Bounding Wrong-Way Risk in CVA Calculation

Paul Glasserman*, Linan Yang†

May 2015

Abstract

A credit valuation adjustment (CVA) is an adjustment applied to the value of a derivative contract or a portfolio of derivatives to account for counterparty credit risk. Measuring CVA requires combining models of market and credit risk to estimate a counterparty's risk of default together with the market value of exposure to the counterparty at default. Wrong-way risk refers to the possibility that a counterparty's likelihood of default increases with the market value of the exposure. We develop a method for bounding wrong-way risk, holding fixed marginal models for market and credit risk and varying the dependence between them. Given simulated paths of the two models, a linear program computes the worst-case CVA. We analyze properties of the solution and prove convergence of the estimated bound as the number of paths increases. The worst case can be overly pessimistic, so we extend the procedure by constraining the deviation of the joint model from a baseline reference model. Measuring the deviation through relative entropy leads to a tractable convex optimization problem that can be solved through the iterative proportional fitting procedure. Here, too, we prove convergence of the resulting estimate of the penalized worst-case CVA and the joint distribution that attains it. We consider extensions with additional constraints and illustrate the method with examples.

Keywords: credit valuation adjustment, counterparty credit risk, robustness, iterative proportional fitting process (IPFP), I-Projection.

1 Introduction

When a firm enters into a swap contract, it is exposed to market risk through changes in market prices and rates that affect the contract's cash flows. It is also exposed to the risk that the party on the other side of the contract may default and fail to make payments due on the transaction. Thus, market risk determines the magnitude of one party's exposure to another, and credit risk determines the likelihood that this exposure will become a loss. Derivatives counterparty risk refers to this combination of market and credit risk, and proper measurement of counterparty risk requires integrating market uncertainty and credit uncertainty.

*Columbia Business School, Columbia University, New York, NY 10027, email: pg20@columbia.edu.

†Industrial Engineering and Operations Research Department, Columbia University, New York, NY 10027; email: ly2220@columbia.edu.

The standard tool for quantifying counterparty risk is the credit valuation adjustment, CVA, which can be thought of as the price of counterparty risk. Suppose firm A has entered into a set of derivative contracts with firm B. From the perspective of firm A, the CVA for this portfolio of derivatives is the difference between the value the portfolio would have if firm B were default-free and the actual value taking into account the credit quality of firm B. More precisely, this is a unilateral CVA; a bilateral CVA adjusts for the credit quality of both firms A and B.

Counterparty risk generally and CVA in particular have taken on heightened importance since the failures of major derivatives dealers Bear Stearns, Lehman Brothers, and AIG Financial Products in 2008. A new CVA-based capital charge for counterparty risk is among the largest changes to capital requirements under Basel III for banks with significant derivatives activity (BCBS [1]). CVA calculations are significant consumers of bank computing resources, typically requiring simulation of all relevant market variables (prices, interest rates, exchanges rates), valuing every derivative at every time step on every path, and integrating these market exposures with a model of credit risk for each counterparty. See Canabarro and Duffie [12] and Gregory [22] for background on industry practice.

Our focus in this paper is on the effect of dependence between market and credit risk. *Wrong-way risk* refers to the possibility that a counterparty will become more likely to default when the market exposure is larger and the impact of the default is greater; in other words, it refers to positive dependence between market and credit risk. Wrong-way risk arises, for example, if one bank sells put options on the stock of another similar bank. The value of the options increases as the price of the other bank's stock falls; this is likely to be a scenario in which the bank that sold the options is also facing financial difficulty and is less likely to be able to make payment on the options. In practice, the sources and nature of wrong-way risk may be less obvious.

Holding fixed the marginal features of market risk and credit risk, greater positive dependence yields a larger CVA. But capturing dependence between market and credit risk is difficult. There is often ample data available for the separate calibration of market and credit models but little if any data for joint calibration. CVA is calculated under a risk-adjusted probability measure, so historical data is not directly applicable. In addition, for their CVA calculations banks often draw on many valuation models developed for trading and hedging specific types of instruments that cannot be easily integrated with a model of counterparty credit risk. CVA computation is much easier if dependence is ignored. Indeed, the Basel III standarized approach for CVA assumes independence and then multiplies the result by a factor of 1.4; this ad hoc factor is intended to correct for several sources of error, including the lack of dependence information.

Models that explicitly describe dependence between market and credit risk include in CVA

calculation include Brigo, Capponi, and Pallavicini [9], Crépey [15], Hull and White [25], and Rosen and Saunders [30]; see Brigo, Morini, and Pallavicini [10] for an extensive overview of modeling approaches. Dependence is usually introduced by correlating default intensities with market risk factors or through a copula. A direct model of dependence is, in principle, the best approach to CVA. However, correlation-based models generally produce weak dependence between market and credit risk, and both techniques are difficult to calibrate.

In this paper, we develop a method to bound the effect of dependence, holding fixed marginal models of market and credit risk. Our approach uses simulated paths that would be needed anyway for a CVA calculation without dependence. Given paths of market exposures and information (simulated or implied from prices) about the distribution of time to the counterparty's default, we show that finding the worst-case CVA is a linear programming problem. The linear program is easy to solve, and it provides a bound on the potential impact of wrong-way risk. We view this in-sample bound based on a finite set of paths as an estimate of the worst-case CVA for a limiting problem and prove convergence of the estimator. The limiting problem is an optimization over probability measures with given marginals. We also show that the LP formulation has additional useful features. It extends naturally to a bilateral CVA calculation, and it allows additional constraints. Moreover, the dual variables associated with constraints on the marginal default time distribution provide useful information for hedging purposes.

The strength of the LP solution is that it yields the largest possible CVA value — the worst possible wrong-way risk — consistent with marginal information about market and credit risk. This is also a shortcoming, as the worst case can be too pessimistic. We therefore extend the method by penalizing or constraining deviations from a nominal reference model. The reference model could be one in which marginals are independent or linked through some simple model of dependence. A large penalty produces a CVA value close to that obtained under the reference model, and with no penalty we recover the LP solution. Varying the penalty parameter allows us to "interpolate" between the reference model and the worst-case joint distribution.

To penalize deviations from the reference model, we use a relative entropy measure between probability distributions, also known as the Kullback-Leibler divergence. Once we add the penalty, finding the worst-case joint distribution is no longer a linear programming problem, but it is still convex. Moreover, the problem has a special structure that allows convenient solution through iterative rescaling of the rows and columns of a matrix. This iterative rescaling projects a starting matrix onto the convex set of joint distributions with given marginals. Here, too, we prove convergence of the in-sample solution to the solution of a limiting problem as the number of paths increases.

The problem of finding extremal joint distributions with given marginals has a long and rich history. It includes the well-known Fréchet bounds in the scalar case and the multivariate generalization of Brenier [8] and Rüschendorf and Rachev [33]; see the books by Rüschendorf [32] and Villani [36] for detailed treatments and historical remarks. In finance, related ideas have been used to find robust or model-free bounds on option prices; see Cox [14] for a survey. In some versions of the robust pricing problem, one observes prices of simple European options and seeks to bound prices of path-dependent or multi-asset options given the European prices, as in Carr, Ellis, and Gupta [13], Brown, Hobson, and Rogers [11], and Tankov [35], among many others. This has motivated the study of martingale optimal transport problems in Dolinsky and Soner [18], Beiglböck and Juillet [2], Henry-Labordère and Touzi [24]. The literature on price bounds focuses on extremal solutions and does constrain or penalize deviations from a baseline model.

Our focus is not on pricing but rather risk measurement. Within the risk measurement literature, questions of joint distributions with given marginals arise in risk aggregation; see, for example, Bernard, Jiang, and Wang [4], Embrechts and Puccetti [20], and Embrechts, Wang, and Wang [21]. A central problem in risk aggregations is finding the worst-case distribution for a sum of random variables, given marginals for the summands.

Our work differs from earlier work in several respects. We focus on CVA, rather than option pricing or risk aggregation. Our marginals may be quite complex and need not be explicitly available — they are implicitly defined through marginal models for market and credit risk. Given the generality of the setting, we do not seek to characterize extremal joint distributions but rather to estimate bounds using samples generated from the marginals. We temper the bounds by constraining deviations from a reference model, drawing on the idea of robustness as developed in economics in Hansen and Sargent [23] and distributional robustness as developed in the optimization literature in Ben-Tal et al. [3] and references there. The methods we develop are easy to implement in practice. The main contribution lies in the formulation and in the convergence analysis. Our general approach to convergence is to use primal and dual optimization problems to get upper and lower bounds.

The rest of the paper is organized as follows. In Section 2, we introduce the problem setting, and in Section 3 we introduce the optimization formulation for the worst case CVA bound and show convergence of the bound estimator. In Section 4, we extend the problem to a robust formulation with a relative entropy constraint, and we provide numerical examples in Section 5. In Section 6, we extend the model further to incorporate expectation constraints.

2 Problem Formulation

Let τ denote the time at which a counterparty defaults, and let $V(\tau)$ denote the value of a swap (or a portfolio of swaps and other derivatives) with that counterparty at the time of its default, discounted to time zero. The swap value could be positive or negative, so the loss at default is the positive part $V^+(\tau)$. The CVA for a time horizon T is the expected exposure at default,

$$\text{CVA} = \mathbb{E}[V^+(\tau)\mathbf{1}\{\tau \le T\}], \tag{2.1}$$

given a joint law for the default time τ and the exposure V^+. Our focus will be on uncertainty around this joint law, but we first provide some additional details on the problem formulation.

CVA is customarily calculated over a finite set of dates $0 = t_0 < t_1 < \cdots < t_d = T < t_{d+1} = \infty$; for example, these may be the payment dates on the underlying contracts. An underlying simulation of market risk factors generates paths of all relevant market variables and is used to generate exposure paths $(V^+(t_1), \ldots, V^+(t_K))$. Calculating these exposures is a demanding task because it requires valuing all instruments in a portfolio with a counterparty in each market scenario at each date. In addition, the calculation of each $V(t_j)$ needs to account for netting and collateral agreements with the counterparty and recovery rates if the counterparty were to default. The method we develop takes these calculations as inputs and assumes the availability of independent copies of the exposure paths. The market risk model implicitly determines the law of $(V^+(t_1), \ldots, V^+(t_d))$, and we denote this law by a probability measure p on $\mathbb{R}^d$.

The distribution of the counterparty's default time τ may be extracted from credit default swap spreads, or it may be the result of a more extensive credit risk model — for example, a stochastic intensity model. In either case, we suppose that a credit risk model fixes the probabilities q_j, $j = 1, \ldots, d$, that default occurs at t_k, or, more precisely that it occurs in the interval $(t_{k-1}, t_k]$.

Let

$$X = (V^+(t_1), \ldots, V^+(t_d)) \text{ and } Y = (\mathbf{1}\{\tau = t_1\}, \ldots, \mathbf{1}\{\tau = t_d\}).$$

The problem of calculating CVA would reduce to the problem of calculating the expectation of the inner product

$$< X, Y >= \sum_{j=1}^{d} V^+(t_j)\mathbf{1}\{\tau = t_j\} = V^+(\tau)\mathbf{1}\{\tau \le T\},$$

if the joint law for X and Y were known. With the marginals fixed but the joint law unknown, we seek to evaluate the *worst-case* CVA, defined by

$$\text{CVA}_* := \sup_{\mu \in \Pi(p,q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} < x, y > d\mu(x, y), \tag{2.2}$$

where $\Pi(p, q)$ denotes the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals p and q.

The characterization of extremal joint distributions with given marginals has a rich history; see Villani [36] and Rüschendorf [32] for recent treatments with extensive historical remarks. In the scalar case $d = 1$, the largest value of (2.2) is attained by the comonotonic construction, which sets $X = F_p^{-1}(U)$ and $Y = F_q^{-1}(U)$, where F_p and F_q are the cumulative distribution functions associated with p and q, and U is uniformly distributed on $[0, 1]$. The smallest value of (2.2) is attined by setting $Y = F_q^{-1}(1 - U)$ instead. In the vector case, a characterization of joint laws maximizing (2.2) has been given by Brenier [8] and Rüschendorf and Rachev [33]. It states that under an optimal coupling, Y is a subgradient of a convex function of X, but this provides more of a theoretical description than a practical characterization. Our setting has the added complication that at least p (and possibly also q) is itself unknown and only implicitly specified through a simulation model.

3 Worst-Case CVA

3.1 Estimation

We develop a simulation procedure to estimate (2.2). As we noted earlier, generating exposure paths is the most demanding part of a CVA calculation. Our approach essentially reuses these paths to bound the potential effect of wrong-way risk at little additional computational cost.

Let $X_1, \ldots, X_N$ be N independent copies of X, and let $Y_1, \ldots, Y_N$ be N independent copies of Y. Denote their empirical measures on $\mathbb{R}^d$ by

$$p_N(\cdot) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{X_i \in \cdot\}, \quad q_N(\cdot) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{Y_i \in \cdot\}, \tag{3.1}$$

For notational simplicity, we will assume that p has no atoms so that, almost surely, there are no repeated values in $X_1, X_2, \ldots$. This allows us to identify the empirical measure p_N on $\mathbb{R}^d$ with the uniform distribution on the set $\{X_1, \ldots, X_N\}$ or on the set of indices $\{1, \ldots, N\}$. The assumption that p has no atoms is without loss of generality because we can expand the dimension of X to include an independent, continuously distributed coordinate X_{d+1} and expand Y by setting $Y_{d+1} \equiv 0$ without changing (2.2).

Observe that Y is supported on the finite set $\{y_1, \ldots, y_{d+1}\}$, with $y_1 = (1, 0, \ldots, 0), \ldots, y_d = (0, 0, \ldots, 1)$, and $y_{d+1} = (0, \ldots, 0)$. Each y_j has probability $q(y_j)$. These probabilities may be known or estimated from simulation of N independent copies $Y_1, \ldots, Y_N$ of Y, in which case we denote the empirical frequency of each y_j by $q_N(y_j)$.

We will put a joint mass function P_{ij}^N on the set of pairs $\{(X_i, y_j), i = 1, \ldots, N, j = 1, \ldots, d+1\}$. We restrict attention to the set $\Pi(p_N, q_N)$ of joint mass functions with marginals p_N and q_N. We

estimate (2.2) using

$$\widehat{\text{CVA}}_* = \max_{P^N \in \Pi(p_N, q_N)} \sum_{i=1}^{N} \sum_{j=1}^{d+1} P_{ij}^N < X_i, y_j > .$$

Finding the worst-case joint distribution is a linear programming problem:

$$\max_{\{P_{ij}\}} \quad \sum_{i=1}^{N} \sum_{j=1}^{d+1} C_{ij} P_{ij}, \tag{3.2}$$

$$\text{subject to} \quad \sum_{j=1}^{d+1} P_{ij} = 1/N, \ i = 1, ..., N, \tag{3.3}$$

$$\sum_{i=1}^{N} P_{ij} = q_N(y_j), \ j = 1, ..., d+1 \quad \text{and} \tag{3.4}$$

$$P_{ij} \geq 0, \ i = 1, ..., N, \ j = 1, ..., d+1, \tag{3.5}$$

with $C_{ij} =< X_i, y_j >$. In particular, this has the structure of a transportation problem, for which efficient algorithms are available, for example a strongly polynomial algorithm; see Kleinschmidt and Schannath [27]. Bilateral CVA, involving the joint distribution of market exposure and the default times of both parties, admits a similar formulation.

3.2 Dual Variables

To formulate the dual problem, let a_i and b_j be dual variables associated with constraints (3.3) and (3.4), respectively. The dual problem is then

$$\min_{a \in \mathbb{R}^N, b \in \mathbb{R}^{d+1}} \quad \sum_{i=1}^{N} a_i/N + \sum_{j=1}^{d+1} b_j q_N(y_j)$$

$$\text{subject to} \quad a_i + b_j \geq C_{ij}, i = 1, ..., N, \ j = 1, ..., d.$$

The dual variables are useful because they measure the sensitivity of the estimated worst-case CVA to the marginal constraints. Consider any vector of perturbations $(\Delta q_1, \ldots, \Delta q_{d+1})$ to the mass function q_N with components that sum to zero. Suppose these perturbations are sufficiently small to leave the dual solution unchanged. Then

$$\Delta \widehat{\text{CVA}}_* = \sum_{j=1}^{d+1} b_j \Delta q_j.$$

In particular, we can calculate the sensitivity of the worst-case CVA to a parallel shift in the credit curve by setting $\Delta q_j = \Delta$, $j = 1, \ldots, d$, and $\Delta q_{d+1} = -d\Delta$, for sufficiently small Δ.

3.3 Convergence as $N \to \infty$

The solution to the linear program provides an estimate $\widehat{\text{CVA}}_*$ based on N simulated paths. But we are ultimately interested in CVA_* in (2.2), the worst-case CVA based on the true marginal laws for market and credit risk, rather than their sample counterparts. We show that our estimate converges to CVA_* almost surely as N increases.

Although in our application Y has finite support, we state the following result more generally. For probability laws p and q on $\mathbb{R}^d$, let p_N and q_N denote the corresponding empirical laws in (3.1). Let $\Pi(p,q)$, $\Pi(p_N,q_N)$, and $\Pi(p_N,q)$ denote the sets of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with the indicated arguments as marginals.

Theorem 3.1. *Let X and Y be d-dimensional random vectors with distributions p and q respectively such that $\int_{\mathbb{R}^d} \|x\|^2 dp(x) < \infty$, and $\int_{\mathbb{R}^d} \|y\|^2 dq(y) < \infty$. Then*

$$\begin{aligned}\lim_{N\to\infty} \sup_{\mu\in\Pi(p_N,q_N)} \int_{\mathbb{R}^d\times\mathbb{R}^d} < x,y > \mu(dx,dy) &= \lim_{N\to\infty} \sup_{\mu\in\Pi(p_N,q)} \int_{\mathbb{R}^d\times\mathbb{R}^d} < x,y > \mu(dx,dy) \\ &= \sup_{\mu\in\Pi(p,q)} \int_{\mathbb{R}^d\times\mathbb{R}^d} < x,y > \mu(dx,dy).\end{aligned}$$

The proof follows from results on optimal transport in Villani [36]; see Appendix A.

4 Robust Formulation with a Relative Entropy Constraint

The linear program (3.2)–(3.5) provides a simple way to bound the impact of wrong-way risk and estimate a worst-case CVA, and Theorem 3.1 establishes the consistency of this estimate as the number of paths grows. An attractive feature of this approach is that it reuses simulated exposure paths that need to be generated anyway to estimate CVA even ignoring wrong-way risk.

A drawback of the bound CVA_* is that it may be too pessimistic: the worst-case joint distribution may be implausible, even if it is theoretically feasible. To address this concern, we extend our analysis and formulate the problem of bounding wrong-way risk as a question of robustness to model uncertainty. By controlling the degree of uncertainty we can temper the bound on wrong-way risk.

4.1 Constrained and Penalized Problems

In this formulation, we start with a reference model for the dependence between the market and credit models and control model uncertainty by constraining deviations from the reference model. To be concrete, we will assume that the reference model takes market and credit risk to be independent, though this is not essential. We use ν to denote the corresponding element of $\Pi(p,q)$ that

makes X and Y independent; in other words,

$$\nu(A \times B) = p(A)q(B),$$

for all measurable $A, B \subseteq \mathbb{R}^d$.

To constrain deviations from the reference model, we need a notion of "distance" between probability measures. Among the many candidates, relative entropy, also known as the Kullback-Leibler divergence, is particularly convenient. For probability measures P and Q on a common measurable space and with $P \gg Q$, define the entropy of Q relative to P to be

$$D(Q|P) = \mathbb{E}_P\left[\frac{dQ}{dP}\ln\left(\frac{dQ}{dP}\right)\right] = \mathbb{E}_Q\left[\ln\left(\frac{dQ}{dP}\right)\right],$$

the subscripts indicating the measure with respect to which the expectation is taken. Relative entropy is frequently used to quantify model uncertainty; see, for example, Hansen and Sargent [23] and Ben-Tal et al. [3]. Relative entropy is not symmetric in its arguments, but this is not necessarily a drawback because we think of the reference model as a favored benchmark. We are interested in the potential impact of deviations from the reference model, but we do not necessarily view nearby alternative models as equally plausible. Relative entropy $D(Q|P)$ is convex in Q, and this will be important for our application. Also, $D(Q|P) = 0$ only if $Q = P$.

To find a tempered worst case for wrong-way risk, we maximize CVA with the marginal models p and q held fixed and with a constraint $\eta > 0$ on the relative entropy divergence from the reference joint model ν:

$$\begin{aligned} \text{CVA}_\eta \quad := \quad & \sup_{\mu\in\Pi(p,q)} \int_{\mathbb{R}^d\times\mathbb{R}^d} < x, y > d\mu(x,y), && (4.1) \\ & \text{subject to} \int \ln(\frac{d\mu}{d\nu})d\mu \le \eta. && (4.2) \end{aligned}$$

At $\eta = 0$, the only feasible solution is the reference model $\mu = \nu$. At $\eta = \infty$, the problem reduces to the worst-case CVA of the previous section. Varying the relative entropy budget η thus controls the degree of model uncertainty or the degree of confidence in the reference model.

We are actually interested in solving this problem for a range of η values to see how the potential impact of wrong-way risk varies with the degree of model uncertainty. For this purpose, it will be convenient to work with a penalty on relative entropy rather than a constraint. The penalty formulation with parameter $\theta > 0$ is as follows:

$$\sup_{\mu\in\Pi(p,q)} \int_{\mathbb{R}^d\times\mathbb{R}^d} < x, y > d\mu(x,y) - \frac{1}{\theta}\int \ln(\frac{d\mu}{d\nu})d\mu. \qquad (4.3)$$

The penalty term subtracted from the linear objective is nonnegative because relative entropy is nonnegative. At $\theta = 0$, the penalty would be infinite unless $\mu = \nu$; at $\theta = \infty$, the penalty drops out and we recover the worst-case linear program of Section 3. A related problem appears in Bosc and Galichon [7], but without a reference model ν. The correspondence between the constrained problem (4.1)–(4.2) and the penalized problem (4.3) is established in the following result, proved in the Appendix B:

Proposition 4.1. *For $\theta > 0$, the optimal solution μ^θ to (4.3) is the optimal solution to (4.1)–(4.2) with*

$$\eta(\theta) = \int \ln(\frac{d\mu^\theta}{d\nu}) d\mu^\theta. \tag{4.4}$$

The mapping from θ to $\eta(\theta)$ is increasing, and $\eta(\theta) \in (0, \eta^*]$ for $\theta \in (0, \infty)$, where η^* is (4.4) evaluated at the solution to (2.2).

In the following, we write CVA_θ instead of $\text{CVA}_{\eta(\theta)}$ for $\theta \in (0, \infty)$. To estimate CVA_θ, we form a sample counterpart, modifying the linear programming formulation (3.2)–(3.5). We denote the finite sample reference joint probabilities by F_{ij}. In the independent case, these are given by $F_{ij} = q_N(y_j)/N$, $i = 1, \ldots, N$, $j = 1, \ldots, d+1$. Let P^θ denote the optimal solution to the following optimization problem:

$$\max_{\{P_{ij}\}} \quad \sum_{i=1}^{N}\sum_{j=1}^{d+1} C_{ij}P_{ij} - \frac{1}{\theta}\sum_{i=1}^{N}\sum_{j=1}^{d+1} P_{ij}\ln\left(\frac{P_{ij}}{F_{ij}}\right) \quad \text{subject to (3.3)-(3.5).} \tag{4.5}$$

We estimate CVA_θ by

$$\widehat{\text{CVA}}_\theta := \sum_{i=1}^{N}\sum_{j=1}^{d+1} C_{ij}P^\theta_{ij}.$$

4.2 Iterative Proportional Fitting Procedure

The penalty problem (4.5) is a convex optimization problem and can be solved using general optimization methods. However, the choice of relative entropy for the penalty leads to a particularly simple and interesting method through the iterative proportional fitting procedure (IPFP). The method dates to Deming and Stephan [17], yet it continues to generate extensions and applications in many areas.

To apply the method in our setting, we use as initial guess the $N \times (d+1)$ matrix M^θ with entries

$$M^\theta_{ij} = \frac{e^{\theta \cdot C_{ij}} \cdot F_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{d+1} e^{\theta \cdot C_{ij}} \cdot F_{ij}}.$$

As before, F_{ij} is the independent joint distribution with prescribed marginals p_N and q_N, which we take as reference model. Each $C_{ij} =< X_i, y_j >$ is the loss on market risk path i if the counterparty defaults at time t_j. With $\theta > 0$, the numerator of M_{ij}^{θ} puts more weight on combinations that produce larger losses. In this sense, M_{ij}^{θ} is designed to emphasize wrong-way risk.

The denominator of M_{ij}^{θ} normalizes the entries to sum to 1, but M^{θ} will not in general have the target marginals. The IPFP algorithm projects a matrix M with positive entries onto the set of joint distribution matrices with marginals p_N and q_N by iteratively renormalizing the rows and columns as follows:

(r) For $i = 1, \ldots, N$ and $j = 1, \ldots, d+1$, set $M_{ij} \leftarrow M_{ij} p_N(i) / \sum_{k=1}^{d+1} M_{ik}$.

(c) For $j = 1, \ldots, d+1$ and $i = 1, \ldots, N$, set $M_{ij} \leftarrow M_{ij} q_N(j) / \sum_{n=1}^{N} M_{nj}$.

This iteration is also known as biproportional scaling, Sinkhorn's algorithm, and the RAS algorithm; see Pukelsheim [29] for an overview of the extensive literature on the theory and application of these methods.

Write $\Phi(M)$ for the result of applying both steps (r) and (c) to M, and write $\Phi^{(n)}$ for the n-fold composition of Φ. For our setting, we need the following result:

Proposition 4.2. *The sequence $\Phi^{(n)}(M^{\theta})$, $n \geq 1$, converges to the solution P^{θ} to (4.5).*

Proof. It follows from Ireland and Kullback [26] that $\Phi^{(n)}(M^{\theta})$ converges to the solution of

$$\min_{P} \quad \sum_{i=1}^{N} \sum_{j=1}^{d+1} P_{ij} \ln \left(\frac{P_{ij}}{M_{ij}^{\theta}} \right) \quad \text{subject to (3.3)-(3.5).}$$

In other words, the IPFP algorithm converges to the feasible matrix (in the sense of (3.3)-(3.5)) that is closest to the initial matrix in the sense of relative entropy. For our particular choice of M^{θ}, this minimization problem has the same solution as the maximization problem

$$\max_{P} \quad \theta \sum_{i=1}^{N} \sum_{j=1}^{d+1} C_{ij} P_{ij} - \sum_{i=1}^{N} \sum_{j=1}^{d+1} P_{ij} \ln(\frac{P_{ij}}{F_{ij}}) - W_{\theta}^{N} \quad \text{subject to (3.3)-(3.5),}$$

with $W_{\theta}^{N} = \ln \left(\sum_{i=1}^{N} \sum_{j=1}^{d+1} e^{\theta \cdot C_{ij}} \cdot F_{ij} \right)$. This follows directly from the definition of M^{θ}. Because W_{θ}^{N} does not depend on P, this maximization problem has the same solution as (4.5). □

To summarize, we start with the reference model F_{ij}, put more weight on adverse outcomes to get M_{ij}^{θ}, and then iteratively renormalize the rows and columns of M^{θ} to match the target marginals. This procedure converges to the penalized worst-case joint distribution defined by (4.5) with penalty parameter θ.

4.3 Convergence as $N \to \infty$

We now formulate a convergence result as the number of paths N increases. As before, let $\Pi(p, q)$ denote the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals p and q. Let p_N, q_N denote the empirical measures in (3.1), and let $\Pi(p_N, q_N)$ denote the set of joint laws with these marginals. The independent joint distributions are $\nu \in \Pi(p, q)$ and $\nu_N \in \Pi(p_N, q_N)$; i.e., $d\nu(x, y) = dp(x)dq(y)$ and $d\nu_N(x, y) = dp_N(x)dq_N(y)$.

Fix $\theta > 0$ and define, for a probability measure μ on $\mathbb{R}^d \times \mathbb{R}^d$,

$$G(\mu, \nu) = \int < x, y > \, d\mu - \frac{1}{\theta} D(\mu|\nu),$$

and define $G(\mu, \nu_N)$ accordingly. To show that our simulation estimate of the penalized worst-case CVA converges to the true value, we need to show that

$$\int < x, y > \, d\mu_N^* \to \int < x, y > \, d\mu^*, \quad a.s. \tag{4.6}$$

where $\mu_N^* \in \Pi(p_N, q_N)$ maximizes $G(\cdot, \nu_N)$ and $\mu^* \in \Pi(p, q)$ maximizes $G(\cdot, \nu)$.

Theorem 4.1. *Suppose the random vectors X and Y satisfy $\mathbb{E}_\nu[e^{\theta < X,Y >}] < \infty$ and that Y has finite support. The following hold as $N \to \infty$.*

(i) $\max_{\mu \in \Pi(p_N, q_N)} G(\mu, \nu_N) \longrightarrow \sup_{\mu \in \Pi(p,q)} G(\mu, \nu)$, *a.s.*

(ii) The maximizer $\mu_N^* \in \Pi(p_N, q_N)$ of $G(\cdot, \nu_N)$ converges weakly to a maximizer $\mu^* \in \Pi(p, q)$ of $G(\cdot, \nu)$.

(iii) The penalized worst-case CVA converges to the true value, a.s.; i.e., (4.6) holds.

The proof is in Appendix C.

5 Examples

5.1 A Gaussian Example

For purposes of illustration we begin with a simple example in which X and Y are scalars and normally distributed. This example is not intended to fit the CVA application but to illustrate some features of the penalty formulation. It also lends itself to a simple comparison with a Gaussian copula, which is another way of introducing dependence with given marginals.

Suppose then that X and Y have the standard normal distribution on $\mathbb{R}$. Paralleling the definition of the matrix M^θ, consider the bivariate density

$$f_0(x, y) = c' e^{\theta xy} p(x) q(y) = c e^{-\frac{1}{2}x^2 - \frac{1}{2}y^2 + \theta xy}, \tag{5.1}$$

where c' and c are normalization constants. This density weights the independent joint density at (x, y) by $\exp(\theta xy)$, so the product xy plays the role that C_{ij} plays in the definition of M^θ.

The reweighting changes the marginals, so now we want to use a continuous version of the IPFP algorithm to project f_0 onto the set of bivariate densities with standard normal marginals. The generalization of the algorithm from matrices to measures has been analyzed in Rüschendorf [31]. The row and column operations become

$$\hat{f}_n(x, y) \leftarrow f_n(x, y)p(x) \Big/ \int f_n(x, y)\, dy$$

and

$$f_{n+1}(x, y) \leftarrow \hat{f}_n(x, y)q(y) \Big/ \int \hat{f}_n(x, y)\, dx\ .$$

An induction argument shows that

$$f_n(x, y) = c_n e^{-\frac{a_n^2}{2}x^2 - \frac{a_n^2}{2}y^2 + \theta xy},$$

for constants c_n and a_n, so each f_n is a bivariate normal density. The a_n satisfy

$$a_n^2 = \left(1 + \frac{\theta^2}{a_{n-1}^2}\right) \to \frac{1}{2} + \frac{1}{2}\sqrt{1 + 4\theta^2}, \quad \text{as } n \to \infty.$$

Some further algebraic simplification then shows that the limit is a bivariate normal density with standard normal marginals and correlation parameter

$$\rho = \frac{2\theta}{1 + \sqrt{1 + 4\theta^2}}, \quad \theta = \frac{\rho}{1 - \rho^2}. \tag{5.2}$$

This is the bivariate distribution with standard normal marginals that maximizes the expectation of XY with a penalty parameter of θ on the deviation from independence as measured by relative entropy.

Observe that $\rho = 0$ when $\theta = 0$; $\rho \to 1$ as $\theta \to \infty$; and $\rho \to -1$ as $\theta \to -\infty$. Because θ penalizes deviations from independence, it controls the strength of the dependence between X and Y. The relationship between ρ and θ allows us to reinterpret the strength of dependence as measured by θ in terms of the correlation parameter ρ. This is somewhat analogous to the role of a correlation parameter in the Gaussian copula, where it measures the strength of dependence but is not literally the correlation between the marginals except when the marginals are normal.

The fact that the IPFP algorithm projects f_0 to a bivariate normal is a specific feature of the weight $\exp(\theta xy)$ in (5.1). For contrast, we consider the weight $\exp(\theta x^2 y)$. The resulting f_0 is no longer integrable for $\theta > 0$, so we work instead with truncated and discretized marginal distributions and apply the IPFP numerically. The result is shown in Figure 1. The resulting

density has nearly standard normal marginals (up to truncation and discretization), but the joint distribution is clearly not bivariate normal.

The dependence illustrated in the figure is beyond the scope of the Gaussian copula because any joint distribution with Gaussian marginals and a Gaussian copula must be Gaussian. This example thus illustrates the broader point that our approach generates a wider range of dependence than can be achieved with a specific type of copula. For examples of wrong-way risk CVA models based on the Gaussian copula, see Brigo et al. [10], Hull and White [25], and Rosen and Saunders [30].

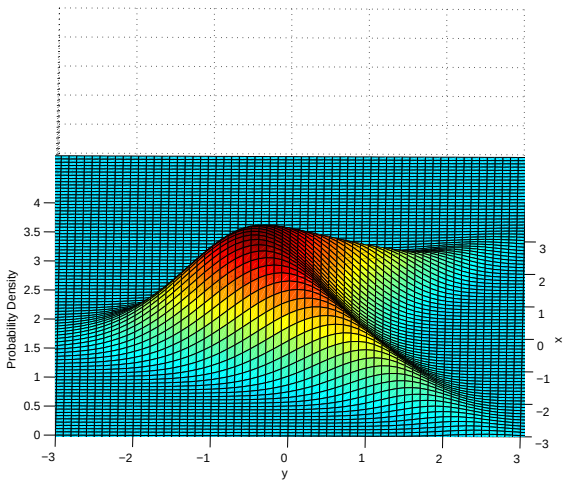

Figure 1: Probability mass of joint truncated and discretized normal random variables X and Y, with $\theta = 1$ and intial weight $\exp(\theta x^2 y)$.

5.2 A Currency Swap Example

In a currency swap between a U.S. bank receiving U.S. dollars and a foreign bank receiving its own currency, the U.S. bank faces wrong-way risk: when the foreign currency depreciates, the exposure of the U.S. bank increases, and the foreign bank's credit quality usually deteriorates as its currency depreciates.[1] Similarly, when a firm borrows money from a bank and posts collateral which is positively correlated with the firm's credit quality, the bank lending the money faces wrong-way risk.

We illustrate our method with a foreign exchange forward, the simplest currency swap that exchanges only the principal, evaluating the CVA from the U.S. dollar receiver's perspective. Let U_t be the number of units of the foreign currency paid in exchange for one U.S. dollar at time t.

[1] Banks writing credit protection on their sovereigns create similar wrong-way risk. Specific cases of this practice are documented in "FVA, correlation, wrong-way risk: EU stress test's hidden gems," *Risk magazine*, Dec 5, 2014.

This exchange rate follows an Ornstein-Uhlenbeck process,

$$U_{t_{j+1}} = U_{t_j} + \kappa(\bar{U} - U_{t_j})(t_{j+1} - t_j) + \sigma(W_{t_{j+1}} - W_{t_j}),$$

where $\bar{U}$ is the long term mean of the exchange rate and W_t is a standard Brownian motion.

Let δ be the U.S. dollar discount rate, N the contract notional (in U.S. dollars) and K the contract forward exchange rate. Let R be the recovery rate at failure of the counterparty. The expected exposure (in U.S. dollars) of this foreign exchange transaction at time t is

$$\mathrm{EE}(t) = \mathbb{E}[e^{-\delta(T-t)} N(U_T - K)/U_T \,|\, U_t].$$

The expected exposure discounted to today and adjusted by the recovery rate is

$$V(t) = e^{-\delta t} R \cdot \mathrm{EE}(t).$$

We take $T = 10$ years, divide time into 20 time steps, and simulate 1000 market scenarios. Expected exposures are adjusted for recoveries and discounted. The mean positive expected exposure is shown in Figure 2. For illustrative purpose, we assume that the counterparty's default time follows an exponential distribution with a constant hazard rate λ. We use the following parameters: $(U_{t_0}, \bar{U}, K, \kappa, \sigma, \lambda, \delta, N) = (1000, 1000, 1000, 0.3, 50, 0.04, 0.03, 10^6)$.

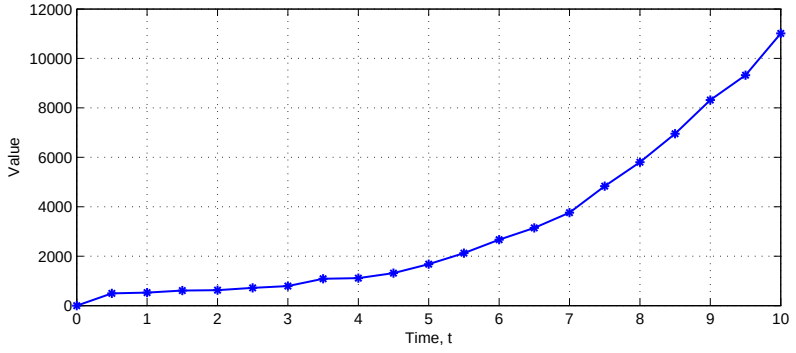

Figure 2: Sample Average Positive Exposure

Figure 3 shows a CVA stress test for wrong-way risk. It plots CVA against the penalty parameter θ. The numbers are normalized by dividing by the independent market-credit risk CVA, so the independent case $\theta = 0$ is presented as 100%. As θ increases, the positive dependence between market and credit risk increases, approaching the worst-case bound, which is over six times as large as the independent CVA. For $\theta < 0$, we have right-way risk, and the CVA bound approaches zero as θ decreases. The parameter θ could be rescaled using the transformation in (5.2) to allow a rough interpretation as a correlation parameter.

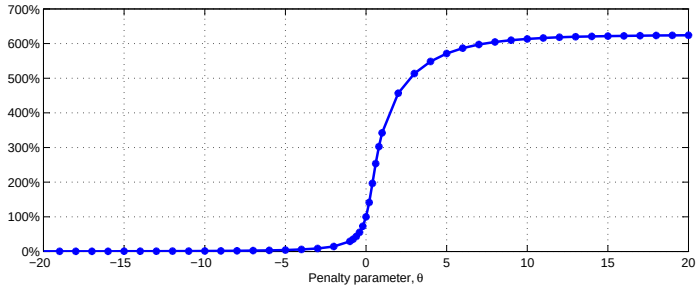

Figure 3: CVA Stress Test

The Gaussian copula provides a simple alternative way to vary dependence and measure wrong-way risk; see Rosen and Saunders [30] for details and applications. Figure 4 shows how wrong-way risk varies in the Gaussian copula model as the correlation parameter ρ varies from -1 to 1. Comparison with Figure 3 shows that constraining dependence to conform to a Gaussian copula significantly underestimates the potential wrong-way risk.

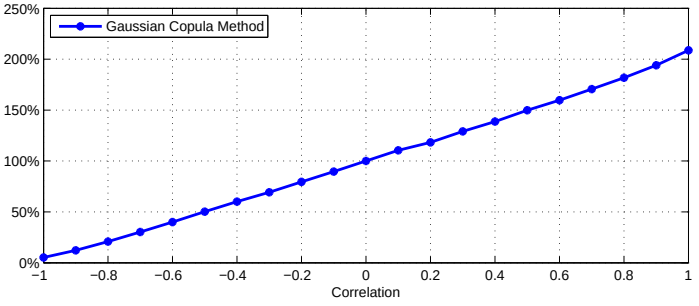

Figure 4: CVA Stress Test by Gaussian Copula Method

In Figure 5, we show the impact of varying the foreign exchange volatility σ, and the counterparty default hazard rate. Increasing either of these parameters shifts the curve up for $\theta > 0$. In other words, increasing the volatility of the market exposure or the level of the credit exposure in this example increases the potential impact of wrong-way risk, relative to the benchmark of independent market and credit risk.

6 Adding Expectation Constraints

When additional information is available, we can often improve our CVA bound by incorporating the information through constraints on the optimization problem. Constraints on expectations are linear constraints on joint distributions and thus particularly convenient in our framework.

Recall that we think of the exposure path X as the output of a simulation of a market model.

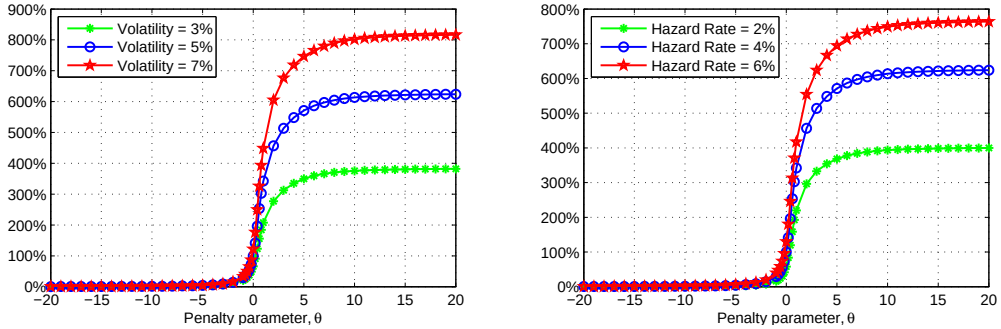

Figure 5: CVA with different volatility and hazard rate

Such a model generates many other market variables, and in specifying the joint distribution between the market and credit models, we may want to add constraints through other variables. Constraints represent relationships between market and credit risk that should be preserved as the joint distribution varies. To incorporate such constraints, we expand the simulation output from X to (X, Z), where the random vector $Z = (Z_1, \ldots, Z_d)$ represents a path of auxiliary variables. The joint law of (X, Z) is determined by the market model. We want to add a constraint of the form $E[Z_\tau \mathbf{1}\{\tau \le t_d\}] = z_0$, for given z_0, when the expectation is taken with respect to the joint law of the market and credit models. This is a constraint on the expectation of $< Z, Y >$.

As a specific illustration, suppose $\tilde{Z}$ is a martingale generated by the market model and we want to impose the constraint $\mathbb{E}[\tilde{Z}_{\tau \wedge t_d}] = z_0$ on the joint law of $\tilde{Z}$ and τ. This is equivalent to the constraint $\mathbb{E}[(\tilde{Z}_{t_d} - \tilde{Z}_\tau)\mathbf{1}\{\tau \le t_d\}] = 0$, so we can define $Z_j = \tilde{Z}_d - \tilde{Z}_j$, $j = 1, \ldots, d$, and then impose the constraint $\mathbb{E}[< Z, Y >] = 0$.

To incorporate constraints, we redefine p to denote the joint law of (X, Z) on $\mathbb{R}^d \times \mathbb{R}^d$; we continue to use q for the marginal law of Y. Let $\Pi(p, q)$ be the set of probability measures on $(\mathbb{R}^d \times \mathbb{R}^d) \times \mathbb{R}^d$ with the specified marginals of (X, Z) and Y. We denote by $h_X(x, z) = x$ and $h_Z(x, z) = z$ the projections of (x, z) to x and z respectively. Set

$$\bar{\Pi}(p, q) = \{\mu \in \Pi(p, q) : \int < h_Z(x, z), y > d\mu((x, z), y) = v_0\}. \tag{6.1}$$

We will assume that $\bar{\Pi}(p, q)$ is nonempty so that the problem is feasible.

Given independent samples (X_i, Z_i), $i = 1, \ldots, N$, let p_N denote their empirical measure. As before q_N denotes the empirical measure for N independent copies of Y. Even if $\bar{\Pi}(p, q)$ is nonempty, we cannot assume that the equality constraint in (6.1) holds for some element of $\Pi(p_N, q_N)$, so for finite N we will need a relaxed formulation. Let $\Pi_\epsilon(p_N, q_N)$ denote the set of joint distributions on $\{((X_i, Z_i), y_j), i = 1, \ldots, N, j = 1, \ldots, d + 1\}$ with marginals p_N and $\tilde{q}$, where

$$\max_{1 \le j \le d+1} |q_N(y_j) - \tilde{q}(y_j)| < \epsilon,$$

and define

$$\bar{\Pi}_\epsilon(p_N, q_N) = \left\{ \mu \in \Pi_\epsilon(p_N, q_N) : \left| \int < h_Z(x,z), y > d\mu((x,z),y) - v_0 \right| < \epsilon \right\}. \tag{6.2}$$

In our convergence analysis, we will let $\epsilon \equiv \epsilon_N$ decrease to zero as N increases.

Let $\nu \in \Pi(p,q)$ denote the independent case $d\nu((x,z),y) = dp(x,z)dq(y)$, and let $\nu_N \in \Pi(p_N, q_N)$ denote the independent case $d\nu_N((x,z),y) = dp_N(x,z)dq_N(y)$. We will assume that v_0 is chosen so that $\nu \in \bar{\Pi}(p,q)$. It then follows that $\nu_N \in \bar{\Pi}_\epsilon(p_N, q_N)$ for all sufficiently large N, for all $\epsilon > 0$.

The worst-case CVA with an auxiliary constraint on Z is

$$c_\infty \quad = \quad \sup_{\mu \in \bar{\Pi}(p,q)} \int_{(\mathbb{R}^d \times \mathbb{R}^d) \times \mathbb{R}^d} < h_X(x,z), y > d\mu((x,z),y) \tag{6.3}$$

The corresponding estimator is

$$c_{N,\epsilon} \quad = \quad \max_{\mu \in \bar{\Pi}_\epsilon(p_N, q_N)} \sum_{i=1}^{N} \sum_{j=1}^{d+1} < X_i, y_j > \mu((X_i, Z_i), y_j). \tag{6.4}$$

This is a linear programming problem: the objective and the constraints are linear in the variables $\mu((X_i, Z_i), y_j)$. The following result establishes convergence of the estimator.

Theorem 6.1. *Suppose the following conditions hold:*

(i) $\int_{\mathbb{R}^d \times \mathbb{R}^d} \|h_X(x,z)\|^2 dp(x,z) < \infty$ *and* $\int_{\mathbb{R}^d \times \mathbb{R}^d} \|h_Z(x,z)\|^2 dp(x,z) < \infty$*;*

(ii) $\bar{\Pi}(p,q)$ *contains the independent joint distribution* ν*.*

Then with $\epsilon_N = 1/N^\alpha$ *for any* $\alpha \in (0, 1/2)$*, the finite sample estimate converges to the constrained worst-case CVA for the limiting problem; i.e.,* $c_{N,\epsilon_N} \to c_\infty$*, a.s.*

We define a penalty formulation with $\theta > 0$ for the limiting problem,

$$\sup_{\mu \in \bar{\Pi}(p,q)} G(\mu, \nu) \quad = \quad \sup_{\mu \in \bar{\Pi}(p,q)} \int_{(\mathbb{R}^d \times \mathbb{R}^d) \times \mathbb{R}^d} < h_X(x,z), y > d\mu((x,z),y) - \frac{1}{\theta} D(\mu|\nu),$$

and with (6.2) for the finite problem,

$$\max_{\mu \in \bar{\Pi}_\epsilon(p_N, q_N)} G(\mu, \nu) \quad = \quad \max_{\mu \in \bar{\Pi}_\epsilon(p_N, q_N)} \sum_{i=1}^{N} \sum_{j=1}^{d+1} < h_X(X_i, Z_i), y_j > \mu_N((X_i, Z_i), y_j) - \frac{1}{\theta} D(\mu_N|\nu_N).$$

The corresponding convergence result given by the following theorem.

Theorem 6.2. *Suppose the following conditions hold:*

(i) $\int_{\mathbb{R}^d\times\mathbb{R}^d} \|h_X(x,z)\|^2 dp(x,z) < \infty$, $\int_{\mathbb{R}^d\times\mathbb{R}^d} \|h_Z(x,z)\|^2 dp(x,z) < \infty$, *and* $\mathbb{E}_\nu[e^{\theta<h_X(X,Z),Y>}] < \infty$;

(ii) $\bar{\Pi}(p,q)$ *contains the independent joint distribution* ν.

Then with $\epsilon_N = 1/N^\alpha$ *for any* $\alpha \in (0,1/2)$, *the following hold,*

(i) $\max_{\mu\in\bar{\Pi}_{\epsilon_N}(p_N,q_N)} G(\mu,\nu_N) \longrightarrow \sup_{\mu\in\bar{\Pi}(p,q)} G(\mu,\nu)$, *a.s.*

(ii) The maximizer $\bar{\mu}_N^* \in \bar{\Pi}_{\epsilon_N}(p_N,q_N)$ *of* $G(\cdot,\nu_N)$ *converges weakly to a maximizer* $\bar{\mu}^* \in \bar{\Pi}(p,q)$ *of* $G(\cdot,\nu)$.

(iii) The penalized worst-case CVA converges to the true value, a.s.; i.e.,

$$\int < x,y > d\bar{\mu}_N^* \to \int < x,y > d\bar{\mu}^*, \quad a.s. \tag{6.5}$$

7 Concluding Remarks

We have focused in this article on the problem of bounding wrong-way risk in CVA calculation, taking the marginal models for market and credit risk as given and varying the dependence between the two. Put more generally, the problem we have addressed is one of bounding the expected inner product between two random vectors with fixed marginals. A key feature of our setting is that these marginals need not be known explicitly. Instead, they are outputs of the simulation of potentially very complex models, of the type used to model asset prices and default times.

Calculating the worst-case bound for the exact marginal distributions is typically infeasible. But using simulated outcomes, the problem reduces to a tractable linear programming problem. We extend this formulation by penalizing deviations from a reference model, which results in a convex optimization problem. In both cases, we prove convergence of the solutions calculated from simulated outcomes to the corresponding solutions using exact distributions as the sample size grows. The approach is sufficiently general and flexible to be applicable to many other settings in which the nature of dependence between different model components is unknown.

A Proof of Theorem 3.1

The Wasserstein metric of order 2 between probability measures p and q on $\mathbb{R}^d$ is $W_2(p,q)$, where

$$\begin{aligned} W_2^2(p,q) &= \inf_{\pi\in\Pi(p,q)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 \, d\pi(x,y) \\ &= \int_{\mathbb{R}^d} \|x\|^2 \, dp(x) + \int_{\mathbb{R}^d} \|y\|^2 \, dq(y) - 2 \sup_{\pi\in\Pi(p,q)} \int_{\mathbb{R}^d\times\mathbb{R}^d} < x,y > d\pi(x,y). \end{aligned} \tag{A.1}$$

The empirical measures p_N and q_N converge weakly to p and q, respectively, a.s., so it follows from Corollary 6.11 of Villani [36] that $W_2^2(p_N, q_N) \to W_2^2(p, q)$, a.s., and $W_2^2(p_N, q) \to W_2^2(p, q)$, a.s. Under the assumed square-integrability conditions, we also have

$$\int_{\mathbb{R}^d} \|x\|^2 \, dp_N(x) \to \int_{\mathbb{R}^d} \|x\|^2 \, dp(x), \quad \text{a.s.},$$

and similarly for q_N. The theorem now follows from (A.1).

B Proof of Proposition 4.1

Problem (4.3) is equivalent to

$$-\inf_{\mu \in \Pi(p,q)} \frac{1}{\theta} \int \ln\Big(\frac{d\mu}{\exp(\theta < x, y >) d\nu}\Big) d\mu. \tag{B.1}$$

Theorem 3 of Rüschendorf and Thomsen [34] implies the existence of a unique optimal solution to (B.1), which we denote by μ^θ.

First we show that μ^θ is optimal for (4.1)–(4.2) with $\eta = \eta(\theta)$. Suppose μ^θ is not optimal, then there exists $\mu^{\eta(\theta)}$ such that

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} < x, y > d\mu^{\eta(\theta)}(x, y) > \int_{\mathbb{R}^d \times \mathbb{R}^d} < x, y > d\mu^{\theta}(x, y),$$

and

$$\int \ln(\frac{d\mu^{\eta(\theta)}}{d\nu}) d\mu^{\eta(\theta)} \leq \int \ln(\frac{d\mu^{\theta}}{d\nu}) d\mu^{\theta}.$$

But then

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} < x, y > d\mu^{\eta(\theta)}(x, y) - \frac{1}{\theta} \int \ln(\frac{d\mu^{\eta(\theta)}}{d\nu}) d\mu^{\eta(\theta)} > \int_{\mathbb{R}^d \times \mathbb{R}^d} < x, y > d\mu^{\theta}(x, y) - \frac{1}{\theta} \int \ln(\frac{d\mu^{\theta}}{d\nu}) d\mu^{\theta},$$

which contradicts the optimality of μ^θ for the penalty problem (4.3).

Next we show that the mapping from θ to $\eta(\theta)$ is increasing. For any $\theta_2 > \theta_1 > 0$, let μ^{θ_1} and μ^{θ_2} denote optimal solution to the penalty problem with θ_1 and θ_2 respectively. If $\mu^{\theta_1} = \mu^{\theta_2}$, then $\eta(\theta_1) = \eta(\theta_2)$. If $\mu^{\theta_1} \neq \mu^{\theta_2}$, then, by unique optimality of μ^{θ_2}, it holds that

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} < x, y > d\mu^{\theta_2}(x, y) - \frac{1}{\theta_2} \int \ln(\frac{d\mu^{\theta_2}}{d\nu}) d\mu^{\theta_2} \; > \; \int_{\mathbb{R}^d \times \mathbb{R}^d} < x, y > d\mu^{\theta_1}(x, y) - \frac{1}{\theta_2} \int \ln(\frac{d\mu^{\theta_1}}{d\nu}) d\mu^{\theta_1}. \tag{B.2}$$

Compare the first term on each side. If $\int_{\mathbb{R}^d \times \mathbb{R}^d} < x, y > d\mu^{\theta_2}(x, y) \leq \int_{\mathbb{R}^d \times \mathbb{R}^d} < x, y > d\mu^{\theta_1}(x, y)$, then $\int \ln(\frac{d\mu^{\theta_2}}{d\nu}) d\mu^{\theta_2} < \int \ln(\frac{d\mu^{\theta_1}}{d\nu}) d\mu^{\theta_1}$ by (B.2). Adding $(\frac{1}{\theta_2} - \frac{1}{\theta_1}) \int \ln(\frac{d\mu^{\theta_2}}{d\nu}) d\mu^{\theta_2}$ to the left side and $(\frac{1}{\theta_2} - \frac{1}{\theta_1}) \int \ln(\frac{d\mu^{\theta_1}}{d\nu}) d\mu^{\theta_1}$ to the right side of (B.2), the sign does not change, which means μ^{θ_2} is

optimal for the penalty problem with θ_1. However that contradicts the unique optimality of μ^{θ_1}. We conclude that

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} < x, y > d\mu^{\theta_2}(x,y) \;>\; \int_{\mathbb{R}^d\times\mathbb{R}^d} < x, y > d\mu^{\theta_1}(x,y).$$

Now compare the second term on each side. If $\int \ln(\frac{d\mu^{\theta_2}}{d\nu})d\mu^{\theta_2} \leq \int \ln(\frac{d\mu^{\theta_1}}{d\nu})d\mu^{\theta_1}$, then the unique optimality of μ^{θ_1} is again contradicted, so we have

$$\eta(\theta_2) \;>\; \eta(\theta_1).$$

Next we show $\eta(\theta) \in (0, \eta^*]$ for $\theta \in (0,\infty)$. Since the relative entropy $\int \ln(\frac{d\mu^{\theta}}{d\nu})d\mu^{\theta}$ is nonnegative and equals 0 only if $\mu^{\theta} = \nu$, we have $\eta(\theta) > 0$ for $\theta > 0$. Let μ^* denote optimal solution to (2.2) and let $\eta^* = \int \ln(\frac{d\mu^*}{d\nu})d\mu^*$. Since problem (2.2) is a relaxation of problem (4.1)–(4.2), we conclude that for all $\theta > 0$,

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} < x, y > d\mu^*(x,y) \geq \int_{\mathbb{R}^d\times\mathbb{R}^d} < x, y > d\mu^{\theta}(x,y). \tag{B.3}$$

Suppose there exists $\theta^* > 0$ such that $\eta(\theta^*) = \int \ln(\frac{d\mu^{\theta^*}}{d\nu})d\mu^{\theta^*} > \eta^*$. By adding $-\frac{1}{\theta^*}\eta^*$ to the left and $-\frac{1}{\theta^*}\eta(\theta^*)$ to the right of (B.3), the inequality does not change, which contradicts the optimality of μ^{θ^*}. Thus $\eta(\theta) \leq \eta^*$.

C Proof of Theorem 4.1

We divide the proof into several parts, starting with the convergence of the objective function value asserted in part (i) of the theorem.

C.1 Convergence of the Optimal Objective Value

We will first show that for any feasible solution to the limiting problem, we can construct a sequence of approximating solutions that approach the limiting objective function from above. To get the reverse inequality we will use a dual formulation of the limiting objective and show that it is approached from below.

Since Y has finite support, we may assume without loss of generality that $q(y_j) > 0$ for all j. If we had $q(y_j) = 0$ for some j, we could reformulate an equivalent problem by removing the marginal constraint on y_j.

Let $\mu \in \Pi(p,q)$ be any feasible solution to the limiting problem. Write $\mu(dx, y) = p(dx)q(y|x)$, and define the following mass function on the pairs (X_i, y_j), $i = 1,\ldots,N$, $j = 1,\ldots,d+1$:

$$\mu_N(X_i, y_j) = \frac{1}{N}q(y_j|X_i). \tag{C.1}$$

If we sum over the y_j for any X_i, we get

$$\sum_{j=1}^{d+1} \mu_N(X_i, y_j) = \frac{1}{N} \sum_{j=1}^{d+1} q(y_j|X_i) = \frac{1}{N}.$$

If we sum over the X_i for any y_j, we get

$$\sum_{i=1}^{N} \mu_N(X_i, y_j) = \frac{1}{N} \sum_{i=1}^{N} q(y_j|X_i) =: \bar{q}_N(y_j).$$

We will not in general have $\bar{q}_N = q_N$, so μ_N is not in general a feasible solution to the finite problem, in the sense that $\mu_N \notin \Pi(p_N, q_N)$. However, by the strong law of large numbers for $\{X_1, X_2, \dots\}$, for each y_j, $j = 1, \dots, d+1$,

$$\bar{q}_N(y_j) = \frac{1}{N} \sum_{i=1}^{N} q(y_j|X_i) \to \int q(y_j|x)\, dp(x) = q(y_j), \text{ a.s.},$$

because $\mu \in \Pi(p, q)$. Also by the strong law of large numbers, we have $q_N(y_j) \to q(y_j)$, a.s. We will therefore consider a relaxed constraint. Let $\Pi_\epsilon(p_N, q_N)$ denote the set of joint distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals p_N and q', where $|q'(y_j) - q_N(y_j)| \le \epsilon$, $j = 1, \dots, d+1$.

Lemma C.1. *As $N \to \infty$,*

$$\underline{\lim}_{N\to\infty} \max_{\mu \in \Pi(p_N, q_N)} G(\mu, \nu_N) \ge \sup_{\mu \in \Pi(p,q)} G(\mu, \nu).$$

Proof: For each N, we are maximizing a concave function over a compact convex set, so the maximum is indeed attained. Write c_N for $\max_{\mu\in\Pi(p_N,q_N)} G(\mu, \nu_N)$ and $c_{N,\varepsilon}$ for $\max_{\mu\in\Pi_\epsilon(p_N,q_N)} G(\mu, \nu_N)$. For any $\mu \in \Pi(p, q)$, define μ_N as in (C.1). Then $\mu_N \in \Pi_\epsilon(p_N, q_N)$ for all sufficiently large N, a.s., and

$$\begin{aligned} c_{N,\varepsilon} &\ge \sum_{i=1}^{N} \sum_{j=1}^{d+1} \mu_N(X_i, y_j) < X_i, y_j > -\frac{1}{\theta} D(\mu_N|\nu_N) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{d+1} \frac{q(y_j|X_i)}{N} < X_i, y_j > -\frac{1}{\theta} \sum_{i=1}^{N} \sum_{j=1}^{d+1} \ln\left(\frac{q(y_j|X_i)/N}{q_N(y_j)/N}\right) \frac{q(y_j|X_i)}{N}. \end{aligned}$$

By the strong law of large numbers, almost surely,

$$\sum_{i=1}^{N} \sum_{j=1}^{d+1} \frac{q(y_j|X_i)}{N} < X_i, y_j > \to \int \sum_{j=1}^{d+1} < x, y_j > q(y_j|x) dp(x) = \int < x, y > d\mu(x, y)$$

and

$$\sum_{i=1}^{N} \sum_{j=1}^{d+1} \ln\left(\frac{q(y_j|X_i)}{q_N(y_j)}\right) \frac{q(y_j|X_i)}{N} \to \int \sum_{j=1}^{d+1} \ln\left(\frac{q(y_j|x)}{q(y_j)}\right) q(y_j|x) dp(x) = \int \ln(\frac{d\mu}{d\nu}) d\mu.$$

Since this holds for any $\mu \in \Pi(p,q)$,

$$\underline{\lim}_{N\to\infty} c_{N,\varepsilon} \geq c_\infty \equiv \sup_{\mu\in\Pi(p,q)} G(\mu,\nu). \tag{C.2}$$

Recall $c_N = \max_{\mu\in\Pi(p_N,q_N)} G(\mu,\nu_N)$. We claim that

$$c_{N,\varepsilon} \leq c_N + \varepsilon K_N, \tag{C.3}$$

for

$$K_N = K_1 \cdot \max_{i=1,\ldots,N} \max_{j=1,\ldots,d+1} |< X_i, y_j >| + \frac{1}{\theta}\cdot K_2,$$

where K_1 and K_2 are constants. We prove (C.3) in Appendix C.3.

Under our assumption that $\mathbb{E}_\nu[\exp(\theta < X, Y >)] < \infty$, the sequence K_N satisfies $K_N/N^\alpha \to 0$, for any $\alpha \in (0, 1/2)$. Set $\varepsilon_N = 1/N^\alpha$ so $\varepsilon_N K_N \to 0$. By the law of the iterated logarithm, with probability 1,

$$\max_{1\leq j\leq d+1} |q_N(y_j) - q(y_j)| < \varepsilon_N/2 \quad \text{and} \quad \max_{1\leq j\leq d+1} |\bar{q}_N(y_j) - q(y_j)| < \varepsilon_N/2$$

for all sufficiently large N, and then

$$\max_{1\leq j\leq d+1} |\bar{q}_N(y_j) - q_N(y_j)| < \varepsilon_N$$

as well. In other words, for any $\mu \in \Pi(p,q)$, we have $\mu_N \in \Pi_{\epsilon_N}(p_N, q_N)$ for all sufficiently large N, a.s. We can therefore strengthen (C.2) to

$$\underline{\lim}_{N\to\infty} c_{N,\varepsilon_N} \geq c_\infty.$$

But

$$\underline{\lim}_{N\to\infty} c_{N,\varepsilon_N} \leq \underline{\lim}_{N\to\infty} c_N + K_N\varepsilon_N = \underline{\lim}_{N\to\infty} c_N.$$

So we have shown that

$$\underline{\lim}_{N\to\infty} c_N \geq c_\infty.$$

□

We now establish the reverse inequality.

Lemma C.2. *As $N \to \infty$,*

$$\overline{\lim}_{N\to\infty} \max_{\mu\in\Pi(p_N,q_N)} G(\mu,\nu_N) \leq \sup_{\mu\in\Pi(p,q)} G(\mu,\nu), \text{ a.s.}$$

Proof: The supremum of $G(\mu,\nu)$ over $\mu \in \Pi(p,q)$ can be written as

$$-\frac{1}{\theta}\inf_{\mu\in\Pi(p,q)}\int \ln\left(\frac{d\mu(x,y)}{\exp\{\theta<x,y>\}d\nu(x,y)}\right)d\mu(x,y) = \frac{1}{\theta}\sup_{\mu\in\Pi(p,q)} -D(\mu|e^{\theta<x,y>}\nu). \tag{C.4}$$

By Theorem 3 of Rüschendorf and Thomsen [34], the optimum in (C.4) is attained at a solution of the form

$$d\mu^*(x,y) = e^{a(x)+b(y)+\theta<x,y>}\,d\nu(x,y), \tag{C.5}$$

for some functions a and b on $\mathbb{R}^d$. Similarly, for finite N, the optimizer of $G(\mu,\nu_N)$ over $\mu \in \Pi(p_N,q_N)$ has the form

$$d\mu_N^*(x,y) = e^{a_N(x)+b_N(y)+\theta<x,y>}\,d\nu_N(x,y).$$

For the rest of the proof, we will work with the formulation in (C.4), omitting the constant factor of $1/\theta$. We will apply a dual formulation of Bhattacharya [5]. To this end, consider the set Π^* of functions $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ of the form $h(x,y) = h_1(x) + h_2(y)$ with

$$\int h_1(x)\,dp(x) + \int h_2(y)\,dq(y) \geq 0.$$

The convex cone Π^* is contained within the dual cone of $\Pi(p,q)$, which is the set of functions $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ that have nonnegative expectations with respect to all $\mu \in \Pi(p,q)$. We consider the dual problem

$$\inf_{h\in\Pi^*} \ln \int e^{h_1(x)+h_2(y)+\theta<x,y>}\,d\nu(x,y).$$

With a and b as in (C.5), set

$$h_1^*(x) = a(x) + c/2, \quad h_2^*(x) = b(x) + c/2,$$

where

$$c = -\int [a(x)+b(y)]\,d\mu^*(x,y).$$

Observe that

$$\int h_1^*(x)\,dp(x) + \int h_2^*(y)\,dq(y) = 0,$$

so this (h_1^*, h_2^*) is dual feasible. Moreover, with this choice of h_1^*, h_2^*, the dual objective function value is

$$\ln \int e^{a(x)+b(x)+c+\theta<x,y>}\,d\nu(x,y) = \ln \int e^c\,d\mu^*(x,y) = c. \tag{C.6}$$

The primal objective in (C.4) evaluated at (C.5) yields

$$-D(\mu^*|e^{\theta<x,y>}\nu) = -\int [a(x)+b(y)]\,d\mu^*(x,y) = c,$$

so the primal and dual objective values agree. It follows from Theorem 2.1 of Bhattacharya [5] that this choice of (h_1^*, h_2^*) is optimal for the dual objective.

Parallel results hold for finite N as well. The maximal value of $G(\cdot, \nu_N)$ is $1/\theta$ times the dual optimum

$$\begin{aligned} c_N &= \inf_{h_1,h_2} \ln \int e^{h_1(x)+h_2(y)+\theta<x,y>}\, d\nu_N, \\ &\text{s.t.} \quad \int h_1(x)\, dp_N(x) + \int h_2(y)\, dq_N(y) \geq 0. \end{aligned} \tag{C.7}$$

For $\varepsilon \geq 0$, define

$$\begin{aligned} c_\infty^\varepsilon &:= \inf_{h_1,h_2} \ln \int e^{h_1(x)+h_2(y)+\theta<x,y>}\, d\nu, \\ &\text{s.t.} \quad \int h_1(x)\, dp(x) + \int h_2(y)\, dq(y) \geq \varepsilon. \end{aligned} \tag{C.8}$$

The infimum is finite because the integral is finite for any constant h_1, h_2. Let $(h_1^\varepsilon, h_2^\varepsilon)$ be feasible for (C.8) and satisfy

$$\ln \int e^{h_1^\varepsilon(x)+h_2^\varepsilon(y)+\theta<x,y>}\, d\nu \leq c_\infty^\varepsilon + \varepsilon.$$

Then, with probability 1,

$$\int h_1^\varepsilon(x)\, dp_N(x) + \int h_2^\varepsilon(y)\, dq_N(y) = \frac{1}{N}\sum_{i=1}^N (h_1^\varepsilon(X_i) + h_2^\varepsilon(Y_i)) \to \int h_1^\varepsilon(x)\, dp(x) + \int h_2^\varepsilon(y)\, dq(y) \geq \varepsilon,$$

so, for all sufficiently large N,

$$\int h_1^\varepsilon(x)\, dp_N(x) + \int h_2^\varepsilon(y)\, dq_N(y) \geq 0.$$

In other words, $(h_1^\varepsilon, h_2^\varepsilon)$ is feasible for (C.7) for all sufficiently large N, so

$$c_N \leq \ln \int e^{h_1^\varepsilon(x)+h_2^\varepsilon(y)+\theta<x,y>}\, d\nu_N \to \ln \int e^{h_1^\varepsilon(x)+h_2^\varepsilon(y)+\theta<x,y>}\, d\nu \leq c_\infty^\varepsilon + \varepsilon.$$

Hence,

$$\overline{\lim_{N\to\infty}}\, c_N \leq c_\infty^\varepsilon + \varepsilon.$$

By construction,

$$\int h_1^*(x)\, dp(x) + \int h_2^*(y)\, dq(y) = 0,$$

so $(h_1^* + \varepsilon/2, h_2^* + \varepsilon/2)$ is feasible for (C.8) and then

$$c_\infty^\varepsilon \leq \ln \int e^{h_1^*(x)+h_2^*(y)+\varepsilon+\theta<x,y>}\, d\nu$$

and

$$\lim_{\varepsilon\downarrow 0} \ln \int e^{h_1^*(x)+h_2^*(y)+\varepsilon+\theta<x,y>}\, d\nu = c,$$

with c as in (C.6). Thus, since $\varepsilon > 0$ can be taken arbitrarily small,

$$\overline{\lim_{N\to\infty}} c_N \leq c.$$

□

Combining Lemmas C.1 and C.2 proves part (i) of the theorem.

C.2 Weak Convergence of Optimal Solutions

Define

$$\Pi^N = \Pi(p,q) \cup \left(\underset{n\geq N}{\cup} \Pi(p_n, q_n) \right)$$

We will show that, almost surely, Π^N is compact (with respect to the topology of weak convergence on $\mathbb{R}^d \times \mathbb{R}^d$) for all sufficiently large N. It will follow that any sequence of optimizers $\{\mu_n^*\}$ is then eventually contained within a compact set, so every subsequence has a convergent subsequence.

Lemma C.3. *Π^N is compact for all sufficiently large N, a.s.*

Proof: By Prohorov's Theorem (Billingsley [6], p.37) the set Π^N is compact if it is uniformly tight, meaning that for all $\varepsilon > 0$ we can find a compact subset A of $\mathbb{R}^d \times \mathbb{R}^d$ such that $\mu(A) \geq 1 - \varepsilon$, for all $\mu \in \Pi^N$. Let A_1, A_2 be compact subsets of $\mathbb{R}^d$ such that

$$P(X \in A_1) = \int_{A_1} dp(x) \geq 1 - \varepsilon/4, \quad P(Y \in A_2) = \int_{A_2} dq(x) \geq 1 - \varepsilon/4.$$

Then, for any $\mu \in \Pi(p,q)$,

$$\int \mathbf{1}_{\{(x,y)\notin A_1\times A_2\}}\, d\mu(x,y) \leq P(X \notin A_1) + P(Y \notin A_2) \leq \varepsilon/2.$$

With probability 1, for all sufficiently large N and $\mu \in \Pi(p_N, q_N)$,

$$\int \mathbf{1}_{\{(x,y)\notin A_1\times A_2\}}\, d\mu(x,y) \leq \frac{1}{N}\sum_{i=1}^{N} \left(\mathbf{1}_{\{X_i\notin A_1\}} + \mathbf{1}_{\{Y_i\notin A_2\}} \right) \leq \varepsilon.$$

Thus, with probability 1, Π^N is uniformly tight for all sufficiently large N, and thus compact. □

The optimizers μ_N^* are contained in the sets $\Pi(p_N, q_N)$, so for all sufficiently large N, the sequence μ_n^*, $n \geq N$, is contained in a compact set Π^N, and then every subsequence has a further subsequence that converges weakly.

Suppose the subsequence $\mu_{n_k}^*$ converges, say $\mu_{n_k}^* \Rightarrow \tilde{\mu}$. The marginals of $\mu_{n_k}^*$ converge to p and q, so $\tilde{\mu} \in \Pi(p,q)$, making $\tilde{\mu}$ feasible for the limiting problem. We claim that it is optimal. We have, a.s.,

$$\int e^{\theta<x,y>}\, d\mu_{n_k}^* \leq \int \sum_{j=1}^{d+1} e^{\theta<x,y_j>}\, dp_{n_k}(x) \to \int \sum_{j=1}^{d+1} e^{\theta<x,y_j>}\, dp(x),$$

by the strong law of large numbers, because the condition $\mathbb{E}_\nu[e^{\theta<x,y>}] < \infty$ implies that the limit is finite. This is then more than sufficient to ensure that

$$\int < x, y > \, d\mu^*_{n_k}(x, y) \to \int < x, y > \, d\tilde{\mu}(x, y). \tag{C.9}$$

Moreover, relative entropy is lower semi-continuous with respect to weak convergence (Dupuis and Ellis [19], Lemma 1.4.3), so

$$D(\tilde{\mu}|\nu) \leq \underline{\lim}_{k\to\infty} D(\mu^*_{n_k}|\nu_{n_k})$$

and then

$$G(\tilde{\mu}, \nu) \geq \overline{\lim}_{k\to\infty} G(\mu^*_{n_k}, \nu_{n_k}) = \sup_{\mu\in\Pi(p,q)} G(\mu, \nu),$$

by part (i) of the theorem. Thus, $\tilde{\mu}$ is optimal. Using the equivalence between the optimization of $G(\cdot, \nu)$ and (C.4), we know from Theorem 3 of Rüschendorf and Thomsen [34] that the maximum is uniquely attained by some μ^*, and thus $\tilde{\mu} = \mu^*$.

We have shown that every subsequence of μ^*_n has a further subsequence that converges to μ^*. It follows that $\mu^*_n \Rightarrow \mu^*$. This proves part (ii) of the theorem. The uniform integrability needed for (4.6) follows as in (C.9), which proves part (iii). □

C.3 Proof of Inequality (C.3)

It remains to prove (C.3). First we construct a feasible solution $\hat{\mu}_N$ of $\max_{\mu\in\Pi(p_N,q_N)} G(\mu, \nu_N)$ by modifying the optimal solution $\mu^*_{N,\varepsilon}$ of the relaxed problem $\max_{\mu\in\Pi_\varepsilon(p_N,q_N)} G(\mu, \nu_N)$. Then we use the difference between $G(\hat{\mu}_N, \nu_N)$ and $G(\mu^*_{N,\varepsilon}, \nu_N)$ to bound the difference between c_N and $c_{N,\varepsilon}$.

Define $\varepsilon^N_j = \sum_{i=1}^N (\mu^*_{N,\varepsilon})_{ij} - q_N(y_j)$, which is the difference between the Y marginal of $\mu^*_{N,\varepsilon}$ and the empirical distribution of Y. Note that $|\varepsilon^N_j| \leq \varepsilon$ for $j = 1, ..., d+1$. We claim that there exists $\{\varepsilon^*_{ij}\}$ for which

$$(\hat{\mu}_N)_{ij} := (\mu^*_{N,\varepsilon})_{ij} - \varepsilon^*_{ij}, \quad i = 1, ..., N \text{ and } j = 1, ..., d+1,$$

satisfies the following conditions:

$$\hat{\mu}_N \in \Pi(p_N, q_N), \tag{C.10}$$

$$\textstyle\sum_{i=1}^N \sum_{j=1}^{d+1} |\varepsilon^*_{ij}| \leq (d+1)\varepsilon, \tag{C.11}$$

$$-C_N \cdot \varepsilon \cdot \tfrac{1}{N} \leq \varepsilon^*_{ij} \leq C_N \cdot \varepsilon \cdot (\mu^*_{N,\varepsilon})_{ij}, \tag{C.12}$$

where $C_N = \max_{j=1,\cdots,d+1}\{1/q_N(y_j)\}$. Since $q(y_j) > 0$ for $j = 1, \cdots, d+1$, we know $q_N(y_j) > 0$ for all j and N large enough, and C_N is well defined.

To see that such $\{\varepsilon^*_{ij}\}$ exist, rearrange $\{\varepsilon^N_j\}$ in descending order $\{\varepsilon^N_{j_k}\}$ for $k = 1, \cdots, d+1$, and let m denote number of nonnegative elements. Note that $\varepsilon^N_{j_k} \geq 0$ for $k = 1, \cdots, m$, and $\varepsilon^N_{j_k} < 0$ for $k = m+1, \cdots, d+1$, and $\sum_{k=1}^m \varepsilon^N_{j_k} = -\sum_{k=m+1}^{d+1} \varepsilon^N_{j_k}$. Let

$$\varepsilon^*_{i,j_k} = \frac{\left(\mu^*_{N,\varepsilon}\right)_{i,j_k}}{\sum_{i=1}^N \left(\mu^*_{N,\varepsilon}\right)_{i,j_k}} \cdot \varepsilon^N_{j_k}$$

for $i = 1, \cdots, N$ and $k = 1, \cdots, m$. Let

$$S_i = \sum_{k=1}^m \varepsilon^*_{i,j_k}$$

for $i = 1, \cdots, N$. Let

$$\varepsilon^*_{i,j_k} = \frac{\varepsilon^N_{j_k}}{|\sum_{l=m+1}^{d+1} \varepsilon^N_{j_l}|} \cdot S_i$$

for $i = 1, \cdots, N$ and $k = m+1, \cdots, d+1$.

We verify (C.10)-(C.12) for $\{\varepsilon^*_{ij}\}$. Since (C.10) is equivalent to $\sum_{i=1}^N \varepsilon^*_{i,j_k} = \varepsilon^N_{j_k}$ for $k = 1, \cdots, d+1$, we know that by construction it holds for $\{\varepsilon^*_{ij}\}$. Next,

$$\begin{aligned}
\sum_{i=1}^N \sum_{j=1}^{d+1} |\varepsilon^*_{ij}| &= \sum_{i=1}^N \sum_{k=1}^{d+1} |\varepsilon^*_{i,j_k}| \\
&= \sum_{i=1}^N \sum_{k=1}^m \varepsilon^*_{i,j_k} - \sum_{i=1}^N \sum_{k=m+1}^{d+1} \varepsilon^*_{i,j_k} \\
&= \sum_{i=1}^N \sum_{k=1}^m \frac{\left(\mu^*_{N,\varepsilon}\right)_{i,j_k}}{\sum_{i=1}^N \left(\mu^*_{N,\varepsilon}\right)_{i,j_k}} \cdot \varepsilon^N_{j_k} - \sum_{i=1}^N \sum_{k=m+1}^{d+1} \frac{\varepsilon^N_{j_k}}{|\sum_{l=m+1}^{d+1} \varepsilon^N_{j_l}|} \cdot S_i \\
&= \sum_{k=1}^m \varepsilon^N_{j_k} - \sum_{k=m+1}^{d+1} \varepsilon^N_{j_k} \\
&\leq (d+1)\varepsilon
\end{aligned}$$

The last equality follows by $\sum_{i=1}^N S_i = \sum_{k=1}^m \varepsilon^N_{j_k} = |\sum_{l=m+1}^{d+1} \varepsilon^N_{j_l}|$. Thus $\{\varepsilon^*_{ij}\}$ satisfy (C.11).

For $k = 1, \cdots, m$,

$$0 \leq \varepsilon^*_{i,j_k} = \frac{\left(\mu^*_{N,\varepsilon}\right)_{i,j_k}}{q_N(y_{j_k}) + \varepsilon^N_{j_k}} \cdot \varepsilon^N_{j_k} \leq \frac{\left(\mu^*_{N,\varepsilon}\right)_{i,j_k}}{q_N(y_{j_k})} \cdot \varepsilon^N_{j_k} \leq C_N \cdot \varepsilon \cdot \left(\mu^*_{N,\varepsilon}\right)_{i,j_k}$$

For $k = m+1, \cdots, d+1$,

$$0 \geq \varepsilon^*_{i,j_k} \geq -S_i \geq -C_N \cdot \varepsilon \cdot \sum_{k=1}^m \left(\mu^*_{N,\varepsilon}\right)_{i,j_k} \geq -C_N \cdot \varepsilon \cdot \frac{1}{N}$$

Thus $\{\varepsilon^*_{ij}\}$ satisfy (C.12).

Because $\hat{\mu}_N$ is feasible but not necessarily optimal, we have

$$G(\hat{\mu}_N, \nu_N) \leq c_N \leq c_{N,\varepsilon}.$$

We will show that

$$c_{N,\varepsilon} - G(\hat{\mu}_N, \nu_N) \leq \varepsilon K_N, \tag{C.13}$$

for

$$K_N = (d+1) \cdot \max_{i=1,\ldots,N} \max_{j=1,\ldots,d+1} | < X_i, y_j > | + \frac{1}{\theta} \cdot K_2,$$

where K_2 is a constant. It then follows that

$$c_{N,\varepsilon} - c_N \leq \varepsilon K_N.$$

To show (C.13), write

$$\begin{aligned} c_{N,\varepsilon} - G(\hat{\mu}_N, \nu_N) &= \Big(\int < x, y > d\mu_{N,\varepsilon} - \int < x, y > d\hat{\mu}_N \Big) \\ &\quad -\frac{1}{\theta} \int \Big(\frac{d\mu_{N,\varepsilon}}{d\nu_N} \ln\big(\frac{d\mu_{N,\varepsilon}}{d\nu_N}\big) - \frac{d\hat{\mu}_N}{d\nu_N} \ln\big(\frac{d\hat{\mu}_N}{d\nu_N}\big) \Big) d\nu_N. \end{aligned}$$

The first part has upper bound

$$(d+1) \cdot \max_{i=1,\ldots,N} \max_{j=1,\ldots,d+1} | < X_i, y_j > | \cdot \varepsilon.$$

Let $x = d\mu_{N,\varepsilon}/d\nu_N$ and $x - \Delta x = d\hat{\mu}_N/d\nu_N$. Drop the factor $-1/\theta$ and rewrite the second part as

follows:

$$
\begin{aligned}
&\int x\ln x-(x-\Delta x)\ln(x-\Delta x)\,d\nu_N\\
&= \int x\ln x - x\ln(x-\Delta x)+\Delta x\ln(x-\Delta x)\,d\nu_N\\
&= \int -x\ln(1-\frac{\Delta x}{x})+\Delta x\ln(x-\Delta x)\,d\nu_N\\
&\geq \int -x\cdot(-\frac{\Delta x}{x})\,d\nu_N+\int \Delta x\ln(x-\Delta x)\,d\nu_N\\
&= \int \Delta x\,d\nu_N+\int \mathbf{1}_{\{\Delta x\geq 0\}}\Delta x\ln(x-\Delta x)\,d\nu_N+\int \mathbf{1}_{\{\Delta x<0\}}\Delta x\ln(x-\Delta x)\,d\nu_N\\
&= 0+\int \mathbf{1}_{\{\Delta x\geq 0\}}\frac{\Delta x}{x-\Delta x}(x-\Delta x)\ln(x-\Delta x)\,d\nu_N+\int \mathbf{1}_{\{\Delta x<0\}}\Delta x\ln(x-\Delta x)\,d\nu_N\\
&\geq \int \mathbf{1}_{\{\Delta x\geq 0\}}\frac{\Delta x}{x-\Delta x}(x-\Delta x-1)\,d\nu_N+\int \mathbf{1}_{\{\Delta x<0\}}\Delta x(x-\Delta x-1)\,d\nu_N\\
&= \int \mathbf{1}_{\{\Delta x\geq 0\}}\Delta x(1-\frac{1}{x-\Delta x})\,d\nu_N+\int \mathbf{1}_{\{\Delta x<0\}}(\Delta x\cdot x-(\Delta x)^2-\Delta x)\,d\nu_N\\
&\geq \int \mathbf{1}_{\{\Delta x\geq 0\}}(-\frac{\Delta x}{x-\Delta x})\,d\nu_N-C_N^2\varepsilon-C_N^3(d+1)\varepsilon^2\\
&\geq \int \mathbf{1}_{\{\Delta x\geq 0\}}(-\frac{C_N\cdot\varepsilon}{(1-C_N\cdot\varepsilon)})\,d\nu_N-C_N^2\varepsilon-C_N^3(d+1)\varepsilon^2\\
&\geq -\frac{C_N}{(1-C_N\cdot\varepsilon)}\cdot\varepsilon-C_N^2\varepsilon-C_N^3(d+1)\varepsilon^2\\
&= -\big(\frac{C_N}{(1-C_N\cdot\varepsilon)}+C_N^2+C_N^3(d+1)\varepsilon\big)\cdot\varepsilon\\
&:= -K_{C_N,\varepsilon}\cdot\varepsilon
\end{aligned}
$$

We explain the inequalities in turn. The first inequality follows from $\ln x\leq x-1$ for $x\geq 0$, and the second inequality follows from both $\ln x\leq x-1$ for $x\geq 0$ and $x\ln x\geq x-1$ for $x\geq 0$. The third inequality follows by dropping a positive term Δx in the first integral and noting that

$$
\begin{aligned}
\int \mathbf{1}_{\{\Delta x<0\}}\Delta x\cdot x\,d\nu_N &= \int \mathbf{1}_{\{d\mu_{N,\varepsilon}-d\hat{\mu}_N<0\}}\frac{d\mu_{N,\varepsilon}-d\hat{\mu}_N}{d\nu_N}\cdot\frac{d\mu_{N,\varepsilon}}{d\nu_N}\,d\nu_N\\
&= \sum_{ij}\frac{\mathbf{1}_{\{\varepsilon_{ij}^*<0\}}\varepsilon_{ij}^*}{\frac{1}{N}q_N(y_j)}\cdot(\mu_{N,\varepsilon})_{ij}\\
&\geq \sum_{ij}-C_N^2\cdot\varepsilon\cdot(\mu_{N,\varepsilon})_{ij}=-C_N^2\cdot\varepsilon,\quad \text{by (C.12)},
\end{aligned}
$$

and

$$\begin{aligned}\int \mathbf{1}_{\{\Delta x<0\}}(-\Delta x)^2\, d\nu_N &\geq -\int \Delta x^2\, d\nu_N = -\sum_{ij} \frac{(\varepsilon^*_{ij})^2}{\frac{1}{N}q_N(y_j)} \\ &\geq -\sum_{ij} \frac{C_N^2(\frac{1}{N})^2\varepsilon^2}{\frac{1}{N}q_N(y_j)}, \quad \text{by (C.12)}, \\ &\geq -C_N^3(d+1)\varepsilon^2.\end{aligned}$$

The fourth inequality holds because $\Delta x \leq x \cdot C_N \cdot \varepsilon$ for $\Delta x \geq 0$, by (C.12).

The coefficient $K_{C_N,\varepsilon}$ is increasing in both C_N and ε. Since $q_N(y_j) \to q(y_j)$ as $N \to \infty$, $C_N \to \max_{j=1,\cdots,d+1}\{1/q(y_j)\}$ as $N \to \infty$, thus we can find a constant $C \geq C_N$ for all N. On the other hand, without loss of generality we can assume that ε is small enough, such that $1-C\cdot\varepsilon > 1/2$, i.e. $\varepsilon < 1/(2C)$. Choose $K_2 = K_{C,1/(2C)}$. Then

$$\int x\ln x - (x-\Delta x)\ln(x-\Delta x)\, d\nu_N \geq -K_2 \cdot \varepsilon$$

for all N and ε small enough. We thus have

$$\begin{aligned}c_{N,\varepsilon} - G(\hat{\mu}_N, \nu_N) &\leq (d+1)\cdot \max_{i=1,\ldots,N}\max_{j=1,\ldots,d+1} |< X_i, y_j >| \cdot \varepsilon + \frac{1}{\theta}K_2 \cdot \varepsilon \\ &= K_N \cdot \varepsilon,\end{aligned}$$

and ((C.13)) is proved.

D Proof of Theorem 6.1

Let $\mu \in \bar{\Pi}(p,q)$ be any feasible solution to the limiting problem. Write $\mu((dx,dz),y) = p(dx,dz)q(y|x,z)$, and define the mass function μ_N on $((X_i,Z_i),y_j)$, $i = 1,\ldots,N$, $j = 1,\ldots,d+1$, by setting

$$\mu_N((X_i,Z_i),y_j) = \frac{1}{N}q(y_j|(X_i,Z_i)).$$

For each y_j, we get the marginal probability

$$\bar{q}_N(y_j) = \sum_{i=1}^N \mu_N((X_i,Z_i),y_j) = \frac{1}{N}\sum_{i=1}^N q(y_j|(X_i,Z_i)).$$

The expectation of $< Z_i, y_j >$ with respect to μ_N is given by

$$\bar{v}_0^N = \sum_{i=1}^N\sum_{j=1}^{d+1} < Z_i, y_j > \mu_N((X_i,Z_i),y_j).$$

By the strong law of large numbers for the i.i.d. sequence (X_i, Z_i), $i = 1, \ldots, N$, we have $(\bar{q}_N(y_1), \ldots, \bar{q}_N(y_m)) \to (q(y_1), \ldots, q(y_m))$, a.s., and also

$$\begin{aligned} \bar{v}_0^N = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d+1} < Z_i, y_j > q(y_j|(X_i, Z_i)) \quad &\to \quad \int \sum_{j=1}^{d+1} < h_Z(x,z), y_j > \, q(y_j|(x,z))\, dp(x,z) \\ &= \quad \int < h_Z(x,z), y > \, d\mu((x,z),y) = v_0, \end{aligned}$$

where v_0 is the value in the constraint (6.1) because $\mu \in \bar{\Pi}(p,q)$. In fact, by the law of the iterated logarithm, if we set $\epsilon_N = 1/N^\alpha$ with $0 < \alpha < 1/2$, then, with probability 1,

$$\max_{1\le j\le d+1} |\bar{q}_N(y_j) - q(y_j)| < \epsilon_N, \quad \max_{1\le j\le d+1} |\bar{q}_N(y_j) - q_N(y_j)| < \epsilon_N$$

and, under our square-integrability condition on Z,

$$|\bar{v}_0^N - v_0| < \epsilon_N,$$

for all sufficiently large N. It follows that $\mu_N \in \bar{\Pi}_{\epsilon_N}(p_N, q_N)$, for all sufficiently large N.

D.1 Upper Bound

Because μ_N is feasible for all sufficiently large N, it provides a lower bound on the optimal value c_{N,ϵ_N} in (6.4),

$$c_{N,\epsilon_N} \ge \sum_{i=1}^{N}\sum_{j=1}^{d+1} \mu_N((X_i, Z_i), y_j) < X_i, y_j > = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d+1} q(y_j|(X_i, Z_i)) < X_i, y_j > .$$

By the strong law of large numbers

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d+1} q(y_j|(X_i, Z_i)) < X_i, y_j > \quad &\to \quad \int_{\mathbb{R}^d\times\mathbb{R}^d} \sum_{j=1}^{d+1} q(y_j|(x,z)) < h_X(x,z), y_j > dp(x,z) \\ &= \quad \int_{(\mathbb{R}^d\times\mathbb{R}^d)\times\mathbb{R}^d} < h_X(x,z), y > d\mu((x,z),y). \end{aligned}$$

So

$$\underline{\lim}_{N\to\infty} \; c_{N,\epsilon_N} \ge \int_{(\mathbb{R}^d\times\mathbb{R}^d)\times\mathbb{R}^d} < h_X(x,z), y > d\mu((x,z),y)$$

And since this holds for any $\mu \in \bar{\Pi}(p,q)$,

$$\underline{\lim}_{N\to\infty} \; c_{N,\epsilon_N} \ge c_\infty. \tag{D.1}$$

D.2 Lower Bound

To prove a lower bound, we formulate a dual problem for the relaxed finite-N problem (6.4) with objective value $d_{N,\epsilon}$, and we formulate a dual for the limiting problem (6.3) with objective value d_∞.

The relaxed finite problem in (6.4) is a linear program. Its dual can be written as

$$d_{N,\epsilon} \equiv \min_{\Phi,\Psi_1,\Psi_2,\xi_1,\xi_2} \{F_N(\Phi,\Psi_1,\Psi_2,\xi_1,\xi_2) + \epsilon K(\Psi_1,\Psi_2,\xi_1,\xi_2)\} \tag{D.2}$$

with

$$F_N(\Phi,\Psi_1,\Psi_2,\xi_1,\xi_2) = \frac{1}{N}\sum_{i=1}^{N}\Phi_i + \sum_{j=1}^{d+1}(\Psi_{1j}+\Psi_{2j})\cdot q_N(y_j) + (\xi_1+\xi_2)v_0$$

and

$$K(\Psi_1,\Psi_2,\xi_1,\xi_2) = \sum_{j=1}^{d+1}(\Psi_{1j}-\Psi_{2j}) + (\xi_1-\xi_2),$$

the infimum taken over $\Phi \in \mathbb{R}$, $\Psi_{1j} \geq 0$, $\Psi_{2j} \leq 0$, $\xi_1 \geq 0$, $\xi_2 \leq 0$, satisfying

$$\Phi_i + \Psi_{1j} + \Psi_{2j} + (\xi_1+\xi_2)\cdot < Z_i, y_j > \geq < X_i, y_j >,$$

for $i = 1,...,N$, and all $j = 1,...,d+1$ with $q_N(y_j) > 0$. We have already seen that problem (6.4) is feasible for all sufficiently large N, and once it is feasible $c_{N,\epsilon} = d_{N,\epsilon}$ by standard linear programming duality.

We define the dual of the limiting problem (6.3) by setting

$$d_\infty = \inf_{\phi,\psi,\xi} F(\phi,\psi,\xi)$$

with

$$F(\phi,\psi,\xi) = \int \phi(x,z)\, dp(x,z) + \sum_{j=1}^{d+1}\psi(y_j)q(y_j) + \xi v_0,$$

the infimum taken over functions $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $\psi : \mathbb{R}^d \to \mathbb{R}$, and a scalar $\xi \in \mathbb{R}$, satisfying, for all (x,z) in the support of p and all y in the support of q,

$$\phi(x,z) + \psi(y) + \xi < z, y > \ \geq \ < x, y >,$$

with $\phi \in L^1(p)$.

For any $\tilde{\epsilon} > 0$, we may pick $\phi_{\tilde{\epsilon}}$, $\psi_{\tilde{\epsilon}}$, and $\xi_{\tilde{\epsilon}}$ feasible for the limiting dual and for which

$$F(\phi_{\tilde{\epsilon}}, \psi_{\tilde{\epsilon}}, \xi_{\tilde{\epsilon}}) \leq d_\infty + \tilde{\epsilon}.$$

We may then define a feasible solution to (D.2) by setting $\Phi_i = \phi_{\tilde{\epsilon}}(X_i, Z_i)$, $\Psi_{1j} = \psi^+_{\tilde{\epsilon}}(y_j)$, $\Psi_{2j} = -\psi^-_{\tilde{\epsilon}}(y_j)$, $\xi_1 = \xi^+_{\tilde{\epsilon}}$, and $\xi_2 = -\xi^-_{\tilde{\epsilon}}$. By the strong law of large numbers, this choice yields

$$F_N(\Phi, \Psi_1, \Psi_2, \xi_1, \xi_2) \to F(\phi_{\tilde{\epsilon}}, \psi_{\tilde{\epsilon}}, \xi_{\tilde{\epsilon}}), \quad \text{a.s.}$$

For any $\bar{\epsilon} > 0$, there is a stochastic $N(\tilde{\epsilon}, \bar{\epsilon})$ such that for all $N > N(\tilde{\epsilon}, \bar{\epsilon})$,

$$F_N(\Phi, \Psi_1, \Psi_2, \xi_1, \xi_2) \leq F(\phi_{\tilde{\epsilon}}, \psi_{\tilde{\epsilon}}, \xi_{\tilde{\epsilon}}) + \bar{\epsilon}, \quad \text{a.s.},$$

and this $N(\tilde{\epsilon}, \bar{\epsilon})$ does not depend on the ϵ that defines the relaxation (D.2). Thus, we have, for all sufficiently large N,

$$d_{N,\epsilon} \leq d_\infty + \tilde{\epsilon} + \bar{\epsilon} + K(\Psi_1, \Psi_2, \xi_1, \xi_2)\epsilon;$$

and, because $N(\tilde{\epsilon}, \bar{\epsilon})$ does not depend on ϵ,

$$d_{N,\epsilon_N} \leq d_\infty + \tilde{\epsilon} + \bar{\epsilon} + K(\Psi_1, \Psi_2, \xi_1, \xi_2)\epsilon_N,$$

for all $N > N(\tilde{\epsilon}, \bar{\epsilon})$, so

$$\overline{\lim_{N\to\infty}} d_{N,\epsilon_N} \leq d_\infty + \tilde{\epsilon} + \bar{\epsilon}.$$

Because $\tilde{\epsilon} > 0$ and $\bar{\epsilon} > 0$ are arbitrary,

$$\overline{\lim_{N\to\infty}} d_{N,\epsilon_N} \leq d_\infty.$$

We have already noted that $d_{N,\epsilon_N} = c_{N,\epsilon_N}$ by ordinary linear programming duality. In Appendix D.3 we show that that

$$d_\infty = c_\infty. \tag{D.3}$$

Thus,

$$\overline{\lim_{N\to\infty}} c_{N,\epsilon_N} = \overline{\lim_{N\to\infty}} d_{N,\epsilon_N} \leq d_\infty = c_\infty,$$

which, together with (D.1) proves the result.

D.3 A Duality Result

In this section, we prove the equality $c_\infty = d_\infty$ used in Appendix D.2. The result follows from Theorem 5.10 of Villani [36], once we show that we can transform the primal problem to an equivalent problem that satisfies the conditions of the theorem. We formulate the equivalent problem using a result of Luenberger [28], for which we adopt his notation.

Let X be the vector space of signed finite measures μ on $(\mathbb{R}^d \times \mathbb{R}^d) \times \mathbb{R}^d$ satisfying

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|u\| \, \mu(dx, \mathbb{R}^d) < \infty.$$

Let $\Omega \subset X$ be the subset of probability measures with marginals p and q, which is a convex set. For $\mu \in X$, let

$$f(\mu) \;=\; \int_{(\mathbb{R}^d \times \mathbb{R}^d)\times \mathbb{R}^d} - < h_X(x,z), y > \, d\mu((x,z),y).$$

Let $G(\cdot)$ be a mapping from X to $\mathbb{R}$ defined by

$$G(\mu) = \int < h_Z(x,z), y > d\mu((x,z),y) - v_0.$$

The primal problem is

$$c_\infty = - \inf_{\mu \in \Omega, G(\mu)=0} f(\mu).$$

Define

$$L(\xi) = \inf_{\mu \in \Omega} \big\{ \int - < h_X(x,z), y > d\mu((x,z),y) + \xi \cdot G(\mu) \big\}. \tag{D.4}$$

Now apply Theorem 1 of Section 8.6 of Luenberger [28] (with the extension in problem 7 of Section 8.8) to conclude that

$$\inf_{\mu \in \Omega, G(\mu)=0} f(\mu) = \max_{\xi \in \mathbb{R}} \; L(\xi),$$

and there exists ξ^* such that $L(\xi^*) = -c_\infty$.

Drop the constant term $-\xi^* \cdot v_0$ in $L(\xi^*)$, and denote it by L^*, so

$$\begin{aligned} L^* \;&=\; \inf_{\mu \in \Pi(p,q)} \int - < h_X(x,z), y > d\mu((x,z),y) + \xi^* \cdot \int < h_Z(x,z), y > d\mu((x,z),y) \\ &=\; \inf_{\mu \in \Pi(p,q)} \int \big(- < h_X(x,z), y > + \xi^* \cdot < h_Z(x,z), y > \big) d\mu((x,z),y) \end{aligned}$$

Define the dual problem DL^*,

$$DL^* \;=\; \sup_{(\phi,\psi) \in L^1(p) \times L^1(q);\; -\phi - \psi \le -c + \xi^* \cdot v} - \int_{\mathbb{R}^d \times \mathbb{R}^d} \phi(x,z) dp(x,z) - \sum_{j=1}^{d+1} \psi(y_j) q(y_j),$$

where $c((x,z),y) =< h_X(x,z), y >$, and $v((x,z),y) =< h_Z(x,z), y >$.

Let $a(x,z) = \frac{1}{2} < (x, \xi^* z), (x, \xi^* z) >$ and $b(y) = \frac{1}{2} < y, y >$. We have

$$- < h_X(x,z), y > + \xi^* \cdot < h_Z(x,z), y > \; \ge \; -a(x,z) - b(y).$$

By condition (i) in Theorem 6.1, $a \in L^1(p)$ and $b \in L^1(q)$. It follows from Theorem 5.10 of Villani [36] that strong duality holds, i.e. $L^* = DL^*$.

Since $L^* < +\infty$ and $- < h_X(x,z), y > + \xi^* \cdot < h_Z(x,z), y > \leq \ a(x,z) + b(y)$, it follows from part (iii) of Theorem 5.10 of Villani [36] that solutions exists for both problems. Let (ϕ^*, ψ^*) denote an optimal solution to DL^*, then (ϕ^*, ψ^*, ξ^*) is a feasible solution to the dual problem

$$d_\infty = \inf_{\phi(x,z)+\psi(y)+\xi v((x,z),y) \geq c((x,z),y)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \phi(x,z) dp(x,z) + \sum_{j=1}^{d+1} \psi(y_j) q(y_j) + \xi v_0.$$

Let d^* denote the objective value by substituting (ϕ^*, ψ^*, ξ^*) in the objective function. Note that $d^* = -DL^* + \xi^* v_0 = -L^* + \xi^* v_0 = c_\infty$, so (ϕ^*, ψ^*, ξ^*) is optimal for the dual problem d_∞, and strong duality holds $d_\infty = c_\infty$.

E Proof of Theorem 6.2

In this section we show the convergence result for the penalty problems with the auxiliary constraints in (6.1) as $N \to \infty$. We start with the convergence of the objective function value asserted in part (i) of the theorem.

E.1 Convergence of the Optimal Objective Value

Let G_∞ denote the optimal value of the penalty limit problem,

$$G_\infty = \sup_{\mu \in \bar{\Pi}(p,q)} G(\mu, \nu) = \sup_{\mu \in \bar{\Pi}(p,q)} \int < x, y > \, d\mu - \frac{1}{\theta} \int \ln(\frac{d\mu}{d\nu}) \, d\mu. \tag{E.1}$$

Let $G_{N,\epsilon}$ be the optimal value of the penalty finite relaxed problem with sample size N,

$$G_{N,\epsilon} = \sup_{\mu \in \bar{\Pi}_\epsilon(p_N, q_N)} G(\mu, \nu_N) = \sup_{\mu \in \bar{\Pi}_\epsilon(p_N, q_N)} \int < x, y > \, d\mu - \frac{1}{\theta} \int \ln(\frac{d\mu}{d\nu_N}) \, d\mu. \tag{E.2}$$

Lemma E.1. $\underline{\lim}_{N \to \infty} G_{N,\epsilon_N} \geq G_\infty$, *for* $\epsilon_N = 1/N^\alpha$ *and* $\alpha \in (0, 1/2)$.

Proof: Let $\mu \in \bar{\Pi}(p,q)$ be any feasible solution to the limiting problem. Define a mass function on the pairs $((X_i, Z_i), y_j)$, $i = 1, \cdots, N$, $j = 1, \cdots, d+1$:

$$\mu_N((X_i, Z_i), y_j) = \frac{1}{N} q(y_j | (X_i, Z_i)).$$

From the argument in Appendix D, we know that $\mu_N \in \bar{\Pi}^N_{\epsilon_N}$, so

$$\begin{aligned}
G_{N,\epsilon_N} &\geq \sum_{i=1}^{N} \sum_{j=1}^{d+1} \mu_N((X_i, Z_i), y_j) < X_i, y_j > - \frac{1}{\theta} D(\mu_N | \nu_N) \\
&= \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d+1} q(y_j | (X_i, Z_i)) < X_i, y_j > - \frac{1}{\theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{d+1} \ln(\frac{q(y_j | (X_i, Z_i))}{q(y_j)}) q(y_j | (X_i, Z_i)) \\
&\to \int < x, y > d\mu((x,z), y) - \frac{1}{\theta} \int \ln(\frac{d\mu}{d\nu}) d\mu,
\end{aligned}$$

the limit following from the strong law of large numbers. Thus, $\underline{\lim}_{N\to\infty} G_{N,\epsilon_N} \geq G_\infty$. □

We have shown that the limiting objective value is a lower bound for the sequence in part (i) of Theorem 6.2. We will use a dual formulation to show the reverse inequality. The argument requires several lemmas.

We reformulate the problem of maximizing $G(\cdot,\nu)$ as

$$-\frac{1}{\theta}\min_{\mu\in\bar{\Pi}(p,q)}\int\left(\frac{d\mu((x,z),y)}{\exp(\theta<x,y>)d\nu((x,z),y)}\right)d\mu((x,z),y) = -\frac{1}{\theta}\min_{\mu\in\bar{\Pi}(p,q)} D(\mu|e^{(\theta<x,y>)}\nu). \quad \text{(E.3)}$$

Dropping the constant factor $-1/\theta$ from (E.3), we get the equivalent problem

$$P_\infty = \min_{\mu\in\bar{\Pi}(p,q)} D(\mu|e^{(\theta<x,y>)}\nu) \quad \text{(E.4)}$$

Define $\bar{\Pi}^*(p,q)$ to be the set of functions $h : (\mathbb{R}^d\times\mathbb{R}^d)\times\mathbb{R}^d \to \mathbb{R}$ of the form

$$h((x,z),y) = h_1(x,z) + h_2(y) + h_3 v((x,z),y) - h_4$$

where

$$v((x,z),y) =< h_Z(x,z), y >=< z, y >,$$

with

$$\int h((x,z),y)d\mu((x,z),y) \geq 0, \quad \text{for all } \mu\in\bar{\Pi}(p,q).$$

Lemma E.2. *Let D_∞ be the dual problem to P_∞, defined as*

$$D_\infty = \inf_{h\in\bar{\Pi}^*(p,q)} \ln\int e^{h((x,z),y)+\theta<x,y>}\, d\nu((x,z),y), \quad \text{(E.5)}$$

The following statements hold:
(i) The optimal solution to the primal problem is

$$d\mu^*((x,z),y) = e^{a(x,z)+b(y)+\xi v((x,z),y)+\theta<x,y>}\, d\nu((x,z),y). \quad \text{(E.6)}$$

(ii) The optimal solution to the dual problem is

$$h^*((x,z),y) = h_1^*(x,z) + h_2^*(y) + h_3^* v((x,z),y) - h_4^*,$$

$$h_1^*(x,z) = a(x,z), \quad h_2^*(x) = b(x), \quad h_3^* = \xi, \quad h_4^* = \int a(x,z) + b(y) + \xi\cdot v((x,z),y)\, d\mu^*((x,z),y).$$

(iii) Strong duality holds, $P_\infty = -D_\infty$.

Proof: Conclusion (i) follows Theorem 3 of Rüschendorf and Thomsen [34].

To apply the dual formulation in Bhattacharya, we consider the set $\bar{\Pi}^*(p,q)$ of functions $h : (\mathbb{R}^d \times \mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ of the form

$$h((x,z),y) = h_1(x,z) + h_2(y) + h_3 v((x,z),y) - h_4$$

with, for any $\mu \in \bar{\Pi}(p,q)$

$$\begin{aligned} \int h((x,z),y) d\mu((x,z),y) &= \int h_1(x,z)\, dp((x,z)) + \int h_2(y)\, dq(y) + h_3 v_0 - h_4 \\ &\geq 0. \end{aligned}$$

Observe that (the convex cone) $\bar{\Pi}^*(p,q)$ is contained within the dual cone of $\bar{\Pi}(p,q)$. We consider the dual problem

$$\inf_{h \in \bar{\Pi}^*} \ln \int e^{h((x,z),y) + \theta <x,y>}\, d\nu((x,z),y).$$

With μ^*, $a(x,z)$, $b(x)$, ξ as in (E.6), set

$$h_1^*(x,z) = a(x,z), \quad h_2^*(x) = b(x), \quad h_3^* = \xi, \quad h_4^* = c \equiv \int a(x,z) + b(y) + \xi \cdot v((x,z),y)\, d\mu^*((x,z),y).$$

Observe that

$$\int h^*((x,z),y) d\mu((x,z),y) = \int h_1^*((x,z))\, dp((x,z)) + \int h_2^*(y)\, dq(y) + h_3^* v_0 - h_4^* = 0,$$

for all $\mu \in \bar{\Pi}(p,q)$, so this (h_1^*, h_2^*, h_3^*, h_4^*) is dual feasible. Moreover, with this choice of h_1^*, h_2^*, h_3^*, h_4^*, the dual objective function value in (E.5) is

$$D_\infty = \ln \int e^{a(x,z) + b(y) + \xi v((x,z),y) - c + \theta <x,y>}\, d\nu((x,z),y) = \ln \int e^{-c}\, d\mu^*(x,y) = -c.$$

The primal objective function value is

$$P_\infty = D(\mu^* | e^{\theta <x,y>} \nu) = \int a(x,z) + b(y) + \xi v((x,z),y)\, d\mu^*((x,z),y) = c.$$

It follows from Theorem 2.1 of Bhattacharya that this choice of $(h_1^*, h_2^*, h_3^*, h_4^*)$ is optimal for the dual problem (E.5), and strong duality holds $P_\infty = -D_\infty$. □

Next we establish a similar result for the discrete problem. Define $\bar{\Pi}^*_{\epsilon_N}(p_N, q_N)$ to be set of functions $h : (\mathbb{R}^d \times \mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ of the form

$$h((x,z),y) = h_1(x,z) + h_2(y) + h_3 v((x,z),y) - h_4$$

with

$$\int h((x,z),y) d\mu_N((x,z),y) \geq 0,$$

for all $\mu_N \in \bar{\Pi}_{\epsilon_N}(p_N, q_N)$.

Lemma E.3. *For the primal problem*

$$P_{N,\epsilon_N} = \min_{\mu \in \bar{\Pi}_{\epsilon_N}(p_N, q_N)} \int \left(\frac{d\mu((x,z),y)}{\exp(\theta < x, y >) d\nu_N((x,z),y)} \right) d\mu((x,z),y), \tag{E.7}$$

define the dual

$$D_{N,\epsilon_N} = \inf_{h \in \bar{\Pi}^*_{\epsilon_N}(p_N, q_N)} \ln \int e^{h((x,z),y) + \theta < x, y >} \, d\nu_N((x,z),y). \tag{E.8}$$

The following statements hold:

(i) The optimal solution to the primal problem takes the form

$$d\mu^*_N((x,z),y) = e^{a^N(x,z) + b_1^N(y) + b_2^N(y) + \xi_1^N v((x,z),y) + \xi_2^N v((x,z),y) + \theta < x, y >} \, d\nu_N((x,z),y),$$

where $b_1^N(y) \le 0$, $b_2^N(y) \ge 0$, $\xi_1^N \le 0$, $\xi_2^N \ge 0$.

(ii) A feasible solution to the dual is $\tilde{h}$,

$$\tilde{h}((x,z),y) = \tilde{h}_1(x,z) + \tilde{h}_2(y) + \tilde{h}_3 v((x,z),y) - \tilde{h}_4, \ \textit{where}$$

$$\tilde{h}_1(x,z) = a(x,z), \quad \tilde{h_2}(x) = b_1(x) + b_2(x), \quad \tilde{h_3} = \xi_1 + \xi_2,$$

$$\tilde{h_4} = \int a(x,z)\, dp_N(x,z) + \int (b_1(y) + b_2(y))\, dq_N(y) + (\xi_1 + \xi_2) v_0 + \sum_{j=1}^{d+1} (b_1(y_j) - b_2(y_j)) \epsilon_N + (\xi_1 - \xi_2) \epsilon_N$$

where $b_1(y) = b(y)^-$, $b_2(y) = b(y)^+$, *and* $\xi_1 = \xi^-$, $\xi_2 = \xi^+$, *for* $a(x,z)$, $b(x)$, ξ *as in (E.6).*

(iii) $D_\infty \ge \overline{\lim_{N \to \infty}} \, D_{N,\epsilon_N}$.

Proof: Conclusion (i) is the discrete form of part (i) in Lemma E.2. For (ii), we consider the dual problem

$$\inf_{h \in \bar{\Pi}^*_{\epsilon_N}(p_N, q_N)} \ln \int e^{h((x,z),y) + \theta < x, y >} \, d\nu_N((x,z),y).$$

Let $\tilde{h_1}(x,z) = a(x,z)$, $\tilde{h_2}(x) = b_1(x) + b_2(x)$, $\tilde{h_3} = \xi_1 + \xi_2$ and

$$\begin{aligned} \tilde{h_4} = \tilde{c} \equiv \quad & \int a(x,z)\, dp_N(x,z) + \int b_1(y) + b_2(y)\, dq_N(y) + (\xi_1 + \xi_2) v_0 \\ & + \sum_{j=1}^{d+1} (b_1(y_j) - b_2(y_j)) \epsilon_N + (\xi_1 - \xi_2) \epsilon_N \end{aligned}$$

where $b_1(y) = b(y)^-$, $b_2(y) = b(y)^+$, and $\xi_1 = \xi^-$, $\xi_2 = \xi^+$, for $a(x,z)$, $b(x)$, ξ as in (E.6).

Notice that

$$\sum_{j=1}^{d+1} (b_1(y_j) - b_2(y_j)) \epsilon_N + (\xi_1 - \xi_2) \epsilon_N \le 0.$$

For any $\mu_N \in \bar{\Pi}_{\epsilon_N}(p_N, q_N)$,

$$\begin{aligned}
\int \tilde{h}(x,y)d\mu_N((x,z),y) &= \int \tilde{h_1}(x,z)\,dp_N(x,z) + \int \big(\tilde{h_2}(y) + \tilde{h_3}v((x,z),y)\big)\,d\mu_N((x,z),y) - \tilde{h_4} \\
&\geq \int \tilde{h_1}(x,z)\,dp_N(x,z) + \int \tilde{h_2}(y)\,dq_N(y) + \tilde{h_3}v_0 - \tilde{h_4} \\
&\quad + \sum_{j=1}^{d+1}(b_1(y_j) - b_2(y_j))\epsilon_N + (\xi_1 - \xi_2)\epsilon_N \\
&= \int a(x,z)\,dp_N(x,z) + \int b_1(y) + b_2(y)\,dq_N(y) + (\xi_1 + \xi_2)v_0 \\
&\quad + \sum_{j=1}^{d+1}(b_1(y_j) - b_2(y_j))\epsilon_N + (\xi_1 - \xi_2)\epsilon_N - \tilde{h_4} \\
&= 0
\end{aligned}$$

so this ($\tilde{h_1}$, $\tilde{h_2}$, $\tilde{h_3}$, $\tilde{h_4}$) is feasible for the dual problem (E.8).

For (iii), let $\tilde{D}_{N,\epsilon_N}$ denote the objective value in (E.8) with solution $\tilde{h}$. Since $\tilde{h}$ is dual feasible,

$$\tilde{D}_{N,\epsilon_N} \geq D_{N,\epsilon_N}$$

We show that $\tilde{D}_{N,\epsilon_N} \to D_\infty$ as $N \to \infty$. Substituting $\tilde{h}$ in (E.8)

$$\begin{aligned}
\tilde{D}_{N,\epsilon_N} &= \ln \sum_{i=1}^{N}\sum_{j=1}^{d+1} \exp\big(a(X_i,Z_i) + b(y_j) + \xi v((X_i,Z_i),y_j) - \frac{1}{N}\sum_{i=1}^{N} a(X_i,Z_i) - \int b(y)\,dq_N(y) \\
&\qquad -\xi v_0 - \sum_{j=1}^{d+1}|b(y_j)|\epsilon_N - |\xi|\epsilon_N + \theta < X_i, y_j > \big) \cdot \frac{1}{N} \cdot q_N(y_j) \\
&= \ln \exp\big(-\frac{1}{N}\sum_{i=1}^{N} a(X_i,Z_i) - \int b(y)\,dq_N(y) - \xi v_0 - \sum_{j=1}^{d+1}|b(y_j)|\epsilon_N - |\xi|\epsilon_N\big) \\
&\quad + \ln \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d+1} \exp\big(a(X_i,Z_i) + b(y_j) + \xi v(X_i,y_j) + \theta < X_i, y_j > \big) q_N(y_j) \\
&= \big(-\frac{1}{N}\sum_{i=1}^{N} a(X_i,Z_i) - \frac{1}{N}\sum_{j=1}^{N} b(Y_j) - \xi v_0 - \sum_{j=1}^{d+1}|b(y_j)|\epsilon_N - |\xi|\epsilon_N\big) \\
&\quad + \ln\big(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \exp(a(X_i,Z_i) + b(Y_j) + \xi v((X_i,Z_i),Y_j) + \theta < X_i, Y_j >)\big) \\
&\to -\int a(x,z)\,dp(x,z) - \int b(y)\,dq(y) - \xi v_0 + \ln \int e^{a(x,z)+b(y)+\xi v((x,z),y)+\theta<x,y>}\,d\nu = D_\infty
\end{aligned}$$

so $D_\infty \geq \overline{\lim_{N\to\infty}} D_{N,\epsilon_N}$. □

Lemma E.4. $G_\infty \geq \limsup_N G_{N,\epsilon_N}$.

Proof: From strong duality of the continuous problem (E.4) and (E.5) in Lemma E.2, we have $G_\infty = -\frac{1}{\theta}P_\infty = \frac{1}{\theta}D_\infty$. By weak duality of the finite relaxed problem (E.7) and (E.8), we have $\frac{1}{\theta}D_{N,\epsilon_N} \geq -\frac{1}{\theta}P_{N,\epsilon_N} = G_{N,\epsilon_N}$. Therefore by Lemma E.3, we have

$$G_\infty = \frac{1}{\theta}D_\infty \geq \limsup_N \frac{1}{\theta}D_{N,\epsilon_N} \geq \limsup_N G_{N,\epsilon_N}.$$

□

Combining Lemma E.1 and Lemma E.4 proves part (i) of Theorem 6.2.

E.2 Weak Convergence of Optimal Solutions

The argument is similar to that of Section C.2. Define

$$\bar{\Pi}^N = \bar{\Pi}(p,q) \cup \Big(\bigcup_{n \geq N} \bar{\Pi}_{\epsilon_n}(p_n, q_n) \Big).$$

By the argument used in Lemma C.3, we have

Lemma E.5. $\bar{\Pi}^N$ *is compact for all sufficiently large N, a.s.*

The optimizers $\bar{\mu}^*_N$ are contained in the sets $\bar{\Pi}_{\epsilon_N}(p_N, q_N)$, so for all sufficiently large N, the sequence $\bar{\mu}^*_n$, $n \geq N$, is contained in a compact set $\bar{\Pi}^N$, and then every subsequence has a further subsequence that converges weakly.

Suppose the subsequence $\bar{\mu}^*_{n_k}$ converges, say $\bar{\mu}^*_{n_k} \Rightarrow \tilde{\mu}$. The marginals of $\bar{\mu}^*_{n_k}$ converge to p and q, and $\lim_{k\to\infty} \int v((x,z),y)\, d\bar{\mu}^*_{n_k} = v_0$, so $\tilde{\mu} \in \bar{\Pi}(p,q)$, making $\tilde{\mu}$ feasible for the limiting problem. We claim that it is optimal. We have, a.s.,

$$\int e^{\theta < h_x(x,z), y >}\, d\bar{\mu}^*_{n_k} \leq \int \sum_{j=1}^{d+1} e^{\theta < h_x(x,z), y_j >}\, dp_{n_k}(x,z) \to \int \sum_{j=1}^{d+1} e^{\theta < h_x(x,z), y_j >}\, dp(x,z),$$

by the strong law of large numbers, because the condition $\mathbb{E}_\nu[e^{\theta < h_x(x,z), y >}] < \infty$ implies that the limit is finite. This is then more than sufficient to ensure that

$$\int < h_x(x,z), y >\, d\bar{\mu}^*_{n_k}((x,z),y) \to \int < h_x(x,z), y >\, d\tilde{\mu}((x,z),y). \tag{E.9}$$

Moreover, relative entropy is lower semi-continuous with respect to weak convergence (Dupuis and Ellis [19], Lemma 1.4.3), so

$$D(\tilde{\mu}|\nu) \leq \varliminf_{k\to\infty} D(\bar{\mu}^*_{n_k}|\nu_{n_k})$$

and then

$$G(\tilde{\mu},\nu) \geq \overline{\lim_{k\to\infty}} G(\bar{\mu}^*_{n_k},\nu_{n_k}) = \sup_{\mu\in\bar{\Pi}(p,q)} G(\mu,\nu),$$

by part (i) of the theorem. Thus, $\tilde{\mu}$ is optimal. Using the equivalence between the optimization of $G(\cdot,\nu)$ and (E.4), we know from Theorem 2.1 of Csiszár [16] that the maximum is uniquely attained by some $\bar{\mu}^*$, and thus $\tilde{\mu} = \bar{\mu}^*$.

We have shown that every subsequence of $\bar{\mu}^*_n$ has a further subsequence that converges to $\bar{\mu}^*$. It follows that $\bar{\mu}^*_n \Rightarrow \bar{\mu}^*$. This proves part (ii) of the theorem. The uniform integrability needed for (6.5) follows as in (E.9), which proves part (iii).

References

[1] Basel Committee on Banking Supervision (2010) Results of the comprehensive quantitative impact study. Bank for International Settlements, Basel, Switzerland.

[2] Beiglböck, M., and Juillet, N. (2012) On a problem of optimal transport under marginal martingale constraints. ArXiv preprint arXiv:1208.1509.

[3] Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59(2), 341–357.

[4] Bernard, C., Jiang, X., and Wang, R. (2014). Risk aggregation with dependence uncertainty. *Insurance: Mathematics and Economics* 54, 93–108.

[5] Bhattacharya, B. (2006). An iterative procedure for general probability measures to obtain I-projections onto intersections of convex sets. *Annals of Statistics* 34(2), 878–902.

[6] Billingsley, P. (1968). *Convergence of probability measures*, Wiley, New York.

[7] Bosc, D., and Galichon, A. (2014) Extreme dependence for multivariate data. *Quantitative Finance* 14(7), 1187–1199.

[8] Brenier, Y. (1991) Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics* 44(4), 375–417.

[9] Brigo, D., Capponi, A., and Pallavicini, A. (2014). Arbitrage-free bilateral counterparty risk valuation under collateralization and application to credit default swaps. *Mathematical Finance* 24(1), 125–146.

[10] Brigo, D., Morini, M., and Pallavicini, A. (2013). *Counterparty Credit Risk, Collateral and Funding*, Wiley, Chichester, U.K.

[11] Brown, H., Hobson, D., and Rogers, L. C. (2001) Robust hedging of barrier options, *Mathematical Finance* 11(3), 285–314.

[12] Canabarro, E., and Duffie, D. (2003). Measuring and marking counterparty risk, in *Asset/Liability Management for Financial Institutions*, Institutional Investor Books.

[13] Carr, P., Ellis, K., and Gupta, V. (1998) Static hedging of exotic options. *Journal of Finance* 53(3), 1165–1190.

[14] Cox, A. (2010) Arbitrage bounds, *Encyclopedia of Quantitative Finance*, Wiley, New York.

[15] Crépey, S. (2012) Bilateral counterparty risk under funding constraints — part II: CVA. *Mathematical Finance.*

[16] Csiszár, I. (1975) I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability* 3, 146-158.

[17] Deming, W.E., and Stephan, F.F. (1940) On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics* 11, 427–444.

[18] Dolinsky, Y., and Soner, H. M. (2012) Martingale optimal transport and robust hedging in continuous time. *Probability Theory and Related Fields*, 1–37.

[19] Dupuis, P., and Ellis, R. S. (1997). *A weak convergence approach to the theory of large deviations.* Wiley, New York.

[20] Embrechts, P. and Puccetti, G. (2006) Bounds for functions of dependent risks. *Finance and Stochastics* 10, 341-352.

[21] Embrechts, P., Wang, B., and Wang, R. (2014) Aggregation-robustness and model uncertainty of regulatory risk measures. Working paper, ETH Zurich, Switzerland.

[22] Gregory, J. (2012). *Counterparty Credit Risk and Credit Value Adjustment: A Continuing Challenge for Global Financial Markets*, Wiley, Chichester, U.K.

[23] Hansen, L. P., and Sargent, T. J. (2008). *Robustness.* Princeton University Press, Princeton, New Jersey.

[24] Henry-Labordère, P., and Touzi, N. (2013) An explicit martingale version of Brenier's theorem. Available at SSRN: http://ssrn.com/abstract=2218488.

[25] Hull, J.C., and White, A. (2012) CVA and wrong way risk. Working paper, Rotman School of Management, University of Toronto, Toronto, Canada.

[26] Ireland, C. T., and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika* 55(1), 179–188.

[27] Kleinschmidt, P., and Schannath, H. (1995). A strongly polynomial algorithm for the transportation problem, *Mathematical Programming* 68, 1-13.

[28] Luenberger, D. (1969). *Optimization by Vector Space Methods*, Wiley, New York.

[29] Pukelsheim, F. (2014). Biproportional scaling of matrices and the iterative proportional fitting procedure. *Annals of Operations Research* 215(1), 269–283.

[30] Rosen, D., and Saunders, D. (2012). CVA the wrong way. *Journal of Risk Management in Financial Institutions* 5(3), 252–272.

[31] Rüschendorf, L. (1995). Convergence of the iterative proportional fitting procedure. *Annals of Statistics* 1160-1174.

[32] Rüschendorf, L. (2013) *Mathematical Risk Analysis*, Springer-Verlag, Berlin.

[33] Rüschendorf, L., and Rachev, S. T. (1990) A characterization of random variables with minimum L^2-distance. *Journal of Multivariate Analysis* 32(1), 48–54.

[34] Rüschendorf, L., and Thomsen, W. (1993). Note on the Schrödinger equation and I-projections. *Statistics and Probability Letters* 17(5), 369–375.

[35] Tankov, P. (2011) Improved Fréchet bounds and model-free pricing of multi-asset options. *Journal of Applied Probability* 48, 389–403.

[36] Villani, C. (2008). *Optimal Transport: Old and New*, Berlin, Springer.