LEARNING ADVANTAGE FUNCTION SURROGATES VIA CONVEX CONJUGATE MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

A policy trained to make per-unit decisions, such as how much of a product to order, tells us the best action but not the cost of deviating from it. That cost, the advantage function, is what downstream systems need when they have to depart from the per-unit optimum: to estimate the impact of ordering more or fewer units, to respond to a one-off vendor discount, or to satisfy a shared capacity constraint. The advantage function is what makes such trade-offs computable, yet pure policy-learning algorithms often produce the policy alone, without a critic that would yield it. We propose a method to train an advantage function surrogate through price perturbations motivated by a Fenchel-Young loss. We describe a two-stage recipe (fit the policy, then the cost model) and illustrate it on a multi-period inventory management problem.

1 INTRODUCTION

Motivation. Consider a policy trained to take optimal decisions at the level of individual products, accounts, or units, such as how many units of a given SKU to buy, learned in a setting where each unit targets its per-unit optimum. Beyond the optimal action, several downstream uses call for an *advantage* function that quantifies the cost of deviating from this optimum: counterfactual impact assessment (the financial effect of buying X more or fewer units), one-off vendor offers (the order quantity trades a time-limited discount against the advantage's curvature around the optimum), and occasional coupling constraints (vendor minimum-order quantities or capacity limits requiring coordinated deviations across products). In each case the policy is unchanged; what is needed is a faithful, typically convex, model of the cost landscape around the optimum.

The under-determinism problem. In an actor-critic algorithm the advantage function is learned alongside the policy, but pure policy methods learn or evaluate it separately, typically by regression on bootstrapped Bellman or Monte-Carlo targets. Differentiable optimization layers Amos & Kolter (2017); Agrawal et al. (2019) suggest learning it jointly, as a by-product of policy learning. This direct approach, however, is under-determined since any advantage representation whose minimum coincides with the optimal action yields the same reward. For example, for a quadratic $A(\mathbf{a};\mathbf{s}) = \frac{1}{2}(\mathbf{a}-\mathbf{a}^*(\mathbf{s}))^T B(\mathbf{a}-\mathbf{a}^*(\mathbf{s}))$ every positive-definite B is indistinguishable so long as the policy $\mathbf{a}^*(\mathbf{s})$ is optimal.

A generic consumption model. For a unified abstraction of the uses above, suppose an external consumer observes state $\mathbf{s}$ and solves

$$\mathbf{a}^{\mathcal{E}}(\mathbf{s}) \in \arg\min_{\mathbf{a}} \left\{ Q(\mathbf{a};\mathbf{s}) + g(\mathbf{a};\mathbf{s}) \right\}, \tag{1}$$

where g is a convex objective specific to the consumer and Q is the action-value function. Setting $g \equiv 0$ recovers the unconstrained optimum, a linear $g(\mathbf{a}) = -\boldsymbol{\lambda}^\top \mathbf{a}$ models a vendor discount or the dualization of a coupling constraint, and a quadratic or coupled g captures more elaborate business rules. We aim for a surrogate of Q, and this surrogate need not approximate Q pointwise, what matters is to produce the right decision under (1). In particular, $\mathbf{a}^{\mathcal{E}}(\mathbf{s}) = \nabla Q^*(-\nabla g(\mathbf{a}^{\mathcal{E}};\mathbf{s});\mathbf{s})$, where Q^* is the (state-conditional) convex conjugate of Q in the action argument. A good surrogate $f_{\boldsymbol{\mu}}$ is thus one such that $\nabla f^*_{\boldsymbol{\mu}}(\boldsymbol{\lambda};\mathbf{s}) \approx \nabla Q^*(\boldsymbol{\lambda};\mathbf{s})$ for dual values $\boldsymbol{\lambda}$ typically encountered in the consumption uses, so that the consumer's actions under the surrogate match those under the true action-value function.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Approach. To resolve the under-determinism we fit a strictly convex advantage surrogate $f_{\boldsymbol{\mu}}(\mathbf{a}; \mathbf{s})$ around the optimal action under a Fenchel-Young loss, trained against a distribution $p(\boldsymbol{\lambda})$ of dual perturbations that simulate the shadow prices the consumption uses induce. The distribution $p(\boldsymbol{\lambda})$ is an *exogenous* design choice, not an empirical distribution to be matched: the policy remains a function of $\mathbf{s}$ alone, optimal in the unconstrained problem, and the role of $p(\boldsymbol{\lambda})$ is to span the range of dual values the surrogate will be queried at. We develop a two-stage procedure (policy first, surrogate second). The surrogate's structure can encode problem knowledge, for example a quadratic when the consumption model solves QPs, or M-convexity Murota (2003) for inventory-placement action-value functions.

2 Preliminaries

2.1 Exo-MDP / Input-Driven MDP

We work in the exo-MDP (or Input-Driven MDP, IDP) framework of Mao et al. (2019); Sinclair et al. (2023), in which uncertainty is driven by an exogenous process independent of the agent's actions. At each epoch $t \in \{0, 1, \ldots, T-1\}$, the state decomposes as $\mathbf{s}_t = (\mathbf{x}_t, \mathbf{y}_t)$, where $\mathbf{x}_t = (\mathbf{o}_{t-H}, \ldots, \mathbf{o}_{t-1})$ is a fixed-length history of exogenous observations, and $\mathbf{y}_t$ the endogenous part of the state. The agent's action $\mathbf{a}_t$ is selected by a deterministic policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$, $\mathbf{a}_t = \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}_t)$. The endogenous state then evolves under a deterministic, differentiable transition $\mathbf{y}_{t+1} = \mathcal{T}(\mathbf{s}_t, \mathbf{o}_t, \mathbf{a}_t)$, and a differentiable per-period cost $c_t(\mathbf{s}_t, \mathbf{a}_t)$ is incurred. The exogenous sequence $\{\mathbf{o}_t\}_t$ is drawn from a distribution $\mathbb{P}$ independent of the agent.

The objective is to minimize the finite-horizon discounted expected cost, or equivalently the long-run average cost in the stationary case,

$$J_T(\boldsymbol{\theta}; \mathbf{y}_0) = \mathbb{E}_{\mathbb{P}}\left[\sum_{t=0}^{T-1} \gamma^t \, c_t(\mathbf{s}_t, \mathbf{a}_t) \,\middle|\, \mathbf{y}_0\right], \qquad J(\boldsymbol{\theta}) = \lim_{T\to\infty} \mathbb{E}_{\mathbb{P}}\left[\tfrac{1}{T}\sum_{t=0}^{T-1} c(\mathbf{s}_t, \mathbf{a}_t)\right].$$

The exogenous structure lets objectives and gradients be evaluated by rolling out against stored historical traces, which underlies the DirectBackprop/HDPO training algorithms Madeka et al. (2022); Alvo et al. (2023) that we rely on for upstream training.

2.2 Policy and Value Functions

We assume that a trained instance of the upstream policy from Section 2.1 is available, namely $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta}$ fixed to its trained value. The value, action-value, and advantage functions of $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ (in cost form) are

$$V_t^{\boldsymbol{\theta}}(\mathbf{s}) \;=\; \mathbb{E}\left[\sum_{\tau=t}^{T-1} \gamma^{\tau-t} \, c_t(\mathbf{s}_\tau, \mathbf{a}_\tau) \,\middle|\, \mathbf{s}_t = \mathbf{s},\; \mathbf{a}_\tau = \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s}_\tau)\right],$$

$$Q_t^{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}) \;=\; c(\mathbf{s}, \mathbf{a}) + \gamma\, \mathbb{E}\big[V_{t+1}^{\boldsymbol{\theta}}(\mathbf{s}_{t+1}) \,\big|\, \mathbf{s}_t = \mathbf{s},\; \mathbf{a}_t = \mathbf{a}\big], \qquad A_t^{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}) \;=\; Q_t^{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}) - V_t^{\boldsymbol{\theta}}(\mathbf{s}).$$

In the infinite-horizon stationary setting (time-homogeneous dynamics and cost), these are independent of t and we drop the subscript, writing $V^{\boldsymbol{\theta}}$, $Q^{\boldsymbol{\theta}}$, $A^{\boldsymbol{\theta}}$. We adopt this stationary convention for the remainder of the paper for ease of exposition.

Under cost minimization, the upstream-optimal action at state $\mathbf{s}$ satisfies

$$A^{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s})) = 0, \qquad A^{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}) \geq 0 \text{ for all } \mathbf{a}.$$

For fitting the surrogate we assume access either to $V^{\boldsymbol{\theta}}$ (via a critic or rollouts) or to one-step transitions $(\mathbf{s}_t, \mathbf{a}_t, c_t, \mathbf{s}_{t+1})$ sampled under $\boldsymbol{\pi}_{\boldsymbol{\theta}}$. When the exo-MDP decomposition is not relevant, we write $\mathbf{s}$ for the full state.

2.3 Convex-Analytic Background

We use several standard facts related to the convex conjugate (Rockafellar, 1997; Blondel et al., 2020, Ch.12, §26). For the convex conjugate $f^*(\boldsymbol{\lambda}) := \sup_{\mathbf{a}}\{\langle \mathbf{a}, \boldsymbol{\lambda}\rangle - f(\mathbf{a})\}$, strict convexity of f makes

∇f^* single-valued and the perturbed minimizer $\arg\min_{\mathbf{a}}\{f(\mathbf{a}) - \boldsymbol{\lambda}^\top \mathbf{a}\}$ equals $\nabla f^*(\boldsymbol{\lambda})$, with the Legendre correspondence $\mathbf{a} = \nabla f^*(\boldsymbol{\lambda}) \Leftrightarrow \boldsymbol{\lambda} = \nabla f(\mathbf{a})$. The Fenchel-Young gap $L^{\mathrm{FY}}(\mathbf{a}, \boldsymbol{\lambda}; f) := f(\mathbf{a}) + f^*(\boldsymbol{\lambda}) - \langle \mathbf{a}, \boldsymbol{\lambda} \rangle \geq 0$ vanishes iff $\boldsymbol{\lambda} = \nabla f(\mathbf{a})$ and can serve as a fitting loss Blondel et al. (2020). For parameterized strictly convex $f_{\boldsymbol{\mu}}$, implicit differentiation of $\nabla_{\mathbf{a}} f_{\boldsymbol{\mu}}(\mathbf{a}) = \boldsymbol{\lambda}$ gives the argmin Jacobian $\frac{\partial \mathbf{a}}{\partial \boldsymbol{\mu}} = -[\nabla^2_{\mathbf{a}} f_{\boldsymbol{\mu}}(\mathbf{a})]^{-1} \nabla_{\boldsymbol{\mu}} \nabla_{\mathbf{a}} f_{\boldsymbol{\mu}}(\mathbf{a})$, the OptNet KKT Jacobian Amos & Kolter (2017) for quadratic $f_{\boldsymbol{\mu}}$ and the layer of Agrawal et al. (2019) in general.

3 The Advantage Surrogate via Convex Conjugate Matching

3.1 Surrogate Class

General parametric family. Let $\mathcal{F} = \{f_{\boldsymbol{\mu}}(\cdot;\cdot) : \boldsymbol{\mu} \in \mathcal{M}\}$ be a family of functions $f_{\boldsymbol{\mu}} : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$, parameterized by $\boldsymbol{\mu} \in \mathcal{M}$. We require $f_{\boldsymbol{\mu}}(\cdot; \mathbf{s})$ to be strongly convex in $\mathbf{a}$ with continuous gradient $\nabla_{\mathbf{a}} f_{\boldsymbol{\mu}}$, and with strong-convexity modulus bounded below uniformly on bounded subsets of $\mathcal{S}$. Under these conditions the conjugate $f^*_{\boldsymbol{\mu}}(\boldsymbol{\lambda}; \mathbf{s}) = \sup_{\mathbf{a}}\{\langle \mathbf{a}, \boldsymbol{\lambda} \rangle - f_{\boldsymbol{\mu}}(\mathbf{a}; \mathbf{s})\}$ is finite everywhere and differentiable in $\boldsymbol{\lambda}$ on $\mathbb{R}^d$.

Anchored parameterization. A useful specialization factorizes f into a *decision* component and a *shape* component:

$$f_{\boldsymbol{\mu},\boldsymbol{\theta}}(\mathbf{a}; \mathbf{s}) = \psi_{\boldsymbol{\mu}}(\mathbf{a} - \mathbf{a}^*(\mathbf{s}; \boldsymbol{\theta});\ \mathbf{s})\,, \tag{2}$$

where $\mathbf{a}^*(\cdot; \boldsymbol{\theta}) : \mathcal{S} \to \mathbb{R}^d$ is the *anchor network* representing the optimal policy, and $\psi_{\boldsymbol{\mu}} : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$ is a *shape function*, strongly convex with continuous gradient in its first argument (we use C^2 when convenient), satisfying

$$\psi_{\boldsymbol{\mu}}(\mathbf{0}; \mathbf{s}) = 0, \qquad \nabla_{\mathbf{u}} \psi_{\boldsymbol{\mu}}(\mathbf{0}; \mathbf{s}) = \mathbf{0} \qquad \text{for every } (\boldsymbol{\mu}, \mathbf{s}).$$

These conditions enforce $f_{\boldsymbol{\mu},\boldsymbol{\theta}}(\mathbf{a}^*; \mathbf{s}) = 0$ and $\nabla_{\mathbf{a}} f_{\boldsymbol{\mu},\boldsymbol{\theta}}(\mathbf{a}^*; \mathbf{s}) = \mathbf{0}$ structurally, and the conjugate separates cleanly as $f^*_{\boldsymbol{\mu},\boldsymbol{\theta}}(\boldsymbol{\lambda}; \mathbf{s}) = \langle \mathbf{a}^*(\mathbf{s}; \boldsymbol{\theta}), \boldsymbol{\lambda} \rangle + \psi^*_{\boldsymbol{\mu}}(\boldsymbol{\lambda}; \mathbf{s})$, so the price-perturbed argmin is $\nabla_{\boldsymbol{\lambda}} f^*_{\boldsymbol{\mu},\boldsymbol{\theta}}(\boldsymbol{\lambda}; \mathbf{s}) = \mathbf{a}^*(\mathbf{s}; \boldsymbol{\theta}) + \nabla_{\boldsymbol{\lambda}} \psi^*_{\boldsymbol{\mu}}(\boldsymbol{\lambda}; \mathbf{s})$. The factorization supports *two-stage* training (anchor first, shape second; Section 4).

Example (quadratic). Taking $\psi_{\boldsymbol{\mu}}(\mathbf{u}; \mathbf{s}) = \frac{1}{2} \mathbf{u}^\top B(\mathbf{s}; \boldsymbol{\mu}) \mathbf{u}$ with $B \succ 0$ recovers the anchored quadratic with closed-form conjugate $\psi^*_{\boldsymbol{\mu}}(\boldsymbol{\lambda}; \mathbf{s}) = \frac{1}{2} \boldsymbol{\lambda}^\top B(\mathbf{s}; \boldsymbol{\mu})^{-1} \boldsymbol{\lambda}$. The same algebraic form, with $(\mu(\mathbf{s}), P(\mathbf{s}))$ in place of $(\mathbf{a}^*, B)$, is the Normalized Advantage Function Q-network of Gu et al. (2016); the differences are in the loss and protocol. NAF fits (μ, P) via MSE on Bellman targets jointly with the value function, whereas we use the FY conjugate-matching loss of Section 3.4 against an externally provided $Q^{\boldsymbol{\theta}}$ with a two-stage protocol that freezes the anchor before fitting $\boldsymbol{\mu}$, and admit a broader class of strongly convex shapes $\psi_{\boldsymbol{\mu}}$ than the quadratic.

3.2 Under-Identification of the Simple Loss

Directly minimizing the upstream cumulative cost with the differentiable-argmin action $\mathbf{a}^\star_{\boldsymbol{\mu}}(\mathbf{s}) := \arg\min_{\mathbf{a}} f_{\boldsymbol{\mu}}(\mathbf{a}; \mathbf{s})$ fails to identify the surrogate beyond its argmin. The naive loss $\mathcal{L}_{\mathrm{naive}}(\boldsymbol{\mu}) := \mathbb{E}[\sum_t \gamma^t c(\mathbf{s}_t, \mathbf{a}^\star_{\boldsymbol{\mu}}(\mathbf{s}_t))]$, with the trajectory rolled out under $\mathbf{a}_t = \mathbf{a}^\star_{\boldsymbol{\mu}}(\mathbf{s}_t)$, depends on $\boldsymbol{\mu}$ only through the argmin map: if $\mathbf{a}^\star_{\boldsymbol{\mu}} = \mathbf{a}^\star_{\boldsymbol{\mu}'}$ $\rho^{\boldsymbol{\theta}}$-almost everywhere then $\mathcal{L}_{\mathrm{naive}}(\boldsymbol{\mu}) = \mathcal{L}_{\mathrm{naive}}(\boldsymbol{\mu}')$. For the anchored parameterization (2) the argmin is $\mathbf{a}^*(\mathbf{s}; \boldsymbol{\theta})$, independent of the curvature $\boldsymbol{\mu}$, so all curvatures minimize $\mathcal{L}_{\mathrm{naive}}$ equally.

3.3 Convex Conjugate Matching

To break this degeneracy, we learn the action-value function through its convex conjugate. Let $Q^{\boldsymbol{\theta}*}(\mathbf{s}, \boldsymbol{\lambda}) = \sup_{\mathbf{a}}\{\boldsymbol{\lambda}^\top \mathbf{a} - Q^{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a})\}$ be the state-conditional conjugate in the action argument, and let the surrogate-induced action be the perturbed argmin $\mathbf{a}(\boldsymbol{\lambda}, \mathbf{s}; \boldsymbol{\mu}) := \arg\min_{\mathbf{a}}\{f_{\boldsymbol{\mu}}(\mathbf{a}; \mathbf{s}) - \boldsymbol{\lambda}^\top \mathbf{a}\} = \nabla f^*_{\boldsymbol{\mu}}(\boldsymbol{\lambda}; \mathbf{s})$, which collapses to the anchor $\mathbf{a}^\star_{\boldsymbol{\mu}}(\mathbf{s})$ at $\boldsymbol{\lambda} = \mathbf{0}$ and traces the dual conjugate of $f_{\boldsymbol{\mu}}$ for $\boldsymbol{\lambda} \neq \mathbf{0}$. The Fenchel-Young inequality applied to this action,

$$Q^{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}(\boldsymbol{\lambda}, \mathbf{s}; \boldsymbol{\mu})) + Q^{\boldsymbol{\theta}*}(\mathbf{s}, \boldsymbol{\lambda}) - \boldsymbol{\lambda}^\top \mathbf{a}(\boldsymbol{\lambda}, \mathbf{s}; \boldsymbol{\mu}) \ \geq\ 0, \tag{3}$$

holds with equality iff the surrogate's action matches the true one, $\mathbf{a}(\boldsymbol{\lambda}, \mathbf{s}; \boldsymbol{\mu}) = \mathbf{a}(\boldsymbol{\lambda}, \mathbf{s}) := \arg\min_{\mathbf{a}}\{Q^{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}) - \boldsymbol{\lambda}^\top \mathbf{a}\}$, equivalently $\nabla f_{\boldsymbol{\mu}}(\mathbf{a}(\boldsymbol{\lambda}, \mathbf{s}; \boldsymbol{\mu}); \mathbf{s}) = \nabla_{\mathbf{a}} Q^{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}(\boldsymbol{\lambda}, \mathbf{s}))$. A surrogate yielding decisions close to the true action-value function thus minimizes the gap in (3). ($Q^{\boldsymbol{\theta}}$ need not be convex in $\mathbf{a}$; the loss below then identifies the lower-semicontinuous convex envelope $Q^{\boldsymbol{\theta}**}(\mathbf{s}, \cdot)$ on $\mathrm{supp}(p)$.)

3.4 Fenchel-Young Loss (Two-Stage)

Assume the upstream policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ and value function $V^{\boldsymbol{\theta}}$ are given (Section 2.2). Let $\rho^{\boldsymbol{\theta}}$ be the state distribution of $\boldsymbol{\pi}_{\boldsymbol{\theta}}$-rollouts and $p(\boldsymbol{\lambda})$ a chosen dual-perturbation distribution. Averaging the FY gap (3) over $\rho^{\boldsymbol{\theta}} \otimes \mathbb{P} \otimes p$ and dropping the $\boldsymbol{\mu}$-independent constant $\mathbb{E}[Q^{\boldsymbol{\theta}*}(\mathbf{s}, \boldsymbol{\lambda})]$, the surrogate is fit by minimizing

$$\begin{aligned} \mathcal{J}(\boldsymbol{\mu}) &:= \mathbb{E}_{\mathbf{s}_t \sim \rho^{\boldsymbol{\theta}},\, \mathbf{o}_t \sim \mathbb{P},\, \boldsymbol{\lambda} \sim p(\boldsymbol{\lambda})} \left[Q^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}(\boldsymbol{\lambda}, \mathbf{s}_t; \boldsymbol{\mu})) \;-\; \boldsymbol{\lambda}^\top \mathbf{a}(\boldsymbol{\lambda}, \mathbf{s}_t; \boldsymbol{\mu}) \right] \\ &= \mathbb{E}_{\mathbf{s}_t \sim \rho^{\boldsymbol{\theta}},\, \mathbf{o}_t \sim \mathbb{P},\, \boldsymbol{\lambda} \sim p(\boldsymbol{\lambda})} \left[c_t(\mathbf{s}_t, \mathbf{a}(\boldsymbol{\lambda}, \mathbf{s}_t; \boldsymbol{\mu})) \;-\; \boldsymbol{\lambda}^\top \mathbf{a}(\boldsymbol{\lambda}, \mathbf{s}_t; \boldsymbol{\mu}) \;+\; \gamma\, V^{\boldsymbol{\theta}}_{t+1}(\mathbf{s}_{t+1}) \right], \end{aligned} \tag{4}$$

the two forms being equal because the one-step Bellman target $c_t + \gamma V^{\boldsymbol{\theta}}(\mathbf{s}_{t+1})$ estimates $Q^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a})$ without a separately trained critic, since the exo-MDP transition is differentiable in $\mathbf{a}$ and $V^{\boldsymbol{\theta}}$ comes from a learned value network or DirectBackprop rollout continuation under $\boldsymbol{\pi}_{\boldsymbol{\theta}}$. The states are drawn from the optimal-policy rollout, keeping the second-stage distribution aligned with deployment.

3.5 First-Order Condition and Policy-Gradient Equivalence

Differentiating (4) through the differentiable exo-MDP transition gives $\nabla_{\boldsymbol{\mu}} \mathcal{J}(\boldsymbol{\mu}) = \mathbb{E}[\frac{\partial \mathbf{a}}{\partial \boldsymbol{\mu}}^\top (\nabla_{\mathbf{a}} Q^{\boldsymbol{\theta}}(\mathbf{s}_t, \mathbf{a}) - \boldsymbol{\lambda})]$. Using $\nabla_{\mathbf{a}} Q^{\boldsymbol{\theta}} = \nabla_{\mathbf{a}} A^{\boldsymbol{\theta}}$ ($V^{\boldsymbol{\theta}}$ is action-independent) and the dual consistency $\boldsymbol{\lambda} = \nabla_{\mathbf{a}} f_{\boldsymbol{\mu}}(\mathbf{a}; \mathbf{s}_t)$ at the argmin, the bracket becomes the advantage residual $\nabla_{\mathbf{a}} A^{\boldsymbol{\theta}} - \nabla_{\mathbf{a}} f_{\boldsymbol{\mu}}$.

Theorem 3.1 (First-order condition)**.** *At every stationary point of* (4)*, with* $\mathbf{a} = \mathbf{a}(\boldsymbol{\lambda}, \mathbf{s}; \boldsymbol{\mu})$*,*

$$\mathbb{E}_{\mathbf{s} \sim \rho^{\boldsymbol{\theta}},\, \boldsymbol{\lambda} \sim p(\boldsymbol{\lambda})} \left[\frac{\partial \mathbf{a}}{\partial \boldsymbol{\mu}}^\top \left(\nabla_{\mathbf{a}} A^{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}) - \nabla_{\mathbf{a}} f_{\boldsymbol{\mu}}(\mathbf{a}; \mathbf{s}) \right) \right] = \mathbf{0}, \tag{5}$$

equivalently with $Q^{\boldsymbol{\theta}}$ *in place of* $A^{\boldsymbol{\theta}}$*.*

This FOC has a policy-gradient reading. Rearranging (5) at $\mathbf{a} = \mathbf{a}(\boldsymbol{\lambda}, \mathbf{s}; \boldsymbol{\mu})$,

$$\mathbb{E}_{\rho^{\boldsymbol{\theta}} \otimes p} \left[\frac{\partial \mathbf{a}}{\partial \boldsymbol{\mu}}^\top \nabla_{\mathbf{a}} Q^{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{a}) \right] \;=\; \mathbb{E}_{\rho^{\boldsymbol{\theta}} \otimes p} \left[\frac{\partial \mathbf{a}}{\partial \boldsymbol{\mu}}^\top \nabla_{\mathbf{a}} f_{\boldsymbol{\mu}}(\mathbf{a}; \mathbf{s}) \right],$$

each side a deterministic policy gradient Silver et al. (2014) through the dual policy: the left with $Q^{\boldsymbol{\theta}}$ as cost, the right with the surrogate $f_{\boldsymbol{\mu}}$. The fitting FOC imposes that the two coincide under $\rho^{\boldsymbol{\theta}} \otimes p$.

Theorem 3.2 (Policy-gradient equivalence)**.** *Let* $\tilde{\boldsymbol{\pi}}_{\boldsymbol{\mu}}(\mathbf{s}, \boldsymbol{\lambda}) := \mathbf{a}(\boldsymbol{\lambda}, \mathbf{s}; \boldsymbol{\mu}) = \nabla f^*_{\boldsymbol{\mu}}(\boldsymbol{\lambda}; \mathbf{s})$ *denote the dual policy induced by* $f_{\boldsymbol{\mu}}$*. For any joint distribution* μ *on* $(\mathbf{s}, \boldsymbol{\lambda})$ *and any cost function* $h(\mathbf{s}, \mathbf{a})$ *differentiable in* $\mathbf{a}$*, define the policy gradient of* h *through* $\tilde{\boldsymbol{\pi}}_{\boldsymbol{\mu}}$ *under* μ *by*

$$\nabla^{\mathrm{PG}}_{\boldsymbol{\mu}} \mathbb{E}_{\mu}[h(\mathbf{s}, \tilde{\boldsymbol{\pi}}_{\boldsymbol{\mu}}(\mathbf{s}, \boldsymbol{\lambda}))] \;:=\; \mathbb{E}_{\mu} \left[\frac{\partial \mathbf{a}}{\partial \boldsymbol{\mu}}^\top \nabla_{\mathbf{a}} h(\mathbf{s}, \mathbf{a}) \right]_{\mathbf{a} = \tilde{\boldsymbol{\pi}}_{\boldsymbol{\mu}}(\mathbf{s}, \boldsymbol{\lambda})},$$

where only the dependence of $\tilde{\boldsymbol{\pi}}_{\boldsymbol{\mu}}$ *on* $\boldsymbol{\mu}$ *is differentiated; any explicit* $\boldsymbol{\mu}$*-dependence of* h *(or of* μ*) is held fixed. Then the stationarity condition* (5) *is equivalent to*

$$\nabla^{\mathrm{PG}}_{\boldsymbol{\mu}} \mathbb{E}_{\rho^{\boldsymbol{\theta}} \otimes p}\left[Q^{\boldsymbol{\theta}}(\mathbf{s}, \tilde{\boldsymbol{\pi}}_{\boldsymbol{\mu}}(\mathbf{s}, \boldsymbol{\lambda})) \right] \;=\; \nabla^{\mathrm{PG}}_{\boldsymbol{\mu}} \mathbb{E}_{\rho^{\boldsymbol{\theta}} \otimes p}[f_{\boldsymbol{\mu}}(\tilde{\boldsymbol{\pi}}_{\boldsymbol{\mu}}(\mathbf{s}, \boldsymbol{\lambda}); \mathbf{s})]\,. \tag{6}$$

The equation asserts that the policy gradient of the true objective, through the upstream action-value $Q^{\boldsymbol{\theta}}$, equals the policy gradient obtained by treating the surrogate $f_{\boldsymbol{\mu}}$ as the cost function. The fitted surrogate is *compatible* with the true advantage in the sense that any policy-improvement step taken through $f_{\boldsymbol{\mu}}$ produces the same direction as one taken through the true advantage.

Relation to compatible function approximation. Theorem 3.2 is reminiscent of, but weaker than, the compatible function approximation theorems of Sutton et al. (1999) (stochastic policies) and Silver et al. (2014) (deterministic policies). Those results give a *structural* condition on the critic parameterization under which the critic-induced policy gradient equals the true policy gradient *at every parameter value*. For instance, Silver et al. (2014) requires $\nabla_{\mathbf{a}} Q^{\mathbf{w}}(\mathbf{s}, \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s})) = \nabla_{\boldsymbol{\theta}} \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{s})^{\top} \mathbf{w}$ together with an MSE fit on $\nabla_{\mathbf{a}} Q$. Theorem 3.2, in contrast, only asserts that the two policy gradients coincide *at stationary points* of the FY-loss FOC (5).

4 Training

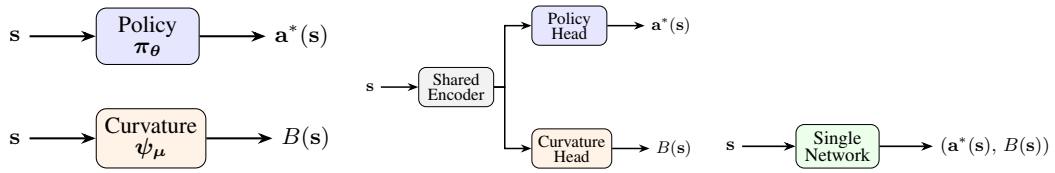
4.1 Procedure

Rollout mechanics. Training relies on differentiable rollouts: a length-T trajectory is rolled out against a sampled exogenous trace $\{\mathbf{o}_t\}$, and the differentiability of the transition $\mathcal{T}$ and cost c in the action allows backpropagation through the whole trajectory. At a perturbation step a dual price $\boldsymbol{\lambda} \sim p(\boldsymbol{\lambda})$ is drawn and the action is the perturbed argmin $\mathbf{a}_t = \nabla f^*_{\boldsymbol{\mu}}(\boldsymbol{\lambda}; \mathbf{s}_t)$, with per-step loss $c(\mathbf{s}_t, \mathbf{a}_t) - \boldsymbol{\lambda}^{\top} \mathbf{a}_t$; at non-perturbation steps the action follows the frozen base policy $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ with loss $c(\mathbf{s}_t, \mathbf{a}_t)$. The perturbed action produces a counterfactual next-state at which the value target $V^{\boldsymbol{\theta}}$ is evaluated, and K common-random-number draws $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K$ per step reduce variance.

Perturbation schedule. Let M be the number of steps between consecutive perturbations. When the value target is estimated by Monte Carlo rollout continuation, later perturbations at $t+M, t+2M, \ldots$ contaminate the step-t target by $\sum_{j \geq 1} \gamma^{jM} \mathbb{E}_{\boldsymbol{\lambda}}[A^{\boldsymbol{\theta}}(\mathbf{s}_{t+jM}, \mathbf{a}(\boldsymbol{\lambda}, \mathbf{s}_{t+jM}; \boldsymbol{\mu}))]$. What matters for the gradient is the action sensitivity $\|\partial \mathbf{s}_{t+jM} / \partial \mathbf{a}_t\|$, which decays once jM exceeds the closed-loop mixing time τ_{mix}: a stabilizing policy drives the state back to target after a perturbation, breaking the sensitivity chain (for our lost-sales example with a base-stock policy, τ_{mix} is a few, e.g. 5, periods). The estimator uses $\lfloor T/M \rfloor \cdot K$ samples per trajectory, so variance scales as $O(M/(TK))$ while the later-perturbation bias is negligible for $M \gg \tau_{\text{mix}}$; a practical default is $M \approx \tau_{\text{mix}}$.

Two-stage procedure. Stage 1 trains $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ to minimize the exo-MDP cumulative cost by any method yielding a deterministic policy and an evaluable $V^{\boldsymbol{\theta}}$; we use DirectBackprop/HDPO Madeka et al. (2022); Alvo et al. (2023), with $V^{\boldsymbol{\theta}}$ from a critic or rollout continuation. Stage 2 fixes $(\boldsymbol{\pi}_{\boldsymbol{\theta}}, V^{\boldsymbol{\theta}})$ and optimizes $\boldsymbol{\mu}$ by SGD on the empirical version of (4), backpropagating through the differentiable argmin and transition. Because Stage 2 only requires a frozen $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ and a way to evaluate $V^{\boldsymbol{\theta}}$, the surrogate may be fitted to any deployed policy without modification.

4.2 Network Architecture and Design Choices



(a) Separate networks (b) Shared backbone (c) Single network

Figure 1: Three architecture options for the advantage surrogate. (a) Separate policy and curvature networks. (b) Shared encoder with specialized heads. (c) Single network outputting both decision and curvature parameters.

The surrogate requires two outputs per state: the anchor $\mathbf{a}^*(\mathbf{s})$ and the curvature parameters $\boldsymbol{\psi}(\mathbf{s})$ (or $B(\mathbf{s})$ for a quadratic). Three architectures are natural (Figure 1): (a) *separate networks* for policy and curvature, the most modular option and the natural choice for two-stage training and for retrofitting surrogates to pre-trained policies; (b) a *shared backbone with specialized heads*, parameter-efficient; and (c) a *single network* outputting both, most compact but fully coupling the two tasks. We use separate networks throughout.

For the surrogate parameterization, the anchored quadratic $\psi_{\boldsymbol{\mu}}(\mathbf{u};\mathbf{s}) = \frac{1}{2}\mathbf{u}^\top B(\mathbf{s};\boldsymbol{\mu})\mathbf{u}$ is the simplest option, although problems with asymmetric or heavy-tailed advantage geometry admit any strictly convex family (piecewise-quadratic, quadratic-plus-ICNN) whose conjugate is evaluable. The perturbation distribution $p(\boldsymbol{\lambda})$ controls the dual-space region on which $f^*_{\boldsymbol{\mu}}$ is identified, and we default to an isotropic Gaussian with σ calibrated to the dual prices the downstream system produces, or matched to the distribution of ∇g if the downstream objective g is known at training time. For quadratic $f_{\boldsymbol{\mu}}$ the perturbed argmin $\nabla f^*_{\boldsymbol{\mu}}(\boldsymbol{\lambda};\mathbf{s}) = \mathbf{a}^*(\mathbf{s}) + B(\mathbf{s})^{-1}\boldsymbol{\lambda}$ is closed-form with an OptNet-style backward pass Amos & Kolter (2017); general strictly convex $f_{\boldsymbol{\mu}}$ use the differentiable layer of Agrawal et al. (2019). When $V^{\boldsymbol{\theta}}$ is a learned critic the Bellman target is one forward pass; without a critic it is a Monte Carlo rollout continuation under $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ Madeka et al. (2022).

5 Experiments

5.1 Setup

We use the exo-MDP formulation of stationary multi-period inventory management with lost sales and zero lead time (Maggiar et al., 2025, Section 4.2): endogenous on-hand inventory y_t, exogenous demand history $\{d_{t+s}\}_{s=1,\dots,H}$, static economic parameters (price p, cost c, holding cost h, lost-sale penalty b), and order action $a_t \in \mathbb{R}_{\geq 0}$, with $y_{t+1} = (y_t + a_t - d_t)^+$ and $c_t(s,a) = ca + b(d_t - y - a)^+ + h(y + a - d_t)^+ - p\min(d_t, y + a)$. The problem admits closed-form optima: the policy is an order-up-to level at a demand quantile,

$$a^*_t = \left(F_t^{-1}\left(\frac{c_u}{c_u + c_o}\right) - y\right)^+ \tag{7}$$

$$\hat{G}_t(a,s) = c_o\mathbb{E}\left[(a + y - D_t)^+\right] + c_u\mathbb{E}\left[(D_t - (a+y))^+\right] + K \tag{8}$$

where $c_u = p - c + b$ and $c_o = h$ are the underage and overage costs, K is a constant, and F_t^{-1} is the inverse cumulative distribution function of the demand distribution D_t. The quantity $\hat{G}_t$ is the single-period newsvendor cost, and the resulting advantage $A^{\boldsymbol{\theta}} = Q^{\boldsymbol{\theta}} - V^{\boldsymbol{\theta}}$ is derived in Appendix A.6.

We compare two anchored surrogates, motivated by the asymmetry of underage versus overage costs in inventory management: the **symmetric quadratic** $f_\mu(a;\mathbf{s}) = \frac{1}{2}B(\mathbf{s})(a - a^*(\mathbf{s}))^2$ and the **asymmetric (piecewise) quadratic** $f_{\boldsymbol{\mu}}(a;\mathbf{s}) = \frac{1}{2}B_+(\mathbf{s})[(a-a^*)^+]^2 + \frac{1}{2}B_-(\mathbf{s})[(a-a^*)^-]^2$, both with positive curvatures and closed-form conjugates. We report two-stage training at perturbation cadences $M \in \{1, 10, 100\}$.

Training. We generate $N_{\text{train}} = 40{,}000$ products using the same procedure as in (Maggiar et al., 2025, Section 4.2) (32 initialization + 100 rollout periods), whose Gamma-distributed demand admit the closed forms (7)–(8). Stage-1 trains the policy network and Stage-2 the curvature network on top of it, both for 500 epochs with separate dilated-CNN policy and curvature heads (full architecture and hyperparameters in Appendix A). The single-period cost $\hat{G}$ is asymptotically linear with slopes $-c_u$ and c_o, which sets the natural scale for the dual prices (a discount is bounded by the margin c_u, a holding or capacity surcharge by c_o); we draw the training perturbation $\lambda_t \sim \frac{1}{2}(c_o - c_u) + \sigma(c_o + c_u)\mathcal{N}(0,1)$ with $\sigma = 0.33$ at every M-th post-initialization period and $\lambda_t = 0$ otherwise, centered and scaled so its bulk falls inside $[-c_u, c_o]$, and cap the surrogate action at five times the mean demand as a soft truncation of the effective support.

Evaluation. We evaluate on $N_{\text{test}} = 100{,}000$ fresh products. The *unconstrained* evaluation rolls the policy at $\boldsymbol{\lambda} \equiv 0$ over 500 periods and regresses the trained order quantities against the non-omniscient closed-form newsvendor optimum (7) (the CDF F_t in (7) fit to the mean and variance of the same H-period demand history the policy sees). The *constrained* evaluation realizes the shared-capacity use case of Section 1: a transient storage-capacity limit on units held, common across products, is dualized to a single scalar shadow price λ that surcharges the inventory position held under the constraint. We deliver it as the dual perturbation $\delta = -\lambda$ (a shared negative price), so the surrogate-induced action is the closed-form perturbed argmin $\nabla f^*_{\boldsymbol{\mu}}(-\lambda;\mathbf{s})$, which orders less. Because the position recovers to the unconstrained optimum as soon as the constraint lifts, one period

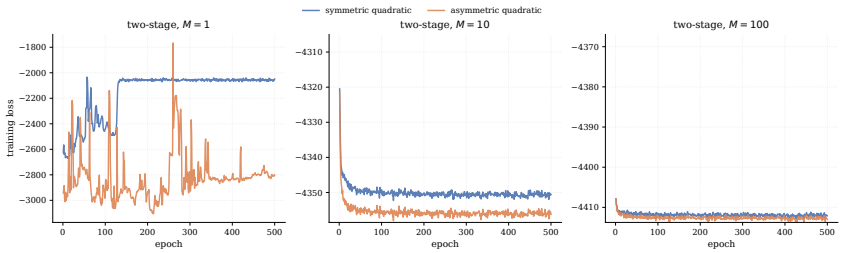


Figure 2: Stage-2 training loss for the symmetric and asymmetric quadratic surrogates, columns $M \in \{1, 10, 100\}$.

captures the effect; we evaluate a single constrained period from a null initial inventory and report both the realized reward $\overline{R}$ and the mean per-product on-hand position, sweeping $\lambda \in \{1, 2, 4, 8, 16\}$.

5.2 Results

5.2.1 Training

Figure 2 reports the Stage-2 rollout loss for $M \in \{1, 10, 100\}$. At $M = 1$ a non-zero $\boldsymbol{\lambda}$ is injected at every rollout step, leaving no buffer over which the inventory dynamics relax, and the loss is correspondingly noisier; stability improves rapidly with M because the anchor head is frozen and only the curvature head co-adapts. Figure 3 shows the surrogate fits at $M = 10$: the two surrogates agree near the anchor where the local quadratic is unconstrained, the asymmetric variant is tighter on the underage tail (low critical ratio) on most products, and neither captures the closed-form's eventual linear asymptotes, as expected.

5.2.2 Unconstrained Evaluation

The unconstrained evaluation rolls out each trained policy at $\boldsymbol{\lambda} \equiv 0$ ($N_{\text{test}} = 100{,}000$ products, 500 periods), reporting the mean per-period reward $\overline{R}$ and the slope and R^2 of regressing trained order quantities against the closed-form non-omniscient optimum (7) over the converged window $t \in [100, 109]$ (slope and R^2 near 1 is the target). At $\boldsymbol{\lambda} = 0$ the curvature head is inactive, so the action reduces to the frozen Stage-1 anchor $\nabla f^*_{\hat{\boldsymbol{\mu}}}(0; \mathbf{s}) = \mathbf{a}^*(\mathbf{s})$ irrespective of the surrogate shape. This evaluation thus exercises the anchor alone; its purpose is to confirm the anchor is correct before the surrogate curvature is tested in the constrained sweep. The anchor reproduces the optimum at slope 0.93, $R^2 = 0.90$, the ceiling achievable from a 32-period history (Table 1).

Table 1: Unconstrained evaluation ($\boldsymbol{\lambda} \equiv 0$). $\overline{R}$ is the mean per-period reward; slope and R^2 regress trained order quantities against the closed-form non-omniscient newsvendor optimum over the window $t \in [100, 109]$.

Mode	$\overline{R}$	Slope	R^2
Stage-1 anchor	4,593.69	0.93	0.90

The trained-vs.-optimal scatter (Figure 4) is a tight diagonal band, confirming the anchor tracks the optimum across the population.

5.2.3 Constrained Evaluation

The constrained evaluation probes the conjugate response to a shared capacity price: the constrained-period action is $\nabla f^*_{\hat{\boldsymbol{\mu}}}(-\lambda; \mathbf{s}) = \mathbf{a}^*(\mathbf{s}) + \nabla \psi^*_{\hat{\boldsymbol{\mu}}}(-\lambda; \mathbf{s})$, which trims the order below the anchor by an amount set by the surrogate's curvature. We compare the trained asymmetric quadratic surrogate against two references: a *passive anchor* that ignores the price and plays $\mathbf{a}^*(\mathbf{s})$ regardless (isolating the value of responding to the constraint), and the *non-omniscient optimal*, the exact penalized-newsvendor order-up-to $z_\lambda = F^{-1}\big((c_u - \lambda)/(c_u + h)\big)$ from the same 32-period demand fit the

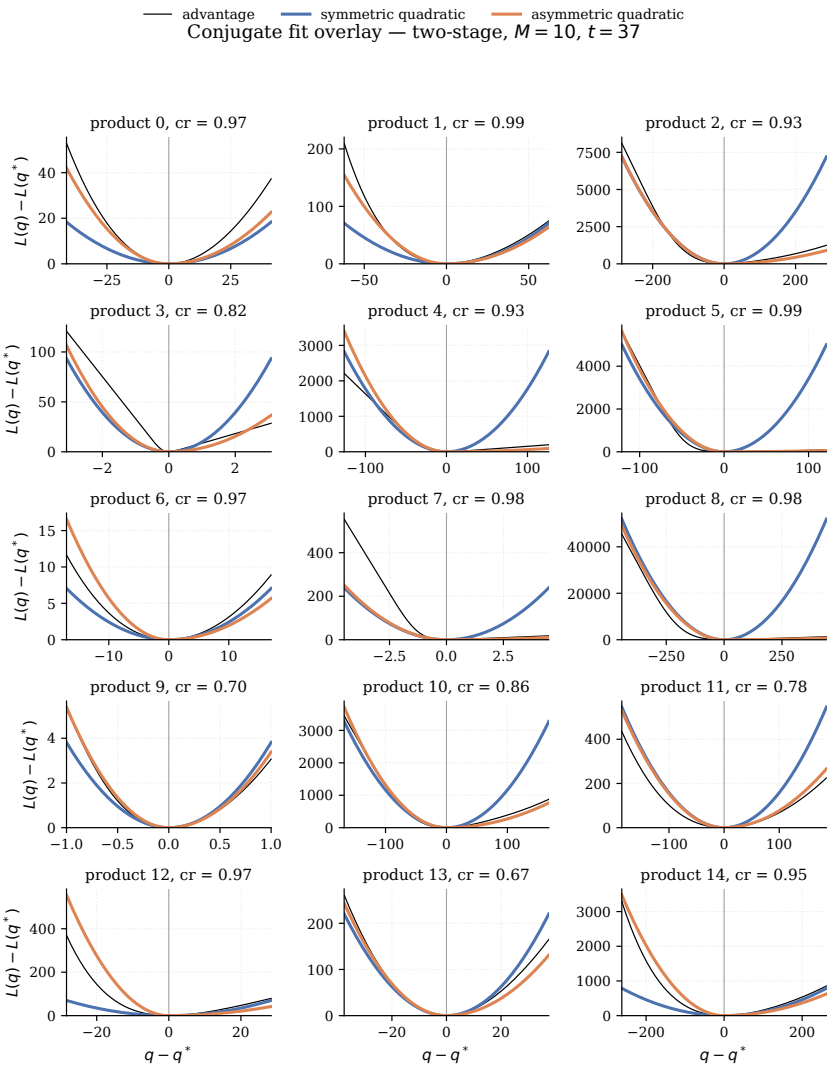


Figure 3: Conjugate fits for the first 15 products of the test set, two-stage training at $M = 10$. Each panel overlays the true one-shot-deviation advantage $A^{\boldsymbol{\theta}}$ (thin black, centered at the anchor; the single-period cost on the underage arm and the multi-period renewal drain on the overage arm, Appendix A.6), with the trained symmetric (blue) and asymmetric (orange) quadratic surrogates. Titles report the product's critical ratio.

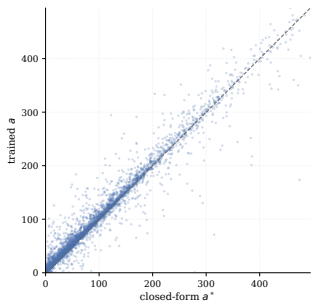


Figure 4: Regressed actions vs. closed-form non-omniscient newsvendor optimum, $\boldsymbol{\lambda} \equiv 0$, periods $t \in [100, 109]$.

policy sees (derived in Appendix A.5). Table 2 reports the mean per-period realized reward and the mean per-product on-hand position; Figure 5 plots both against λ.

The surrogate tracks the optimal response closely and stays far above the passive anchor. As λ rises, the passive anchor holds a fixed position and pays a linearly growing surcharge, while the surrogate draws its held inventory down toward the optimal reduction, recovering most of the achievable reward. The learned conjugate thus lets an unchanged upstream policy honor a shared capacity price it was never trained against, deferring the buy on the products where the marginal value of the held unit is lowest.

Table 2: Shared-capacity evaluation. For each shadow price λ: mean per-period realized reward $\overline{R}$ and mean per-product on-hand position.

	$\lambda = 0$	$\lambda = 1$	$\lambda = 2$	$\lambda = 4$	$\lambda = 8$	$\lambda = 16$
Mean per-period reward $\overline{R}$						
passive anchor	4,598.3	4,438.3	4,278.3	3,958.4	3,318.6	2,038.9
asymmetric quadratic	4,598.3	4,438.4	4,286.8	4,005.6	3,516.9	2,752.6
non-omniscient optimal	4,611.7	4,445.5	4,293.8	4,018.3	3,551.7	2,842.8
Mean on-hand position						
passive anchor	160.0	160.0	160.0	160.0	160.0	160.0
asymmetric quadratic	160.0	154.1	148.3	136.9	117.3	89.6
non-omniscient optimal	174.7	155.9	144.7	128.4	105.0	75.5

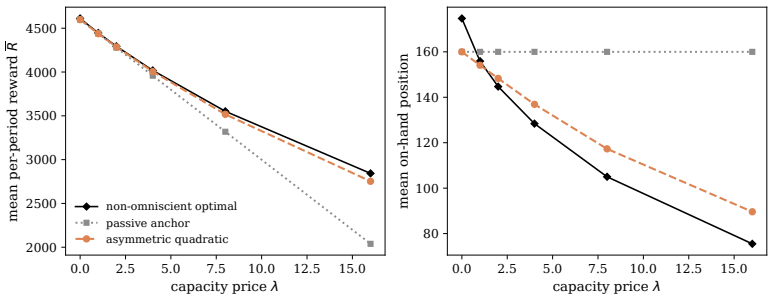


Figure 5: Shared-capacity evaluation vs. the capacity price λ.

6 Conclusion

We formulated structured advantage-function distillation as a Fenchel-Young matching problem: a strictly convex surrogate $f_{\boldsymbol{\mu}}$ is fit so that its conjugate $f^*_{\boldsymbol{\mu}}$ tracks the conjugate of the true action-value function on the support of an exogenous dual-perturbation distribution, approximating Q where downstream consumers query it rather than pointwise. This removes the under-identification of naïve differentiable-argmin training, makes the perturbation distribution a design parameter spanning the dual-price range of interest, and makes the surrogate parameterization a modeling lever; the anchored factorization $f_{\boldsymbol{\mu},\boldsymbol{\theta}} = \psi_{\boldsymbol{\mu}}(\mathbf{a} - \mathbf{a}^*(\mathbf{s};\boldsymbol{\theta});\mathbf{s})$ separates policy from curvature. The two-stage protocol, which fits the surrogate to a frozen upstream policy, solves the FY problem by construction. A natural extension is to consider surrogates with built-in structural constraints, such as M-convexity for inventory-placement problems, where the closed-form conjugate machinery should generalize provided the minimizers admit a tractable parameterization.

References

Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J Zico Kolter. Differentiable convex optimization layers. *Advances in neural information processing systems*, 32, 2019.

Matias Alvo, Daniel Russo, Yash Kanoria, and Minuk Lee. Deep reinforcement learning for inventory networks: Toward reliable policy optimization, 2023. arXiv preprint 2306.11246.

Brandon Amos and J. Zico Kolter. OptNet: Differentiable optimization as a layer in neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 136–145. PMLR, 2017.

Mathieu Blondel, André F. T. Martins, and Vlad Niculae. Learning with Fenchel–Young losses. *Journal of Machine Learning Research (JMLR)*, 21(35):1–69, 2020.

William Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. John Wiley & Sons, New York, 2nd edition, 1971.

Shixiang Gu, Timothy Lillicrap, Ilya Sutskever, and Sergey Levine. Continuous deep q-learning with model-based acceleration. In *International Conference on Machine Learning*, pp. 2829–2838. PMLR, 2016.

Dhruv Madeka, Kari Torkkola, Carson Eisenach, Anna Luo, Dean P. Foster, and Sham M. Kakade. Deep inventory management, 2022. arXiv preprint 2210.03137.

Alvaro Maggiar, Sohrab Andaz, Akhil Bagaria, Carson Eisenach, Dean Foster, Omer Gottesman, and Dominique Perrault-Joncas. Structure-informed deep reinforcement learning for inventory management, 2025. arXiv preprint 2507.22040.

Hongzi Mao, Shaileshh Bojja Venkatakrishnan, Malte Schwarzkopf, and Mohammad Alizadeh. Variance reduction for reinforcement learning in input-driven environments. In *International Conference on Learning Representations (ICLR)*, 2019.

Kazuo Murota. *Discrete Convex Analysis*. Discrete Mathematics and Applications. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2003.

R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 32 of *Proceedings of Machine Learning Research*, pp. 387–395. PMLR, 2014.

Sean R. Sinclair, Felipe Frujeri, Ching-An Cheng, Luke Marshall, Hugo Barbalho, Jingling Li, Jennifer Neville, Ishai Menache, and Adith Swaminathan. Hindsight learning for MDPs with exogenous inputs. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31877–31914. PMLR, 2023.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Paul H. Zipkin. *Foundations of Inventory Management*. McGraw-Hill, Boston, 2000.

A Experimental Details

This appendix documents the data-generating process, network architecture, optimizer, and surrogate parameterizations used in the experiments of Section 5, in enough detail to reproduce the reported numbers.

A.1 Data-Generating Process

Each product is a tuple of static economic parameters (p, c, h, b) paired with a Gamma demand process D_t. Parameters are sampled independently across products as

$$\begin{aligned} \bar{d} &\sim \text{Exp}(\text{mean} = 100), & \text{CV} &\sim \text{U}(0, 1), \\ p &\sim \text{Exp}(\text{mean} = 100), & c \mid p &\sim \text{U}(0, p), \\ h &\sim \text{Exp}(\text{mean} = 5), & b &\sim \text{U}(0, 10). \end{aligned}$$

Demand at every period is drawn i.i.d. from $\text{Gamma}(\alpha, \beta)$ with $\alpha = 1/\text{CV}^2$ and scale $\beta = \bar{d} \cdot \text{CV}^2$, so that $\mathbb{E}[D_t] = \bar{d}$ and $\text{Var}(D_t) = \bar{d}^2\text{CV}^2$. The static parameters are constant across periods within a product. The training set has $N_{\text{train}} = 40{,}000$ products generated over $32 + 100 = 132$ periods (32 used for initialization, 100 for the rollout). The test set has $N_{\text{test}} = 100{,}000$ products generated over $32 + 500 = 532$ periods.

The endogenous on-hand inventory is initialized at $y_0 \sim 1.5\,\bar{d}_{\text{init}} \cdot \text{U}(0, 1)$, where $\bar{d}_{\text{init}}$ is the mean of the first 32 demand observations of the product, and is then evolved deterministically by $y_{t+1} = (y_t + a_t - d_t)^+$.

A.2 Network Architecture

We detail the shared head architecture used for the three options of Section 4.2 (Figure 1). The policy and curvature heads share the same architecture but live in separate networks (no parameter sharing). Each head is a temporal-convolutional encoder followed by a small MLP decoder, applied to per-product features:

- *Series features.* The history of demand over a backward window $H = 32$, fed as a single-channel time series.
- *Static features.* Five scalars: (p, c, h, b, y_t) broadcast across the convolutional output.
- *Encoder.* A dilated causal convolutional stack with dilations $\{1, 2, 4, 8, 16\}$ (i.e. 2^i for $i = 0, \ldots, \log_2 H - 1$), output channel width 8, two encoding layers of hidden width 32, embedding dimension 32.
- *Decoder.* A linear MLP (no hidden layers) of width 32, mapping the encoder embedding to the head's output.

The policy head outputs a single scalar passed through a softplus to enforce non-negativity, giving the anchor $a^*(\mathbf{s}; \boldsymbol{\theta}) \geq 0$. The curvature head outputs $d_{\boldsymbol{\mu}}$ scalars, each passed through $\text{ELU}(\cdot) + 1 + 10^{-4}$, where $d_{\boldsymbol{\mu}}$ is the surrogate's parameter dimension. The decoder's last linear layer is initialized with zero weights and bias 0.5, so that at initialization every curvature output is approximately $\text{ELU}(0.5) + 1 \approx 0.97$. This puts both heads in a well-conditioned regime at the start of training.

A.3 Surrogate Parameterizations

Both surrogates are anchored (the surrogate's argmin is the policy network's anchor $a^*(\mathbf{s}; \boldsymbol{\theta})$); we describe them as functions of the deviation $x = a - a^*(\mathbf{s}; \boldsymbol{\theta})$.

- *Symmetric quadratic* ($d_{\boldsymbol{\mu}} = 1$). $f_{\boldsymbol{\mu}}(x) = \frac{1}{2}B(\mathbf{s})\,x^2$, with conjugate response $\nabla f^*_{\boldsymbol{\mu}}(\lambda) = \lambda/B$.
- *Asymmetric quadratic* ($d_{\boldsymbol{\mu}} = 2$). $f_{\boldsymbol{\mu}}(x) = \frac{1}{2}B_+(\mathbf{s})\,[x_+]^2 + \frac{1}{2}B_-(\mathbf{s})\,[x_-]^2$, with curvature head outputs $(B_+(\mathbf{s}), B_-(\mathbf{s}))$ and conjugate response $\nabla f^*_{\boldsymbol{\mu}}(\lambda) = \lambda/B_+$ for $\lambda > 0$ and λ/B_- otherwise.

Each curvature scalar is clamped to $\geq 10^{-8}$ before use. The action passed to the cost function is then $a_t = \text{clip}(a^*(\mathbf{s}; \boldsymbol{\theta}) + \nabla f^*_{\boldsymbol{\mu}}(\lambda_t; \mathbf{s});\ 0,\ 5\,a^*(\mathbf{s}; \boldsymbol{\theta}))$. The upper cap at $5\,a^*$ mirrors the soft truncation discussion of Section 5 and bounds the action's response to dual prices outside $[-c_u, c_o]$.

A.4 Training Loop

Both stages share the following optimizer and rollout settings:

- *Optimizer.* AdamW with learning rate 3×10^{-3} on the active parameter group, gradient-norm clipping at 1.0, no weight decay tuning beyond AdamW defaults.
- *Rollout.* Reset period 32; rollout length 100 periods of differentiable backpropagation, with terminal value $V_T(s) = y_T \cdot c$ added to the discounted return. Discount factor $\gamma = 1$. Mini-batches of 2,000 products, drawn without replacement.
- *Epochs.* 500 epochs for every cell. EMA shadow weights start at epoch 10 with decay 0.995; evaluation always uses the EMA snapshot of the final epoch.
- *Parameter groups.* Stage-1 trains the policy head only ($\lambda \equiv 0$ during the rollout). Stage-2 trains the curvature head only with the Stage-1 policy frozen by exclusion from the optimizer's parameter list.

The Stage-1 EMA snapshot is reused as the policy initialization for both Stage-2 surrogates, so the unconstrained policy is held identical across rows in Tables 1–2.

A.5 Shared-Capacity Constrained Example

This appendix specifies the shared-capacity constrained evaluation of Section 5.2.3 and derives its exact optimal benchmark.

Setting. A transient storage-capacity limit is common to all products: in the constrained period a bound $\sum_j z_t^j \le C_t$ caps the total inventory held, where $z = y + a$ is the post-order position. Dualizing this coupling constraint attaches a single scalar shadow price $\lambda \ge 0$, identical across products, that surcharges the held position z (the dual term is $-\lambda z$ in the per-period reward). The price is shared, and products respond heterogeneously only through their own economics. We evaluate a single constrained period from a null initial inventory $y = 0$ because the position recovers to the unconstrained base stock as soon as the constraint lifts, thus one period suffices to capture the effect.

Exact optimal order-up-to. With purchase cost charged at order, inventory carried at holding cost h, and terminal salvage at cost, the effective per-unit overage in the single-period newsvendor is h and the underage is $c_u = p - c + b$ (as in (8)). Adding the shared surcharge $-\lambda z$ makes the constrained-period objective

$$\max_{z \ge 0} \left\{ \; - c_u \, \mathbb{E}[(D - z)^+] - h \, \mathbb{E}[(z - D)^+] - \lambda z \; \right\},$$

whose first-order condition $c_u - (c_u + h)F(z) - \lambda = 0$ gives the penalized order-up-to

$$z_\lambda = F^{-1}\left(\frac{c_u - \lambda}{c_u + h} \right)^+, \tag{9}$$

where the critical ratio is clipped at zero: when $\lambda \ge c_u$ the marginal held unit cannot cover the surcharge and the optimum is not to buy ($z_\lambda = 0$). The *non-omniscient optimal* of Table 2 evaluates (9) with F the Gamma fit to the product's 32-period demand history (the same information the policy sees).

Surrogate response and delivery. The surrogate's constrained-period action is the anchored perturbed argmin $\mathbf{a}^*(\mathbf{s}) + \nabla \psi^*_{\boldsymbol{\mu}}(-\lambda; \mathbf{s})$, delivered by injecting the shared dual value $\lambda_t = -\lambda$ at the constrained period (zero otherwise). The negative price drives the response down the underage arm (λ / B_- for the asymmetric quadratic), trimming the order below the anchor. The passive anchor plays $\mathbf{a}^*(\mathbf{s})$ and pays $-\lambda\, z$ on its unadjusted position, isolating the value of responding to the price.

A.6 True Advantage of the Lost-Sales Base-Stock Policy

This appendix derives the true advantage $A^{\boldsymbol{\theta}}(\mathbf{s}, a) = Q^{\boldsymbol{\theta}}(\mathbf{s}, a) - V^{\boldsymbol{\theta}}(\mathbf{s})$ of the base-stock policy in the lost-sales, zero-lead world of Section 5.1, and shows how it differs from the single-period newsvendor cost $\hat{G}$ of (8). Minimizing $\hat{G}$ recovers the correct stationary base-stock level Zipkin (2000), but the advantage measures something different: the cost of a *one-shot* deviation that perturbs the order-up-to level in a single period and then reverts to the optimum thereafter. As we show below,

$A^{\boldsymbol{\theta}}$ therefore coincides with $\hat{G}$ only where a deviation carries no multi-period consequence. Write the deviation from the optimal order-up-to level as $e = a - a^*$ (equivalently the post-order position is $z^* + e$ with $z^* = F^{-1}(c_u/(c_u + c_o))$), and let $g(e) = \hat{G}(z^* + e) - \hat{G}(z^*)$ be the single-period cost deviation, so $g(0) = 0$, $g'(0) = 0$, with asymptotic slopes $-c_u$ (under) and c_o (over).

Mechanism. By definition $Q^{\boldsymbol{\theta}}(\mathbf{s}, a) = c(\mathbf{s}, a) + \gamma\, \mathbb{E}[V^{\boldsymbol{\theta}}((z - D)^+)]$ with $z = y + a$. The continuation value $V^{\boldsymbol{\theta}}(y')$ is linear in y' for $y' \le z^*$: the policy is already at its optimum, so only the purchase term $-c\, y'$ carried by a lower starting inventory survives, and this linear piece combines with $c(\mathbf{s}, a)$ to give exactly $\hat{G}$. For $y' > z^*$ the non-negativity of orders ($a \ge 0$) prevents the policy from ordering down, so it is stuck above target and pays excess holding until demand drains the surplus; $V^{\boldsymbol{\theta}}$ bends strictly convex above the linear extrapolation. A deviation with $z \le z^*$ keeps $y' = (z - D)^+ \le z^*$ in every future period, so it stays on the linear branch and $\hat{G}$ is exact; a deviation with $z > z^*$ can leave $y' > z^*$, entering the convex branch and adding a multi-period cost on top of $\hat{G}$.

Underage arm ($e < 0$). An under-order is corrected by the very next reorder (a single period), and lost sales leave no carryover, so the advantage is exactly the single-period cost,

$$A^-(e) = \hat{G}(z^* + e) - \hat{G}(z^*) = g(e),$$

Here $\hat{G}$ coincides with $A^{\boldsymbol{\theta}}$.

Overage arm ($e > 0$). The excess e can only be worked off by cumulative demand $S_j = \sum_{i<j} D_{t_0+i}$. Summing the single-period overage cost over the excess still outstanding at each future period gives the renewal drain

$$A^+(e) = \sum_{j \ge 0} \mathbb{E}\big[\, g\big((e - S_j)^+\big)\,\big], \tag{10}$$

a genuinely multi-period cost (roughly e/μ periods) that grows super-linearly. By the elementary renewal theorem the renewal function satisfies $M(x) \to x/\mu$ as $x \to \infty$ Feller (1971), which yields the asymptotic quadratic limit $A^+(e) \to \frac{c_o}{2\mu}\, e^2$; a compact operating-range approximation is $A^+(e) \approx \beta\, g(e) + \frac{1}{\mu} \int_0^e g(u)\, du$ with $\beta = \frac{1}{2}(1 + \mathrm{cv}^2)$.