ESG-Constrained Bayesian Optimization of Deep Reinforcement Learning Portfolio Agents

A soft–hard constraint framework

Madrid, May 2026

Eduardo C. Garrido-Merchán **María Coronado Vaca**

Quantitative Methods Dept. · Finance Dept.

Universidad Pontificia Comillas, ICADE · IIT

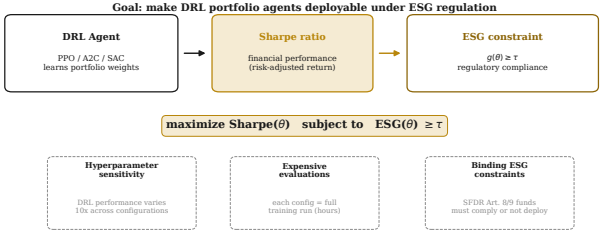
ecgarrido@comillas.edu · maria.coronado@comillas.edu

RCEA International Conference 2026 · Madrid, 25–27 May

COMILLAS
UNIVERSIDAD PONTIFICIA

Outline.

DRL agents need ESG compliance to be deployable in regulated funds.



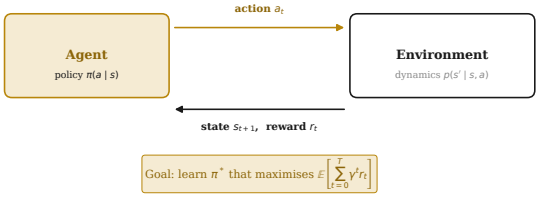
The goal. DRL agents learn portfolio allocations that maximize risk-adjusted returns (Sharpe ratio). We want them to *also* satisfy ESG constraints.

Three obstacles. (1) DRL performance varies by an order of magnitude across hyperparameter configurations. (2) Each configuration requires a full training run. (3) EU SFDR (Art. 8/9) makes ESG compliance a *hard* deployment requirement.

What we need. A sample-efficient optimizer that jointly maximizes Sharpe and guarantees $g(\theta) \geq \tau$.

Visual. the constrained optimization problem: maximize Sharpe subject to ESG $\geq$ τ, under hyperparameter sensitivity and expensive evaluations. Quantitative. three obstacles to deploying DRL in regulated funds. Take-away. making DRL portfolio agents ESG-compliant requires constrained hyperparameter optimization.

Reinforcement learning: an agent learns by trial and error.



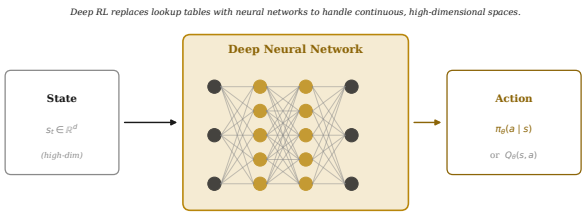
Setup. An *agent* interacts with an *environment* in discrete time steps. At each step t, it observes a state s_t, takes an action a_t, and receives a scalar reward r_t.

Objective. Find a policy $\pi^*(a \mid s)$ that maximises cumulative discounted reward $\mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_t\right]$.

Key property. No labelled examples needed. The agent discovers good behaviour purely through interaction.

Visual. agent–environment loop with state, action, and reward signals. Quantitative. policy π maps states to actions; γ discounts future rewards. Take-away. RL is trial-and-error sequential decision-making.

Deep RL replaces tables with neural networks for high-dimensional control.



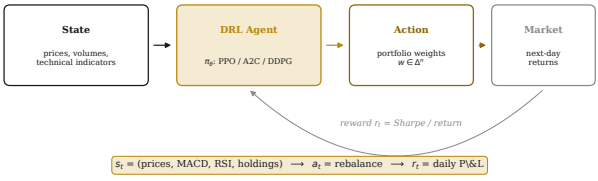
The scalability bottleneck. Tabular RL (Q-tables) works only when states and actions are discrete and small.

Deep RL solution. Parametrise $\pi_\theta(a \mid s)$ or $Q_\theta(s, a)$ with a deep neural network trained via gradient descent on the RL objective.

Algorithms. PPO, A2C, SAC, DDPG, TD3 differ in how they collect data and compute gradients, but all share the same core: a neural network that maps high-dimensional observations to continuous actions.

Visual. state vector feeds into a multi-layer network that outputs actions. Quantitative. neural networks handle $\mathbb{R}^d$ state spaces that tables cannot. Take-away. deep RL scales RL to real-world, high-dimensional problems.

In finance, a DRL agent learns to rebalance a portfolio daily.



Standard DRL algorithms (PPO, A2C, SAC) treat portfolio allocation as a continuous control problem.

State. Prices, volumes, technical indicators (MACD, RSI), and current holdings for n assets.

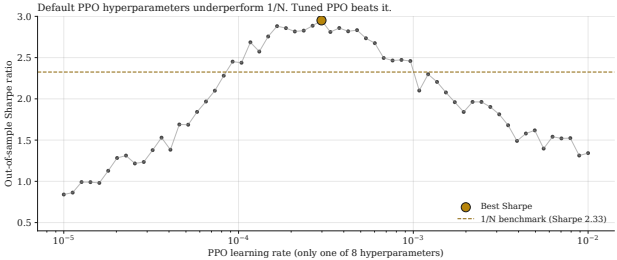
Action. Portfolio weights $w \in \Delta^n$ (the n-simplex): how much capital to allocate to each stock.

Reward. Daily profit-and-loss, Sharpe ratio, or risk-adjusted return over a rolling window.

Promise. DRL agents can adapt non-linearly to market regimes, outperforming static rules. But they are notoriously sensitive to their hyperparameters.

Visual. DRL portfolio loop: state (market data) ⟶ agent ⟶ action (weights) ⟶ reward (P&L). **Quantitative.** daily rebalancing across n assets. **Take-away.** DRL treats portfolio allocation as a continuous control problem.

DRL portfolio agents are powerful but extremely hyperparameter-sensitive.



Setup. PPO agent over 28 DJIA stocks; 8-dim hyperparameter space.

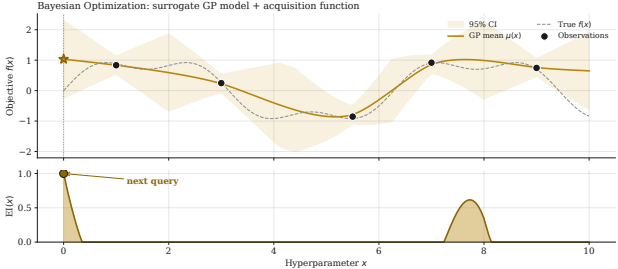
What we observe. Default PPO (Sharpe **1.30**) *loses* to the parameter-free 1/N benchmark (Sharpe **2.33**). Tuned PPO (Sharpe **2.63**) beats it by 13.3%.

Implication. Hyperparameter optimization is not optional. But each evaluation is a full DRL training run, so grid search is infeasible.

We need. A sample-efficient optimizer with the ESG constraint baked into the search.

Visual. Sharpe ratio across PPO learning rates (1 of 8 hyperparameters). Quantitative. default 1.30 vs. 1/N 2.33 vs. tuned 2.63. Take-away. tuning is decisive for DRL portfolio performance.

Bayesian optimization finds good hyperparameters in very few evaluations.



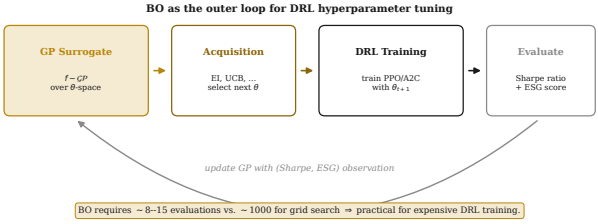
Surrogate. A Gaussian process (GP) models the unknown objective $f(\theta)$ from past evaluations, providing a predictive mean $\mu(\theta)$ and uncertainty $\sigma(\theta)$.

Acquisition function. Expected Improvement (EI) balances exploitation (high μ) with exploration (high σ). The next evaluation is $\theta_{t+1} = \arg\max \mathrm{EI}(\theta)$.

Loop. Evaluate $f(\theta_{t+1})$, update the GP, repeat. Typically 8–15 iterations suffice where grid search needs thousands.

Visual. GP posterior (top) and EI acquisition (bottom); star marks next query. Quantitative. 8–15 BO iterations replace ∼1000 grid-search evaluations. Take-away. BO is the standard for expensive black-box optimization.

BO wraps DRL training as an outer optimization loop.



Outer loop (BO). Selects the next hyperparameter configuration θ_{t+1} (learning rate, batch size, network width, etc.).

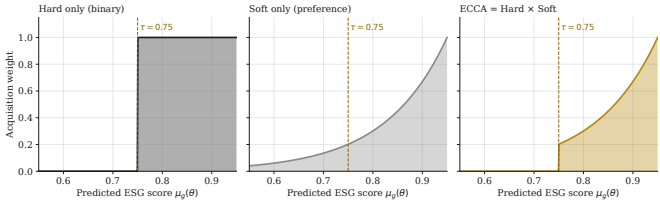
Inner loop (DRL). Trains a PPO/A2C/SAC agent with θ_{t+1}, backtests on held-out data, and returns the Sharpe ratio and ESG score.

The missing piece. Standard BO optimizes Sharpe but ignores ESG. We need to incorporate the ESG constraint $g(\theta) \geq \tau$ directly into the acquisition function.

Notation. We write τ for the ESG threshold: the minimum acceptable ESG score for regulatory compliance.

Visual. GP surrogate → acquisition → DRL training → evaluate → update GP. Quantitative. each BO iteration = one full DRL training run. Take-away. BO makes DRL hyperparameter tuning tractable; ESG is the missing constraint.

Hard-only ESG constraints treat the margin above τ indifferently.



State of the art. Constrained BO (Gelbart 2014; Gardner 2014) multiplies the acquisition by $P(g(\theta) \geq \tau)$.

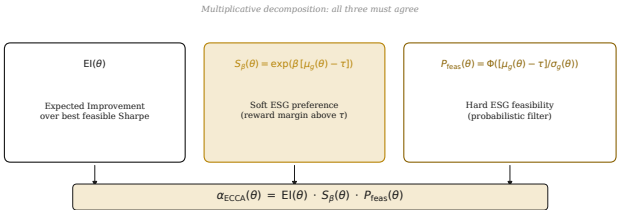
Recall: τ is the ESG threshold, the minimum ESG score a portfolio must achieve for regulatory compliance.

Hidden assumption. A portfolio at ESG = 0.75 is treated identically to one at 0.92. The optimizer is *indifferent* to the margin above τ.

Our claim. ESG is dual: a hard regulatory floor *and* a soft preference among feasible solutions.

Visual. hard (binary), soft (exponential), and ECCA (product) as functions of μ_g. Quantitative. soft shapes ranking inside the feasible set. Take-away. the soft term carries the preference signal hard-only methods discard.

ECCA decomposes the ESG constraint into hard × soft × EI.



GP surrogates on Sharpe $f \sim \mathcal{GP}$ and on ESG $g \sim \mathcal{GP}$ (Matérn 5/2, ARD).

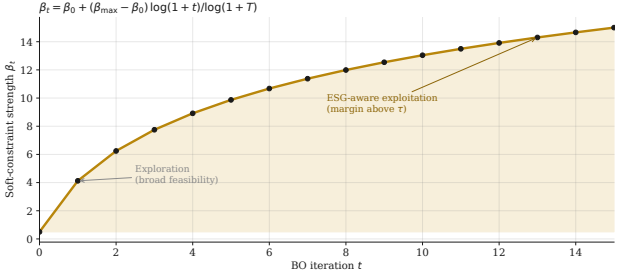
$$\alpha_{\text{ECCA}}(\theta) = \text{EI}(\theta) \cdot S_\beta(\theta) \cdot P_{\text{feas}}(\theta).$$

Read as. A candidate must be financially promising, ESG-feasible, *and* rewarded for exceeding the threshold.

Why multiplicative. Composition with P_{feas} retains the Bayesian-decision interpretation; the soft term adds a continuous preference inside the feasible region.

Visual. three multiplicative factors and the resulting ECCA acquisition. **Quantitative.** β adaptive over T iterations; τ user-specified. **Take-away.** ECCA = EI gated by feasibility and steered by ESG margin.

An adaptive β_t moves the optimizer from exploration to ESG exploitation.



$$\beta_t = \beta_0 + (\beta_{\max} - \beta_0)\,\frac{\log(1+t)}{\log(1+T)}.$$

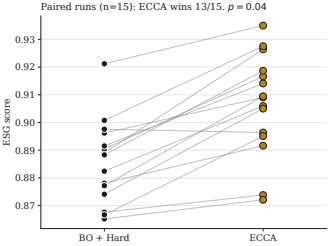
Early ($\beta_t \approx 0.5$). Soft term close to unity; GPs map the landscape without ESG bias.

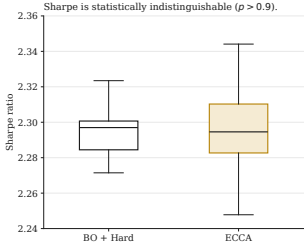
Late ($\beta_t \to 15$). Steeply rewards higher ESG margin once feasibility is mapped.

Why logarithmic. Matches the rate at which the posterior over $g(\theta)$ contracts.

Visual. β_t trajectory with $\beta_0 = 0.5$, $\beta_{\max} = 15$, $T = 15$. Quantitative. exploration phase $t \leq 3$, exploitation $t \geq 8$. Take-away. a single schedule replaces manual tuning of soft-vs-hard trade-off.

On a controlled landscape, ECCA lifts ESG by +0.029 at zero Sharpe cost.





Design. 2D ridge with constant Sharpe (≈ 2.3); ESG varies $0.84 \rightarrow 0.93$; $\tau = 0.75$, entire ridge feasible. Any ESG gain must come from the *soft* term.

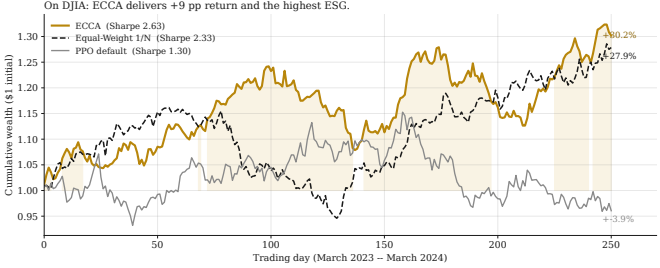
Result. 15 paired runs, 30 BO iterations.

	Sharpe	ESG
BO+Hard	2.299	0.889
ECCA	2.297	**0.918**
Δ	-0.002	**+0.029**
p	0.93	**0.04**

Cohen's $d \approx 1.2$. ECCA wins 13/15.

Visual. paired ESG dots and Sharpe boxplots across 15 seeds. **Quantitative.** +0.029 ESG, p = 0.04; Sharpe gap p > 0.9. **Take-away.** soft term buys ESG margin without spending Sharpe.

On DJIA, ECCA beats the 1/N benchmark by 13.3% in Sharpe with higher ESG.



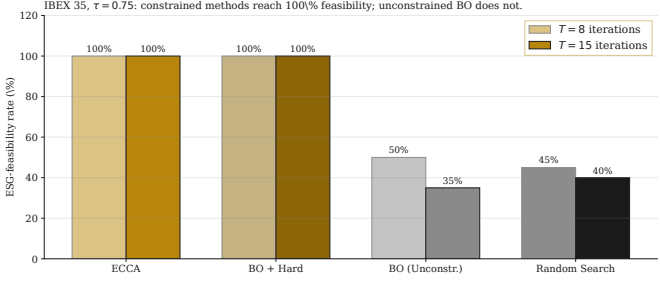
Setup. 28 DJIA stocks, daily OHLCV 2018–2024, ESG from MSCI; $\tau = 0.50$, test 03/2023–03/2024.

	Sharpe	ESG
1/N	2.33	0.653
PPO default	1.30	0.665
ECCA	**2.63**	**0.672**

Reads as. \$1M → \$1.347M in one year vs. \$1.257M for 1/N. ESG also strictly improves. Lowest drawdown (−7.73%).

Visual. cumulative wealth on \$1 over 250 trading days. Quantitative. ECCA +34.7%, 1/N +25.7%, PPO default +14.1%. Take-away. ECCA wins Sharpe, return, drawdown, and ESG simultaneously.

Constrained methods achieve 100% feasibility; unconstrained BO fails half the time.



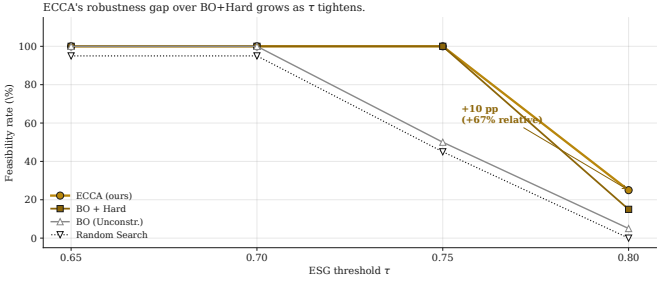
Setup. 35 IBEX stocks, 10 seeds, 2 budgets (T=8 and T=15 BO iterations), 4 methods $\Rightarrow$ 480 total runs. $\tau = 0.75$.

The feasibility gap. **ECCA** and **BO+Hard** deliver **100%** ESG-feasible portfolios. **BO (Unc.)** fails in 50–65% of runs.

For regulated funds. An Article 8/9 fund cannot deploy an infeasible allocation regardless of Sharpe. Unconstrained BO is non-deployable.

Visual. feasibility rate per method at $T = 8$ and $T = 15$. **Quantitative.** ECCA & BO+Hard 100%; BO unconstr. 35–50%. **Take-away.** constraint mechanisms are essential under any binding ESG mandate.

ECCA's robustness over BO+Hard *grows* as the ESG threshold tightens.



Sweep. $\tau \in \{0.65, 0.70, 0.75, 0.80\}$, 80 runs per cell.

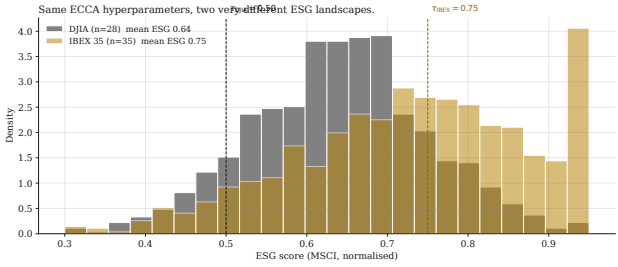
Permissive ($\tau \leq 0.70$). All methods feasible; ESG is free.

Binding ($\tau = 0.75$). **ECCA** and **BO+Hard** stay at 100%; **BO (Unc.)** collapses to 50%.

Boundary ($\tau = 0.80$). **ECCA** reaches **25%** vs. **15%** for **BO+Hard**. That is +10 pp, or +67% relative.

Visual. feasibility curves across four ESG thresholds. Quantitative. ECCA 25% vs. Hard 15% at $\tau = 0.80$. Take-away. soft weighting acts as a search prior when feasibility is rare.

ECCA generalizes across markets without per-market recalibration.



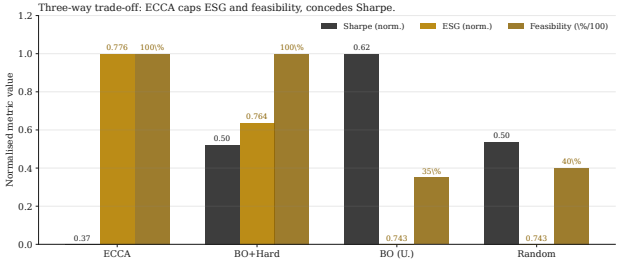
Two markets, very different ESG landscapes.
DJIA: mean ESG 0.64, $\tau = 0.50$.
IBEX 35: mean ESG 0.75, $\tau = 0.75$.

Same hyperparameters. $\beta_0 = 0.5$, $\beta_{\max} = 15$, Matérn 5/2 + ARD. No per-market tuning.

Same qualitative result. Highest ESG, highest feasibility, competitive Sharpe in both markets. The soft–hard decomposition transfers without recalibration.

Visual. ESG distributions for DJIA and IBEX 35 with their respective thresholds. Quantitative. means 0.64 vs. 0.75; identical ECCA hyperparameters. Take-away. soft–hard split transfers across markets without retuning.

Three findings and one take-home message.



Finding 1. Constraint mechanisms are *essential*: unconstrained BO fails ESG in 50–65% of runs.

Finding 2. The soft term has *zero financial cost* vs. hard-only ($p > 0.2$ on Sharpe).

Finding 3. The soft term delivers measurable ESG benefits (+0.012 above **BO+Hard**) and 67% greater resilience under tight constraints.

Take-home. Article 8/9 funds gain full compliance and ESG margin at negligible Sharpe cost. The framework requires no per-market recalibration.

Visual. normalised Sharpe, ESG, and feasibility for the four methods. Quantitative. ECCA caps ESG & feasibility; concedes Sharpe non-significantly. Take-away. ECCA delivers compliance, ESG margin, and competitive returns.

Conclusions.

Contributions

ECCA acquisition function: a principled soft–hard multiplicative decomposition that rewards ESG margin above τ without discarding feasibility guarantees.

Adaptive β_t schedule: logarithmic annealing from exploration to ESG exploitation, requiring no manual tuning.

Empirical validation: synthetic benchmark (15 seeds), DJIA (28 stocks), and IBEX 35 (480 runs, 4 thresholds) confirm that ECCA achieves the highest ESG with no significant Sharpe loss.

Future work

Multi-objective extension (Pareto front over Sharpe and ESG). Dynamic τ adapting to evolving regulation. Transfer learning across markets and time periods.

	Sharpe	ESG feas.
ECCA	2.63	100%
BO+Hard	2.55	100%
BO (Unc.)	2.71	35–50%
Random	2.40	40–45%

ECCA = EI
× soft ESG
× hard feasibility

Thank you. Questions?

ECCA = EI · soft ESG · hard feasibility

Eduardo C. Garrido-Merchán · **María Coronado Vaca**

Universidad Pontificia Comillas, Madrid

ecgarrido@comillas.edu · maria.coronado@comillas.edu

RCEA International Conference 2026 · Madrid, 25–27 May