Discussion: sampling vs. quantile approach. We describe the numerical penalization method being currently used, with emphasis on the sampling of points for penalization and discuss an alternative option to consider. To approximate a solution of equation 2.6 with constraint 2.7, we want to minimize

$$\sum_i \int \phi_i \mu_i + \sum_i \int \psi_i d\nu_i + \int b_\gamma \left(c - \varphi\right) d\theta$$

where b_γ is a penalization function defined as

$$b_\gamma \left(t\right) = \frac{1}{\gamma} \frac{1}{2} \left[\left(\gamma t\right)^+\right]^2$$

Computationally, each function ϕ_i, ψ_i, h_i is implemented as a neural network that takes any floating-point number (ϕ_i, ψ_i) or vector (h_i) as input. The sum of integrals above becomes a discrete sum (to be detailed later), and is referenced to a gradient descent algorithm as a loss function (that is, the one to be minimized numerically). The proposed alternative is to work on the quantile domain. Let F_i, G_i be the cumulative distribution functions of x_i, y_i (assumed to be known) and define

$$\begin{aligned}
\hat{x}_i &= F\left(x_i\right) \\
\hat{y}_i &= G\left(y_i\right) \\
\hat{\phi}_i\left(\hat{y}_i\right) &= \phi_i\left(y_i\right) \\
\hat{\psi}_i\left(\hat{y}_i\right) &= \psi_i\left(y_i\right) \\
\hat{h}_i\left(\hat{x}\right) &= h_i\left(x\right)
\end{aligned}$$

Now each $\hat{\phi}_i, \hat{\psi}_i, \hat{h}_i$ is implemented by a neural network instead of their primitives. Instead of random sampling, we use a grid of quantiles. Our marginals are uniform discrete probabilities depending on the choice of grid size. The sampling measure $\hat{\theta}$ is a discrete coupling of uniform discrete marginals. Some advantages of this "quantile" approach are

- It spreads the area covered by the training evenly. This is critical in the high dimension case where much of the area can be left uncovered with random sampling.
- The neural network will potentially perform better in the interval $[0, 1]$ than in the line. A good reason to believe that is the tail noise observed in the normal-marginals example. To be checked.
- Potential to find ways to update of the sampling measure $\hat{\theta}$.

Approximations to the penalization integral: description and comparison. We approximate the penalty term $\varepsilon = \int b_\gamma \left(c - \varphi\right) d\theta$ with a sum, as described below.

1. Sampling. Let $\mathcal{X}_i = \left\{x_i^1, \ldots x_i^n\right\}, \mathcal{Y}_i = \left\{y_i^1, \ldots y_i^n\right\}$ be (simulated) samples, $i = 1, \ldots, d$. Note 1: the marginal probabilities are expressed in the samples; their construction is the only place where the marginal probabilities are ever considered. Note 2: In the empirical example, even though we are using actual market data, the method first builds marginal distributions of random, future prices and then generates a sample from those distributions. Hence, the final sample is simulated in any case.

   - Current method.
     First step is to build a common sample Θ_1 of $2d$-dimension points. In the independent case, we wrap the points arbitrarily as (for instance)

     $$\Theta_1 = \left\{\left(x_1^1, \ldots, x_d^1, y_1^1, \ldots, y_d^1\right), \ldots, \left(x_1^n, \ldots, x_d^n, y_1^n, \ldots, y_d^n\right)\right\}$$

     In the monotone coupling case, the x_i^k's are first ordered as $x_i^{(1)}, \ldots, x_i^{(n)}$ and then coupled in order, while the y_i^k's are coupled arbitrarily, like

     $$\Theta_1 = \left\{\left(x_1^{(1)}, \ldots, x_d^{(1)}, y_1^1, \ldots, y_d^1\right), \ldots, \left(x_1^{(n)}, \ldots, x_d^{(n)}, y_1^n, \ldots, y_d^n\right)\right\}$$

     The penalty is given by

     $$\varepsilon_1 = \frac{1}{|\Theta_1|} \sum_{(x,y)\in\Theta_1} b_\gamma \left(c\left(x,y\right) - \varphi\left(x,y\right)\right)$$

   - Separated sampling. The common sample Θ_2 is the product of the marginal samples. In the independent coupling case, the penalty is calculated as

     $$\varepsilon_2 = \frac{1}{\prod_i \prod_j |\mathcal{X}_i| \, |Y_i|} \sum_{x_1\in\mathcal{X}_1} \cdots \sum_{x_d\in\mathcal{X}_d} \sum_{y_1\in\mathcal{Y}_1} \cdots \sum_{y_d\in\mathcal{Y}_d} b_\gamma \left(c\left(x,y\right) - \varphi\left(x,y\right)\right)$$

     In the monotone case, the choice of x can be seen as a function of x_1, therefore we can write

     $$\varepsilon_2 = \frac{1}{|\mathcal{X}_1| \prod_j |Y_i|} \sum_{x_1\in\mathcal{X}_1} \sum_{y_1\in\mathcal{Y}_1} \cdots \sum_{y_d\in\mathcal{Y}_d} b_\gamma \left(c\left(x\left(x_1\right), y\right) - \varphi\left(x\left(x_1\right), y\right)\right)$$

     To keep the complexity comparable, we must have either $\prod_i \prod_j |\mathcal{X}_i| \, |Y_i| = n^{2d}$ (in the independent case) or $|\mathcal{X}_1| \prod_j |Y_i| = n^{d+1}$ (in the monotone coupling case) in the same scale as $|\Theta_1|$ in the previous method. Thus, suppose we used $|\Theta_1| = 1M$ there. Here, in the independent case, we should use a comparable $n = (1M)^{1/4} \sim 32$ for $d = 2$ and $n = (1M)^{1/6} = 10$ for $d = 3$. In the monotone coupling case, we should use $n = (1M)^{1/3} = 100$ for $d = 2$ and $n = (1M)^{1/4} \sim 32$ for $d = 3$.

Notice that, provided the sample sets have the same size, there is no difference between the two methods just discussed in terms of computational effort (other than possibly memory allocation). But in the separated sampling case, given that marginal sizes are small specially in higher dimension, we may have some areas of the marginal distributions being scarcely covered throughout Θ.

2. Quantile method. The points $(\hat{x}, \hat{y})$ are taken from a grid in the 1-hypercube. So, for example, if the grid granularity is 10, we divide each marginal of the cube into 10 intervals, and pick the middle points

$$\hat{x}_i \in \hat{\mathcal{X}}_i = \{0.05, \ldots, 0.95\}$$
$$\hat{y}_i \in \hat{\mathcal{Y}}_i = \{0.05, \ldots, 0.95\}$$

The points $(\hat{x}, \hat{y})$ are attributed the discrete probability $\hat{\theta}$ with uniform marginals. The penalty is given by

$$\varepsilon_3 = \frac{1}{\prod_i \prod_j \left|\hat{\mathcal{X}}_i\right| \left|\hat{Y}_i\right|} \sum_{\hat{x}_1 \in \hat{\mathcal{X}}_1} \cdots \sum_{\hat{x}_d \in \hat{\mathcal{X}}_d} \sum_{\hat{y}_1 \in \hat{\mathcal{Y}}_1} \cdots \sum_{\hat{y}_d \in \hat{\mathcal{Y}}_d} \hat{\theta}\left(\hat{x}, \hat{y}\right) b_\gamma \left(c\left(x, y\right) - \varphi\left(x, y\right)\right)$$

Notice that the concept of sampling measure has entered the formula through the probability $\hat{\theta}$, while previously it only influenced the construction of the sample Θ. In the simplest, independent case, $\hat{\theta}$ is the independent coupling, which is the uniform probability $\hat{\theta}\left(\hat{x}, \hat{y}\right) = \frac{1}{n^{2d}}$. In the positive monotone coupling case, only points satisfying $\hat{x}_i = \hat{x}_j \forall i, j$ have positive probability. So, if the y-side of $\hat{\theta}$ is independent, we set

$$\hat{\theta}\left(\hat{x}, \hat{y}\right) = \begin{cases} \frac{1}{n^{d+1}} & \hat{x}_i = \hat{x}_j \forall i, j \\ 0 & \text{otherwise} \end{cases}$$

There could be interesting cases where additional information is available, giving rise to a more elaborated and efficient $\hat{\theta}$. One topic to investigate is whether it is possible to update $\hat{\theta}$ with information from $\hat{\varphi}$ in the same fashion as in formula 2.6 of Eckstein and Kupper (2021).

Note: There might be some literature about similar quantile-based methods. In their brief introduction to Example 4.6, Eckstein and Kupper (2021) mention the following papers, not yet analyzed.

- Embrechts, P., Puccetti, G., Rüschendorf, L. "Model uncertainty and VaR aggregation." (2013, J. Bank. Financ.)
- Puccetti, G., Rüschendorf, L. "Computation of sharp bounds on the distribution of a function of dependent risks." (J. Comput. Appl. Math., 2012)