Distributionally Robust Dynamic Resource Provisioning under Service Level Agreement

Runyu Tang

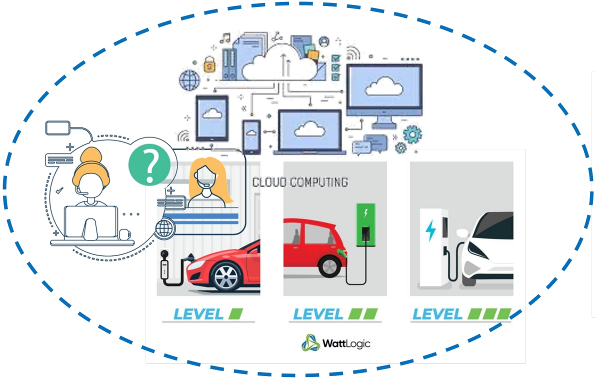
Xi'an Jiaotong Unviversity

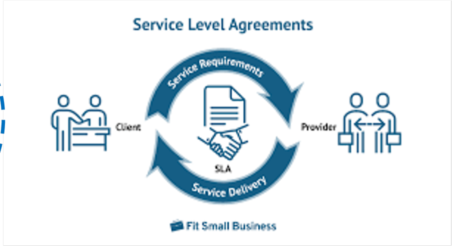
with Yong Liang

Dec. 10, 2022 ISCOM

Background

Service level agreement, SLA





Focused metric: **service availability**

Background

SLA examples in a cloud computing case:

Monthly Uptime Percentage	Service Credit Percentage
Less than 99.9% but greater than or equal to 99.0%	10%
Less than 99.0% but greater than or equal to 95.0%	25%
Less than 95.0%	100%

k-fault tolerance

To overcome servers' failure and provide high-quality service:

- **fault tolerance system**

The minimum server configuration for the service can still be satisfied when k hosting servers concurrently fail. [Zhou et al., 2017, Yuan et al., 2018, Guo et al., 2019]

Main trade-off:

with more backup server

- the likelihood of SLA violation $\downarrow$
- the cost of servers $\uparrow$

Trade-off

SLA violation cost v.s. cost of back-up servers

$$\min_{x} \underbrace{hx}_{\text{Holding cost}} + \underbrace{c\xi_x}_{\text{Penalty cost}}$$

Estimation of the distribution of servers' downtime. [Du et al., 2015, Guo et al., 2020]

Trade-off

SLA violation cost v.s. cost of back-up servers

$$\min_{x} \quad \underbrace{hx}_{\text{Holding cost}} + \underbrace{c\xi_x}_{\text{Penalty cost}}$$

Estimation of the distribution of servers' downtime. [Du et al., 2015, Guo et al., 2020]

The distributionally robust version:

$$\min_{x} \quad hx + \max_{\mu \in \mathscr{F}_x} \mathbb{E}^{\mathbb{P}}[c\xi_x]$$

Dynamic adjustment

Technology advancement: IoT, virtual machines...

As the **cumulative system downtime** randomly grows with the progression of service in a contracted period, the service providers can take advantage of the **observed downtime information** to make dynamic decisions on backup deployment.

$$\underbrace{\text{decision } (x_1)}_{\text{stage 1}} \rightarrow \underbrace{\text{observation } (s_2) \rightarrow \text{decision}(x_2)}_{\text{stage 2}} \rightarrow \ldots$$

$$\rightarrow \underbrace{\text{observation } (s_t) \rightarrow \text{decision}(x_t)}_{\text{stage t}} \rightarrow \ldots$$

$$\rightarrow \underbrace{\text{observation } (s_T) \rightarrow \text{decision}(x_T)}_{\text{stage T}}$$

Robust Dynamic Programming

Literature

SLA related

- inventory SLA [Katok et al., 2008, Liang and Atkins, 2013, Jiang et al., 2019]
- cloud SLA [Passacantando et al., 2016, Guo et al., 2019]

Robust related

- Uncertainty set (rectangularity): [Nilim and Ghaoui, 2005, Iyengar, 2005, Wiesemann et al., 2013, Mannor et al., 2016, Goyal and Grand-Clément, 2021]
- Linear adjusted strategy: [Ben-Tal et al., 2005, Bertsimas et al., 2010, Bertsimas and Goyal, 2012, Bertsimas et al., 2019]
- Approximate Robust DP: [Petrik, 2012, Petrik and Subramanian, 2014, Lim and Autef, 2019, Yu and Shen, 2020]

Our problem: finite decision space and a continuous state space.
We develop convexified surrogates with performance guarantees.

Model

- discrete time: T periods
- state: cumulative system downtime s_t
- action: backup server number $x_t \in \mathscr{A}$
- ambiguity set: $\mathscr{F}(x)$

The ambiguity set can be constructed using 1-norm Wasserstein distance:

$$\mathscr{F}(x) = \{\mathbb{Q} \in \mathcal{P}(\Xi) | W_1(\mathbb{Q}, \hat{\mathbb{P}}_{N_t^x}) \leq \theta\},$$

where

$$W_1(\mathbb{Q}_1, \mathbb{Q}_2) := \inf_{\pi \in \Pi(\mathbb{Q}_1, \mathbb{Q}_2)} \int \|\xi_1 - \xi_2\| \pi(\mathrm{d}\xi_1, \mathrm{d}\xi_2).$$

Model

Distributionally Robust Dynamic Programming

$$(\textbf{DRDP}) \quad V_t(s_t) = \min_{x_t} \max_{\mathbb{P} \in \mathscr{F}(x_t)} hx_t + \mathbb{E}^{\mathbb{P}} \left[\delta \left(s_t + \xi(x_t) - \max\{s_t,\ b\} \right) + \rho V_{t+1}(s_{t+1}) \right],$$

$$s_{t+1} = s_t + \xi_{x_T},$$

$$V_{T+1}(s) = 0, \quad \forall s$$

where $b = (1-\alpha)T$ is the acceptable downtime in SLA.

Last-period problem

LP reformulation [Kuhn et al., 2019]

$$
\begin{aligned}
V_T(s_T) = \min_{x_T,\gamma,\boldsymbol{r},\boldsymbol{u}} \quad & hx_T + \gamma\theta + \frac{1}{N_T^x}\sum_{i=1}^{N_T^x} r_i \\
\text{s.t.} \quad & c(s_T - b)^- + c\hat{\xi}_i^{xT} + u_{i1}(\tau - \hat{\xi}_i^{xT}) \le r_i, \quad \forall i \le N_T^x \\
& c\hat{\xi}_i^{xT} + u_{i2}(\tau - \hat{\xi}_i^{xT}) \le r_i, \quad \forall i \le N_T^x \\
& |u_{i1} - c| \le \gamma, \quad \forall i \le N_T^x \\
& |u_{i2}| \le \gamma, \quad \forall i \le N_T^x \\
& x_T \in \mathcal{X}, \gamma \in \mathbb{R}, r_i \in \mathbb{R}, u_{i1}, u_{i2} \ge 0 \quad \forall i \le N_T^x.
\end{aligned}
$$

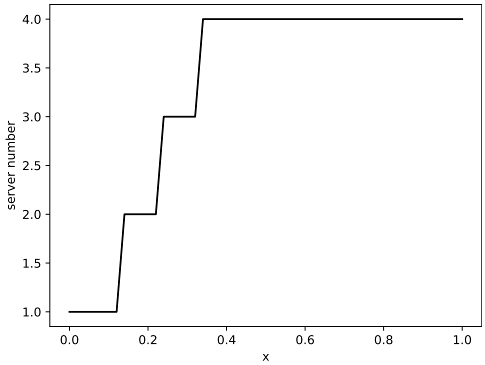
where $\hat{\xi}_i^x$ is the historical downtime data with x backup servers.
The last period problem is a **finite-dimension LP**.

Properties

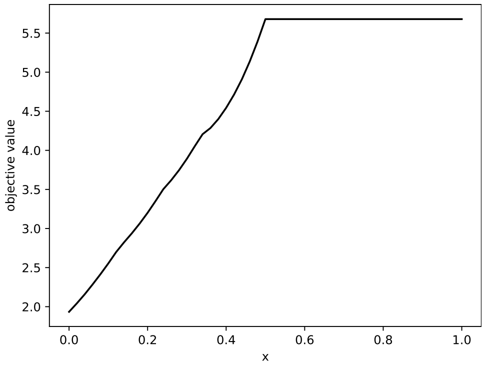
$$V_T(s_T) = \begin{cases} \min_{x_T} hx_T + \max_{\mathcal{P}\in\mathscr{F}_{x_T}} \int_{\mathcal{P}} \delta\left(\xi(x_T)\right) \mathrm{d}\xi_{x_T}, & \text{if } s_T \geq b, \\ \min_{x_T} hx_T + \max_{\mathcal{P}\in\mathscr{F}_{x_T}} \left(\int_{\mathcal{P}} \delta\left(\xi(x_T) + s_T - b\right) \mathrm{d}\xi_{x_T}\right), & \text{otherwise.} \end{cases}$$

- When $s_T < b$, for a given x, $V_T(s_T; x)$ is piece-wise linear and convex increasing in s_T.
- When $s_T \geq b$, for a given x, $V_T(s_T; x)$ is a constant.

A Numerical Example



Optimal backup servers number



Optimal objective value

Moiving forward

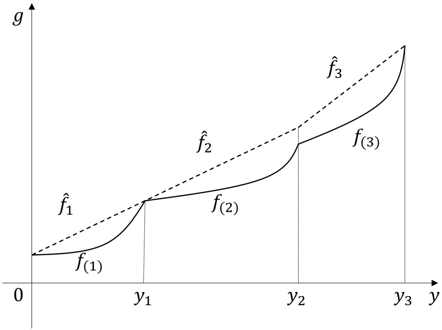
For period $t < T$, we have

$$V_t(s_t) = \min_{x_t} hx_t + \max_{\mathcal{P} \in \mathscr{F}_{x_t}} \int_{\mathcal{P}} [\delta\left(\xi(x_t) + s_t - \max\{s_t,\, b\}\right) + V_{t+1}(s_t + \xi(x_t))] \, \mathrm{d}\xi_{x_t}.$$

Let $L_t(s_t, \xi(x_t)) := \delta\left(\xi(x_t) + s_t - b\right) + V_{t+1}(s_t + \xi(x_t))$ denote the integrand when $s_t < b$.

Then $L_t(s_t, \xi)$ is generally nonconvex in ξ, which prevents us from applying the LP reformulation.

Convexified surrogates



Proposition 1.

The linear approximation error is bounded:

$$\|\hat{g} - g\|_\infty = \max_y \{\hat{g}(y) - g(y)\} = l\epsilon/4.$$

Moving forward

$$
\begin{aligned}
\hat{L}_t(s_t, \xi_t) &= \delta(s_t + \xi_t - \max\{b, s_t\}) + \hat{V}_{t+1}(s_t + \xi_t) \\
&= \begin{cases} \max\limits_{n \in [m]} \{\hat{c}_n(s_t + \xi_t) + \hat{d}_n\}, & \text{if } s_t + \xi_t \le b, \\ c\xi_t + c(s_t - b)^- + \bar{V}_{t+1}, & \text{otherwise.} \end{cases} \\
&:= \max_{\kappa \le 1+m} \{\tilde{c}_\kappa \xi_t + \tilde{d}_\kappa\},
\end{aligned}
$$

After applying the approximation, the problem in each period is always a finite-dimension (parametric) LP:

Radius Adjustment

Theorems 3.4 and 3.5 of [Mohajerin Esfahani and Kuhn, 2018].

Lemma 1.

Assume that the true distribution $\mathbb{Q}$ *is light tailed, then, for any* $\beta \in (0,1]$, *there exist constants* $c_1, c_2 > 0$ *such that* $P\left\{W_1(\mathbb{Q}, \hat{\mathbb{P}}_N) \leq \eta_N\right\} \geq 1-\beta$ *holds as long as*

$$\eta_N(\beta) = \begin{cases} \left(\frac{\log(c_1/\beta)}{c_2 N}\right)^{1/2}, & \text{if } N \geq \frac{\log(c_1/\beta)}{c_2}, \\ \left(\frac{\log(c_1/\beta)}{c_2 N}\right)^{1/\alpha}, & \text{if } N < \frac{\log(c_1/\beta)}{c_2}. \end{cases} \quad (1)$$

Additionally, the finite sample guarantee holds as follows:

$$P\{V_t^{\mathbb{Q}}(s_t;\Theta_t) \leq \hat{V}_t(s_t;\Theta_t)\} \geq (1-\beta)^{T-t+1}, \quad \forall\ 0 \leq t \leq T, \quad (2)$$

where $\Theta_t = \{\eta_N(\beta), \eta_N(\beta), \ldots, \eta_N(\beta)\}$.

State and stage dependent radius adjustment

Proposition 2.

At stage t, if we choose $\tilde{\beta}_t(s_t)$ such that $D_t\left(\tilde{\beta}_t(s_t); s_t\right) = 0$ for any state $s_t < b$, then the following inequality holds

$$P\{V_t^{\mathbb{Q}}(s_t; \tilde{\Theta}_t(s_t)) \leq \hat{V}_t(b; \bar{\Theta}_t)\} \geq (1 - \bar{\beta})^{T-t+1}, \quad \forall s_t \in [0, b] \text{ and } 0 \leq t \leq T, \quad (3)$$

and the out-of-sample performance under different states is upper bounded by $\hat{V}_t(b; \bar{\Theta}_t)$ with a probability no lower than $(1 - \bar{\beta})^{T-t+1}$.

Proposition 3.

At the same stage t, the confidence level $\tilde{\beta}_t(s_t)$ is nonincreasing in s_t. Under the same cumulative service shortages state s, the confidence level $\tilde{\beta}_t(s)$ is nondecreasing in t.

Adaptive radius adjustment

The core idea:
If the realized cumulative costs are lower than the expected costs up to t, then the supplier could act more adventurously by choosing a smaller radius to construct the ambiguity set, thereby leading to less-conservative resource provisioning decisions while maintaining the same confidence regarding the maximum expected total costs across the entire planning horizon.

We choose an adjusted confidence level $\tilde{\beta}$ that satisfies the following equation:

$$(1-\tilde{\beta})(1-\bar{\beta})^{T-t}\hat{V}_t(s_t;\tilde{\Theta}_t)+(\tilde{\beta}-\bar{\beta})(1-\bar{\beta})^{T-t}\bar{\bar{V}}_t-\rho^{-t}(1-\bar{\beta})^{T-t+1}\left(\hat{V}_0(0;\bar{\Theta})-\check{U}_t\right)=0.$$

A cloud computing example

- $n = 100$ virtual machines (VMs)
- $T = 30$ stages
- $\alpha = 99\%$ SLA guarantee
- $h/c = 0.3$ holding/penalty cost

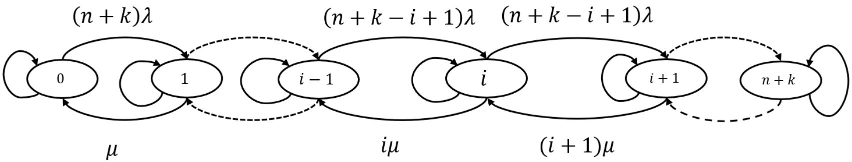
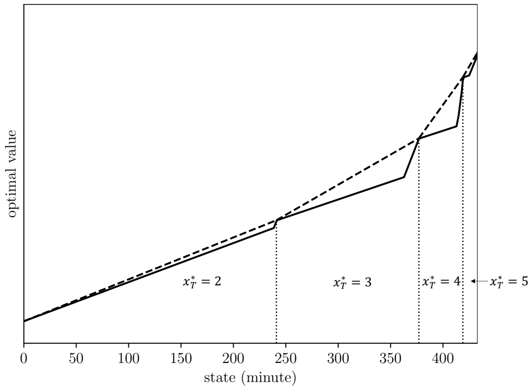
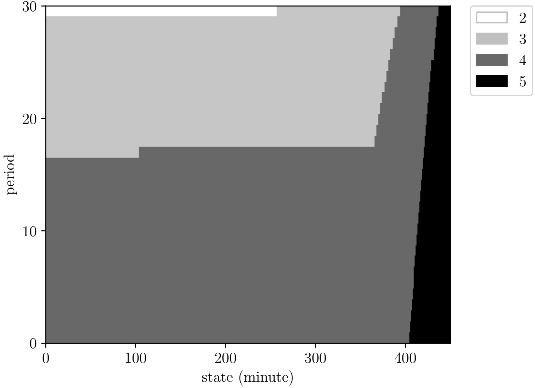


Illustration of servers' up and down state transitions

A cloud computing example

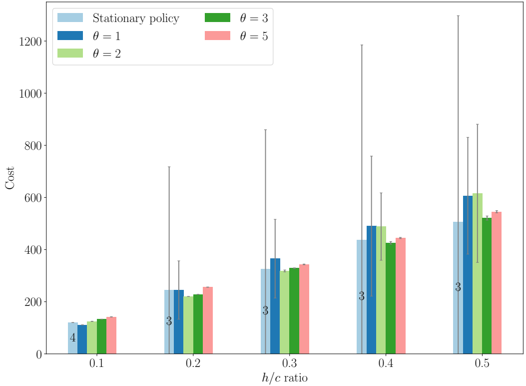


Convexification for the last-stage value function

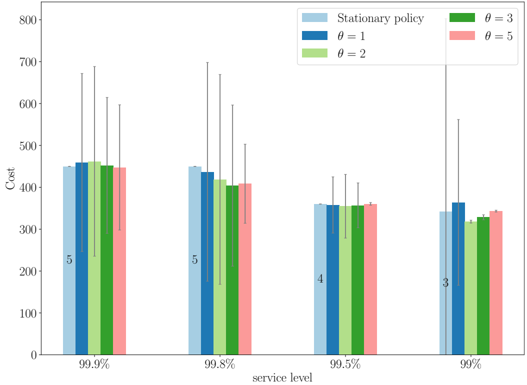


Service provisioning policy for the whole contract period

Sensitivity analysis



Cost performance under different h/c ratios



Cost performance under different service levels

Sensitivity analyses for the DRDP policies

Radius adjustment

Different θ for ambiguity sets under different x while keeping β unchanged
$\rightarrow$ by **Bootstrapping**.

Cost performance under different β

Policy	AveCost	StdCost	AveDown	StdDown	Improvement
best fixed $\theta = 2$	317.96	1.86	226.83	60.98	—
$\beta = 1$	427.18	450.65	412.29	16.49	-25.57%
$\beta = 0.8$	315.31	3.40	183.75	50.80	0.84%
$\beta = 0.6$	327.54	2.83	153.24	44.30	-2.92%
$\beta = 0.4$	339.14	3.00	126.44	38.56	-6.24%
$\beta = 0.2$	360.00	0.00	75.68	28.60	-11.68%

Radius adjustment

Cost performance with state- and stage-dependent radius adjustments

Policy	AveCost	StdCost	AveDown	StdDown	Improvement
$\beta = 0.8$ w/o RA	315.31	3.40147	183.75	50.79951	—
$\beta = 0.8$	302.30	3.00	216.26	54.25	4.13%
$\beta = 0.6$	291.70	4.02	239.98	58.93	7.49%
$\beta = 0.4$	306.23	3.97	207.34	52.61	2.88%
$\beta = 0.2$	324.26	3.81	163.15	44.37	-2.84%

Radius adjustment

Cost performance with adaptive radius adjustment

Policy	AveCost	StdCost	AveDown	StdDown	Improvement
$\beta = 0.6$ w/o ARA	291.70	4.02	239.98	58.93	—
$\beta = 0.8$	274.47	4.73	278.18	72.71	5.91%
$\beta = 0.6$	273.33	4.68	305.58	54.45	6.30%
$\beta = 0.4$	272.28	3.84	298.89	58.00	6.45%
$\beta = 0.2$	273.42	5.42	306.54	48.54	6.27%

Insights

- The DRDP framework helps generate **cost-efficient** dynamic resource provisioning policies, which outperform the best static policies in both average and variance of the out-of-sample performance.
- Introducing **a small amount of robustness** in the DRDP framework can bring substantial performance improvements.
- In the dynamic setting, applying our radius adjustment approaches, which assign different Wasserstein radii depending on the states, stages, and cumulative cost performances, can achieve **better out-of-sample performances**.
- Adaptive radius adjustment is relatively robust in terms of ensuring **less reliance on the choice of β**. In other words, implementing adaptive radius adjustment offsets the over-conservativeness brought about by the supplier using an unnecessarily small confidence level or an excessively large Wasserstein ambiguity set.

Main takeaway:

- Wasserstein-based distributionally robust dynamic programming.
- Solution approaches using convexified surrogates.
- Radius adjustments.
- Application to cloud computing services.

Thank you!