in a market with multiple dealers having potentially conflicting allocation preferences for their shared trades. We compare the sum of individually optimal costs (assuming each dealer makes its individually optimal allocations), the optimal systemwide cost (assuming coordination among dealers), and costs under a sequential protocol, in which dealers make allocation decisions in order of market power. We use the structure of cost attribution vectors to bound cost differences across these scenarios.

To position our work relative to the literature on combinatorial portfolio optimization, we emphasize that our starting point (1) is a problem of optimal allocation rather than selection of assets. Combinatorial portfolio optimization addresses the problem of selecting an optimal portfolio of assets with discrete costs or constraints arising, for example, through cardinality constraints (Bertsimas and Shioda 2009,

Gao and Li 2013), fixed transaction costs (Lobo et al. 2007), logical constraints (Cornuejols and Tütüncü 2007), or indivisibility of assets (Sirignano et al. 2016). In our setting, the set of assets is already given; the problem is to divide the overall portfolio into subportfolios. Whereas an investment manager picks assets to maximize risk-adjusted returns, we have in mind the perspective of a dealer that enters into swaps or other trades as a service to clients or to offset its risk; the dealer is not entering into trades to speculate on the direction of the market. In addition, most of the literature on combinatorial portfolio optimization focuses on computational efficiency, whereas our focus is on the structure of optimal allocations, both for a single dealer in isolation and for multiple interacting dealers. We will see that analyzing this structure provides useful insight in studying multidealer risk allocation problems. Computationally efficient submodular optimization has been studied extensively, as we discuss in Section 3.6. Indeed, without taking advantage of submodularity, (1) would typically be intractable, as it optimizes over $2^{|S|}$ subsets.

Section 2 develops conditions for submodularity of portfolio risk measures, with particular emphasis on standard deviation. Section 3 analyzes the optimal allocation decision for a single dealer. Section 4 introduces the effect of counterparty risk between a pair of dealers, and Section 5 examines systemwide optimality with multiple dealers. Most proofs are deferred to an appendix.

2. Submodular Risk Measures

In our motivating application, a dealer has a fixed set of trades to allocate, and we represent these trades through a set $S = \{1, \dots, N\}$ indexing jointly distributed random variables $X_1, \dots, X_N$. Interpret each X_i as the change in value of a derivatives contract over a period of 1–10 days. The dealer posts collateral (to a CCP or to another derivatives counterparty) as a backstop against possible changes in the value of the contract over this time interval, which is known as the margin period of risk. The amount of collateral required is based on a measure of risk of the dealer's portfolio of trades with the counterparty; for example, the collateral required may be a multiple of the portfolio standard deviation. The X_i are often assumed to have mean zero: derivatives are often priced so that this holds and, more broadly, over a short time horizon the expected change in market prices is typically negligibly small compared to the volatility in prices. However, this assumption is not needed for our analysis.

A margin function (or risk measure) F assigns a margin requirement $F(A)$ to any subset $A \subseteq S$ of trades. The margin function is submodular if

$$F(A \cap B) + F(A \cup B) \le F(A) + F(B), \quad \forall A, B \subseteq S. \tag{2}$$

This condition is equivalent to the requirement that

$$F(A \cup \{i,j\}) - F(A \cup \{i\}) \le F(A \cup \{j\}) - F(A), \quad \forall A \subseteq S, \forall i, j \in S \backslash A, i \ne j, \tag{3}$$

where $S\backslash A$ denotes the complement of A in S. In other words, the incremental margin required by adding trade j to the portfolio is reduced by the addition of trade i to the portfolio. In (3), we have implicitly identified the set of random variables S with the set of indices $\{1, 2, \dots, N\}$, a simplification we use throughout.

2.1. Submodularity of Standard Deviation: Basic Properties

Let Σ denote the covariance matrix of the trades $X_1, \dots, X_N$. For any $A \subseteq S$, let

$$\sigma(A) = \text{Standard Deviation}\left(\sum_{i \in A} X_i\right).$$

There is a natural correspondence between the subsets of S and the vertices of the hypercube $[0,1]^N$, in which we identify $A \subseteq S$ with the vector $x \in \{0,1\}^N$ satisfying $x_i = 1$ if $i \in A$ and $x_i = 0$ if $i \notin A$. We sometimes write this vector as x_A. An individual element $i \in S$ is identified with the unit vector e_i. We therefore also write

$$\sigma(x) = \sqrt{x^\top \Sigma x} \quad \text{and} \quad \sigma(A) = \sqrt{x_A^\top \Sigma x_A}. \tag{4}$$

With this notation, condition (3) for submodularity becomes

$$\sigma(x + e_i + e_j) - \sigma(x + e_i) \le \sigma(x + e_j) - \sigma(x), \quad \forall x, x + e_i + e_j \in \{0,1\}^N. \tag{5}$$

For brevity, we call a covariance matrix Σ submodular if (5) holds with σ as defined in (4).

When there is little chance of confusion, we also use the notation σ_i^2 for the ith diagonal entry Σ_{ii} of Σ; that is, $\sigma_i = \sigma(e_i)$. If we write D_σ for the diagonal matrix with the σ_i on the diagonal, then we may represent Σ as

$$\Sigma = D_\sigma R D_\sigma,$$

where R is a correlation matrix. If Σ is positive definite, then R is uniquely determined. For $0 \le \lambda \le 1$, let

$$\Sigma_\lambda = D_\sigma(\lambda R + (1-\lambda)I)D_\sigma,$$

where I is the $N \times N$ identity matrix. Our first result shows that a variety of correlation patterns are compatible with submodularity:

Proposition 2.1. *Let Σ be a covariance matrix.*

i. *If Σ is diagonal (i.e., $R = I$), then Σ is submodular.*

ii. *If R is the matrix of all 1s, then Σ is submodular.*

iii. *For any correlation matrix R, there is a $\lambda' > 0$ such that Σ_λ is submodular for all $0 \le \lambda < \lambda'$.*

Proposition 2.1 shows that submodularity of Σ does not have an immediate connection with the strength or sign of pairwise correlations: both the absence of correlation and perfect correlation yield submodularity, as do perturbations in the direction of any correlation matrix.

An immediate consequence of part (i) of the proposition is that every covariance matrix is similar to a submodular covariance matrix. In particular, if we can split and recombine the original assets into portfolios using eigenvectors of Σ for portfolio weights, the resulting portfolios will be uncorrelated and their covariance matrix therefore submodular. But this transformation is not in general possible if the individual assets are indivisible, as we assume. (For a discussion of risk allocation problems with divisible risks, see Embrechts et al. (2018) and the many references cited there.)

The indivisibility of trades means that we treat $\sigma(\cdot)$ as a function on the vertices of the unit hypercube. Such a function can be submodular even if its natural extension to the interior of the unit hypercube is not, and indeed this distinction will be important in light of the following result.

Proposition 2.2. *Suppose Σ is positive definite, and let $\sigma(w) = \sqrt{w^\top \Sigma w}$, for all $w \in [0,1]^N$. Then σ is submodular throughout $[0,1]^N$ if and only if either Σ is diagonal or $N = 2$.*

Although we assume that trades are indivisible, a dealer may choose to bundle certain subsets of trades, meaning that the entire bundle will always be assigned to the same portfolio. Such a constraint may arise for example for groups of trades associated with a single client, or because some trades are used to hedge other trades. Bundling trades i and j means replacing the original random variables X_i and X_j with their sum $X_i + X_j$. The effect of bundling on the covariance matrix is to replace Σ with $P^\top \Sigma P$, where P is an $N \times K$ matrix whose columns are orthogonal elements of $\{0,1\}^N$. We might expect that even if Σ is not submodular, it may become submodular through some bundling of trades. This procedure leads to the following observation:

Proposition 2.3. *If Σ is submodular, then so is any bundling $P^\top \Sigma P$. Every covariance matrix admits a submodular bundling $P^\top \Sigma P$, for some $K \geq 2$.*

2.2. Conditions on Variance

A sufficient condition for submodularity of standard deviation can be formulated through submodularity and monotonicity of variance. Write σ_{ij} for the ij-entry of Σ

Proposition 2.4. *Suppose Σ satisfies the following two conditions:*

(i) $\sigma_{ij} \leq 0$, *for all distinct* $i, j = 1, \ldots, N$;

(ii) $\sigma_i^2 \geq -2 \sum_{j \neq i} \sigma_{ij}$, *for all* $i = 1, \ldots, N$.

Then Σ is submodular, and $\sigma(\cdot)$ is monotone increasing on $\{0,1\}^N$.

Without the factor of 2, and under the sign restriction in (i), condition (ii) would be the familiar diagonal dominance condition for positive semidefiniteness of Σ. In particular, a matrix that satisfies (i) and (ii) is an M-matrix by condition (M_{35}) of Berman and Plemmons (1979, p. 137).

2.3. Generalized Exchangeability

The random variables $X_1, \ldots, X_N$ are called exchangeable if their joint distribution is invariant under permutations of the random variables. The covariance matrix of exchangeable random variables takes the form

$$\Sigma = \begin{pmatrix} \sigma^2 & \sigma^2\rho & \cdots & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 & & \sigma^2\rho \\ \vdots & & \ddots & \vdots \\ \sigma^2\rho & \sigma^2\rho & \cdots & \sigma^2 \end{pmatrix}, \quad -1/(N-1) \leq \rho \leq 1, \tag{6}$$

and Σ is positive definite if the bounds on ρ are strict. For $\rho < 0$ and sufficiently close to zero, this matrix satisfies the conditions in Proposition 2.4. In fact, submodularity holds for all feasible ρ:

Proposition 2.5. *The covariance matrix of exchangeable random variables is submodular.*

We will establish this result as a corollary to a more general condition. For any $v \in \mathbb{R}^N$, let $\mathsf{diag}(v)$ denote the $N \times N$ diagonal matrix with the entries of v on the diagonal. Write $|v|$ for the sum of the absolute values of the entries of v. Write $\mathbb{R}_+$ for the nonnegative elements of $\mathbb{R}$, and write $\mathbb{R}_{++}$ for the strictly positive elements.

Proposition 2.6. (i) *Suppose that for some $v \in \mathbb{R}^N_{++}$ and some $a \in \mathbb{R}$,*

$$\Sigma = \mathsf{diag}(v) + avv^\top. \tag{7}$$

If Σ is positive semidefinite then it is submodular. (ii) *Suppose that for some $v, w \in \mathbb{R}^N_+$ and some $a, b \geq 0$*

$$\Sigma = \mathsf{diag}(v + w) + avv^\top + bww^\top. \tag{8}$$

If

$$4a(|w| + b|w|^2) \leq 1 \quad \text{and} \quad 4b(|v| + a|v|^2) \leq 1, \tag{9}$$

then Σ is submodular.

The exchangeable case becomes a special case of this result (for $\rho \neq 1$) by taking $a = \rho/(1-\rho)^2\sigma^2$ and $v_i = (1-\rho)\sigma^2$, $i = 1, \ldots, N$, in (7). The case $\rho = 1$ is covered by Proposition 2.1.

The conditions in Proposition 2.6 take advantage of the fact that we require submodularity only on the

vertices of the hypercube. In particular, if (7) holds then, for $x \in \{0,1\}^N$,

$$\sigma^2(x) = \sum_i v_i x_i^2 + a(v^\top x)^2 = (v^\top x) + a(v^\top x)^2, \tag{10}$$

so σ depends on x only through $v^\top x$.

Submodularity is also preserved by diagonal deviations from exchangeability that satisfy the bound in the following result.

Proposition 2.7. *For $\xi \in \mathbb{R}_+^N$, let $\Sigma_\xi = \mathsf{diag}(\xi) + \Sigma$, with Σ as in (6). Then Σ_ξ is submodular if either* (i) $-1/(2N-1) \le \rho \le 0$ *or* (ii) $0 < \rho < 1$ *and*

$$|\xi| \equiv \sum_{i=1}^N \xi_i \le \frac{\sigma^2}{4}\frac{(1-\rho)^2}{\rho}. \tag{11}$$

Covariance matrices are often specified through factor models to simplify and regularize estimation. In financial applications, a small number of factors often explain most of the observed correlation. To express the matrices of this section as factor models, consider a general single-factor model

$$X_i = b_i Z + c_i \epsilon_i, \quad i = 1, \dots, N, \tag{12}$$

in which $Z, \epsilon_1, \dots, \epsilon_N$ are uncorrelated, each with unit variance. The resulting covariance matrix Σ has $\Sigma_{ii} = b_i^2 + c_i^2$ and $\Sigma_{ij} = b_i b_j$, $j \ne i$. The exchangeable case corresponds to fixing the same (b_i, c_i) for all $i = 1, \dots, N$. The case of (7) with $a > 0$ corresponds to setting $b_i = v_i\sqrt{a}$ and $c_i = \sqrt{v_i}$, $i = 1, \dots, N$.

The general single-factor model (12) need not yield submodularity. For example, with $N = 3$, $b_1 = b_2 = 1$, $b_3 = 2$, and $c_1 = c_2 = c_3 = 1$, we get

$$\Sigma = \begin{pmatrix} 2 & 1 & 2 \\ 1 & 2 & 2 \\ 2 & 2 & 5 \end{pmatrix},$$

and then

$$\begin{aligned} \sigma(0,1,0) + \sigma(1,1,1) &= \sqrt{2} + \sqrt{19} > \sqrt{6} + \sqrt{11} \\ &= \sigma(1,1,0) + \sigma(0,1,1), \end{aligned}$$

so submodularity fails.

Covariance matrices are often approximated or regularized using principal components analysis. A spectral decomposition yields

$$\Sigma = \lambda_1 v_1 v_1^\top + \cdots + \lambda_N v_N v_N^\top,$$

where $\lambda_1 \ge \cdots \ge \lambda_N$ are the eigenvalues of Σ and $v_1, \dots, v_N$ are corresponding orthonormal eigenvectors. Keeping just the first $k \le N$ terms in this representation yields a rank-k approximation to Σ. In financial data, the first principal component often dominates, meaning that λ_1 is much larger than the other eigenvalues. The first principal component often represents a market factor, for which all components of the eigenvector v_1 are positive, meaning that all securities have positive exposure to the market factor.

Proposition 2.8. *If $\Sigma = \lambda v v^\top$, where $\lambda > 0$ and all entries of v have the same sign, then Σ is submodular.*

This follows directly from $(x^\top \Sigma x)^{1/2} = \lambda(x^\top v v^\top x)^{1/2} = \lambda|x^\top v|$.

2.4. Correlation Conditions

For positive definite Σ and vectors $x, y \in \mathbb{R}^N \setminus \{\mathbf{0}\}$, define the correlation between x and y as

$$\rho(x,y) = \frac{x^\top \Sigma y}{\sigma(x)\sigma(y)}.$$

Our next result formulates conditions for submodularity in terms of correlations.

Proposition 2.9. *Suppose the covariance matrix Σ is positive definite. Then Σ is submodular under either of the following conditions.*

(i) *For any distinct $x, y, w \in \{0,1\}^N \setminus \{\mathbf{0}\}$ and any $t \in [0,1]$,*

$$\rho(x + y + tw, w) \le \rho(x + tw, w).$$

(ii) *For all $x, y, w, \in \{0,1\}^N \setminus \{\mathbf{0}\}$ with $x + y + w \in \{0,1\}^N \setminus \{\mathbf{0}\}$,*

$$\rho(x + y + w, w) \le \rho(x, w). \tag{13}$$

Moreover, if for all $w \in (0,1)^N$ and all distinct $i, j \in \{1, \dots, N\}$,

$$\rho(e_i, e_j) \le 2\rho(e_i, w)\rho(e_j, w), \tag{14}$$

then $\sigma(\cdot)$ is logsubmodular, in the sense that

$$\sigma(x \wedge y)\sigma(x \vee y) \le \sigma(x)\sigma(y),$$

for all $x, y \in \mathbb{R}^N$. In particular, (14) holds if Σ is exchangeable with $\rho \ge 1/2$ in (6).

We will encounter applications of logsubmodularity later. It is also useful through the following connection, which shows in particular that a strong version of logsubmodularity yields submodularity.

Proposition 2.10. *If $\sigma(\cdot)$ is submodular and monotone increasing or decreasing on $\{0,1\}^N$, then it is logsubmodular on $\{0,1\}^N$. If all off-diagonal elements of Σ are nonnegative and*

$$\sigma_{ij} + \sigma(x)\sigma(x + e_i + e_j) \le \sigma(x + e_i)\sigma(x + e_j), \tag{15}$$

for all $x, x + e_i + e_j \in \{0,1\}^N$, then $\sigma(\cdot)$ is submodular and monotone increasing on $\{0,1\}^N$.

Proof. The first assertion follows from lemma 2.6.6 of Topkis (1998). For the second assertion, we need to verify (5). If all correlations are nonnegative, $\sigma(\cdot)$ is monotone increasing, and in this case (5) holds if and

only if the squares of the two sides of (5) are ordered the same way. This condition is implied by (15). To see this, rearrange (5) as

$$\sigma(x+e_i+e_j)+\sigma(x)\le\sigma(x+e_j)+\sigma(x+e_i),$$

and note that

$$\sigma^2(x+e_i+e_j)=x^\top\Sigma x+\sigma_i^2+\sigma_j^2+2\sigma_{ij}+2x^\top\Sigma e_i+2x^\top\Sigma e_j.$$

Squaring both sides of the above inequality and using the above equality and simple algebra gives (15). □

2.5. Conditional Covariance Matrices

Suppose $X_1,\dots,X_N$ are jointly normal with positive definite covariance matrix Σ, and consider the effect of conditioning on a subset of these variables. Without loss of generality, we may consider the case of conditioning on the last $N-k$ of the variables, $X_{k+1},\dots,X_k$. Write Σ in block form as

$$\Sigma=\begin{pmatrix}\Sigma_{11}&\Sigma_{12}\\ \Sigma_{21}&\Sigma_{22}\end{pmatrix},$$

where Σ_{11} is $k\times k$. Then the conditional covariance matrix is given by the Schur complement

$$\Sigma_{1|2}=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

The conditional covariance matrix is the relevant tool for measuring the risk in $X_1,\dots,X_k$ conditional on stressed values of $X_{k+1},\dots,X_N$. The following result shows that conditioning often preserves submodularity:

Proposition 2.11. *If Σ satisfies the conditions of Proposition 2.4, 2.5, or 2.6(i), then $\Sigma_{1|2}$ is submodular.*

2.6. A Submodularity Ratio

Das and Kempe (2011) find that certain greedy algorithms that yield optimal solutions for submodular functions perform well on functions that are close to being submodular. To measure if a function is close to being submodular, they introduced the notion of a submodularity ratio. In this spirit, we define

$$\gamma=\max_{x<y,y+e_i\in\{0,1\}^N}\frac{\sigma(x)+\sigma(y+e_i)}{\sigma(x+e_i)+\sigma(y)}. \tag{16}$$

Then $\sigma(\cdot)$ is submodular on $\{0,1\}^N$ if and only if $\gamma\le 1$, and $\gamma>1$ provides a measure of how far σ deviates from submodularity. Our next result provides a simple upper bound on γ.

Proposition 2.12. *If $\lambda_{\max}$ and $\lambda_{\min}$ denote the largest and smallest eigenvalues of the positive definite covariance matrix Σ, then*

$$\gamma\le\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\,\frac{\sqrt{N-2}+\sqrt{N}}{2\sqrt{N-1}}\le\sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}. \tag{17}$$

Thus, we can bound deviation from submodularity for an arbitrary covariance matrix with knowledge of its eigenvalues.

2.7. Other Risk Measures

We have thus far focused on submodularity of standard deviation. In this section, we briefly consider other risk measures.

Additive Measures. If a separate margin requirement is imposed on each trade, without consideration of portfolio diversification, the total margin requirement for a set of trades $A\subseteq S$ would take the form $\sum_{i\in A}c_i$, for some trade-specific charges c_i. Any such measure is submodular. Under current rules for bilateral trading, separate margin requirements are calculated by asset category—interest-rate, credit, equity, and commodity derivatives—with no diversification recognized across categories; see the discussion in Ghamami and Glasserman (2017). If the margin requirement in each category is submodular, then the sum across categories is also submodular.

Downside Risk. By the expected downside risk for a random variable X we mean $E[\max\{0,X\}]$; recall that each X_i represents a potential loss. If $X_1,\dots,X_N$ are jointly normal with zero mean, then

$$d(A)=E\left[\max\left\{0,\sum_{i\in A}X_i\right\}\right]=\frac{\sigma(A)}{\sqrt{2\pi}},$$

so submodularity of $d(\cdot)$ is equivalent to submodularity of $\sigma(\cdot)$.

Value-at-Risk and Expected Shortfall. The value-at-risk at tail probability $0<\alpha<1/2$ for a portfolio of trades A is given by

$$\mathsf{VaR}_\alpha(A)=\inf\left\{u\in\mathbb{R}:P\left(\sum_{i\in A}X_i>u\right)\le\alpha\right\}.$$

The widely used SPAN (standard portfolio analysis of risk) methodology developed by the Chicago Mercantile Exchange for margin requirements is based on a portfolio's value-at-risk. The corresponding expected shortfall is

$$\mathsf{ES}_\alpha(A)=\frac{1}{\alpha}\int_{1-\alpha}^1\mathsf{VaR}_p(A)\,dp.$$

These measures simplify for elliptical distributions. The vector $X = (X_1, \ldots, X_N)^\top$ of trades has an elliptical distribution if it can be represented as $\mu + RMU$, where μ is a vector of means, U is uniformly distributed on the sphere $\{u \in \mathbb{R}^k : \|u\| = 1\}$, for some $k \le N$, M is a fixed $N \times k$ matrix, and R is a scalar random variable independent of U. The elliptical distributions include, among many other examples, the multivariate normal and multivariate t distributions.

In the elliptical case,

$$\mathsf{VaR}_\alpha(A) = x_A^\top \mu + \sigma(A) k_\alpha, \quad \sigma(A) = \sqrt{x_A^\top M M^\top x_A},$$

where k_α does not depend on A; see theorem 6.8 of McNeil et al. (2005). Thus, VaR_α is submodular precisely if the matrix $MM^\top$ is submodular. Expected shortfall inherits submodularity from value-at-risk.

Exponential Utility. Hedging errors in markets with transactions costs are often evaluated using exponential utility (following Hodges and Neuberger 1989) or its corresponding risk premium. In the multivariate normal case with mean zero and covariance Σ, this yields the risk measure, for some parameter $\gamma > 0$,

$$F(A) = \frac{1}{\gamma} \log E\left[\exp\left(-\gamma \sum_{i \in A} X_i\right)\right] = \frac{\gamma}{2}\sigma^2(A) = \frac{\gamma}{2} x_A \Sigma x_A.$$

The variance function $\sigma^2(\cdot)$ is submodular on $\{0,1\}^N$ if and only if all off-diagonal entries of Σ are negative (Murota 2003, proposition 2.6); that is, if and only if all pairwise correlations between trades are negative.

Entropy. Entropy measures are sometimes used to quantify dispersion in asset returns; see, for example, Philippatos and Wilson (1972), Backus et al. (2014), and, in the setting of CCP risk management, De Genaro (2016). Suppose that $X_1, \ldots, X_N$ have finite support, and for any $A \subseteq 1, \ldots, N$, let f_A denote the probability mass function of the vector X_A formed by those X_i with $i \in A$. Define the entropy $\mathcal{H}(A) = -E[\log f_A(X_A)]$. Fujishige (1978) showed that $\mathcal{H}$ is increasing ($A \subseteq B \Rightarrow \mathcal{H}(A) \le \mathcal{H}(B)$) and submodular, without restrictions on the dependence among the X_i. For continuous random variables, the corresponding differential entropy is not automatically submodular. However, in the special case of multivariate normal $X_1, \ldots, X_N$, entropy simplifies to

$$\mathcal{H}(A) = \frac{1}{2}(|A|(1 + \log 2\pi) + \log\det(\Sigma_A)), \tag{18}$$

where Σ_A is covariance matrix of X_i, $i \in A$. The mapping $A \mapsto \log\det(\Sigma_A)$ is submodular (see Fan 1968 and references there), as is $A \mapsto |A|$, so $\mathcal{H}$ is submodular.

An alternative risk measure sets $h(A)$ equal to the entropy of the single random variable $\sum_{i \in A} X_i$. For jointly normal $X_1, \ldots, X_N$, the sum is again normal and $h(A)$ simplifies to $1 + \log(2\pi\sigma(A))$. Thus, $h(\cdot)$ is submodular if $\sigma(\cdot)$ is logsubmodular, a case considered in Proposition 2.9.

3. Risk Allocation: A Single Dealer

We now return to the problem we introduced at the outset in (1). A dealer has a fixed set of trades S to either clear through a CCP or manage through the noncleared market. The subset $A \subseteq S$ of trades cleared through the CCP incur a margin charge $F(A)$, and the remaining set of trades incur a margin charge of $G(S \backslash A)$. The dealer would like to choose the set of cleared trades A to minimize the total margin charge in (1).

3.1. Polymatroid Structure

A function f on the lattice of subsets 2^S is call *normalized* if $f(\emptyset) = 0$. We will assume that F and G are normalized, so that no margin charge applies where no trades are allocated, and we will also assume that they are nonnegative. For a normalized submodular function f, define the *submodular polyhedron*

$$P(f) = \left\{x \in \mathbb{R}^N : \sum_{i \in A} x_i \le f(A), \forall A \subseteq S\right\}, \tag{19}$$

recalling that $N = |S|$, and its *base* polyhedron,

$$B(f) = \left\{x \in P(f) : \sum_{i \in S} x_i = f(S)\right\}. \tag{20}$$

The *polymatroid* associated with f is the intersection of $P(f)$ with the nonnegative orthant,

$$P_+(f) = \left\{x \in \mathbb{R}_+^N : \sum_{i \in A} x_i \le f(A), \forall A \subseteq S\right\}. \tag{21}$$

If f is monotone increasing as well as normalized and submodular, then it is called a polymatroid rank function, and it is uniquely determined by $P_+(f)$; see corollary 44.3f of Schrijver (2003). As explained in section 46.2 of Schrijver (2003), theorem 35 of Edmonds (1970) solves problem (1) in the case where both F and G are polymatroid rank functions, which yields the polymatroid intersection problem. We will use the extension to normalized submodular functions (the submodular intersection problem) in Fujishige (2005), theorem 4.9, and Schrijver (2003), corollary 46.1b, as follows.

Proposition 3.1. *If F and G are normalized and submodular, then*

$$\min_{A \subseteq S}\{F(A) + G(S \backslash A)\} = \max\left\{\sum_{i=1}^N x_i \wedge y_i : x \in B(F), y \in B(G)\right\}. \tag{22}$$

If x and y solve the problem on the right, then $A_0 = \{i : x_i < y_i\}$ and $A_1 = \{i : x_i \le y_i\}$ are optimal for the problem on the left.

The result in (22) is commonly formulated with $x \in P_+(F)$ (or $P(F)$) and $y \in P_+(G)$ (or $P(G)$). For any $x \in P(F)$ and $y \in P(G)$, we can find $\bar{x} \ge x$ in $B(F)$ and $\bar{y} \ge y$ in $B(G)$ and thus $\sum_i \bar{x}_i \wedge \bar{y}_i \ge \sum_i x_i \wedge y_i$. Optimality of x and y then implies optimality of $\bar{x}$ and $\bar{y}$, so we may restrict the optimization to the bases. The optimality of A_0 and A_1 follows from lemma 7.4 of Fujishige (2005). If $x_i \ne y_i$ for all $i = 1, \dots, N$, then $A_0 = A_1$ minimizes the left side of (22) uniquely.

To interpret Proposition 3.1, the result can be fruitfully connected to cooperative game theory. In particular, the literature on convex games deals with the problem of allocating a supermodular value function among multiple players, which can be recast as a problem of allocating submodular costs; see Shapley (1971) and Topkis (1998).

To exploit this connection, consider the problem of decomposing the total cost $F(S)$ of clearing trades through the CCP into charges attributable to individual trades.[1] An attribution is a vector $x \in \mathbb{R}^N$, with the interpretation that x_i is the margin charge attributed to trade i, $i = 1, \dots, N$. It is reasonable to require that no subset of trades A be attributed a total margin charge greater than the charge $F(A)$ incurred in clearing just those trades. The attributions in $P(F)$ are the ones that satisfy this condition, and those in $B(F)$ fully decompose the total charge $F(S)$. The attributions in $B(F)$ are the *core* attributions (or, more conventionally, allocations) in cooperative game theory.

With this background, we can interpret Proposition 3.1 as follows. The dealer chooses an attribution $x \in B(F)$ and an attribution $y \in B(G)$. Each trade i is allocated to the channel for which its attributed charge (x_i or y_i) is smaller. A rule that maximizes the total attributed margin charge $\sum_i x_i \wedge y_i$ is optimal. In most applications of convex games, cost decomposition is the main objective; see, for example, Anily and Haviv (2007). In our setting, it is an intermediate step in characterizing the optimal allocation of trades to portfolios.

The optimal attribution vectors x and y are nevertheless of independent interest. Suppose, for example, that different trades are initiated by different trading desks or different trading units within the dealer's firm. In addition to finding the minimum cost set of trades A to clear through the CCP and trades $S \backslash A$ to trade bilaterally, the dealer needs to charge back the total cost $F(A) + G(S \backslash A)$ to the individual trading units. If the pair (x, y) is optimal for (22), then attributing a cost of $x_i \wedge y_i$ to trade i, $i = 1, \dots, N$, fully decomposes the total cost, taking into account the full portfolio of trades S *and* the optimal split of trades between the cleared and bilateral markets.

A core attribution can be constructed as follows. Consider any permutation $i_1, \dots, i_N$ of the trades, and allocate to each trade its incremental margin charge under this permutation:

$$\begin{aligned} x_{i_1} &= F(\{i_1\}),\ x_{i_2} = F(\{i_1, i_2\}) - F(\{i_1\}),\ \dots,\ x_{i_N} \\ &= F(S) - F(S \backslash \{i_N\}). \end{aligned} \tag{23}$$

Under this rule, the sum of x_i telescopes, so the full amount $F(S)$ is decomposed. Shapley (1971) showed that this attribution is in the core; in fact, x is an extreme point of the convex set $B(F)$, and all extreme points of $B(F)$ are of this form. Taking the equally weighted average over all attributions (23) yields the Shapley value.

3.2. Examples

To illustrate (22) when F and G are defined by standard deviations, let

$$\Sigma_F = \begin{pmatrix} \sigma_{F,1}^2 & \rho_F \sigma_{F,1} \sigma_{F,2} \\ \rho_F \sigma_{F,1} \sigma_{F,2} & \sigma_{F,2}^2 \end{pmatrix} \quad \text{and} \quad \Sigma_G = \begin{pmatrix} \sigma_{G,1}^2 & \rho_G \sigma_{G,1} \sigma_{G,2} \\ \rho_G \sigma_{G,1} \sigma_{G,2} & \sigma_{G,2}^2 \end{pmatrix}, \tag{24}$$

with base-case parameters $\sigma_{F,1} = 1$, $\sigma_{F,2} = 2$, $\rho_F = 0$, and $\sigma_{G,1} = \sigma_{G,2} = 1.5$ and $\rho_G = -0.4$. The associated standard deviation functions are monotone and submodular, so they define polymatroid rank functions. Panel (a) of Figure 1 shows the corresponding polymatroids for F (solid) and G (dashed). The base polyhedron for each is the diagonal segment in the upper right. The figure shows an optimal pair of bases x (filled circle) and y (open circle). Both coordinates of y are dominated by the coordinates of x, so the optimal solution allocates both trades to G. In panels (b)–(d), we vary parameters as indicated in the captions.

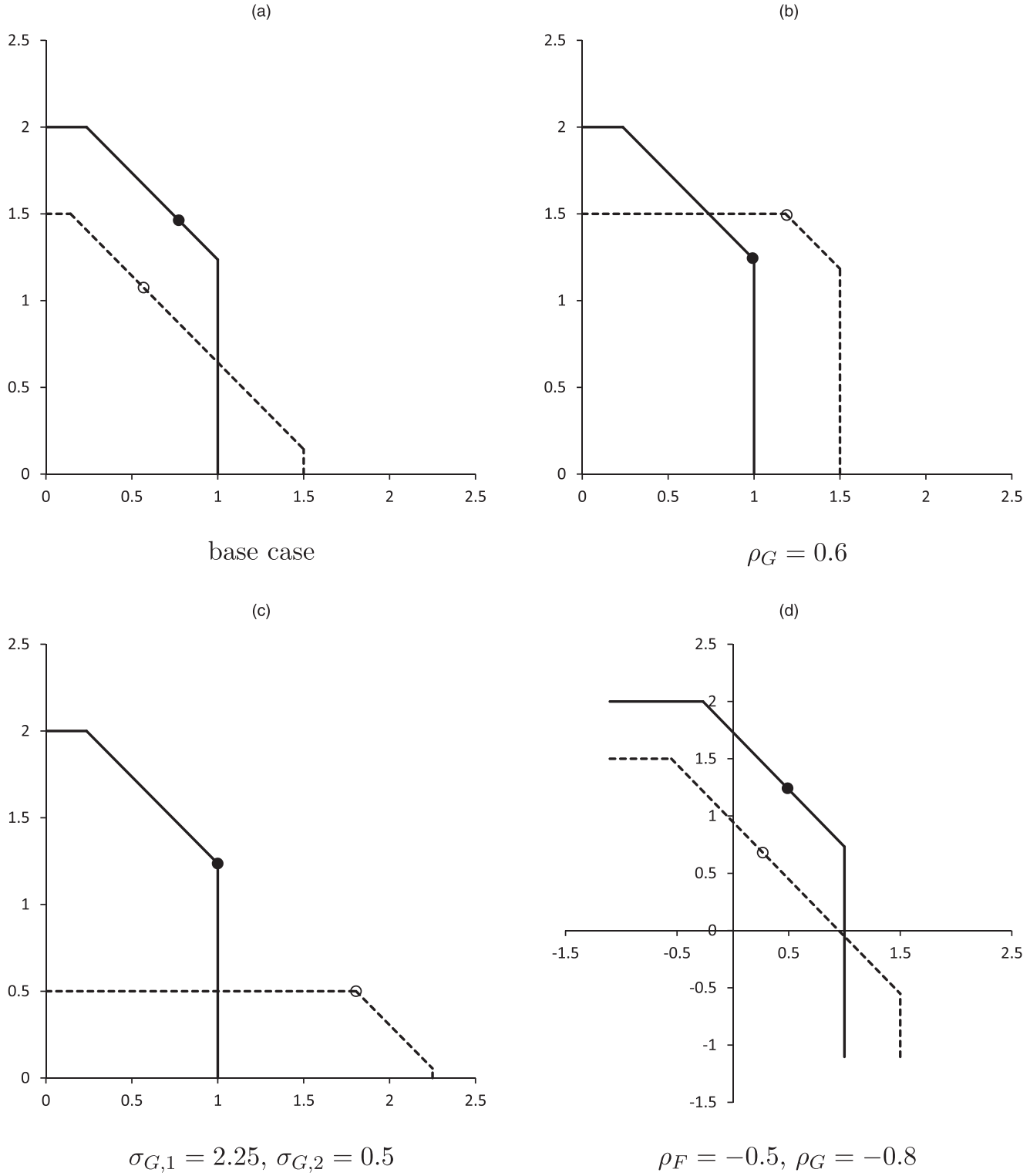
In panel (b), the optimal solution allocates both trades to F, in panel (d) it allocates both to G, and in (c) it allocates one trade to each. With the parameters of panel (d), the standard deviation functions are no longer monotone, though they are still submodular. The figure in this case shows $P(F)$ and $P(G)$, rather than $P_+(F)$ and $P_+(G)$.

Figure 2 illustrates the case of

$$\Sigma_F = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1.5 & 0 \\ 0 & 0 & 0.5 \end{pmatrix}, \quad \Sigma_G = \begin{pmatrix} 0.25 & -0.1 & 0 \\ -0.1 & 1 & 0 \\ 0 & 0 & 6.25 \end{pmatrix}. \tag{25}$$

Both satisfy conditions for monotonicity (Proposition 2.4) as well as submodularity, so their standard deviation

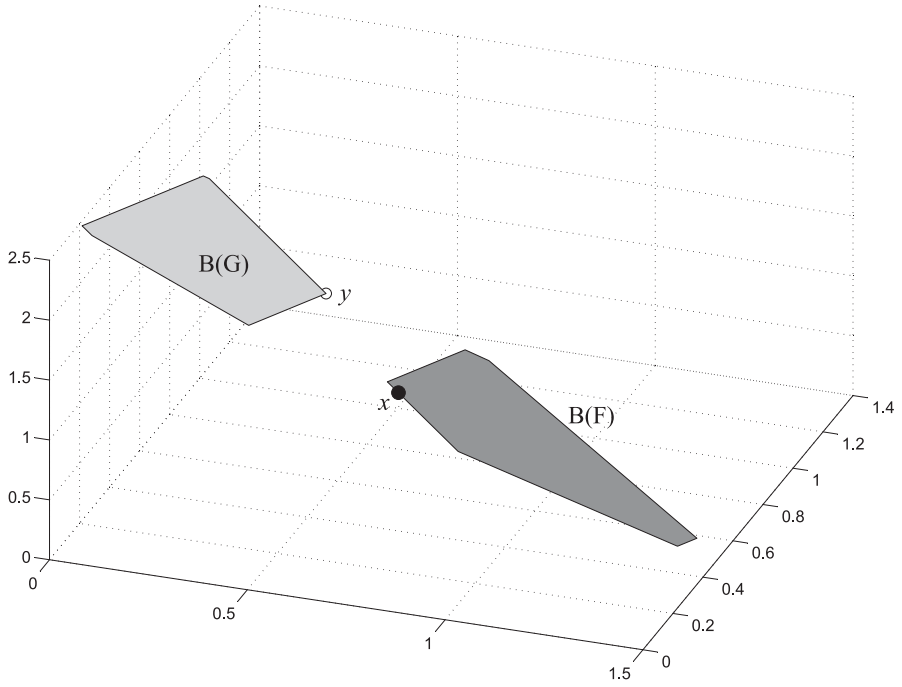
Figure 1. Optimal (x, y) Pairs for F and G Defined by (24)



Notes. Parameters in (a) are $\sigma_{F,1} = 1$, $\sigma_{F,2} = 2$, $\rho_F = 0$, $\sigma_{G,1} = \sigma_{G,2} = 1.5$ and $\rho_G = -0.4$. Parameter changes are indicated in the panel captions.

functions are polymatroid rank functions and we may restrict their bases to the positive orthant. The bases are illustrated in Figure 2, with each defined by six inequalities and lying in the hyperplanes $x_1 + x_2 + x_3 = (\mathbf{1}^\top \Sigma_F \mathbf{1})^{1/2} = 2$ and $y_1 + y_2 + y_3 = (\mathbf{1}^\top \Sigma_G \mathbf{1})^{1/2} = \sqrt{7.3}$, where $\mathbf{1}$ denotes a vector of 1s.

Figure 2. Base Polyhedra and Optimal Attribution Vectors for Example (25)



It is easy to see from the two covariance matrices that the optimal solution allocates trades 1 and 2 to G and trade 3 to F, at a cost of

$$((e_1+e_2)^\top \Sigma_G (e_1+e_2))^{1/2} + (e_3^\top \Sigma_F e_3)^{1/2} = \sqrt{1.05} + \sqrt{0.5}.$$

The figure shows optimal base vectors $x \approx (0.63, 2-\sqrt{0.5}-0.63, \sqrt{0.5})$ and $y = (0.5, \sqrt{1.05}-0.5, \sqrt{7.3}-\sqrt{1.05})$. These yield

$$\begin{aligned}&(x_1 \wedge y_1) + (x_2 \wedge y_2) + (x_3 \wedge y_3)\\ &\quad = 0.5 + (\sqrt{1.05}-0.5) + \sqrt{0.5} = \sqrt{1.05} + \sqrt{0.5}.\end{aligned}$$

3.3. Optimal Attributions

Based on a result of Murota (1988), optimal vectors x and y can be found as the "most similar" attributions in $B(F)$ and $B(G)$. More precisely, suppose the elements of $B(F)$ and $B(G)$ have strictly positive entries, so every trade receives a positive margin charge in every attribution. For any strictly convex and continuously differentiable $\phi : (0,\infty) \mapsto \mathbb{R}$, and any positive vectors $x, y \in \mathbb{R}^{|V|}_{++}$, define the ϕ-divergence

$$D_\phi(x,y) = \sum_{i \in S} x_i \phi(y_i/x_i).$$

Any such ϕ-divergence provides a measure of dissimilarity between x and y. It follows from the main theorem in Murota (1988) that if (x^*, y^*) minimizes some $D_\phi(x,y)$ over $x \in B(F)$ and $y \in B(G)$, then (x^*, y^*) solves (22). This then provides an alternative interpretation of Proposition 3.1: the optimal allocation between the cleared and bilateral markets is achieved through the cost attributions that minimize the divergence between the attributions to each trade in the two channels. The solutions shown in Figures 1 and 2 minimize D_ϕ with $\phi(t) = -\log t$, which corresponds to the Kullback–Leibler divergence for probability distributions.

3.4. Standard Deviation: Euler Decomposition

If F measures portfolio standard deviation with respect to a covariance matrix Σ_F, then the total risk $F(S) = \sigma(S)$ can be decomposed using the Euler rule (as in, for example, Denault 2001)

$$z_i = \partial_i \sigma(\mathbf{1}) = e_i^\top \Sigma_F \mathbf{1}/\sigma(\mathbf{1}), \quad i = 1, \ldots, N.$$

As before, $\mathbf{1}$ is a vector of 1s. In differentiating σ we have implicitly extended it to the function $y \mapsto (y^\top \Sigma_F y)^{1/2}$ on $\mathbb{R}^N$. It is immediate that

$$\sum_{i=1}^N z_i = \mathbf{1}^\top \Sigma_F \mathbf{1}/\sigma(\mathbf{1}) = \sigma(\mathbf{1}) \equiv \sigma(S),$$

so this rule does indeed decompose the total standard deviation. (If Σ_F is diagonally dominant, then its row sums are nonnegative and $z_i \geq 0$.) For any $A \subseteq S$, the Cauchy–Schwarz inequality yields

$$\sum_{i \in A} z_i = x_A^\top \Sigma_F \mathbf{1}/\sigma(\mathbf{1}) \leq \sqrt{(x_A^\top \Sigma_F x_A)(\mathbf{1}^\top \Sigma_F \mathbf{1})}/\sigma(\mathbf{1}) = \sigma(A),$$

which proves that $z \in B(F)$.

It follows that when F and G are both defined by standard deviations (with respect to distinct covariance

matrices Σ_F and Σ_G), a feasible solution to the right side of (22) can be obtained by calculating the Euler decompositions for each and then allocating each trade based on the lower of the two attributions.

The Euler decomposition is not an arbitrary feasible solution, for the following reason. Aubin (1981) investigates a "fuzzy core" based on extending the conditions in (20) to hold with f evaluated throughout the unit hypercube, and not just at its vertices. It follows from his proposition 2.1 that for a convex and positively homogeneous risk measure (such as standard deviation), the Euler decomposition is the unique element of the fuzzy core.

We close this section with a different application of the Euler decomposition. Goemans et al. (2009) show that any monotone submodular function f on 2^S can be approximated by a root-linear function, in the sense that there are positive constants $c_1, \ldots, c_N$ for which

$$\sqrt{\sum_{i\in A} c_i} \le f(A) \le O(\sqrt{N}\log N)\sqrt{\sum_{i\in A} c_i},$$

for all $A \subseteq S$. In our setting, the approximating root-linear function can be interpreted as a standard deviation with respect to a diagonal covariance matrix, with values c_i on the diagonal. In particular, the approximating function is submodular; see Proposition 2.1. The algorithm for computing the c_i in Goemans et al. (2009) is quite involved. In our setting, we can take them proportional to the Euler decomposition values and get a sharper upper bound:

Proposition 3.2. *Suppose Σ satisfies the conditions in Proposition 2.4, and set $c_i = e_i^\top \Sigma \mathbf{1}$. Then*

$$\sqrt{\sum_{i\in A} c_i} \le \sigma(A) \le \sqrt{2\sum_{i\in A} c_i}.$$

Proof. Because the off-diagonal entries of Σ are negative, we have, for any $A \subseteq S$,

$$\sum_{i\in A} c_i = e_A^\top \Sigma \mathbf{1} = \sigma^2(A) + e_A^\top \Sigma(\mathbf{1} - e_A) \le \sigma^2(A).$$

The conditions in Proposition 2.4 also imply that

$$c_i = \sigma_i^2 + \sum_{j\neq i} \sigma_{ij} \ge \sigma_i^2/2,$$

and then

$$\sigma^2(A) \le \sum_{i\in A} \sigma_i^2 \le 2\sum_{i\in A} c_i. \quad \square$$

3.5. Multiple Channels

A natural extension of our basic allocation problem (1) would generalize the problem to

$$\min_{A_1,\ldots,A_m} F_1(A_1) + \cdots + F_m(A_m), \tag{26}$$

where the minimum is taken over disjoint sets $A_1, \ldots, A_m$ whose union is S. The problem can be approached recursively. Define the *infimal convolution* of set functions F and G on subsets of S to be

$$(F\square G)(B) = \min_{A\subseteq B} F(A) + G(B\backslash A), \quad B \subseteq S;$$

then problem (1) is the problem of evaluating $(F\square G)(S)$. Solving (26) similarly reduces to evaluating $F_1\square\cdots\square F_m$ at S. This function can be calculated recursively as $G_{m-1}\square F_m$, with $G_1 = F_1$ and $G_n = G_{n-1}\square F_n$, $n = 2, \ldots, m-1$. In other words, the multiple channel allocation problem (26) reduces to a sequence of nested two-channel allocation problems.

If all but two of the functions F_i in (26) are additive risk measures (in the sense of Section 2.7), then our earlier results extend to this setting, because the infimal convolution of a submodular function and an additive function is submodular; see, e.g., p. 47 of Fujishige (2005). If $F_1, \ldots, F_{m-2}$ are additive, then G_{m-1} is submodular, and the calculation of (26) reduces to evaluating $(G_{m-1}\square F_m)(S)$, with G_{m-1} and F_m submodular, which is the problem in (1).

This approach does not extend to arbitrary submodular F_i in (26) because the infimal convolution of submodular functions is not guaranteed to be submodular. One approach to (26) would be to approximate G_n at each step by a submodular function, for instance along the lines discussed in Sections 2.6 and 3.4. Chapters 6–8 of Murota (2003) discuss special classes of functions, their infimal convolutions, and their connection to submodularity and base polyhedra. It may also be possible to study the allocation problem (26) by imposing these more complex conditions on the F_i.

3.6. Algorithms and Efficiency

Our analysis focuses on the structural properties of the allocation problem (1) and its extensions with more than one dealer, but in this section we comment on computational considerations. The general problem of computationally efficient minimization of submodular functions has been the subject of extensive study. For fixed S, the mapping $A \mapsto F(A) + G(S\backslash A)$, $A \subseteq S$, is submodular when F and G are submodular, so (1) can be viewed as a special case of the general problem.

Grötschel et al. (1981) developed the first polynomial time algorithm for general submodular function minimization, but their algorithm, based on the ellipsoid method, is generally viewed as impractically slow and complex. Subsequent theoretical work has focused on developing strongly polynomial algorithms that are fully combinatorial, meaning that they use addition, subtraction, and logical comparisons but not multiplication or division. Details of several combinatorial algorithms are provided in the books by Fujishige (2005, chapter 14) and Schrijver (2003,

chapter 45). Table 1 in McCormick (2005) compares the theoretical computational complexity of seven algorithms, four of which achieve strongly polynomial running times. That comparison was published before Iwata and Orlin (2009), the theoretically fastest algorithm currently known.

For a problem with the structure of (1), an optimal solution can also be found using the Fujishige–Wolfe (1970) minimum-norm algorithm. Let $H(A) = [F(A) + G(S\backslash A)] - G(S)$. Suppose z^* minimizes the Euclidean norm $\|z\|$ over $z \in B(H)$. Fujishige (2005, lemma 7.4) shows that if z^* solves this quadratic program, then $\{i : z_i^* < 0\}$ and $\{i : z_i^* \le 0\}$ minimize H. (In our setting, each z_i^* can be interpreted as the additional cost of clearing trade i through the CCP rather than trading it bilaterally.)

Fujishige et al. (2006) undertake a computational study to compare the running times of several algorithms. They find that minimizing submodular functions with $N = |S|$ around 100–200 (which would correspond to the number of trades in our setting) takes one to three seconds using the minimum-norm algorithm and one to two minutes using combinatorial algorithms. The minimum-norm algorithm solves problems with $N =$ 1,000 in six minutes, which are intractable for the combinatorial algorithms. The minimum-norm algorithm retains a large computational advantage on small problems as well. More recently, Chakrabarty et al. (2014) prove the first pseudo-polynomial guarantee for the minimum-norm algorithm. In their computational tests, the algorithm outperforms the Iwata–Orlin (Iwata and Orlin 2009) algorithm for problems of all sizes.

These comparisons are all based on algorithms that exploit submodularity. We can also ask how these run times compare with more generic methods that do not take advantage of submodularity. Problem (1) can be formulated as a linear program with an exponential number of constraints of the form in (19). It can also be treated as a discrete optimization problem with an exponential number of variables, noting that (1) optimizes over all subsets of S.

Given the size of these problem formulations, even the implementation of generic methods presents a challenge. Orso et al. (2015) propose and test large-scale optimization methods for minimizing submodular functions using linear programming or mixed-integer linear programming. Their tests include a problem of the form in (1). For problems of size 100–400, they report computing times of 16–60 minutes, indicating that even state-of-the-art general purpose optimization methods are not competitive with methods that exploit submodularity.

4. Adding Counterparty Risk

We now suppose that S consists of all trades between two dealers. If X_i is the loss to one dealer on trade i, then $-X_i$ is the loss to the other dealer. If the dealers face the same collateral cost functions F and G on trade sets $A \subseteq S$, then they will agree on how to allocate the trades between the two trading channels. However, the dealers may face different costs. In particular, a dealer with a higher credit rating may demand more collateral from a dealer with a lower credit rating than it offers in return. If G measures capital and collateral costs for bilateral trades, then the two dealers would face different G functions. They could also face different F functions, where F measures the collateral costs of trading through a CCP, because of trades with other parties. For example, the collateral cost of allocating trades $A \subseteq S$ to the CCP could be $F(\tilde{A} \cup A)$, where $\tilde{A}$ is the set of trades with other parties that the dealer has allocated to the CCP. Because $\tilde{A}$ varies by dealer, so does the function $A \mapsto F(\tilde{A} \cup A)$, even if the CCP applies the same F to all dealers.

4.1. A Cost Comparison Condition

To focus on differences in credit quality, we will suppose that the two dealers face costs

$$h_1(A) = F(A) + G_1(S\backslash A), \quad h_\theta(A) = F(A) + G_\theta(S\backslash A),$$

for some normalized submodular functions G_1 and G_θ on 2^S. Following Fujishige and Nagano (2009), the *strong map* relation $G_\theta \to G_1$ means that for all $A \subseteq B \subseteq S$,

$$G_\theta(B) - G_\theta(A) \ge G_1(B) - G_1(A). \tag{27}$$

Clearly, $G_\theta \to G_1$ implies $h_1 \to h_\theta$.

The strong map relation $h_1 \to h_\theta$ allows the following comparison of optimal solutions. Suppose A_1 minimizes h_1 and A_θ minimizes h_θ. Then $A_1 \cap A_\theta$ also minimizes h_1, and $A_1 \cup A_\theta$ also minimizes h_θ. In particular, if either h_1 or h_θ has a unique minimizer, then $A_1 \subseteq A_\theta$. These statements follow from theorem 2.8.1 of Topkis (1998).

From the perspective of a regulator, the strong map relation (27) shows how to increase collateral costs in the bilateral market to incentivize central clearing (a question studied empirically in Ghamami and Glasserman 2017): replacing G_1 with G_θ would lead a dealer to allocate more trades to F.

In our two-dealer setting, the comparison makes precise the idea that when the bilateral costs for the two dealers satisfy (27), the dealer facing costs h_θ would like to clear at least as many trades through the CCP as a counterparty with costs h_1. An arrangement in which either party can force central clearing favors dealer 1; an arrangement in which both parties have to agree before a trade is centrally cleared favors the other dealer.

In fact, more can be said about the relationship between the optimal choices of the two counterparties using the notion of *universal bases* developed for parametric submodular intersection problems (Nakamura 1988,

Fujishige and Nagano 2009). These results are easiest to describe in the linear case $G_\theta = \theta G_1$, which satisfies (27) for $\theta \geq 1$ if G is monotone increasing. It follows from theorem 4.1 of Fujishige and Nagano (2009) that there exist $x \in B(F)$ and $y \in B(G_1)$ such that

$$\min_{A \subseteq S}\{F(A) + \theta G(S \backslash A)\} = \sum_{i=1}^{N} \min(x_i, \theta y_i), \quad \text{for all } \theta \geq 1. \tag{28}$$

The key point of this result is that the bases x and y do not depend on the parameter θ: this is the sense in which they are universal. In particular, dealers with cost functions h_1 and h_θ can use the same x and y to decompose their collateral charges, as in (22). Dealer θ associates a collateral cost of θy_i for trading bilaterally, whereas dealer 1 associates a collateral cost of y_i for the same trade. If the cost functions are increasing as well as submodular, universal bases x and y can be found as maximally similar elements of $B(F)$ and $B(G_1)$, in the sense of Murota (1988) discussed in Section 3.3.

The difference $(\theta - 1)y_i$ provides a rigorous foundation for computing a *credit valuation adjustment*, or CVA. In the bilateral market, derivative values are adjusted (relative to theoretical, default-free values) to reflect the risk that the counterparty to the trade will default and fail to make promised payments on the contract; see Gregory (2015) for background. With all else equal, the same trade will be worth less to the more creditworthy party—the party that faces the greater risk from default of its counterparty. CVA is often calculated or decomposed into trade-by-trade adjustments, despite the fact that in practice the adjustment for each trade depends on the full portfolio of trades. The decomposition in (28) provides a mechanism for computing trade-level CVA consistent with an overall portfolio of trades, and consistent for both parties.

These properties extend to nonlinear parametric families of normalized submodular functions (F_t, G_t), $t \in \mathbb{R}$, on 2^S satisfying $F_{t_1} \to F_{t_2}$ and $G_{t_2} \to G_{t_1}$ whenever $t_1 < t_2$ (see Fujishige and Nagano 2009), so we next examine when these relations hold for risk measures.

4.2. Verifying the Comparison Condition

We close this section by examining conditions for the strong map relation in (27) to hold with specific cost functions, starting with portfolio standard deviation.

Proposition 4.1. *Suppose σ and $\tilde{\sigma}$ are the standard deviation functions on 2^S associated with the covariance matrices Σ and $\tilde{\Sigma}$. Then*

$$\tilde{\sigma}(B) - \tilde{\sigma}(A) \geq \sigma(B) - \sigma(A) \tag{29}$$

for all $A \subseteq B \subseteq S$ under any of the following conditions:

(i) $\tilde{\Sigma} = \theta\Sigma$, *for some $\theta \geq 1$, with Σ satisfying conditions* (i) *and* (ii) *of Proposition 2.4*;

(ii) $\tilde{\Sigma}$ *and* Σ *are exchangeable, as in* (6), *with $\tilde{\sigma} \geq \sigma$ and $\tilde{\rho} \geq \rho \geq -1/(2N-1)$ or with $\tilde{\sigma} = \sigma$ and $\tilde{\rho} \geq \rho \geq -1/(N-1)$*;

(iii) $\tilde{\Sigma}$ *and* Σ *have the form in* (7), *with $\tilde{a} \geq a$ and $\tilde{v} = v$.*

For the case of the Gaussian entropy function in (18), we compare functions $\mathcal{H}$ and $\tilde{\mathcal{H}}$ of the form in (18) based on positive definite matrices Σ and $\tilde{\Sigma}$. We write Σ^{-1}_{jj} for the jth diagonal entry of Σ^{-1}. The following result is of independent interest beyond our setting.

Proposition 4.2. (i) *If $\Sigma^{-1}_{jj} \leq 1$, for all $j = 1, \ldots, N$, then $\mathcal{H} : 2^S \mapsto \mathbb{R}_+$ is increasing.* (ii) *If, in addition, $\tilde{\Sigma}^{-1}_{jj} \leq 1$ and $\tilde{\Sigma}^{-1}_{jj}\Sigma_{jj} \leq 1$, for all $j = 1, \ldots, N$, then $\tilde{\mathcal{H}} \to \mathcal{H}$.*

5. Systemwide Optimization

We now consider a market with K dealers and a set of trades $S = \{1, \ldots, N\}$ among these dealers. We denote by A_k, $k = 1, \ldots, K$, the set of trades in which dealer k participates, and we denote by $D_i \subseteq \{1, \ldots, K\}$ the set of dealers participating in trade i, $i = 1, \ldots, N$. Two dealers participate in each trade, so we have

$$\bigcup_{k=1}^{K} A_k = S, \quad \sum_{k=1}^{K} 1\{i \in A_k\} = \sum_{k=1}^{K} 1\{k \in D_i\} = 2 \quad \forall i \in S.$$

Write F_k and G_k for dealer k's margin functions, defined on all subsets of A_k, which are dealer k's trades. We extend these functions to all subsets of S by setting, for any $B \subseteq S$,

$$F_k(B) = F_k(B \cap A_k), \quad G_k(B) = G_k(B \cap A_k);$$

in other words, dealer k incurs margin charges only on the trades in which it participates. We assume that F_k and G_k are normalized and submodular on 2^{A_k}, and the same therefore holds for their extensions to 2^S. The functions

$$F = \sum_{k=1}^{K} F_k, \quad G = \sum_{k=1}^{K} G_k$$

are then also normalized and submodular.

For each $k = 1, \ldots, K$, $B(F_k)$ and $B(G_k)$ are subsets of $\mathbb{R}^N$ (rather than $\mathbb{R}^{|A_k|}$) because we have extended F_k and G_k to all subsets of S. For any $x \in B(F_k)$, we claim that $x_i = 0$ if $i \notin A_k$. By definition, $x_i \leq F_k(\{i\}) = F_k(\{i\} \cap A_k) = 0$ if $i \notin A_k$. But for any base we have $\sum_i x_i = F_k(S) = F_k(A_k)$, so

$$F_k(A_k) - \sum_{i \notin A_k} x_i = \sum_{j \in A_k} x_j \leq F(A_k),$$

which can hold only if $x_i = 0$ for all $i \notin A_k$. Similarly, $y_i = 0$ for $y \in B(G_k)$ and $i \notin A_k$.

If dealer k could make its allocation decision in isolation, it would incur an optimal cost of

$$c_k = \min_{A\subseteq A_k} F_k(A) + G_k(A_k\backslash A) = \max_{x^k\in B(F_k), y^k\in B(G_k)} \sum_{i\in A_k} x_i^k \wedge y_i^k. \tag{30}$$

Write c_{tot} for the sum $c_1 + \cdots + c_K$ of these individual costs. This quantity is not in general a feasible systemwide cost because dealers' individually optimal allocation decisions may be incompatible with each other. That is, if dealers k and ℓ share trade i ($i\in A_k\cap A_\ell$), dealer k's optimal solution may allocate trade i to F_k, whereas dealer ℓ's optimal solution may allocate the same trade to G_ℓ, in which case their optimal allocations are not simultaneously feasible. The systemwide optimum solves

$$c_{sys} = \min_{A\subseteq S} F(A) + G(S\backslash A),$$

and can be characterized as follows.

Lemma 5.1. *The systemwide optimal cost satisfies*

$$c_{sys} = \max\left\{\sum_{i=1}^N \left(\sum_{k\in D_i} x_i^k\right) \wedge \left(\sum_{k\in D_i} y_i^k\right) : x^k\in B(F_k), \right.$$
$$\left. y^k\in B(G_k), k=1,\dots,K\right\}. \tag{31}$$

Proof. We know from Proposition 3.1 that

$$c_{sys} = \max\left\{\sum_{i=1}^N x_i\wedge y_i : x\in B(F), y\in B(G)\right\}. \tag{32}$$

It follows from Fujishige (2005, p. 142) that

$$B(F) = \sum_{k=1}^K B(F_k) \equiv \{x^1+\cdots+x^K : x^k\in B(F_k), \quad k=1,\dots,K\}, \tag{33}$$

and similarly $B(G) = \sum_k B(G_k)$. We therefore have

$$c_{sys} = \max\left\{\sum_{i=1}^N \left(\sum_{k=1}^K x_i^k\right) \wedge \left(\sum_{k=1}^K y_i^k\right) : x^k\in B(F_k), \right.$$
$$\left. y^k\in B(G_k), k=1,\dots,K\right\}. \tag{34}$$

We have already shown that $x_i^k = y_i^k = 0$ unless $k\in D_i$, so the result follows. □

This lemma suggests a mechanism for attributing the optimal systemwide cost to individual trades: the cost attributed to trade i is

$$\left(\sum_{k\in D_i} x_i^k\right) \wedge \left(\sum_{k\in D_i} y_i^k\right),$$

where x^k and y^k, $k=1,\dots,K$, achieve the optimum in (31).

The dealers can cooperate to achieve the systemwide optimum by submitting their trade sets A_k and margin functions F_k and G_k to a central planner. The planner announces charges

$$\frac{1}{2}\sum_{k\in D_i} x_i^k \quad \text{and} \quad \frac{1}{2}\sum_{k\in D_i} y_i^k \tag{35}$$

for clearing or not clearing trade i. Based on these charges, each dealer makes its own decisions and contributes the corresponding margin charge. Under this mechanism, each dealer would make the systemwide optimal allocation, and the total margin charges collected throughout the system would sum to c_{sys}. Starting from an arbitrary configuration with a cost $c > c_{sys}$, the difference $c - c_{sys}$ could be used to create incentives for individual dealers to make systemwide optimal allocation decisions. This type of multilateral coordination has a precedent in the OTC derivatives market: in the related setting of trade compression, dealers cooperate by submitting information about their derivatives positions to a third party, which finds cycles in the network of contracts that can be eliminated without changing dealers' net positions. See, for example, Vause (2010).

The coordination required to achieve a systemwide optimal solution may nevertheless be difficult to achieve. In practice, conflicting preferences between counterparties over trade allocation are resolved through negotiation. A model of these negotiations would go beyond the scope of our investigation, so we consider a simple case: dealers make their allocation decisions sequentially in the order $k=1,\dots,K$, which we take to be their ranking by market power.

Let $C_k = A_1\cup\cdots\cup A_k$ denote the cumulative set of trades in which the first k dealers participate. The first dealer solves (30), clears a set $B_1\subseteq A_1$ of trades, and incurs a cost $\bar{c}_1 = c_1$. Once dealer k has made its allocation decision, $k=1,\dots,K-2$, dealer $k+1$ solves

$$\bar{c}_{k+1} = \min_{B_{k+1}\subseteq A_{k+1}\backslash C_k} F_{k+1}(B_{k+1}\cup B_k\cup\cdots\cup B_1)$$
$$+ G_{k+1}((C_k\cup A_{k+1})\backslash(B_{k+1}\cup B_k\cup\cdots\cup B_1)).$$

The process terminates at the first k_o for which $C_{k_o} = S$, which occurs at $k_o\le K-1$. If $k\ge k_o$, then no trades remain to be allocated by dealer $k+1$, and $\bar{c}_{k+1}$ is evaluated with $B_{k+1} = \emptyset$. The total cost under this protocol is $c_{seq} = \bar{c}_1 + \cdots + \bar{c}_K$.

For any $C\subseteq S$ and $k=1,\dots,K$, define

$$\delta_k(C) = \max\left\{\sum_{i\in C\cap A_k} |x_i^k - y_i^k| : x^k\in B(F_k),\ y^k\in B(G_k)\right\}. \tag{36}$$

We will use the δ_k to bound c_{seq}, with the following interpretation. The allocation of trades in C is constrained and cannot be chosen by dealer k. If x_i and y_i measure cost attributions for trades $i \in C$, then the additional cost faced by the dealer as a result of having the allocation of i fixed should be bounded by $|x_i - y_i|$. Summing over $i \in C$ bounds the additional cost due to having all trades in C constrained. More precisely, to get a bound we need to take the maximum over cost attributions $x \in B(F_k)$ and $y \in B(G_k)$. For any normalized submodular functions f and g on 2^S, let

$$\Delta(f,g) = \min_{x \in B(f), y \in B(g)} \frac{1}{2} \sum_{i=1}^{N} |x_i - y_i|.$$

Recall from (33) that every $x \in B(F)$ has a representation as $x^1 + \cdots + x^K$, with $x^k \in B(F_k)$, and every $y \in B(G)$ similarly has a representation as $y^1 + \cdots + y^K$, with $y^k \in B(G_k)$. Define

$$D(F,G) = \min \left\{ \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{K} |x_i^k - y_i^k| : x \in B(F), y \in B(G) \right\}, \tag{37}$$

noting that the sum over k can be limited to $k \in D_i$. In light of the discussion surrounding (35), we interpret $D(F,G)$ as the potential additional cost, over all trades and dealers, resulting from deviations from the systemwide cost attributions $x \in B(F)$ and $y \in B(G)$.

Proposition 5.1. *The systemwide optimal cost satisfies*

$$c_{tot} \le c_{sys} = c_{tot} + \sum_{k=1}^{K} \Delta(F_k, G_k) - \Delta(F,G) \le c_{tot} + D(F,G).$$

The cost under the sequential solution satisfies

$$c_{tot} \le c_{seq} \le c_{tot} + \sum_{k=2}^{K} \delta_k(C_{k-1}).$$

Combining the inequalities yields an upper bound on $c_{seq} - c_{sys} \ge 0$.

This result bounds systemwide cost differences using measures of the deviation between the cost functions F and G. The first comparison shows how much larger the optimal systemwide cost may be than the sum of the individually optimal costs c_{tot}. This sum is in general infeasible because different parties to a trade may have incompatible preferences for allocating the trade to one channel or another. These conflicts are captured by the individual deviations $\Delta(F_k, G_k)$ and ultimately $D(F,G)$. The second statement specializes to the case of sequential decisions among dealers. In this setting, we know that each dealer is constrained by the allocations of dealers earlier in the sequence, and this information is captured in the deviation measures $\delta_k(C_{k-1})$.

6. Concluding Remarks

Motivated by changes in the over-the-counter derivatives market, we have investigated the optimal allocation of trades to portfolios to minimize total risk-based costs. These costs represent capital or collateral requirements associated with a portfolio's risk. We have focused on risk-based costs that are submodular functions of a set of trades. Submodularity reflects diversification benefits, because it implies that the incremental risk from adding one asset to a portfolio decreases with the addition of another asset to the portfolio.

We have provided conditions under which familiar measures of risk are in fact submodular, with particular emphasis on portfolio standard deviation. With these conditions in place, we can draw on the classical work of Edmonds (1970) on polymatroid intersection and its extensions to characterize optimal allocations. The solution decomposes the total risk-based cost of a portfolio into amounts attributable to each trade. An optimal allocation assigns each trade to the portfolio in which it has the lowest cost attribution.

Using this framework, we have analyzed conflicting allocation decisions by the parties to a set of trades. Each trade involves two dealers, and the dealers need to make consistent allocation decisions, but their preferences may differ because they face different costs or because they have different trades with other dealers. Optimal attribution vectors yield trade-specific valuation adjustments to reconcile conflicting preferences between two parties. We also analyzed total systemwide costs in a market with multiple dealers. We have compared decentralized costs, systemwide optimal costs, and costs under a sequential protocol in which dealers make allocation decisions in order of market power.

An important topic for further investigation is understanding the dynamics of conflicting allocation decisions in a market with multiple dealers. The cases we consider—fully decentralized, fully centralized, and sequential—simplify a much more complex process of negotiation, coordination, and competition. These dynamics have broader implications for the financial system. In particular, they influence the split of trades between the cleared and bilateral markets, the split of trades across multiple clearinghouses, and the systemwide demand for collateral, all of which have been regulatory concerns associated with the postcrisis reforms of the derivatives markets.

Acknowledgments

The authors thank the referees and associate editor for helpful comments and suggestions.

Appendix A. Proofs for Section 2

Proof of Proposition 2.1. (i) In the diagonal case, for any $y \in \{0,1\}^N$ we have

$$\sigma^2(y) = \sum_{i:y_i=1} \sigma_i^2.$$

Suppose $\sigma_i^2 > 0$ for all i. For any $x, x + e_i + e_j \in \{0,1\}^N$, let

$$a_1 = \sigma^2(x),\ a_2 = \sigma^2(x + e_j),\ a_3 = \sigma^2(x + e_i),\ a_4 = \sigma^2(x + e_i + e_j). \tag{A.1}$$

Without loss of generality, suppose $a_2 \le a_3$. Then

$$0 < a_1 < a_2 \le a_3 < a_4 \quad \text{and} \quad a_1 + a_4 = a_2 + a_3. \tag{A.2}$$

By the strict concavity of the square root function, (A.2) implies

$$\sqrt{a_4} - \sqrt{a_3} < \sqrt{a_2} - \sqrt{a_1}, \tag{A.3}$$

which is condition (3) for submodularity, with strict inequality. If some diagonal entries of Σ are 0, then $\sigma(\cdot)$ is the limit of submodular functions (defined by perturbing these diagonal entries) and is therefore submodular.

(ii) If all correlations are equal to 1, then

$$\sigma(x + e_i + e_j) = \sigma(x + e_i) + \sigma(e_j) = \sigma(x) + \sigma(e_i) + \sigma(e_j),$$

and (3) again holds.

For case (iii), note that the a_k defined in (A.1) are continuous in λ. The strict inequality in (A.3) therefore continues to hold in a right neighborhood of $\lambda = 0$, which is to say that Σ_λ is submodular in a right neighborhood of $\lambda = 0$. □

Proof of Proposition 2.2. By continuity, σ is submodular on $[0,1]^N$ if and only if it is submodular on $(0,1)^N$, and this holds if and only if all its mixed second derivatives satisfy $\partial_{w_i}\partial_{w_j}\sigma(w) \le 0$, $i \ne j$, for all $w \in (0,1)^N$. We have

$$\partial_{w_i}\sigma(w) = \frac{e_i^\top \Sigma w}{\sigma(w)}$$

and

$$\partial_{w_j}\partial_{w_i}\sigma(w) = \frac{e_i^\top \Sigma e_j}{\sigma(w)} - \frac{(e_i^\top \Sigma w)(e_j^\top \Sigma w)}{\sigma(w)^3}.$$

So $\partial_{w_j}\partial_{w_i}\sigma(w) \le 0$ if and only if

$$(w^\top \Sigma w)(e_i^\top \Sigma e_j) \le (e_i^\top \Sigma w)(e_j^\top \Sigma w). \tag{A.4}$$

Straightforward algebra confirms that this condition holds throughout the unit square if $N = 2$. For diagonal Σ, the left side of (A.4) is zero and the right side is nonnegative for any $w \in (0,1)^N$. It only remains to show that the diagonal condition is necessary if $N \ge 3$.

By continuity, if (A.4) holds throughout $(0,1)^N$ then it holds throughout $[0,1]^N$. Set $w = e_i + \epsilon e_k$, $k \ne i, j$, and set

$$\begin{aligned} d(\epsilon) &= (w^\top \Sigma w)(e_i^\top \Sigma e_j) - (e_i^\top \Sigma w)(e_j^\top \Sigma w) \\ &= 2(e_i^\top \Sigma e_k)(e_i^\top \Sigma e_j)\epsilon - [(e_i^\top \Sigma e_i)(e_j^\top \Sigma e_k) \\ &\quad + (e_i^\top \Sigma e_k)(e_i^\top \Sigma e_j)]\epsilon + O(\epsilon^2). \end{aligned}$$

Then $d(0) = 0$ and (A.4) implies that $d(\epsilon) \le 0$, for all $\epsilon \in [0,1]$, so $d'(0) \le 0$. But applying (A.4) with $w = e_i$ and e_k in place of e_i implies that

$$d'(0) = (e_i^\top \Sigma e_k)(e_i^\top \Sigma e_j) - (e_i^\top \Sigma e_i)(e_j^\top \Sigma e_k) \ge 0,$$

so we must in fact have $d'(0) = 0$. Dividing through by $\sigma^2(e_i)\sigma(e_j)\sigma(e_k)$, we can rewrite the equation $d'(0) = 0$ in terms of correlations as $R_{ik}R_{ij} = R_{jk}$. Because i, j, and k are arbitrary distinct indices, we can use the corresponding identity to replace R_{ij} and get $R_{ik}(R_{ik}R_{jk}) = R_{jk}$. If Σ has full rank, then $R_{ik}^2 < 1$, so we must have $R_{jk} = 0$. For this to hold for all distinct j and k, Σ must be diagonal. □

Proof of Proposition 2.3. Write $\sigma_P(\cdot)$ for standard deviation with respect to the $K \times K$ bundled covariance matrix $P^\top \Sigma P$. For any $x, y \in \{0,1\}^K$, orthogonality of the columns of P implies that $Px, Py \in \{0,1\}^N$, so submodularity of Σ yields

$$\sigma(Px \wedge Py) + \sigma(Px \vee Py) \le \sigma(Px) + \sigma(Py);$$

this is a restatement of (2), taking $A = \{i : (Px)_i = 1\}$, $B = \{i : (Py)_i = 1\}$. Orthogonality of the columns of P also yields $Px \wedge Py = P(x \wedge y)$ and $Px \vee Py = P(x \vee y)$, so we have

$$\sigma(P(x \wedge y)) + \sigma(P(x \vee y)) \le \sigma(Px) + \sigma(Py),$$

which is to say that

$$\sigma_P(x \wedge y) + \sigma_P(x \vee y) \le \sigma_P(x) + \sigma_P(y),$$

and this is what we need to conclude that $P^\top \Sigma P$ is submodular. For the second statement, we know from Proposition 2.2 that any bundling with $K = 2$ is submodular. □

Proof of Proposition 2.4. We will show that the variance function σ^2 on $\{0,1\}^N$ is submodular and increasing. That (i) implies submodularity of variance is easy to see directly and is proved in Murota (2003, proposition 2.6, p. 44). For monotonicity, suppose $w, x, w + x \in \{0,1\}^N$. Then

$$\begin{aligned} \sigma^2(w + x) - \sigma^2(w) &= (w + x)^\top \Sigma(w + x) - w^\top \Sigma w \\ &= x^\top \Sigma(x + 2w) \\ &= \sum_i x_i \left(\sigma_i^2 + \sum_{j \ne i} (x_j + 2w_j)\sigma_{ij} \right), \end{aligned}$$

where the last equality uses the fact that $x_i^2 = x_i$ and $x_i w_i = 0$. For each i,

$$\sigma_i^2 + \sum_{j \ne i} (x_j + 2w_j)\sigma_{ij} \ge \sigma_i^2 + 2\sum_{j \ne i} \sigma_{ij} \ge 0,$$

using condition (i) for the first inequality and condition (ii) for the second. Thus, $\sigma^2(w + x) \ge \sigma^2(x)$.

An increasing concave function of an increasing submodular function is increasing and submodular, so the standard deviation inherits these properties from the variance. More explicitly, suppose $A, B \subseteq S$ with $\sigma(A) \le \sigma(B)$. Then monotonicity yields

$$\sigma^2(A \cap B) \le \sigma^2(A) \le \sigma^2(B) \le \sigma^2(A \cup B)$$

and submodularity adds

$$\sigma^2(A) - \sigma^2(A \cap B) \ge \sigma^2(A \cup B) - \sigma^2(B).$$

For any four positive numbers satisfying these inequalities, their square roots satisfy

$$\sigma(A) - \sigma(A \cap B) \ge \sigma(A \cup B) - \sigma(B),$$

because the square root function is increasing and concave. □

Proof of Proposition 2.6. Under condition (7), we may use (10) to write $\sigma(x) = g(v^\top x)$, where $g(s) = \sqrt{s + as^2}$ is concave on $[0,\infty)$. If $x + e_i + e_j \in \{0,1\}^N$, $i \neq j$, let $s_1 = v^\top x$, $s_2 = v^\top(x + e_i)$, $s_3 = v^\top(x + e_j)$, and $s_4 = v^\top(x + e_i + e_j)$. Without loss of generality, suppose $s_2 \le s_3$. As $v_i, v_j \ge 0$, we have $s_1 \le s_2$ and $s_3 \le s_4$, and we also have $s_2 - s_1 = s_4 - s_3$. Concavity of g therefore implies that $g(s_4) - g(s_3) \le g(s_2) - g(s_1)$, which reduces to (5).

Now let

$$g(s,t) = \sqrt{s + as^2 + t + bt^2},$$

and observe that if Σ has the form in (8), then $\sigma(x) = g(v^\top x, w^\top x)$. We claim that the function g is *directionally concave* on $[0,\infty)\times[0,\infty)$, meaning that it is concave in each argument and submodular (Shaked and Shanthikumar 2007, p. 335). To establish this property, it suffices to show that all second derivatives of g are nonpositive. For the mixed derivatives we have

$$\partial_t\partial_s g(s,t) = -\frac{(1+2as)(1+2bt)}{4g(s,t)^3} \le 0.$$

For concavity in each argument, consider the function $u \mapsto \sqrt{c + u + au^2}$. This function is concave in u if $4ac \le 1$. So, for $g(s,t)$ to be concave in s for fixed t, we need $4a(t + bt^2) \le 1$. As x ranges over the vertices of the unit hypercube, $w^\top x$ is bounded by $|w|$, so it suffices to satisfy the inequality at $t = |w|$, which is the first condition in (9). Concavity of $g(s,t)$ in t similarly follows from the second condition in (9).

Let $s_1, \dots, s_4$ be as before, and set $t_1 = w^\top x$, $t_2 = w^\top(x + e_i)$, $t_3 = w^\top(x + e_j)$, and $t_4 = w^\top(x + e_i + e_j)$. Then $t_1 \le \min(t_2, t_3) \le \max(t_2, t_3) \le t_4$, and $(s_1, t_1) + (s_4, t_4) = (s_2, t_2) + (s_3, t_3)$. Directional concavity of g yields (Shaked and Shanthikumar 2007, p. 335), $g(s_4, t_4) - g(s_3, t_3) \le g(s_2, t_2) - g(s_1, t_1)$, which reduces to (5). □

Proof of Proposition 2.7. Let $c = \sigma^2\rho$, the off-diagonal value in (6). To show submodularity of Σ_ξ, it suffices to show that

$$\begin{aligned}&\left(n\sigma^2 + n(n-1)c + \sum_{i=1}^n \xi_i\right)^{1/2} \\ &\quad - \left((n-1)\sigma^2 + (n-1)(n-2)c + \sum_{i=1}^{n-1}\xi_i\right)^{1/2} \\ &\qquad \le \left((n-1)\sigma^2 + (n-1)(n-2)c + \sum_{i=1}^{n-2}\xi_i + \xi_n\right)^{1/2} \\ &\quad - \left((n-2)\sigma^2 + (n-2)(n-3)c + \sum_{i=1}^{n-2}\xi_i\right)^{1/2} \qquad \text{(A.5)}\end{aligned}$$

for $3 \le n \le N$. The first term on the left side can be rearranged as

$$\left(cn^2 + (\sigma^2 - c)n + \sum_{i=1}^n \xi_i\right)^{1/2}.$$

So, (A.5) holds if $f(x,y) = \sqrt{cx^2 + (\sigma^2 - c)x + y}$ is directionally concave for $(x,y) \in [3,N]\times[0,|\xi|]$. For the mixed derivatives to be nonpositive, we need $\sigma^2 + c(2x-1) \ge 0$, and this holds for all $3 \le x \le N$, if $\rho \ge -1/(2N-1)$. The function f is clearly concave in y. Concavity in x requires $(\sigma^2 - c)^2 \ge 4cy$. This condition holds when $c \le 0$. For $c > 0$, dividing both sides by σ^4 shows that this condition follows from (11). □

Proof of Proposition 2.9. (i) We need to show that for any pairwise orthogonal $x, y, w \in \{0,1\}^N$,

$$\sigma(x + y + w) - \sigma(x + y) \le \sigma(x + w) - \sigma(x). \qquad \text{(A.6)}$$

This inequality holds trivially if either $w = \mathbf{0}$ or $y = \mathbf{0}$; if $x = \mathbf{0}$, the inequality reads

$$\sigma(y + w) \le \sigma(y) + \sigma(w),$$

which holds for all y, w. So, it suffices to restrict attention to $x, y, w \in \{0,1\}^N\backslash\{\mathbf{0}\}$. By writing

$$\sigma(x + y + w) - \sigma(x + y) = \int_0^1 \partial_t\sigma(x + y + tw)\,dt$$

and

$$\sigma(x + w) - \sigma(x) = \int_0^1 \partial_t\sigma(x + tw)\,dt,$$

we find that it suffices to show that $\partial_t\sigma(x + y + tw) \le \partial_t\sigma(x + tw)$, for all $t \in [0,1]$. For any $z \in \{0,1\}^N\backslash\{\mathbf{0}\}$, differentiation yields

$$\begin{aligned}\partial_t\sigma(z + tw) &= \frac{\partial_t[(z + tw)^\top\Sigma(z + tw)]}{2\sigma(z + tw)} \\ &= \frac{2z^\top\Sigma w + 2tw^\top\Sigma w}{2\sigma(z + tw)} \\ &= \frac{(z + tw)^\top\Sigma w}{\sigma(z + tw)} = \rho(z + tw, w)\sigma(w).\end{aligned}$$

So, if condition (i) holds, then

$$\begin{aligned}\partial_t\sigma(x + y + tw) &= \rho(x + y + tw, w)\sigma(w) \le \rho(x + tw, w)\sigma(w) \\ &= \partial_t\sigma(x + tw).\end{aligned}$$

(ii) Write $\nabla\sigma(x)$ for the gradient of σ at $x \in \mathbb{R}^N\backslash\{\mathbf{0}\}$. As a mapping from $\mathbb{R}^N$ to $\mathbb{R}$, σ is convex, so for any $x, y \in \mathbb{R}^N\backslash\{\mathbf{0}\}$, we have $\sigma(y) \ge \sigma(x) + \nabla\sigma(x)\cdot(y - x)$. In particular,

$$\sigma(x + w) - \sigma(x) \ge \nabla\sigma(x)\cdot w,$$

and

$$\sigma(x + y + w) - \sigma(x + y) \le \nabla\sigma(x + y + w)\cdot w.$$

At any $x \neq \mathbf{0}$, $\nabla\sigma(x) = x^\top\Sigma\backslash\sigma(x)$. Condition (13) implies

$$\frac{(x + y + w)^\top\Sigma w}{\sigma(x + y + w)} \le \frac{x^\top\Sigma w}{\sigma(x)},$$

which we can rewrite as $\nabla\sigma(x + y + w)\cdot w \le \nabla\sigma(x)\cdot w$, from which we get (A.6).

For logsubmodularity, let $g(w) = (1/2)\log\sigma^2(w)$. Then

$$\partial_{w_j}\partial_{w_i} g(w) = \frac{\sigma^2(w)\partial_{w_j}\partial_{w_i}\sigma^2(w) - \partial_{w_i}\sigma^2(w)\partial_{w_j}\sigma^2(w)}{2\sigma^4(w)}.$$

So g is submodular if the numerator on the right is nonpositive, which holds if

$$(w^\top\Sigma w)(e_i^\top\Sigma e_j) \le 2(e_i^\top\Sigma w)(e_j^\top\Sigma w). \qquad \text{(A.7)}$$

Dividing both sides by $\sigma^2(w)\sigma(e_i)\sigma(e_j)$ yields the result. For an exchangeable Σ as in (6), we may take $\sigma = 1$. The left side of (A.7) can be written as

$$\rho x_i^2 + \rho x_j^2 + 2\rho^2 x_i x_j + 2\rho^2(x_i + x_j)\sum_{k\neq i,j} x_k + \rho\sum_{k\neq i,j} x_k^2 + \rho^2\sum_{k,l\neq i,j} x_k x_l \tag{A.8}$$

with distinct k and l in the last term. The right side of (A.7) can be written as

$$2\rho x_i^2 + 2\rho x_j^2 + 2(\rho^2+1)x_i x_j + 2(\rho+\rho^2)(x_i+x_j)\sum_{k\neq i,j} x_k + 2\rho^2\sum_{k\neq i,j} x_k^2 + 2\rho^2\sum_{k,l\neq i,j} x_k x_l \tag{A.9}$$

with distinct k and l in the last term. By subtracting (A.8) from (A.9), it is not difficult to see that (A.7) holds when $\rho \geq 1/2$. □

Proof of Proposition 2.11. Suppose Σ satisfies the conditions in Proposition 2.4. As noted following the statement of that proposition, this makes Σ an M-matrix. It then follows by theorem 2.4 (p. 140) of Berman and Plemmons (1979) that all entries of Σ_{22}^{-1} are nonnegative. All entries of Σ_{12} and Σ_{21} are nonpositive, so all entries of $\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ are nonnegative, and all off-diagonal entries of $\Sigma_{1|2}$ are nonpositive.

Suppose we condition on a single variable, which we may assume to be X_N. Then the entries of $\Sigma_{1|2}$ take the form

$$\Sigma_{1|2}(i,i) = \sigma_i^2 - \frac{\sigma_{iN}^2}{\sigma_N^2}, \quad \Sigma_{1|2}(i,j) = \sigma_{ij} - \frac{\sigma_{iN}\sigma_{jN}}{\sigma_N^2}, \quad i,j = 1,\dots,N-1,\ j\neq i.$$

We verify condition (ii) of Proposition 2.4 with $i = 1$ as follows:

$$\begin{aligned}
-2\sum_{j=2}^{N-1}\Sigma_{1|2}(j,1) &= -2\sum_{j=2}^{N-1}\sigma_{1j} + 2\frac{\sigma_{1N}}{\sigma_N^2}\sum_{j=2}^{N-1}\sigma_{jN} \\
&= -2\sum_{j=2}^{N}\sigma_{1j} + 2\sigma_{1N} + 2\frac{\sigma_{1N}}{\sigma_N^2}\sum_{j=1}^{N-1}\sigma_{jN} - 2\frac{\sigma_{1N}^2}{\sigma_N^2} \\
&\leq \sigma_1^2 + 2\sigma_{1N} - \frac{\sigma_{1N}}{\sigma_N^2}\sigma_N^2 - 2\frac{\sigma_{1N}^2}{\sigma_N^2} \\
&\leq \sigma_1^2 - 2\frac{\sigma_{1N}^2}{\sigma_N^2} \\
&\leq \sigma_1^2 - \frac{\sigma_{1N}^2}{\sigma_N^2} = \Sigma_{1|2}(1,1),
\end{aligned}$$

where the first inequality applies condition (ii) of Proposition 2.4 to Σ with $i = 1$ and $i = N$, recalling that $\sigma_{1N} \leq 0$. As the index $i = 1$ is arbitrary, the condition holds for all $i = 1,\dots,N-1$, and $\Sigma_{1|2}$ satisfies the conditions of Proposition 2.4. Proceeding by induction, the conditions are preserved when we condition on any arbitrary subset of $X_1,\dots,X_N$.

Now suppose Σ has the form in (7), and suppose we condition on X_N. Let $v_{1|2}$ denote the truncation of v to its first $N-1$ elements. We can write Σ as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & av_N v_{1|2} \\ av_N v_{1|2}^\top & v_N + av_N^2 \end{pmatrix}, \quad \Sigma_{11} = \mathsf{diag}(v_{1|2}) + av_{1|2}v_{1|2}^\top.$$

Then

$$\Sigma_{1|2} = \Sigma_{11} - \frac{a^2 v_N}{1+av_N}v_{1|2}v_{1|2}^\top = \mathsf{diag}(v_{1|2}) + \frac{a}{1+av_N}v_{1|2}v_{1|2}^\top.$$

The structure in Proposition 2.6(i) is thus preserved by conditioning, and submodularity is therefore also preserved. The same holds for (6) as a special case of (7). □

Proof of Proposition 2.12. For any $v\in\mathbb{R}^n$, we have $\lambda_{\min}\|v\|^2 \leq v^\top\Sigma v \leq \lambda_{\max}\|v\|^2$. Thus,

$$\sigma(x) + \sigma(y+e_i) \leq \sqrt{\lambda_{\max}}(\|x\| + \sqrt{\|y\|^2+1}),$$

and

$$\sigma(x+e_i) + \sigma(y) \geq \sqrt{\lambda_{\min}}(\sqrt{\|x\|^2+1} + \|y\|).$$

We therefore have

$$\gamma \leq \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\cdot\frac{\|x\| + \sqrt{\|y\|^2+1}}{\sqrt{\|x\|^2+1}+\|y\|} = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}\,f(\|y\|^2+1, \|y\|^2-\|x\|^2+1),$$

where

$$f(t,s) = \frac{\sqrt{t-s}+\sqrt{t}}{\sqrt{t-s+1}+\sqrt{t-1}}, \quad 3\leq t\leq N, \quad 2\leq s\leq t-1.$$

Differentiation shows that f is an increasing function of t and a decreasing function of s, so its maximum is attained at $t = N$, $s = 2$. Making these substitutions in f yields (17). □

B. Proofs for Sections 4 and 5

Proof of Proposition 4.1. Under the conditions of Proposition 2.4, $A\subseteq B$ implies $\sigma(A)\leq\sigma(B)$, and (29) follows. For case (ii), we have $\sigma(A) = \sigma\sqrt{|A| + |A|(|A|-1)\rho}$. If $\tilde\rho \geq \rho \geq -1/(2N-1)$, then $\sigma(\cdot)$ and $\tilde\sigma(\cdot)$ are monotone increasing, so $\tilde\sigma(B) - \tilde\sigma(A) \geq (\sigma/\tilde\sigma)[\tilde\sigma(B) - \tilde\sigma(A)]$; in other words, it suffices to verify (29) when $\tilde\sigma = \sigma$. In this case, (29) states that for the function $(m,\rho)\mapsto\sqrt{m+m(m-1)\rho}$, with $m = 1,\dots,N$ and $-1/(N-1) < \rho < 1$, differences in m are increasing in ρ. If we extend this function to $m\in[1,N]$, the mixed second derivative of the extension is positive on $[1,N]\times(-1/(N-1),1)$, so the result follows.

In case (iii), we can extend $\sigma(\cdot)$ to a function of $x\in[0,1]^N$ and $a\geq 0$, with $v\in\mathbb{R}_+^N$ fixed, by setting $g(x,a) = \sqrt{v^\top x + a(v^\top x)^2}$. Differentiation shows that derivatives of this function with respect to each x_i, $i = 1,\dots,N$, are increasing in a. This implies that, for any $x, x+e_i\in[0,1]^N$, the difference $\sigma(x+e_i) - \sigma(x)$ is increasing in a, from which (29) follows. □

Proof of Proposition 4.2. The mapping $A\mapsto|A|$ is increasing and nonnegative. It follows from theorem 1 of Friedland (2013) that under the condition in (i), $A\mapsto\log\det\Sigma_A$ is increasing and nonnegative (taking $\det\Sigma_\emptyset = 1$), so the same holds for $\mathcal{H}$. For (ii), we will use the Hadamard inequality $\det\Sigma_{A\cup j} \leq \Sigma_{jj}\det\Sigma_A$, $A\subseteq S$, $j\notin A$. We will also use the Jacobi identity

$$\frac{\det\Sigma_A}{\det\Sigma} = \det\Sigma_{\bar A}^{-1},$$

where $\Sigma_{\bar A}^{-1}$ is the submatrix of Σ^{-1} formed by the rows and columns i, $i\notin A$; see equation (12) of Brualdi and

Schneider (1983). For any $A \subseteq S$ and $j \notin A$, the Hadamard inequality and the assumption in (ii) yield

$$\frac{\det \Sigma_{A\cup j}}{\det \Sigma_A} \le \Sigma_{jj} \le \frac{1}{\tilde{\Sigma}^{-1}_{jj}}.$$

Applying the Jacobi identity and then the Hadamard inequality we get

$$\frac{\det \tilde{\Sigma}_{A\cup j}}{\det \tilde{\Sigma}_A} = \frac{\det \tilde{\Sigma}_{A\cup j}/\det \tilde{\Sigma}}{\det \tilde{\Sigma}_A/\det \tilde{\Sigma}} = \frac{\det \tilde{\Sigma}^{-1}_{A\cup j}}{\det \tilde{\Sigma}^{-1}_{A}} \ge \frac{\det \tilde{\Sigma}^{-1}_{A\cup j}}{\tilde{\Sigma}^{-1}_{jj} \det \tilde{\Sigma}^{-1}_{A\cup j}} = \frac{1}{\tilde{\Sigma}^{-1}_{jj}},$$

so we have shown that $\log\det \tilde{\Sigma}_{A\cup j} - \log\det \tilde{\Sigma}_A \ge \log\det \Sigma_{A\cup j} - \log\det \Sigma_A$. By applying this inequality repeatedly, it follows that for any $A \subseteq B \subseteq S$ $\log\det \tilde{\Sigma}_B - \log\det \tilde{\Sigma}_A \ge \log\det \Sigma_B - \log\det \Sigma_A$, which is precisely the relation in (27) applied to the log det function with respect to Σ and $\tilde{\Sigma}$. It follows from (18) that $\tilde{\mathcal{H}} \to \mathcal{H}$. □

Proof of Proposition 5.1. The lower bounds hold by definition so we just prove the upper bounds. For any real numbers a and b, we have $2(a \wedge b) + |a - b| = a + b$. It follows that for any $x \in B(F)$ and $y \in B(G)$,

$$\begin{aligned} 2\sum_{i=1}^N x_i \wedge y_i &= \sum_{i=1}^N x_i + \sum_{i=1}^N y_i - \sum_{i=1}^N |x_i - y_i| \\ &= F(S) + G(S) - \sum_{i=1}^N |x_i - y_i|, \end{aligned}$$

so we can rewrite (32) as

$$\begin{aligned} c_{sys} &= \frac{F(S)+G(S)}{2} - \frac{1}{2} \min_{x\in B(F), y\in B(G)} \sum_{i=1}^N |x_i - y_i| \\ &= \frac{F(S)+G(S)}{2} - \Delta(F,G). \end{aligned}$$

We can similarly write (30) as

$$c_k = \frac{F_k(S) + G_k(S)}{2} - \Delta(F_k, G_k).$$

Summing over k and subtracting the sum from c_{sys} yields the first result. Also, $\Delta(F,G) \ge 0$, and

$$\sum_{k=1}^K \Delta(F_k, G_k) \le \min_{x\in B(F), y\in B(G)} \frac{1}{2} \sum_{k=1}^K \sum_{i=1}^N |x_i^k - y_i^k| = D(F,G).$$

To bound c_{seq}, we need to investigate the optimal allocation decision for a dealer once some of the dealer's trades have already been allocated by other dealers. To examine this problem generically, let f and g be normalized submodular functions on 2^S. Consider a constrained allocation problem that requires allocation of a set C_1 to f and a set C_2 to g. The following lemma compares optimal costs with and without these constraints.

Lemma B.1. *Let f and g be normalized submodular functions on 2^S. Let C_1 and C_2 be disjoint subsets of S, and let $C = C_1 \cup C_2$ and $S_C = S\backslash C$. Then*

$$\begin{aligned} &\min_{A\subseteq S_C} f(A\cup C_1) + g((S_C\backslash A)\cup C_2) \qquad \text{(B.1)} \\ &\le \min_{A\subseteq S}\{f(A) + g(S\backslash A)\} \\ &\quad + \max\left\{\sum_{i\in C_1\cup C_2} |x_i - y_i|, x\in B(f), y\in B(g)\right\}. \qquad \text{(B.2)} \end{aligned}$$

Proof. For $A \subseteq S\backslash C$, let

$$f_{C_1}(A) = f(A\cup C_1) - f(C_1), \quad g_{C_2}(A) = g(A\cup C_2) - g(C_2).$$

Then f_{C_1} and g_{C_2} are normalized submodular functions on 2^{S_C}, $S_C = S\backslash C$, and we may write (B.1) as

$$\begin{aligned} &f(C_1) + g(C_2) + \min_{A\subseteq S_C} f_{C_1}(A) + g_{C_2}(S_C\backslash A) \\ &\quad = f(C_1) + g(C_2) + \max\left\{\sum_{i\in S_C} x_i \wedge y_i,\ x\in B(f_{C_1}), y\in B(g_{C_2})\right\}. \end{aligned} \qquad \text{(B.3)}$$

Without loss of generality, we may take $C_1 = \{1,\dots,n_1\}$, $S_C = \{n_1+1,\dots,n_2\}$, and $C_2 = \{n_2+1,\dots,N\}$, for some $0 \le n_1 \le n_2 \le N$. An element of $B(f_{C_1})$ or $B(g_{C_2})$ has dimension $|S_C| = n_2 - n_1$, and an element of $B(f)$ or $B(g)$ has dimension $|C_1| + |S_C| + |C_2| = |S| = N$. We can extend a base vector of f_{C_1} (or g_{C_2}) to a base vector of f (or g) as follows:

Lemma B.2. *For any $\tilde{x} \in B(f_{C_1})$ there is an $x \in B(f)$ with $x_i = \tilde{x}_i$, $i \in S_C$, and $\sum_{i\in C_1} x_i = f(C_1)$.*

Proof. The proof uses the following property from Shapley (1971), theorem 3: For any normalized submodular F on 2^S, a vector $x \in \mathbb{R}^N$ is a vertex of $B(F)$ if and only if it has the form in (23) for some permutation $i_1,\dots,i_N$ of the indices $1,\dots,N$. We will expand $\tilde{x}$ to a vector $(x_1,\dots,x_{n_1},\tilde{x},x_{n_2+1},\dots,x_N)$, so we index the elements of $\tilde{x}$ by $i = n_1+1,\dots,n_2$. If $\tilde{x}$ is a vertex of $B(f_{C_1})$, it has the form in (23) for some permutation $i_{n_1+1},\dots,i_{n_2}$ of the indices $n_1+1,\dots,n_2$, with F replaced by f_{C_1}. Let x be the extreme point of $B(f)$ defined by the permutation $1,\dots,n_1,i_{n_1+1},\dots,i_{n_2},n_2+1,\dots,N$. Then $\sum_{i\in C_1} x_i = f(C_1)$. Moreover, for any $j = n_1+1,\dots,n_2$, $x_j = f(C_1\cup\{i_{n_1+1},\dots,i_j\}) - f(C_1\cup\{i_{n_1+1},\dots,i_{j-1}\})$, and therefore $x_j = f_{C_1}(\{i_{n_1+1},\dots,i_j\}) - f_{C_1}(\{i_{n_1+1},\dots,i_{j-1}\}) = \tilde{x}_j$. In other words, we have constructed an element of $B(f)$ that coincides with $\tilde{x}$ on S_C. If $\tilde{x}$ is not a vertex of $B(f_{C_1})$, we may write it as a convex combination of vertices and extend each vertex to an element of $B(f)$. The corresponding convex combination of these extensions is an element of $B(f)$ that agrees with $\tilde{x}$ on the elements of S_C. □

The corresponding property holds for g_{C_2}, so it follows from this lemma that (B.3) is bounded above by

$$\begin{aligned} &\max\left\{\sum_{i\in C_1} x_i + \sum_{i\in C_2} y_i + \sum_{i\in S_C} x_i\wedge y_i,\ x\in B(f), y\in B(g)\right\} \\ &\le \max\left\{\sum_{i\in S} x_i\wedge y_i,\ x\in B(f), y\in B(g)\right\} \\ &\quad + \max\left\{\sum_{i\in C_1} x_i + \sum_{i\in C_2} y_i - \sum_{i\in C_1\cup C_2} x_i\wedge y_i,\ x\in B(f), y\in B(g)\right\} \\ &= \max\left\{\sum_{i\in S} x_i\wedge y_i,\ x\in B(f), y\in B(g)\right\} + \max\left\{\sum_{i\in C_1}(x_i - y_i)^+ \right. \\ &\quad \left. + \sum_{i\in C_2}(y_i - x_i)^+,\ x\in B(f), y\in B(g)\right\}, \end{aligned}$$

which is bounded above by (B.2). □

We now return to the proof of Proposition 5.1 and consider the constrained optimization problem of dealer k, once the trades in C_{k-1} have been allocated. Applying (B.2), we get

$\bar{c}_k \le c_k + \delta_k(C_{k-1})$, and summing over $k = 2, \ldots, K$ concludes the proof. □

Endnote

[1] We use the term *allocation* for the decision in (1) to assign trades to one portfolio or another. To avoid confusion, in breaking down a total charge $F(S)$ into a sum of trade-specific charges, we will refer to a *decomposition* or *attribution* of costs, rather than an allocation. The allocation decision in (1) seeks to minimize total cost, whereas a cost decomposition or attribution keeps the total cost fixed.

References

Andersen L, Duffie D, Song Y (2019) Funding value adjustments. *J. Finance* 74(1):145–192.

Anily S, Haviv M (2007) The cost allocation problem for the first order interaction joint replenishment model. *Oper. Res.* 55(2):292–302.

Artzner P, Delbaen F, Eber JM, Heath D (1999) Coherent measures of risk. *Math. Finance* 9(3):203–228.

Aubin J-P (1981) Cooperative fuzzy games. *Math. Oper. Res.* 6(1):1–13.

Backus D, Chernov M, Zin S (2014) Sources of entropy in representative agent models. *J. Finance* 69(1):51–99.

Berman A, Plemmons RJ (1979) *Nonnegative Matrices in the Mathematical Sciences* (Academic Press, New York).

Bertsimas D, Shioda R (2009) Algorithm for cardinality-constrained portfolio optimization. *Comput. Optim. Appl.* 43(1):1–22.

Brualdi RA, Schneider H (1983) Determinantal identities: Gauss, Schur, Cauchy, Sylvester, Kronecker, Jacobi, Binet, Laplace, Muir, and Cayley. *Linear Algebra Appl.* 52/53:768–791.

Chakrabarty D, Jain P, Kothari P (2014) Provable submodular minimization using Wolfe's algorithm, *Adv. Neural Inform. Processing Systems* 27:802–809.

Cornuejols G, Tütüncü R (2007) *Optimization Methods in Finance* (Cambridge University Press, Cambridge, UK).

Das A, Kempe D (2011) Submodular meets Spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. *Proc. 28th Internat. Conf. Machine Learn. (ICML-11), Bellevue, WA*, 1057–1064.

Denault M (2001) Coherent allocation of risk capital. *J. Risk* 4(1):1–34.

De Genaro A (2016) Systematic multi-period stress scenarios with an application to CCP risk management. *J. Banking Finance* 67:119–134.

Duffie D, Scheicher M, Vuillemey G (2015) Central clearing and collateral demand. *J. Financial Econom.* 116(2):237–256.

Edmonds J (1970) Submodular functions, matroids, and certain polyhedra. Guy R, Hanani H, Sauer N, Schönheim J, eds. *Combinatorial Structures and Their Applications* (Gordon and Breach, New York).

Embrechts P, Liu H, Wang R (2018) Quantile-based risk sharing. *Oper. Res.* 66(4):893–1188.

Fan K (1968) An inequality for subadditive functions on a distributive lattice, with application to determinantal inequalities. *Linear Algebra Appl.* 1(1):33–38.

Friedland S (2013) Nonnegative definite hermitian matrices with increasing principal minors. *Special Matrices* 1:1–2.

Fujishige S (1978) Polymatroid dependence structure of a set of random variables. *Inform. Control* 39(1):55–72.

Fujishige S (1980) Lexicographically optimal base of a polymatroid with respect to a weight vector. *Math. Oper. Res.* 5(2):186–196.

Fujishige S (2005) *Submodular Functions and Optimization*, 2nd ed. (Elsevier, Amsterdam).

Fujishige S, Nagano K (2009) A structure theory for the parametric submodular intersection problem. *Math. Oper. Res.* 34(3):513–521.

Fujishige S, Hayashi T, Isotani S (2006) *The Minimum-Norm-Point Algorithm Applied to Submodular Function Minimization and Linear Programming* (Research Institute for Mathematical Sciences, Kyoto University, Kyoto, Japan).

Gao J, Li D (2013) Optimal cardinality constrained portfolio selection, *Oper. Res.* 61(3):745–761.

Ghamami S, Glasserman P (2017) Does OTC derivatives reform incentivize central clearing? *J. Financial Intermediation*, 32:76–87.

Goemans MX, Harvey NJA, Iwata S, Mirrokni V (2009) Approximating submodular functions everywhere. Working paper, Massachusetts Institute of Technology, Cambridge.

Gregory J (2015) *The XVA Challenge: Counterparty Credit Risk, Funding, Collateral, and Capital*, 3rd ed. (John Wiley & Sons, New York).

Grötschel M, Lovász L., Schrijver A (1981) The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica* 1(2):169–197.

Heller D, Vause N (2012) Collateral requirements for mandatory central clearing of over-the-counter derivatives. BIS Working Paper 373, Bank for International Settlements, Basel, Switzerland.

Hodges SD, Neuberger A (1989) Optimal replication of contingent claims under transactions costs. *J. Futures Markets* 8: 223–242.

Iancu DA, Trichakis N (2014) Fairness and efficiency in multiportfolio optimization. *Oper. Res.* 62(6):1285–1301.

Iwata S, Orlin JB (2009) A simple combinatorial algorithm for submodular function minimization. *Proc. 20th Annual ACM-SIAM Sympos. Discrete Algorithms* (Society for Industrial and Applied Mathematics, Philadelphia), 1230–1237.

Lobo M, Fazel M, Boyd S (2007) Portfolio optimization with linear and fixed transaction costs *Ann. Oper. Res.* 152(1):341–365.

McCormick ST (2005) Submodular function minimization. Aardal K, Nemhauser GL, Weismantel R, eds. *Discrete Optimization*, Handbooks in Operations Research and Management Science, vol. 12 (Elsevier, Amsterdam), 321–391.

McNeil AJ, Frey R, Embrechts P (2005) *Quantitative Risk Management* (Princeton University Press, Princeton, NJ).

Murota K (1988) Note on the universal bases of a pair of polymatroids. *J. Oper. Res. Soc. Japan* 31(4):565–572.

Murota K (2003) *Discrete Convex Analysis* (Society for Industrial and Applied Mathematics, Philadelphia).

Nakamura M (1988) Structural theorems for submodular functions, polymatroids and polymatroid intersections. *Graphs Combinatorics* 4(1):257–284.

Orso A, Lee J, Shen S (2015) Submodular minimization in the context of modern LP and MILP methods and solvers. *Internat. Sympos. Experiment. Algorithms* (Springer, Cham, Switzerland), 193–204.

Philippatos GC, Wilson CJ (1972) Entropy, market risk, and the selection of efficient portfolios. *Appl. Econom.* 4(3):209–220.

Schneider M, Lillo F (2019) Cross-impact and no-dynamic-arbitrage. *Quant. Finance* 19(1):137–154.

Schrijver A (2003) *Combinatorial Optimization* (Springer, Berlin).

Shaked M, Shanthikumar JG (2007) *Stochastic Orders* (Springer, New York).

Shapley LS (1971) Cores of convex games. *Internat. J. Game Theory* 1(1): 11–26.

Sidanius C, Zikes F (2012) OTC derivatives reform and collateral demand impact. Financial Stability Paper 18, Bank of England, London.

Sirignano J, Tsoukalas G, Giesecke K (2016) Large-scale loan portfolio selection. *Oper. Res.* 64(6):1239–1255.

Topkis DM (1998) *Supermodularity and Complementarity* (Princeton University Press, Princeton, NJ).

Tsoukalas G, Wang J, Giesecke K (2017) Dynamic portfolio execution. *Management Sci.*, ePub ahead of print October 27, https://doi.org/10.1287/mnsc.2017.2865.

Vause N (2010) Counterparty risk and contract volumes in the credit default swap market. *BIS Quart. Rev.* (December):59–69.