A scalable algorithm for sparse portfolio selection

Bertsimas, Dimitris, and Ryan Cory-Wright. "A scalable algorithm for sparse portfolio selection." INFORMS Journal on Computing 34.3 (2022): 1489-1511.

UESTC

Background and Motivation

Since the Nobel-prize-winning work of Markowitz (1952), the mean–variance model

$$\min_{x \in \mathbb{R}^n_+} \ \frac{\sigma}{2} x^\top \Sigma x \ - \ \mu^\top x \quad \text{s.t. } \mathbf{e}^\top x = 1 \tag{1}$$

has become the foundation of modern portfolio theory.

- x = portfolio weights (long-only);
- $x^\top \Sigma x$ = portfolio variance (risk);
- $-\mu^\top x$ = negative expected return (to be minimized);
- $\sigma > 0$ = risk–return trade-off parameter.

However, real-world portfolio construction involves many additional requirements:

- sector and industry allocation bounds;
- factor exposure and style constraints (value, momentum, beta, etc.);
- cardinality constraints limiting the number of active positions.

These constraints interact in nontrivial ways, causing Markowitz portfolios—which are typically dense—to become impractical. This gap motivates scalable methods for solving realistic sparse portfolio selection problems.

UESTC

Realistic Portfolio Constraints

In real-world portfolio management, the classical Markowitz model is insufficient: it produces dense portfolios and ignores practical implementability constraints.

Bienstock (1996) formalized two major classes of constraints required in practice:

(i) Linear exposure constraints

$$l \leq Ax \leq u, \quad A \in \mathbb{R}^{m \times n},\ l, u \in \mathbb{R}^m,$$

capturing industry allocations, factor exposures, regulatory limits, and risk budgets.

(ii) Cardinality constraint

$$\|x\|_0 \leq k \ll n,$$

restricting the number of active holdings to improve interpretability, reduce monitoring/transaction costs, and avoid portfolios with many tiny positions.

These constraints greatly increase the structural complexity of the problem and challenge classical convex optimization approaches.

UESTC

Realistic Sparse Portfolio Model

Combining Markowitz's risk–return tradeoff with realistic exposure and sparsity constraints yields the mixed-integer nonlinear program:

$$\begin{aligned} \min_{x \in \mathbb{R}^n_+} \quad & \frac{\sigma}{2} x^\top \Sigma x - \mu^\top x \\ \text{s.t.} \quad & l \le Ax \le u, \\ & \mathbf{e}^\top x = 1, \\ & \|x\|_0 \le k. \end{aligned} \tag{2}$$

- The sparsity constraint makes the feasible set combinatorial.
- Selecting which k assets to hold is itself NP-hard.
- Even when A is absent and Σ is diagonal, the problem remains NP-hard.

UESTC

Introduce Binary Variables z_i

Represent the sparsity constraint

$$\|x\|_0 \le k$$

in a way that optimization solvers can handle.

1. ℓ_0 is nonconvex and non-differentiable
   - $\|x\|_0$ counts the number of nonzero positions.
   - Optimization algorithms cannot directly handle this combinatorial constraint.
2. Sparsity is inherently a "selection" problem

$$z_i = \begin{cases} 1, & \text{asset } i \text{ is selected,} \\ 0, & x_i = 0. \end{cases}$$

Binary variables explicitly encode which assets are active.

3. Converts ℓ_0 into linear constraints

$$x_i = 0 \text{ if } z_i = 0, \qquad \sum_{i=1}^{n} z_i = k.$$

This turns sparsity into a standard mixed-integer structure.

4. Enables scalable algorithms
   - x handles continuous optimization (a convex QP).
   - z handles combinatorial asset selection.

UESTC

Realistic Sparse Portfolio Model

Introduce binary variables z_i indicating whether asset i is active:

$$x_i = 0 \ \text{ if } \ z_i = 0.$$

The model becomes:

$$\begin{aligned} \min_{z\in\{0,1\}^n,\ x\in\mathbb{R}^n_+} \quad & \frac{\sigma}{2} x^\top \Sigma x - \mu^\top x \\ \text{s.t.} \quad & l \le Ax \le u, \\ & \mathbf{e}^\top x = 1, \\ & x_i = 0 \ \text{ if } \ z_i = 0, \ \forall i, \\ & \sum_{i=1}^{n} z_i = k. \end{aligned} \tag{3}$$

This MIQP is exact, but existing methods fail to scale to practical problem sizes (typically $500 \le n \le 3{,}200$).

UESTC

Solve the sparse MIQP

$$\min_{x,z} \frac{\sigma}{2} x^\top \Sigma x - \mu^\top x \quad \text{s.t. } x_i = 0 \text{ if } z_i = 0, \ \sum z_i = k, \ z_i \in \{0,1\}.$$

Big-M reformulation:

$$0 \le x_i \le M z_i.$$

Branch-and-Bound procedure:

1. Relax $z \in \{0,1\}^n$ to $0 \le z_i \le 1$. Solve the continuous QP $\rightarrow$ gives a lower bound.
2. If the relaxation yields integer z, update the current upper bound.
3. Otherwise, pick a fractional z_j and branch:

$$z_j = 0 \quad \text{or} \quad z_j = 1.$$

4. Prune nodes whose relaxation value $\ge$ current upper bound.

UESTC

Big-M Branch-and-Bound

Weakness of big-M relaxation:

$$0 \le x_i \le M z_i, \quad 0 \le z_i \le 1.$$

When M is large, fractional z_i allow unrealistically large x_i, giving:

$$\text{relaxation objective} \ll \text{true integer objective}.$$

Consequences:

- Relaxation lower bounds are far too optimistic.
- Very few nodes can be pruned.
- B&B tree grows exponentially.

Performance in literature:

- Bienstock (1996): solved up to ~ 50 assets.
- Bertsimas–Shioda (2009): ~ 200 assets.
- Gao & Li (2013): improved relaxations up to ~ 300 assets.

Conclusion: Big-M B&B is too weak to solve realistic $n \ge 500$ sparse portfolios.

UESTC

Outer Approximation

Outer Approximation (OA) is a classical algorithm for solving mixed-integer nonlinear optimization (MINLO) problems of the form:

$$\min_{x \in X,\ y \in Y} f(x, y) \quad \text{s.t.} \quad g(x, y) \leq 0,$$

where

- x are continuous variables,
- y are integer or binary variables,
- for fixed y, the problem in x is a convex optimization problem.

Key Idea of OA:

1. Fix an integer vector y_t and solve the continuous subproblem to obtain $(x_t, f(y_t))$.
2. Compute a subgradient (or KKT multiplier) to construct a first-order linear underestimator:

$$f(y) \ \geq \ f(y_t) + g_t^\top (y - y_t).$$

3. Add this linear inequality (a cut) to a master mixed-integer linear program (MILP).
4. Solve the MILP to obtain a new integer point y_{t+1}.
5. Repeat until the MILP lower bound matches the nonlinear upper bound.

OA alternates between:

(1) Solving convex NLP subproblems in x, (2) Solving MILPs in y.

UESTC

Ridge-Regularized Sparse Portfolio Model

Motivation for Regularization. The MIQP formulation of sparse portfolio selection (Problem (3)) is exact, but remains computationally intractable for realistic dimensions ($500 \leq n \leq 3{,}200$). Two fundamental difficulties arise:

- the cardinality constraint makes the feasible set combinatorial;
- the quadratic objective is not strongly convex, leading to weak relaxations and slow convergence.

To overcome these barriers, the paper introduces a ridge regularization term:

$$\frac{1}{2\gamma}\|x\|^2, \gamma > 0,$$

resulting in the regularized sparse portfolio problem:

$$\begin{aligned} \min_{x \in \mathbb{R}^n_+} \quad & \frac{\sigma}{2} x^\top \Sigma x + \frac{1}{2\gamma}\|x\|^2 - \mu^\top x \\ \text{s.t.} \quad & l \leq Ax \leq u, \mathbf{e}^\top x = 1, \|x\|_0 \leq k. \end{aligned} \tag{4}$$

- It makes the objective strongly convex, improving conditioning and stability.
- It produces much tighter conic relaxations; the duality gap shrinks as $\gamma \to 0$.
- For sufficiently small γ, Problem (4) recovers optimal supports of the original Problem (2).

UESTC

Ridge-Regularized Sparse Portfolio Selection Problem Is Hard

Ridge-Regularized Sparse Portfolio Selection

$$\begin{aligned} \min_{x \in \mathbb{R}^n_+} \quad & \frac{\sigma}{2} x^\top \Sigma x + \frac{1}{2\gamma} \|x\|^2 - \mu^\top x \\ \text{s.t.} \quad & l \le Ax \le u, \;\; \mathbf{e}^\top x = 1, \;\; \|x\|_0 \le k. \end{aligned} \tag{4}$$

Ridge-Regularized Sparse Portfolio Selection Problem Is Hard is difficult:

- The ℓ_0 constraint makes the feasible set combinatorial.
- For each support choice, solving the continuous subproblem requires a strongly convex QP with additional linear exposure constraints.
- Exhaustive enumeration of supports is computationally impossible for realistic sizes ($500 \le n \le 3{,}200$).
- Standard MIQP methods (branch-and-bound, decomposition) do not scale to this dimension.

Key idea of the paper: Rewrite Problem (4) as a minimization over the support vector z,

$$\min_{z \in Z^n_k} f(z),$$

where $f(z)$ is defined by solving a strongly convex QP in x.
However, $f(z)$ is not explicitly available. Theorem 1 shows that $f(z)$ admits a saddle-point dual representation:

$$f(z) = \max_{\alpha, w, \beta_l, \beta_u, \lambda} \Big[-\tfrac{1}{2} \alpha^\top \alpha - \tfrac{\gamma}{2} \sum_i z_i w_i^2 + y^\top \alpha + \beta_l^\top l - \beta_u^\top u + \lambda \Big]$$

subject to linear constraints.

- Theorem 1 proves that $f(z)$ is convex in the binary variable z.
- This enables the use of Outer Approximation (OA), which iteratively builds linear underestimators of $f(z)$ and solves a MILP master problem.
- Algorithm 1 applies OA to obtain a scalable method that reaches certifiable optimality.

UESTC

Theorem Derivation: Reformulating Problem (4)

We begin with the ridge-regularized sparse portfolio problem:

$$\min_{x \geq 0} \ \frac{\sigma}{2} x^\top \Sigma x + \frac{1}{2\gamma}\|x\|^2 - \mu^\top x, \qquad \text{s.t. } l \leq Ax \leq u,\ e^\top x = 1,\ \|x\|_0 \leq k.$$

Scaling the objective. Multiplying the entire objective by any positive constant does not change the minimizer. Hence, we rescale the risk term and work with the equivalent form

$$\frac{1}{2} x^\top \Sigma x.$$

Decomposition of the covariance matrix. Since $\Sigma \succeq 0$ with rank r, there exists a matrix $X \in \mathbb{R}^{r \times n}$ such that

$$\Sigma = X^\top X.$$

This allows us to rewrite the quadratic term as

$$\frac{1}{2} x^\top \Sigma x = \frac{1}{2}\|Xx\|^2,$$

revealing that the Markowitz risk is the squared norm of the transformed weights.

Resulting objective. After substitution, the objective becomes a sum of two squared norms and a linear term:

$$\frac{1}{2\gamma}\|x\|^2 \ + \ \frac{1}{2}\|Xx\|^2 \ - \ \mu^\top x.$$

UESTC

Theorem Derivation: Matching the Linear Term

Our goal is to rewrite the quadratic–linear expression

$$\frac{1}{2}\|Xx\|^2 - \mu^\top x$$

in the regression-like form

$$\frac{1}{2}\|Xx - y\|^2 + d^\top x + C,$$

which will later allow us to complete the square and derive the dual representation of $f(z)$.
Step 1: Expand the square.

$$\frac{1}{2}\|Xx - y\|^2 = \frac{1}{2}\|Xx\|^2 - y^\top Xx + \frac{1}{2}\|y\|^2.$$

Step 2: Match the linear term in x. We want the term involving x to satisfy

$$-y^\top Xx = -(X^\top y)^\top x \stackrel{!}{=} -\mu^\top x.$$

Thus we require

$$X^\top y = \mu. \tag{6}$$

Step 3: Solve for the vector y. Since $\Sigma = X^\top X \succeq 0$ has rank r, the matrix $XX^\top$ is invertible on its range. Hence the equation $X^\top y = \mu$ admits the minimum-norm solution

$$y := (XX^\top)^{-1} X\mu.$$

This choice of y is precisely the orthogonal projection of the return vector μ onto the subspace range$(X^\top)$, ensuring that the rewritten objective is algebraically equivalent to the original one.

Theorem Derivation: Completing the Linear Term

Decomposing μ into a projection and an orthogonal remainder.
From the identity

$$X^\top y = X^\top (XX^\top)^{-1} X\mu,$$

we express μ as

$$\mu = X^\top (XX^\top)^{-1} X\mu + \Big(I - X^\top (XX^\top)^{-1} X\Big)\mu = X^\top y + d,$$

where the residual component is defined by

$$\boxed{d := (X^\top (XX^\top)^{-1} X - I)\mu.} \tag{7}$$

Thus the linear term satisfies

$$-\mu^\top x = -(X^\top y)^\top x - d^\top x.$$

Substitute into the expanded square. Using

$$\frac{1}{2}\|Xx - y\|^2 = \frac{1}{2}\|Xx\|^2 - y^\top Xx + \frac{1}{2}\|y\|^2,$$

we obtain

$$\frac{1}{2}\|Xx\|^2 - \mu^\top x = \frac{1}{2}\|Xx - y\|^2 + d^\top x - \frac{1}{2}\|y\|^2.$$

The constant term $\frac{1}{2}\|y\|^2$ can be dropped since it does not affect the minimizer.

Regression-equivalent formulation. The objective of Problem (4) can now be written as

$$\boxed{\begin{array}{ll} \min\limits_{x \geq 0} & \dfrac{1}{2\gamma}\|x\|^2 + \dfrac{1}{2}\|Xx - y\|^2 + d^\top x \\ \text{s.t.} & l \leq Ax \leq u,\ e^\top x = 1,\ \|x\|_0 \leq k. \end{array}} \tag{8}$$

This transforms the risk–return objective into a regularized regression loss, which is crucial for deriving the dual

Theorem Derivation: Introducing Support Variables

Based on Problem (8):

$$\begin{aligned} \min_{x\geq 0} \quad & \frac{1}{2\gamma}\|x\|^2 + \frac{1}{2}\|Xx - y\|^2 + d^\top x \\ \text{s.t.} \quad & l \leq Ax \leq u, \quad e^\top x = 1, \quad \|x\|_0 \leq k. \end{aligned} \tag{8}$$

Replace the nonconvex cardinality constraint

$$\|x\|_0 \leq k$$

with an explicit support-selection variable $z \in \{0,1\}^n$.

Define the diagonal matrix

$$Z := \mathrm{Diag}(z_1, \ldots, z_n), \qquad (Zx)_i = z_i x_i.$$

Key principle: sparsity as selection.

$$z_i = 0 \implies x_i = 0.$$

Therefore, all expressions involving x must be rewritten in terms of Zx so that the support encoded by z is enforced automatically.
This transforms Problem (8) into an optimization problem over the support pattern z and the continuous weights x, preparing the ground for dual reformulation and outer-approximation.

UESTC

Theorem Derivation: Reformulation with Zx

Apply the substitution $x \mapsto Zx$ to each component of (8):

- Regression term:
$$\frac{1}{2}\|Xx - y\|^2 \implies \frac{1}{2}\|X(Zx) - y\|^2.$$
- Linear term:
$$d^\top x \implies d^\top (Zx).$$
- Linear constraints:
$$l \le Ax \le u \implies l \le AZx \le u,$$
$$e^\top x = 1 \implies e^\top Zx = 1, \qquad Zx \ge 0.$$
- Ridge regularization:
$$\frac{1}{2\gamma}\|x\|^2$$
is left unchanged. This term is strongly convex and ensures that when $z_i = 0$, the optimizer will set $x_i = 0$ without explicitly applying Zx to it.

Thus, for a given support vector z, the inner optimization becomes

$$\boxed{f(z) = \min_{x \in \mathbb{R}^n} \left\{ \frac{1}{2\gamma} x^\top x + \frac{1}{2}\|XZx - y\|^2 + d^\top Zx \right\}}$$

$$\text{s.t. } l \le AZx \le u,\ e^\top Zx = 1,\ Zx \ge 0.$$

The overall sparse portfolio problem is now written as a two-level optimization:

$$\boxed{\min_{z \in Z_k^n} f(z),}$$

an exact reformulation of Problem (8), and a crucial starting point for dualization and the outer-approximation algorithm used later.

UESTC

Theorem Derivation: Outer and Inner Problems

Outer problem: We reformulate sparse portfolio selection as an optimization over the support vector:

$$\min_{z \in Z_k^n} f(z), \tag{9}$$

where

$$Z_k^n := \{\, z \in \{0,1\}^n : e^\top z = k \,\}$$

represents all binary vectors selecting exactly k assets.
Inner problem: definition of $f(z)$. For a fixed support pattern z, let $Z = Diag(z)$. The continuous optimization over portfolio weights becomes:

$$f(z) := \min_{x \in \mathbb{R}^n} \left[\frac{1}{2\gamma} x^\top x + \frac{1}{2} \|XZx - y\|^2 + d^\top Zx \right] s.t., l \le AZx \le u, e^\top Zx = 1, Zx \ge 0.$$

- z encodes the sparsity (which assets may be held).
- Zx enforces that inactive assets must have zero weight.
- For fixed z, the inner problem is a strongly convex quadratic program in x, ensuring a unique minimizer and enabling dual reformulation.

UESTC

Theorem Derivation: Introducing an Auxiliary Variable

To prepare for dualization, we rewrite the quadratic regression term

$$\frac{1}{2}\|XZx - y\|^2$$

in a form that separates linear constraints from the squared norm.

Introduce an auxiliary variable

$$r := y - XZx \in \mathbb{R}^r,$$

so that

$$XZx - y = -r, \|XZx - y\|^2 = \|r\|^2.$$

This allows the quadratic expression to be replaced by a linear constraint and a norm term, a standard approach for obtaining a clean Lagrangian and dual formulation.

Inner problem for fixed z. With this substitution, the definition of $f(z)$ becomes

$$\min_{x\in\mathbb{R}^n,\ r\in\mathbb{R}^r} \quad \frac{1}{2\gamma}\|x\|^2 + \frac{1}{2}\|r\|^2 + d^\top Zx$$

$$\text{s.t.} y - XZx = r, AZx \geq l, \quad AZx \leq u, e^\top Zx = 1, \quad Zx \geq 0.$$

Putting the outer problem back in. We obtain the joint optimization

$$\boxed{\min_{z\in Z_k^n,\ x\in\mathbb{R}^n,\ r\in\mathbb{R}^r} \left(\frac{1}{2\gamma}\|x\|^2 + \frac{1}{2}\|r\|^2 + d^\top Zx\right)}$$

subject to

$$y - XZx = r\ [\alpha], \quad AZx \geq l\ [\beta_l], \quad AZx \leq u\ [\beta_u],$$
$$e^\top Zx = 1\ [\lambda], \qquad Zx \geq 0\ [\pi].$$

(13)

Theorem Derivation: Lagrangian of the Inner Problem

Fix a feasible support vector $z \in Z_k^n$ and let $Z = \mathrm{Diag}(z)$. The Lagrangian associated with the inner problem in (13) is

$$\begin{aligned}\mathcal{L}(x, r; \alpha, \beta_l, \beta_u, \lambda, \pi) = \frac{1}{2\gamma} x^\top x + \frac{1}{2} r^\top r + d^\top Z x \\ + \alpha^\top (y - XZx - r) \\ - \beta_l^\top (AZx - l) + \beta_u^\top (AZx - u) \\ - \lambda(e^\top Z x - 1) - \pi^\top Z x,\end{aligned}$$

where the dual variables satisfy

$$\alpha \in \mathbb{R}^r, \qquad \beta_l, \beta_u \in \mathbb{R}_+^m, \qquad \lambda \in \mathbb{R}, \qquad \pi \in \mathbb{R}_+^n.$$

Primal–dual structure.

- For a fixed support z, the inner optimization defining $f(z)$ is

$$f(z) = \min_{x,r} \max_{\alpha,\beta_l,\beta_u,\lambda,\pi} \mathcal{L}(x, r; \alpha, \beta_l, \beta_u, \lambda, \pi).$$

- The objective is strongly convex in (x, r) and all constraints are affine; hence the primal is a strictly convex feasible QP.
- By standard convex duality theory, strong duality holds whenever the primal is feasible. Therefore, we may exchange the order of minimization and maximization:

$$\boxed{f(z) = \max_{\alpha,\beta_l,\beta_u,\lambda,\pi} \min_{x,r} \mathcal{L}(x, r; \alpha, \beta_l, \beta_u, \lambda, \pi).}$$

- This dual representation is the key step that leads to the saddle-point structure in Theorem 1.

UESTC

Theorem Derivation: Stationarity Conditions

To evaluate

$$\min_{x,r} \mathcal{L}(x, r; \alpha, \beta_l, \beta_u, \lambda, \pi)$$

for fixed dual variables, we impose first-order optimality with respect to (x, r).

Derivative with respect to r.

$$\frac{\partial \mathcal{L}}{\partial r} = r - \alpha = 0 \quad \Longrightarrow \quad \boxed{r = \alpha}.$$

Derivative with respect to x.
Collect all terms in $\mathcal{L}$ that depend on x:

$$\frac{1}{2\gamma} x^\top x + d^\top Z x - \alpha^\top X Z x - \beta_l^\top A Z x + \beta_u^\top A Z x - \lambda\, e^\top Z x - \pi^\top Z x.$$

Taking the gradient gives

$$\nabla_x \mathcal{L} = \frac{1}{\gamma} x + Z\Big(d - X^\top \alpha - A^\top(\beta_l - \beta_u) - \lambda e - \pi\Big) = 0.$$

Define the vector

$$g := d - X^\top \alpha - A^\top(\beta_l - \beta_u) - \lambda e - \pi.$$

Then the stationarity condition becomes

$$x = -\gamma\, Z g.$$

Introducing the reparameterization

$$w := -g$$

yields the compact expression

$$\boxed{x = \gamma\, Z w.}$$

This identity is crucial: it generates the dual quadratic term $-\frac{\gamma}{2}\sum_i z_i w_i^2$, and transfers the primal nonnegativity constraint $Zx \geq 0$ into a simple sign condition on w.

UESTC

Theorem Derivation: Substituting Stationarity Conditions

We substitute the stationarity conditions

$$r = \alpha, \qquad x = \gamma Z w$$

into the Lagrangian to obtain the dual function for a fixed support z.

1. Terms involving r.

$$\frac{1}{2} r^\top r - \alpha^\top r = \frac{1}{2}\|\alpha\|^2 - \alpha^\top \alpha = -\frac{1}{2}\|\alpha\|^2.$$

The term $\alpha^\top y$ remains in the objective.

2. Terms involving x.
The ridge term becomes

$$\frac{1}{2\gamma}\|x\|^2 = \frac{1}{2\gamma}\|\gamma Z w\|^2 = \frac{\gamma}{2} w^\top Z^2 w = \frac{\gamma}{2}\sum_i z_i w_i^2.$$

All linear terms involving x

$$d^\top Z x - \alpha^\top X Z x - \beta_l^\top A Z x + \beta_u^\top A Z x - \lambda e^\top Z x - \pi^\top Z x$$

combine with the ridge term. Using the definition of w, these contributions simplify to

$$-\frac{\gamma}{2}\sum_i z_i w_i^2.$$

This is the origin of the key term appearing in Theorem 1.

3. Constraint-dependent terms. The remaining terms independent of x, r are

$$\beta_l^\top l - \beta_u^\top u + \lambda.$$

The multiplier π associated with the nonnegativity constraint $Zx \geq 0$ can be eliminated, yielding the dual feasibility condition

$$w \;\geq\; X^\top \alpha + A^\top(\beta_l - \beta_u) + \lambda e - d \qquad \text{(componentwise)}.$$

4. Dual representation of $f(z)$. Thus, for any fixed support vector z,

Theorem Derivation: Final Saddle-Point Formulation

Recall that in Problem (9), the support vector is restricted to

$$Z_k^n = \{\, z \in \{0,1\}^n : e^\top z = k \,\}.$$

Hence every feasible z satisfies

$$z_i^2 = z_i \qquad \forall i,$$

and therefore

$$-\frac{\gamma}{2}\sum_i z_i^2 w_i^2 = -\frac{\gamma}{2}\sum_i z_i w_i^2.$$

Substituting this into the dual expression of $f(z)$, we obtain

$$f(z) = \max_{\alpha, w, \beta_l, \beta_u, \lambda} \left\{ -\tfrac{1}{2}\alpha^\top \alpha - \tfrac{\gamma}{2}\sum_i z_i w_i^2 + y^\top \alpha + \beta_l^\top l - \beta_u^\top u + \lambda \right\}$$

subject to the dual-feasibility constraint

$$w \;\geq\; X^\top \alpha + A^\top(\beta_l - \beta_u) + \lambda e - d.$$

By strong duality of the inner convex problem, this dual optimal value equals the primal value $f(z)$. Thus Problem (9),

$$\min_{z \in Z_k^n} f(z),$$

is equivalently written as the saddle-point problem

$$\boxed{\min_{z \in Z_k^n} \max_{\alpha, w, \beta_l, \beta_u, \lambda} \left[-\tfrac{1}{2}\alpha^\top \alpha - \tfrac{\gamma}{2}\sum_i z_i w_i^2 + y^\top \alpha + \beta_l^\top l - \beta_u^\top u + \lambda \right]}$$

subject to

UESTC

Convexity of $f(z)$

Consider the dual representation of $f(z)$ in Theorem 1:

$$f(z) = \max_{\alpha, w, \beta_l, \beta_u, \lambda} \phi(z; \alpha, w, \beta_l, \beta_u, \lambda),$$

where

$$\phi(z; \cdot) = -\frac{1}{2}\alpha^\top \alpha - \frac{\gamma}{2}\sum_i z_i w_i^2 + y^\top \alpha + \beta_l^\top l - \beta_u^\top u + \lambda.$$

For fixed $(\alpha, w, \beta_l, \beta_u, \lambda)$, the dependence on z is

$$\phi(z; \cdot) = \left(-\frac{\gamma}{2}\sum_i w_i^2 z_i\right) + (\text{constant in } z),$$

which is linear in z.

Therefore, $f(z)$ is the pointwise supremum of a family of linear functions of z:

$$f(z) = \sup_{\alpha, w, \beta_l, \beta_u, \lambda} \{\text{linear function in } z\}.$$

Conclusion:
The pointwise supremum of linear (affine) functions is convex, hence

$$\boxed{f(z) \text{ is convex in } z \text{ (on the convex hull of } Z_k^n).}$$

Moreover, from Corollary 1, an optimal dual solution $w^*(z)$ yields a subgradient

$$g_{z,i} = -\frac{\gamma}{2}\left(w_i^*(z)\right)^2 \in \partial f(z),$$

which will be used to construct cutting planes.

UESTC

Outer Approximation

We now face the discrete convex optimization problem

$$\min_{z \in Z_k^n} f(z),$$

where

- Z_k^n is a finite set of binary vectors,
- $f(z)$ is convex (when extended to $conv(Z_k^n)$),
- we can evaluate both $f(z)$ and a subgradient $g_z \in \partial f(z)$ via the dual problem in Theorem 1.

For a convex function f, at any point $\hat{z}$ and any subgradient $g_{\hat{z}}$, the first-order underestimator

$$f(z) \geq f(\hat{z}) + g_{\hat{z}}^\top (z - \hat{z}) \tag{16}$$

is a supporting hyperplane of the epigraph of f.

Outer Approximation (OA):

- Use the inequalities
$$\theta \geq f(z_t) + g_{z_t}^\top (z - z_t)$$
as cutting planes that underapproximate $f(z)$ from below by a piecewise-linear function.
- Solve a sequence of mixed-integer linear problems in (z, θ), adding new cuts iteratively.

Since Z_k^n is finite and f is convex, the OA method converges to an optimal solution of $\min_{z \in Z_k^n} f(z)$ in a finite number of iterations.

Lagrangian Dual Formulation

Substituting the aforementioned expressions for x, r into $\mathcal{L}$ then defines the Lagrangian dual, where we eliminate π (by replacing it with a nonnegativity constraint) and introduce w such that $x := \gamma Z w$ for brevity. The Lagrangian dual reveals that for any z such that Problem (10) is feasible,
Form 1 of Lagrangian Dual:

$$f(z) = \max_{\substack{\alpha \in \mathbb{R}^m, w \in \mathbb{R}^n \\ \beta_l \in \mathbb{R}^p_+, \beta_u \in \mathbb{R}^p_+, \lambda \in \mathbb{R}}} \left\{ -\frac{1}{2}\alpha^\top \alpha - \frac{\gamma}{2} w^\top Z^\top Z w + y^\top \alpha + \beta_l^\top l - \beta_u^\top u + \lambda \right\}$$

Subject to: $w \geq X^\top \alpha + A^\top (\beta_l - \beta_u) + \lambda e - d$
Moreover, at binary points z, $z_i^2 = z_i$, and, therefore, the previous problem is equivalent to solving
Form 2 (Binary z Simplification):

$$f(z) = \max_{\substack{\alpha \in \mathbb{R}^m, w \in \mathbb{R}^n \\ \beta_l \in \mathbb{R}^p_+, \beta_u \in \mathbb{R}^p_+, \lambda \in \mathbb{R}}} \left\{ -\frac{1}{2}\alpha^\top \alpha - \frac{\gamma}{2} \sum_i z_i w_i^2 + y^\top \alpha + \beta_l^\top l - \beta_u^\top u + \lambda \right\}$$

Subject to: $w \geq X^\top \alpha + A^\top (\beta_l - \beta_u) + \lambda e - d$
$Z^\top Z = Z$ (since z is binary: $z_i^2 = z_i \implies Z = \mathrm{Diag}(z_i) = \mathrm{Diag}(z_i^2) = Z^2$)
Corollary 1 (Subgradient of $f(z)$):
Let $w^*(z)$ be an optimal choice of w for a particular subset of securities z. Then, a valid subgradient $g_z \in \partial f(z)$ has components:

$$g_{z,i} = -\frac{\gamma}{2} w_i^*(z)^2 \tag{14}$$

First-Order Underestimator (Link to Master Problem): Corollary 1 implies $f(z)$ has a valid first-order lower bound (critical for linear cuts):

$$f(z) \geq f(\hat{z}) + g_{\hat{z}}^\top (z - \hat{z}) \tag{16}$$

Outer Approximation (OA) Algorithm: Iterative Logic

OA splits the problem into Subproblem (solve for x given z) and Master Problem (MILP) (solve for z via linear cuts):

1. Initialization: Generate initial z^0 (k-sparse 0-1 vector), set $t = 0$, convergence threshold $\epsilon > 0$.
2. Subproblem: Fix z^t, solve Lagrangian function to get $f(z^t)$ and subgradient $g^t = -\frac{\gamma}{2}(w^{*t})^2$.
3. Master Problem (MILP): Build MILP with linear cuts, solve to get new candidate z^{t+1}.
4. Convergence Check: If $|f(z^{t+1}) - \theta^*| < \epsilon$, stop; else $t = t + 1$, return to Step 2.

Core of OA: Linear cuts shrink the feasible region of z iteratively, until optimal z is found.

UESTC

Objective Function:

$$\min_{z,\theta} \theta$$

Constraint 1: Linear Cut Constraints

$$\theta \geq f(z^t) + (g^t)^\top (z - z^t) \quad \forall t = 0, 1, ..., t_{\text{current}}$$

Linear Cut

- "Fence" to shrink the feasible region of z: excludes z with $f(z) > \theta$.
- Raw materials from subproblem: $f(z^t)$ (optimal value of subproblem) + g^t (subgradient of $f(z)$ at z^t).
- Geometric meaning: Tangent line of $f(z)$ at z^t (lower envelope of $f(z)$).

Constraint 2: k-Sparse Constraint

$$e^\top z \leq k \quad (e = \text{all-ones vector})$$

Constraint 3: Binary Constraint

$$z_i \in \{0, 1\} \quad \forall i = 1, ..., n$$

UESTC

Master Problem: Complete Mathematical Form

$$
\begin{cases}
\min_{z,\theta} \theta & \\
\text{s.t.} \quad & \theta \geq f(z^0) + (g^0)^\top (z - z^0) \\
& \theta \geq f(z^1) + (g^1)^\top (z - z^1) \\
& \vdots \\
& \theta \geq f(z^t) + (g^t)^\top (z - z^t) \quad \forall t = 0, ..., t_{\text{current}} \\
& e^\top z \leq k \\
& z_i \in \{0, 1\} \quad \forall i \\
& (\text{Optional}) \sum_{i:\mu_i \geq \bar{r}} z_i \geq 1 \quad \text{(Feasibility constraint)}
\end{cases}
$$

Optional Feasibility Constraint: Ensures selected assets meet minimum return requirement (avoids invalid z).

UESTC

How MILP Solvers Solve the Master Problem?

Use solvers (Gurobi/CPLEX) with Branch-and-Bound Algorithm (core logic):

1. Relaxation: Ignore binary constraint ($z_i \in [0, 1]$), solve Linear Programming (LP) to get fractional z (e.g., $z = [0.8, 0.3, 1, ...]$) and θ^*.
2. Branching: For non-0/1 z_i (e.g., $z_1 = 0.8$), split into two subproblems: $z_1 = 0$ and $z_1 = 1$.
3. Pruning: Discard subproblems with $\theta >$ current minimum θ (no need to explore).
4. Termination: Stop when all subproblems yield 0-1 z; select z with smallest θ as z^{t+1}.

UESTC

Role of Linear Cuts: Shrink Search Space

Intuitive Example (n=1000, k=5):

- Initial feasible region: $\binom{1000}{5} \approx 8.3 \times 10^{12}$ candidate z.
- After 1 cut: Exclude z with $f(z) > \theta^0 \rightarrow \approx 1 \times 10^{12}$ candidates left.
- After 10 cuts: Exclude most invalid $z \rightarrow \approx 1000$ candidates left.

Core Value of Linear Cuts:

- Use subgradient information (from Lagrangian function) to exclude non-optimal z.
- Convert non-convex $f(z)$ into linear approximations (MILP solvable).

UESTC

Algorithm 2 (A Discrete First-Order Heuristic)

$t \leftarrow 1$

$z_1 \leftarrow$ randomly generated k-sparse binary vector.

$w^\star \leftarrow \mathbf{0}$.

while $z_t \neq z_{t-1}$ **and** $t < T$ **do**

Set w_t optimal solution to:

$$\max_{\substack{\alpha\in\mathbb{R}^r, w\in\mathbb{R}^n,\\ \beta_l,\beta_u\in\mathbb{R}^n_+, \lambda\in\mathbb{R}}} \; -\frac{1}{2}\alpha^\top\alpha - \frac{\gamma}{2}\sum_i z_{t,i} w_i^2 + y^\top\alpha + \beta_l^\top l - \beta_u^\top u + \lambda$$

$$\text{s.t.} \quad w \geq X^\top\alpha + A^\top(\beta_l - \beta_u) + \lambda e - d.$$

Average multipliers via $w^\star \leftarrow \frac{1}{t} w_t + \frac{t-1}{t} w^\star$.

Set $g_{z,i} = \frac{-\gamma}{2} w_i^{\star 2} \quad \forall i \in [n]$, $x_{t,i} = \gamma w_i^\star \quad \forall i \in [n] : z_{t,i} = 1$,

$z_{t+1} = \arg\min_{z\in\mathcal{Z}_k^n} \|z - x_t + \frac{1}{L} g_{z_t}\|_2^2$

$t \leftarrow t+1$

end while

return z_t

UESTC

From First-Order to Second-Order Approximation

1. First-Order Approximation (Limitations for Discrete z) For convex $f(z)$, the standard first-order approximation is:

$$f(z) \approx f(z_{\text{old}}) + g_{z_{\text{old}}}^{\top}(z - z_{\text{old}})$$

Limitation: Linear in z —for discrete k-sparse $z \in Z^k$ (binary, $\sum z_i = k$):

- No "stabilization" term (prone to jumping to suboptimal regions);
- Fails to leverage the Lipschitz continuity of g_z (weaker approximation).

2. Why Quadratic (Second-Order) Approximation? We use the Lipschitz condition ($\|g_{z_1} - g_{z_2}\|_2 \leq L\|z_1 - z_2\|_2$) to add a quadratic regularization term, which:

- Incorporates bounds on subgradient variation (more accurate approximation);
- Converts the linear form to a norm-squared (easy to optimize for discrete z).

3. The Conversion Logic Combine first-order information + Lipschitz regularization:

$$\underbrace{f(z) \approx f(z_{\text{old}}) + g_{z_{\text{old}}}^{\top}(z - z_{\text{old}})}_{1^{\text{st}}-order} \xrightarrow{\text{Add quadratic term}} \underbrace{f(z) \approx \left\| z - \left(x^*(z_{\text{old}}) - \frac{1}{L} g_{z_{\text{old}}} \right) \right\|_2^2}_{2^{\text{nd}}-order(quadratic)}$$

- $x^*(z_{\text{old}})$: Optimal portfolio weights (links z to continuous x via $x = \gamma Z w$); - $\frac{1}{L} g_{z_{\text{old}}}$: Scales subgradient by Lipschitz constant (controls approximation strength).

UESTC

For convex $f(z)$, the subgradient inequality gives a lower bound of $f(z)$:

$$f(z) \geq f(z_{\text{old}}) + g_{\text{old}}^{\top}(z - z_{\text{old}})$$

where $g_{\text{old}} = g_{z_{\text{old}}}$ (subgradient of f at z_{old}, from Corollary 1).
The L-Lipschitz continuity of g_z states:

$$\|g_z - g_{\text{old}}\|_2 \leq L\|z - z_{\text{old}}\|_2$$

For convex $f(z)$, this condition lets us derive an upper bound of $f(z)$:

$$f(z) \leq f(z_{\text{old}}) + g_{\text{old}}^{\top}(z - z_{\text{old}}) + \frac{L}{2}\|z - z_{\text{old}}\|_2^2$$

The quadratic term $\frac{L}{2}\|z - z_{\text{old}}\|_2^2$ bounds the error from subgradient variation.
To find a good z (minimizing $f(z)$), we minimize the upper bound :

$$\arg\min_z f(z) \approx \arg\min_z \left[f(z_{\text{old}}) + g_{\text{old}}^{\top}(z - z_{\text{old}}) + \frac{L}{2}\|z - z_{\text{old}}\|_2^2 \right]$$

The constant $f(z_{\text{old}})$ can be dropped, so we simplify to:

$$\arg\min_z \left[g_{\text{old}}^{\top}(z - z_{\text{old}}) + \frac{L}{2}\|z - z_{\text{old}}\|_2^2 \right]$$

Substitute $x_{\text{old}}^* = x^*(z_{\text{old}})$ (optimal x for z_{old}, $x_{\text{old}}^* = \gamma Z_{\text{old}} w_{\text{old}}^*$) and rearrange terms:

$$\text{Above minimization} = \arg\min_z \left\| z - \left(x_{\text{old}}^* - \frac{1}{L} g_{\text{old}} \right) \right\|_2^2$$

UESTC

We need to minimize this quadratic function of z:

$$\mathcal{J}(z) = g_{\text{old}}^{\top}(z - z_{\text{old}}) + \frac{L}{2}\|z - z_{\text{old}}\|_2^2 \tag{3}$$

Expand All Terms Using $\|v\|_2^2 = v^{\top}v$ First, expand the norm term:

$$\|z - z_{\text{old}}\|_2^2 = (z - z_{\text{old}})^{\top}(z - z_{\text{old}}) = z^{\top}z - 2z_{\text{old}}^{\top}z + z_{\text{old}}^{\top}z_{\text{old}}$$

Substitute back into $\mathcal{J}(z)$:

$$\mathcal{J}(z) = g_{\text{old}}^{\top}z - g_{\text{old}}^{\top}z_{\text{old}} + \frac{L}{2}\left(z^{\top}z - 2z_{\text{old}}^{\top}z + z_{\text{old}}^{\top}z_{\text{old}}\right)$$

Distribute the $\frac{L}{2}$:

$$\mathcal{J}(z) = g_{\text{old}}^{\top}z - g_{\text{old}}^{\top}z_{\text{old}} + \frac{L}{2}z^{\top}z - Lz_{\text{old}}^{\top}z + \frac{L}{2}z_{\text{old}}^{\top}z_{\text{old}} \tag{3a}$$

Group Terms by Powers of z Rearrange (3a) to separate terms with $z^{\top}z$ (quadratic), z (linear), and constants (no z):

$$\mathcal{J}(z) = \underbrace{\frac{L}{2}z^{\top}z}_{\text{Quadratic term (degree 2)}} + \underbrace{(g_{\text{old}} - Lz_{\text{old}})^{\top}z}_{\text{Linear term (degree 1)}} + \underbrace{\left(-g_{\text{old}}^{\top}z_{\text{old}} + \frac{L}{2}z_{\text{old}}^{\top}z_{\text{old}}\right)}_{\text{Constant term (degree 0)}}$$

UESTC

We want to rewrite the quadratic+linear terms as $\frac{L}{2}\|z-a\|_2^2$ (standard norm form). Recall:

$$\|z-a\|_2^2 = z^\top z - 2a^\top z + a^\top a$$

Multiply both sides by $\frac{L}{2}$:

$$\frac{L}{2}\|z-a\|_2^2 = \frac{L}{2}z^\top z - La^\top z + \frac{L}{2}a^\top a \tag{C1}$$

Compare (C1) with the quadratic+linear terms of $\mathcal{J}(z)$:

$$\frac{L}{2}z^\top z + (g_{\text{old}} - Lz_{\text{old}})^\top z \quad \equiv \quad \frac{L}{2}z^\top z - La^\top z$$

Equate the coefficients of z (linear term):

$$g_{\text{old}} - Lz_{\text{old}} = -La \implies a = z_{\text{old}} - \frac{1}{L}g_{\text{old}}$$

Plug $a = z_{\text{old}} - \frac{1}{L}g_{\text{old}}$ into (C1):

$$\text{Quadratic+linear terms} = \frac{L}{2}\left\|z - \left(z_{\text{old}} - \frac{1}{L}g_{\text{old}}\right)\right\|_2^2 - \frac{L}{2}a^\top a$$

The $-\frac{L}{2}a^\top a$ is a constant (no z) $\rightarrow$ ignore for minimization. Thus:

$$\arg\min_z \mathcal{J}(z) = \arg\min_z \left\|z - \left(z_{\text{old}} - \frac{1}{L}g_{\text{old}}\right)\right\|_2^2$$

As z_{old} (discrete) and $x^*_{\text{old}} = \gamma Z_{\text{old}} w^*_{\text{old}}$ (continuous) share the same sparse structure (and x^*_{old} is algorithm-friendly), replace z_{old} with x^*_{old}:

$$\arg\min_z \mathcal{J}(z) = \arg\min_z \left\|z - \left(x^*_{\text{old}} - \frac{1}{L}g_{\text{old}}\right)\right\|_2^2$$