Peak-Shaving via Contextual Markov Decision Processes

Constrained Optimization with
Time Series Forecasting

Sarah Trausner, Simon Hirlaender

March 30, 2026

IDA Lab — University of Salzburg

Motivation

The General Problem

What if you operate in an environment that

changes *and is* **not under your full control***?*

The Setting

- Changing exogenous part → **forecast**
- **Uncertain** controllable variables → **forecast**
- **Finite horizon**

The Challenge

- Hard **safety constraint**
- Uncertainty in *both* the environment **and** your own actions
- **Maximise performance** while **staying safe**

This Problem Is Everywhere

	Exogenous Drift	Safety Constraint	Uncertain Controls
Beam orbit	Temperature	Beam within bounds	Magnet hysteresis
Energy mgmt.	Other machines	15-min mean $\leq \theta$	Load power $p_j \pm \sigma_j$

The **physics changes** — temperature, magnets, plasma, loads —
but the **mathematical structure is identical**.

Overview: The Peak-Shaving Problem

- Electricity billing: includes availability price - highest mean over a fixed **time-interval**
- Peaks in the load $\rightarrow$ inflate mean values $\rightarrow$ higher costs
- **Goal:** Apply dynamic asset-switching on top of an unchangeable baseload to keep the total mean power tightly below a threshold θ.

$$P_{\text{total}}(t) = \underbrace{P_{\text{base}}(t)}_{\text{stochastic baseload}} + \underbrace{\mathbf{b}(a_t)^\top \mathbf{p}}_{\text{controllable assets}}$$

Core Goal

Only pay for the energy that we are actually able to use up and not overpay because of a few high peaks.

Model predictive control

The entire optimisation problem (forecasting, estimating and control) is solved in each step and only the first action is applied.

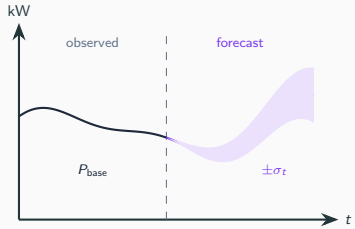
Modelling

The Baseload as External Context

- **Baseload** $P_{\text{base}}(t)$: external $\rightarrow$ prediction
- We forecast it using Bayesian Structural Time Series:

$$P_{\text{base}}(t) \sim \mathcal{N}\left(\mu_t,\ \sigma^2_{\text{base},t}\right)$$

- The forecast provides both a **mean** μ_t and a **variance** σ_t^2 — quantifying our uncertainty
- Controllable load powers p_j are also estimated with variance σ_j^2



From MDP to Constrained MDP (CMDP)

Standard MDP

Single objective — maximise cumulative reward:

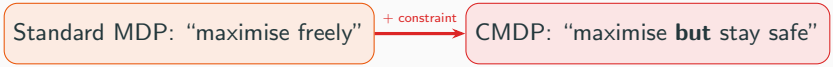
$$\max_{\pi} \mathbb{E}_{\pi}[\sum_t r(s_t, a_t)]$$

No side constraints.

Constrained MDP (CMDP)

Adds a **constraint** on cumulative cost:

$$\max_{\pi} \mathbb{E}_{\pi}[\sum_t r] \quad \text{s.t.} \quad \mathbb{E}_{\pi}[\sum_t c_k] \leq d_k$$



unconstrained: trivially “turn everything on” → violates threshold → constraint makes the problem non-trivial

The Problem as a Contextual CMDP (CCMDP)

Formal Definition

State:	$s_t = (t,\ C_t)$	$C_t = \sum_{\tau=1}^{t} P_{\text{total}}(\tau)$ (cumulative load)
Action:	$a_t \in \mathcal{A}_{\text{safe}} \subseteq \{0, \ldots, 2^N - 1\}$	binary ON/OFF allocation
Transition:	$C_{t+1} = C_t + P_{\text{base}}(t) + \ell(a_t)$	$\ell(a) = \mathbf{b}(a)^\top \mathbf{p}$
Objective:	$\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{H} \ell(a_t)\right]$	maximise production
Constraint:	$\Pr(C_H > \theta) \leq \varepsilon$	limit violation probability

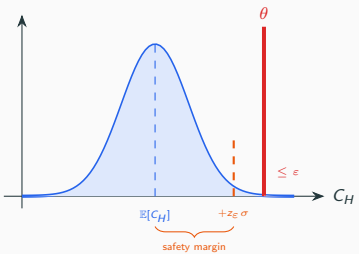
"Contextual": The baseload $P_{\text{base}}(t)$ acts as an *exogenous context* — it affects transitions but is not controlled by the agent.

The Chance Constraint

Action-dependent variance:

$$\mathrm{Var}(C_H) = \sum_t \left(\sigma^2_{\mathrm{base},t} + \sum_{j \in \mathrm{ON}(a_t)} \sigma_j^2 \right)$$

Turning on *uncertain* loads increases total noise.



Deterministic equivalent (Gaussian assumption):

$$\boxed{\mathbb{E}[C_H] + z_\varepsilon \cdot \sqrt{\mathrm{Var}(C_H)} \leq \theta}$$

where $z_\varepsilon = \Phi^{-1}(1-\varepsilon)$ (Ex.: $z_{0.05} = 1.645$)

Full derivation in the appendix.

Solution Approaches

Approach 1: Deterministic DP ($\sigma = 0$)

Assumption: Perfect forecasts - no uncertainty. Transitions are deterministic.

Bellman recursion (backward induction):

$$V_t(C_t) = \min_{a_t \in \mathcal{A}_{\text{safe}}} V_{t+1}(C_t + \ell(a_t) + \mu_t)$$

Terminal cost:

$$g_t(C_t) = \begin{cases} -C_t & C_t \leq \theta \\ \lambda(C_t - \theta) & C_t > \theta \end{cases}$$

with $\lambda = 1000$ (catastrophic penalty).

How it works

1. At $t = H$: evaluate terminal cost
2. Step backwards: for each C_t, pick the action minimising V_{t+1}
3. At $t = 0$: read off optimal action

✓ Exact for deterministic case
✓ Fast
✗ Ignores all uncertainty

Approaches 2 & 3: Stochastic and Robust DP

Both use the **1D state** C_t but handle noise differently:

Stochastic DP (risk-neutral)

Risk-neutral: Focuses purely on the average expected outcome, ignoring variance.

Gauss-Hermite ($K=5$): Approximates the integral over the normal distribution using 5 weighted sample points.

$$V_t(C_t) = \min_{a_t} \sum_{k=1}^{K} w_k \, V_{t+1}(\cdots + \sigma \cdot x_k)$$

✓ Accounts for noise *on average*

✗ Ignores **tail events** (extreme, low-probability outcomes at the edges of the distribution)

Robust DP (worst-case)

Assumes nature acts adversarially.

The max calculation: It evaluates all possible noise deviations δ and assumes the one that maximizes future cost will occur.

$$V_t(C_t) = \min_{a_t} \max_{\delta \in [-3\sigma, +3\sigma]} V_{t+1}(\cdots + \delta)$$

✓ Robust to adversarial noise

✗ Far too conservative for MPC, as the worst-case rarely happens continuously

Both use penalty-based constraint handling (approximate). Neither can guarantee $\Pr[\text{violation}] \leq 5\%$.

Approach 4: Exact Chance-Constrained DP (SOCP / CC-DP)

Key idea: Augment the state to 2D — track cumulative variance alongside cumulative load:

$$\boxed{s_t = (C_t,\ V_t)} \quad \text{where} \quad V_{t+1} = V_t + \sigma^2_{\text{base},t} + \sum_{j \in \text{ON}(a_t)} \sigma_j^2$$

Terminal cost evaluates the chance constraint **exactly**:

$$g_H(C_H, V_H) = \begin{cases} -C_H & \text{if } C_H + z_\varepsilon \sqrt{V_H} \le \theta \\ \lambda \cdot (C_H + z_\varepsilon \sqrt{V_H} - \theta) & \text{otherwise} \end{cases}$$

✓ **SOCP:** Assumes noise is perfectly symmetric (fast).

✓ **CC-DP:** Handles skewed, unpredictable noise by checking multiple scenarios (robust).

✓ ε has a **true probabilistic meaning.**

✗ **Curse of Dimensionality:**
Because we must discretely evaluate a 2D grid, a state space of 500×500 creates 250**k** **discrete points** that must be computed at every timestep!

Approach 5: Reinforcement Learning (DQN)

DQN (Deep Q-Network): Uses a neural network to map continuous states to the long-term expected reward $Q(s, a)$ of taking action a, completely avoiding transition matrices.

The Q-Learning Update Formula:
Shifts the old prediction towards the immediate reward r plus the max discounted future reward.

$$Q(s,a) \leftarrow \underbrace{Q(s,a)}_{\text{old estimate}} + \alpha \Big[\underbrace{r + \gamma \max_{a'} Q(s',a')}_{\text{target value}} - Q(s,a) \Big]$$

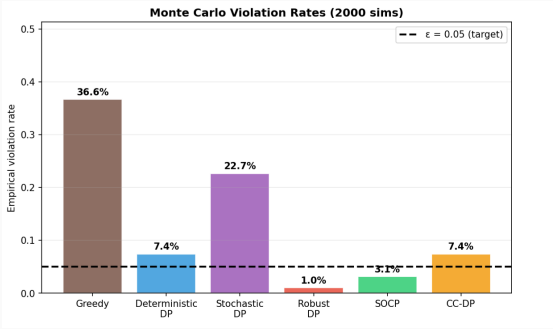
Input	$(t/T,\ C_t,\ P_{\text{base}},\ \theta)$
Network	2×128 ReLU
Output	Q-values for all 2^N actions

When RL becomes necessary

- $N \gg 8$ loads (2^{20} actions)
- Continuous state spaces
- Unknown dynamics
- Multi-day horizons

For $N = 4$, DP is already exact and faster.

Experimental Results



In this scenario (high noise, tight θ), the **risk posture** cleanly diverges:

- **Det. DP (7.4% violations)**
  Ignores uncertainty ($\sigma = 0$), failing to uphold the strict safety constraint.
- **Stoch. DP (22.7% violations)**
  Severe failure under noise; its risk-neutral approach ignores tail risks.
- **Robust DP (1.0% violations)**
  Assumes worst-case ($\pm 3\sigma$) and acts overly cautious, heavily sacrificing production.
- **SOCP / CC-DP (3.1% violations)**
  Tracks variance (V_t). It is well-calibrated against $\varepsilon = 5\%$, safely avoiding unnecessary conservatism.

SOCP: Second-Order Cone Programming (evaluates the chance constraint analytically by assuming symmetric noise).

CC-DP: Chance-Constrained DP

Summary

The Complete Solver Landscape

Solver	State	Uncertainty Treatment	Type	Exact CC?
Greedy Heuristic	—	Myopic threshold proximity	Heuristic	No
Deterministic DP	C_t	Ignored ($\sigma = 0$)	Penalty	No
Stochastic DP	C_t	GH quadrature $K = 5$ (risk-neutral)	Penalty	No
Robust DP	C_t	Worst-case $\pm\kappa\sigma$ (minimax)	Penalty	No
SOCP	(C_t, V_t)	Analytic $z_\varepsilon\sqrt{V}$	Augmented	Yes
CC-DP	(C_t, V_t)	GH quadrature $K = 7$ on 2D grid	Augmented	Yes
RL (DQN)	Continuous	Learned via experience	Model-Free	No

Conservatism ordering:

$$\underbrace{\text{Robust}}_{\text{most cautious}} \geq \underbrace{\text{CC-DP / SOCP}}_{\text{calibrated}} \geq \underbrace{\text{Stochastic}}_{\text{risk-neutral}} \geq \underbrace{\text{Deterministic}}_{\text{optimistic}}$$

Recommendation: SOCP for production (principled + fast), RL when $N \gg 8$.

Summary & Outlook

Summary

- We discussed **CC-MDP**
- A hierarchy of solvers reveals the trade-off between conservatism and safety calibration:
- RL (DQN) becomes necessary when $N \gg 8$ or dynamics are unknown

Outlook

- **Scaling:** Extend CC-DP to $N > 8$ loads via RL
- **Safe RL:** Combine DQN with Lagrangian / CPO methods
- **Sim-to-real transfer:** Validate on real industrial consumption data

Key insight: Tracking uncertainty *explicitly* in the state (C_t, V_t) is the minimal sufficient step to achieve principled chance-constraint satisfaction.

Thank you!

Questions?

Appendix: Derivation of the Deterministic Equivalent

1. **Gaussian sum:** $C_H \sim \mathcal{N}(\mathbb{E}[C_H],\ \mathrm{Var}(C_H))$
2. **Standardise:** $C_H = \mathbb{E}[C_H] + Z \cdot \sqrt{\mathrm{Var}(C_H)}, \quad Z \sim \mathcal{N}(0,1)$
3. **Transform the chance constraint:**

$$\Pr[C_H > \theta] \le \varepsilon$$

$$\Pr\left[Z > \frac{\theta - \mathbb{E}[C_H]}{\sqrt{\mathrm{Var}(C_H)}}\right] \le \varepsilon$$

$$\frac{\theta - \mathbb{E}[C_H]}{\sqrt{\mathrm{Var}(C_H)}} \ge z_\varepsilon = \Phi^{-1}(1-\varepsilon)$$

4. **Rearrange:**

$$\boxed{\mathbb{E}[C_H] + z_\varepsilon \cdot \sqrt{\mathrm{Var}(C_H)} \le \theta}$$

$z_{0.05} = 1.645$ (5% violation), $z_{0.01} = 2.326$ (1% violation)