Distributionally Robust Optimization over Time

Kevin-Martin Aigner*[1], Andreas Bärmann[1], Kristin Braun[1], Frauke Liers[1], Sebastian Pokutta[2], Oskar Schneider[3], Kartikey Sharma[2] and Sebastian Tschuppik[1]

[1]Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany
[2]Zuse Institute Berlin, Berlin, Germany
[3]Fraunhofer Institute for Integrated Circuits IIS, Nürnberg, Germany

26th April 2023

Appendix

Here we present the proofs omitted from the main body of the paper as well as additional numerical experiments.

A.1 Proofs of Theorem 1 and Theorem 2

In the following, we prove that the sequence of solutions to the problem (DRO_t) converge to the true stochastic optimization (SO) problem. These results are an adaptation of the proofs presented in Mohajerin, Esfahani and Kuhn (2018) for our setting.

Lemma 3 *Given the true distribution p^* and the ambiguity set $\mathcal{P}_t$ at any time t, we have*

$$\mathbb{P}\left[p^* \in \mathcal{P}_t\right] \geq 1 - \delta_t \quad \textit{for all } t = 1, \dots, T.$$

Proof: This is true by the construction of the set $\mathcal{P}_t$, which is such that it contains the true distribution with probability at least $1 - \delta_t$. □

Lemma 4 (Finite sample guarantee) *Given a solution x_t to the problem (DRO_t), we prove that*

$$\mathbb{P}\left[\mathbb{E}_{s\sim p^*}[f(x_t, s)] \leq \widehat{J}_t\right] \geq 1 - \delta_t \textit{ for all } t = 1, \dots, T.$$

Proof: From Lemma 3, we know that $p^* \in \mathcal{P}_t$ with probability at least $1 - \delta_t$. Thus, we have

$$\mathbb{E}_{s\sim p^*}[f(x_t, s)] \leq \max_{p \in \mathcal{P}_t} \mathbb{E}_{s\sim p}[f(x_t, s)],$$

with probability at least $1 - \delta_t$.

The right hand side (RHS) term in the above equation is the definition of $\widehat{J}_t$. Thus,

$$\mathbb{E}_{s\sim p^*}[f(x_t, s)] \leq \widehat{J}_t,$$

with probability of at least $1 - \delta_t$. □

Lemma 5 (Borel-Cantelli Lemma) *Let $E_1, E_2, \dots$ be a sequence of events. If $\sum_{i=1}^{\infty} P(E_i) < \infty$ then*

$$P[\textit{an infinite number of } E_i \textit{ occur}] = 0.$$

*Corresponding author. E-mail address: kevin-martin.aigner@fau.de

Lemma 6 (Convergence of Distributions) *Given the ambiguity set $\mathcal{P}_t$, we prove that*

$$\lim_{t\to\infty} \sup_{p\in\mathcal{P}_t} \|p-p^*\|_2 = 0 \textit{ with probability } 1.$$

Proof: From Lemma 9, 10 and 11 we know that for any of the three given types of ambiguity sets there exists a function $r(t)$ which satisfies

$$\mathbb{P}\left[\sup_{p\in\mathcal{P}_t} \|p-p^*\|_2 \le r(t)\right] \ge 1-\delta_t,$$

and $\lim_{t\to\infty} r(t) = 0$.
This means that

$$\mathbb{P}\left[\sup_{p\in\mathcal{P}_t} \|p-p^*\|_2 - r(t) > 0\right] \le \delta_t.$$

By construction, it follows that $\sum_{t=1}^{\infty}\delta_t < \infty$ (as $\delta_t = \frac{6\delta}{\pi^2 t^2}$). Then the Borel-Cantelli Lemma 5 implies that

$$\mathbb{P}\left[\lim_{t\to\infty} \sup_{p\in\mathcal{P}_t} \|p-p^*\|_2 - r(t) \le 0\right] = 1.$$

Since $\lim_{t\to\infty} r(t) = 0$ and $\|p-p^*\|_2 \ge 0$, this means that

$$\lim_{t\to\infty} \sup_{p\in\mathcal{P}_t} \|p-p^*\|_2 = 0 \text{ with probability } 1.$$

□

Proof of Theorem 1: We know that $x_t \in \mathcal{X}$ and $J^* \le \mathbb{E}_{s\sim p^*}[f(x_t,s)]$ as x_t is a suboptimal solution. Applying Lemma 4, we obtain

$$\mathbb{P}\left[J^* \le \mathbb{E}_{s\sim p^*}[f(x_t,s)] \le \widehat{J}_t\right] \ge \mathbb{P}\left[p^* \in \mathcal{P}_t\right] \ge 1-\delta_t.$$

Since $\sum_{t=1}^{\infty}\delta_t < \infty$, by the Borel-Cantelli lemma,

$$\mathbb{P}\left[J^* \le \lim_{t\to\infty} \mathbb{E}_{s\sim p^*}[f(x_t,s)\right] \le \lim_{t\to\infty} \widehat{J}_t] = 1.$$

Let $\gamma \ge 0$. Since $\mathcal{X}$ is compact, there exists a γ-optimal solution x^γ to the stochastic problem, i.e.,

$$\mathbb{E}_{s\sim p^*}[f(x^\gamma,s)] \le J^* + \gamma.$$

Let $p_t^\gamma \in \mathcal{P}_t$ be a γ-optimal distribution to x^γ, i.e.

$$\sup_{p\in\mathcal{P}_t} \mathbb{E}_{s\sim p}[f(x^\gamma,s)] \le \mathbb{E}_{s\sim p_t^\gamma}[f(x^\gamma,s)] + \gamma.$$

Then we can write

$$\begin{aligned}
&\limsup_{t\to\infty} \widehat{J}_t \\
\le &\limsup_{t\to\infty} \sup_{p\in\mathcal{P}_t} \mathbb{E}_{s\sim p}[f(x^\gamma,s)] \\
\le &\limsup_{t\to\infty} \mathbb{E}_{s\sim p_t^\gamma}[f(x^\gamma,s)] + \gamma \\
= &\limsup_{t\to\infty} \mathbb{E}_{s\sim p^*}[f(x^\gamma,s)] + \sum_{s\in\mathcal{S}} f(x^\gamma,s)(p_{st}^\gamma - p_s^*) + \gamma \\
\le &\limsup_{t\to\infty} \mathbb{E}_{s\sim p^*}[f(x^\gamma,s)] + G\|p_t^\gamma - p^*\|_2 + \gamma \\
= &\mathbb{E}_{s\sim p^*}[f(x^\gamma,s)] + \gamma \quad \text{w.p. } 1 \\
= &J^* + 2\gamma \quad \text{w.p. } 1,
\end{aligned}$$

where the first inequality holds because of the definition of $\widehat{J}_t$, the second inequality holds because of the definition of p_t^γ, the first equality holds as we add and subtract p^* and the third inequality holds as $|f(x,s)| \leq G$ for all $(x,s) \in \mathcal{X} \times \mathcal{S}$ by assumption. The final two equalities hold because of Lemma 6 and the definition of x^γ respectively.

With this we conclude that $\limsup_{t\to\infty} \widehat{J}_t \leq J^*$. Along with the earlier assertion of $J^* \leq \lim_{t\to\infty} \widehat{J}_t$, we can now complete the proof that $\widehat{J}_t \to J^*$ via the sandwich argument. □

Proof of Theorem 2: Let $\{s_t\}_{t=1}^\infty$ be any sequence of scenario realizations such that $\lim_{t\to\infty} \widehat{J}_t = J^*$. By Theorem 1, we have $J^* \leq \mathbb{E}_{s\sim p^*}[f(x_t,s)] \leq \widehat{J}_t$ with probability 1. By the same theorem, we also know that $\lim_{t\to\infty} \widehat{J}_t = J^*$ w.p. 1. Then, we can write

$$\liminf_{t\to\infty} \mathbb{E}_{s\sim p^*}[f(x_t,s)] \leq \liminf_{t\to\infty} \widehat{J}_t = J^*. \tag{6}$$

Consider any limit point of the sequence $\{x_t\}_{t=1}^\infty$. Since the set $\mathcal{X}$ is compact, then there exists a limit point of $\{x_t\}_{t=1}^\infty$ which lies in $\mathcal{X}$. WLOG let x^* be that point and
$\liminf_{t\to\infty} x_t = x^*$.
Then we have

$$\begin{aligned} J^* &\leq \mathbb{E}_{s\sim p^*}[f(x^*,s)] \\ &= \mathbb{E}_{s\sim p^*}[\liminf_{t\to\infty} f(x_t,s)] \\ &= \sum_{s\in\mathcal{S}} \liminf_{t\to\infty} f(x_t,s)p_s^* \\ &= \liminf_{t\to\infty} \sum_{s\in\mathcal{S}} f(x_t,s)p_s^* \leq J^*, \end{aligned}$$

where the first inequality holds because $x^* \in \mathcal{X}$, the second inequality holds as $\liminf_{t\to\infty} x_t = x^*$ and because $f(x,s)$ is continuous in x. The second equality exploits that $\mathcal{S}$ is finite and the final inequality holds because of (6). Thus, we have $\mathbb{E}_{s\sim p^*}[f(x^*,s)] = J^*$ which completes the proof. □

A.2 Proofs of Dynamic regret bounds

In order to prove the dynamic regret bound of Theorem 3, we first show that the ambiguity sets are shrinking at a rate of $\mathcal{O}(\sqrt{\log T/T})$ and contain the true data generating distribution with a high confidence. The latter is stated in the following Lemma.

A crucial point for a shrinking dynamic regret bound is that the ambiguity sets are shrinking over time. Our ambiguity sets are constructed with increasing confidence probabilities for the multinomial distribution. We show that the ambiguity sets are shrinking even though the confidence $1-\delta_t$ is increasing ($\delta_t = \frac{6\delta}{\pi^2 t^2}$).

Lemma 7 *For the upper $(1-\frac{\delta_t}{2})$-percentile $z_{\frac{\delta_t}{2}}$ of the standard normal distribution with confidence update*
$\delta_t := \frac{6\delta}{\pi^2 t^2}$ *and $\delta \in (0,1)$, it follows that*

$$z_{\frac{\delta_t}{2}}^2 \leq 4\log(\pi t),$$

for all rounds $t = 1, ..., T$.

Proof: By the definition of the standard normal distribution, for the upper percentile $1-\frac{\delta_t}{2}$, we have the Gaussian tail bound

$$
\begin{aligned}
& 1-\frac{\delta_t}{2} \leq e^{-\frac{1}{2} z^2_{\frac{\delta_t}{2}}} \\
\Longrightarrow \quad & \log(1-\frac{\delta_t}{2}) \leq -\frac{z^2_{\frac{\delta_t}{2}}}{2} \\
\Longrightarrow \quad & z^2_{\frac{\delta_t}{2}} \leq -2\log(1-\frac{\delta_t}{2}) \\
& = -2\log\left(\frac{2\pi^2 t^2 - 6\delta}{2\pi^2 t^2}\right) \\
& = 2\log\left(\frac{\pi^2 t^2}{\pi^2 t^2 - 3\delta}\right),
\end{aligned}
$$

for all rounds $t = 1, ..., T$. Since $\pi^2 t^2 - 3\delta \geq 1$ for all rounds $t = 1, ..., T$ and $\delta \in (0,1)$ (try $t = 1$ and $\delta = 1$), we are able write

$$
z^2_{\frac{\delta_t}{2}} \leq 2\log(\pi^2 t^2) \leq 4\log(\pi t),
$$

for all $t = 1, ..., T$. □

We now prove that the ambiguity sets $\mathcal{P}_t$ shrink at a rate of $\mathcal{O}(\sqrt{\log t/t})$ even as the confidence requirement $1-\delta_t$ increases.

Here, we provide a proof for ambiguity sets as defined by confidence intervals. The proofs for the other sets are provided in the electronic companion.

Lemma 8 (Confidence Interval Sets) *The ambiguity sets $\mathcal{P}_t$ derived from (1) with confidence update $\delta_t := \frac{6\delta}{\pi^2 t^2}$ and $\delta \in (0,1)$ for all rounds $t = 1, ..., T$ fulfill*

$$
\sup_{x\in\mathcal{P}_0, y\in\mathcal{P}_1} \|x-y\| \leq \sqrt{16|\mathcal{S}|\log\pi} \quad \text{and} \quad \sup_{x\in\mathcal{P}_{t-1}, y\in\mathcal{P}_t} \|x-y\| \leq \frac{\sqrt{16|\mathcal{S}|\log(\pi(t-1))}}{\sqrt{t-1}},
$$

for all $t = 2, ..., T$ with a probability of at least $1-\delta$.

Proof: As we know from Lemma 1 that $p^* \in \bigcap_{t=0,...,T} \mathcal{P}_t$ with a probability of at least $1-\delta$, we can compute for $t = 1$:

$$
\begin{aligned}
\sup_{x\in\mathcal{P}_0, y\in\mathcal{P}_1} \|x-y\| &= \sup_{x\in\mathcal{P}_0, y\in\mathcal{P}_1} \|x-p^*+p^*-y\| \\
&\leq \sup_{x\in\mathcal{P}_0} \|x-p^*\| + \sup_{y\in\mathcal{P}_1} \|p^*-y\| \\
&\leq \sup_{x,p\in\mathcal{P}_0} \|x-p\| + \sup_{p,y\in\mathcal{P}_1} \|p-y\| \\
&\leq \sqrt{2} + \sqrt{4|\mathcal{S}|\log\pi} \leq \sqrt{16|\mathcal{S}|\log\pi},
\end{aligned}
$$

with a probability of at least $1-\delta$ because of $\sup_{x,y\in\mathcal{P}_0} \|x-y\| = \sqrt{(1-0)^2+(0-1)^2} = \sqrt{2}$ and Lemma 9.

Similarly for $t = 2, ..., T$:

$$
\begin{aligned}
\sup_{x\in\mathcal{P}_{t-1}, y\in\mathcal{P}_t} \|x-y\| &= \sup_{x\in\mathcal{P}_{t-1}, y\in\mathcal{P}_t} \|x-p^*+p^*-y\| \\
&\leq \sup_{x\in\mathcal{P}_{t-1}} \|x-p^*\| + \sup_{y\in\mathcal{P}_t} \|p^*-y\| \\
&\leq \sup_{x,p\in\mathcal{P}_{t-1}} \|x-p\| + \sup_{p,y\in\mathcal{P}_t} \|p-y\| \\
&\leq 2\frac{\sqrt{4|\mathcal{S}|\log(\pi(t-1))}}{\sqrt{t-1}},
\end{aligned}
$$

with a probability of at least $1-\delta$. $\square$

Lemma 9 *The ambiguity sets $\mathcal{P}_t$ derived from (1) with confidence update $\delta_t := \frac{6\delta}{\pi^2 t^2}$ and $\delta \in (0,1)$ for all rounds $t = 1, ..., T$ fulfill*

$$\sup_{x,y\in\mathcal{P}_t} \|x-y\| \leq \frac{\sqrt{4|\mathcal{S}|\log(\pi t)}}{\sqrt{t}}.$$

Proof: We can compute using Lemma 7:

$$\sup_{x,y\in\mathcal{P}_t} \|x-y\|^2 \leq \sum_{k=1}^{|\mathcal{S}|}(u_{kt}-l_{kt})^2 = \sum_{k=1}^{|\mathcal{S}|} \frac{z^2_{\frac{\delta_t}{2}}}{t} \leq \frac{4|\mathcal{S}|\log(\pi t)}{t}.$$

$\square$

Finally, In the following result, we prove that the distance between elements from consecutive ℓ_2-norm and kernel based ambiguity sets $\mathcal{P}_{t-1}$ and $\mathcal{P}_t$ also shrinks as measured in the ℓ_2-norm while providing tighter guarantees of inclusion on the true distribution..

Lemma 10 (ℓ_2-norm Sets) *Given an ambiguity set of the from $\mathcal{P}_t = \{p \in \mathcal{P} \mid \|p-\hat{p}\|_2 \leq \epsilon_t\}$ with $\epsilon_t := \sqrt{\frac{2|\mathcal{S}|\log 2/\delta_t}{t}}$ and $\delta_t = \frac{6\delta}{\pi^2 t^2}$, then for $|\mathcal{S}| \geq 2$ we have*

$$\sup_{x\in\mathcal{P}_0,y\in\mathcal{P}_1} \|x-y\|_2 \leq 4\sqrt{|\mathcal{S}|\log(\pi/\sqrt{3\delta})} \text{ and } \sup_{x\in\mathcal{P}_{t-1},y\in\mathcal{P}_t} \|x-y\|_2 \leq 4\sqrt{\frac{|\mathcal{S}|\log(\pi(t-1)/\sqrt{3\delta})}{t-1}},$$

with probability at least $1-\delta$.

Proof: For the case $t=1$ we have

$$\begin{aligned}\sup_{x\in\mathcal{P}_0,y\in\mathcal{P}_1} \|x-y\|_2 &\leq \sup_{x\in\mathcal{P}_0}\|x-p^*\|_2 + \sup_{y\in\mathcal{P}_1}\|p^*-y\|_2 \\ &\leq 2+2\sqrt{|\mathcal{S}|\log(\pi/\sqrt{3\delta})} \\ &\leq 4\sqrt{|\mathcal{S}|\log(\pi/\sqrt{3\delta})}.\end{aligned}$$

The last inequality occurs because $\sqrt{|\mathcal{S}|\log(\pi/\sqrt{3\delta})} > 1$ for $|\mathcal{S}| \geq 2$. Now for the case $t > 1$, given the true distribution p^*, we can write,

$$\begin{aligned}\|x-y\|_2 &\leq \|x-p^*\|_2 + \|p^*-y\|_2 \\ &\leq 2\epsilon_{t-1}.\end{aligned}$$

Thus,

$$\sup_{x\in\mathcal{P}_{t-1},y\in\mathcal{P}_t} \|x-y\|_2 \leq 4\sqrt{\frac{|\mathcal{S}|\log(\pi(t-1)/\sqrt{3\delta})}{t-1}}.$$

Here, the first inequality arises from triangle inequality. The second inequality is due to the fact that the true distribution is contained inside all sets $\mathcal{P}_t$ for $t = 1, \ldots, T$ with probability at least $1-\delta$. $\square$

Lemma 11 (Kernel Based Sets) *Given an ambiguity set of the form $\mathcal{P}_t = \{p \in \mathcal{P} \mid \|p-\hat{p}\|_M \leq \epsilon_t\}$ with $\epsilon_t := \frac{\sqrt{C}}{\sqrt{t}}(2+\sqrt{2\log(1/\delta_t)})$ with $\delta_t = \frac{6\delta}{\pi^2 t^2}$ we have for $t \geq 2$,*

$$\sup_{x\in\mathcal{P}_{t-1},y\in\mathcal{P}_t} \|x-y\|_2 \leq \frac{8\sqrt{C}}{\lambda\sqrt{t-1}}(\sqrt{\log(\pi t/\sqrt{6\delta})}),$$

$$\text{and } \sup_{x\in\mathcal{P}_0,y\in\mathcal{P}_1} \|x-y\|_2 \leq 2+\frac{4\sqrt{C}}{\lambda} \quad \text{for } t=1,$$

with probability at least $1-\delta$.

Proof: For $t = 1$, we have

$$\begin{aligned}\|x-y\|_2 &\leq \|x-p^*\|_2 + \|p^*-y\|_2 \\ &\leq 2 + \frac{1}{\lambda}\|p^*-y\|_M \\ &\leq 2 + \frac{4\sqrt{C}}{\lambda}.\end{aligned}$$

The second inequality comes from the definition of the set $\mathcal{P}_0$ and the definition of the norm $\|\cdot\|_M$.

For the case $t \geq 2$, we get.

$$\begin{aligned}\|x-y\|_2 &\leq \|x-p^*\|_2 + \|p^*-y\|_2 \\ &\leq \frac{1}{\lambda}(\|x-p^*\|_M + \|p^*-y\|_M) \\ &\leq \frac{4}{\lambda}\frac{\sqrt{C}}{\sqrt{t-1}}(1+\sqrt{\log(\pi(t-1)/\sqrt{6\delta})}) \\ &\leq \frac{4}{\lambda}\frac{\sqrt{C}}{\sqrt{t-1}}(1+\sqrt{\log(\pi t/\sqrt{6\delta})}).\end{aligned}$$

Here, the first inequality comes from the triangle inequality, the second from the fact that $\sqrt{x^\top M x} \geq \lambda\|x\|_2$ for any positive definite matrix M with the minimum eigen value λ. The final inequality arises from the construction of the ambiguity set which contains the true distribution with probability at least $1-\delta$. Now note that $\pi^2 t^2/6\delta > 3$ for $t \geq 2$. Thus, we can write

$$\|x-y\|_2 \leq \frac{8}{\lambda}\frac{\sqrt{C}}{\sqrt{t-1}}\sqrt{\log(\pi t/\sqrt{6\delta})}.$$

□

The properties of the confidence interval ambiguity sets mentioned previously and those of the ℓ_2-norm and kernel based sets prove in Lemmas 10 and 11 along with the cumulative path length bounds from Theorem 5 enable us to prove the following shrinking dynamic regret.

Theorem 4 (Dynamic regret bound) *Let $f : \mathcal{X} \times \mathcal{S} \to \mathbb{R}$ be uniformly bounded, i.e., for all $(x,s) \in \mathcal{X} \times \mathcal{S}$, a constant $G > 0$ exists such that $|f(x,s)| \leq G$. Let $\eta := \sqrt{\frac{3+2h'(T)}{TG^2|\mathcal{S}|}}$ with $\sum_{t=2}^{T}\|p-q\| \leq h'(T)$ for $p \in \mathcal{P}_{t-1}$ and $q \in \mathcal{P}_t$*
The output $(x_1, ..., x_T)$ from Algorithm with confidence update $\delta_t := \frac{6\delta}{\pi^2 t^2}$ and $\delta \in (0,1)$ fulfills

$$\frac{1}{T}\sum_{t=1}^{T}\left(\max_{p\in\mathcal{P}_t}\mathbb{E}_{s\sim p}\left[f(x_t,s)\right] - \min_{x\in\mathcal{X}}\max_{p\in\mathcal{P}_t}\mathbb{E}_{s\sim p}\left[f(x,s)\right]\right) \leq G\sqrt{\frac{3|\mathcal{S}|+2|\mathcal{S}|h'(T)}{T}} + \frac{2G}{T},$$

with probability at least $1-\delta$.

Proof of Theorem 4: Define $g_t(p) := -\mathbb{E}_{s\sim p}\left[f(x_t,s)\right]$. A gradient descent iteration is given by

$$p_{t+1} = \arg\min_{p\in\mathcal{P}_t} \;\; \langle \eta\nabla g_t(p_t), p\rangle + \frac{1}{2}\|p-p_t\|^2,$$

with optimality criteria $\langle \eta\nabla g_t(p_t), u_t - p_{t+1}\rangle + \langle p_{t+1}-p_t, u_t-p_{t+1}\rangle \geq 0$ for all $u_t \in \mathcal{P}_t$. Classical theory for gradient descent yields

$$\langle \eta\nabla g_t(p_t), p_t - u_t\rangle \leq \frac{1}{2}\|p_t-u_t\|^2 - \frac{1}{2}\|p_{t+1}-u_t\|^2 + \frac{\eta^2}{2}\|\nabla g_t(p_t)\|^2.$$

Summation over rounds $t = 1, ..., T$ results in the following inequality for all $u_t \in \mathcal{P}_t$:

$$\sum_{t=1}^{T} \langle \eta \nabla g_t(p_t), p_t - u_t \rangle \leq \sum_{t=1}^{T} \frac{\eta^2}{2} \|\nabla g_t(p_t)\|^2. + \sum_{t=1}^{T} \left(\frac{1}{2}\|p_t - u_t\|^2 - \frac{1}{2}\|p_{t+1} - u_t\|^2 \right). \quad (7)$$

Next, we rearrange the terms on the RHS as

$$\begin{aligned}
&\sum_{t=1}^{T} \left(\frac{1}{2}\|p_t - u_t\|^2 - \frac{1}{2}\|p_{t+1} - u_t\|^2 \right) \\
&= \frac{1}{2} \sum_{t=1}^{T} \left(\|p_t\|^2 - \|p_{t+1}\|^2 \right) + \frac{1}{2} \sum_{t=1}^{T} 2\left(p_{t+1} - p_t\right) \cdot u_t \\
&= \frac{1}{2} \sum_{t=1}^{T} \left(\|p_t\|^2 - \|p_{t+1}\|^2 \right) + \sum_{t=1}^{T} \left(p_{t+1} - p_t\right) \cdot u_t.
\end{aligned}$$

Now consider the expression $\sum_{t=1}^{T} (p_{t+1} - p_t) \cdot u_t$. We can write it as

$$\begin{aligned}
&\sum_{t=1}^{T} \left(p_{t+1} - p_t\right) \cdot u_t \\
&= (p_2 - p_1) \cdot u_1 + (p_3 - p_2) \cdot u_2 + \cdots + (p_{T+1} - p_T) \cdot u_T \\
&= p_{T+1} \cdot u_T - p_1 \cdot u_1 + \sum_{t=2}^{T} (u_{t-1} - u_t) \cdot p_t.
\end{aligned}$$

For the other expression we have

$$\frac{1}{2} \sum_{t=1}^{T} \left(\|p_t\|^2 - \|p_{t+1}\|^2 \right) = \frac{1}{2}\|p_1\|^2 - \frac{1}{2}\|p_{T+1}\|^2.$$

This then allows us to write

$$\begin{aligned}
\sum_{t=1}^{T} \langle \eta \nabla g_t(p_t), p_t - u_t \rangle &\leq \sum_{t=1}^{T} \frac{\eta^2}{2} \|\nabla g_t(p_t)\|^2 + \frac{1}{2}\|p_1\|^2 \\
&- \frac{1}{2}\|p_{T+1}\|^2 + p_{T+1} \cdot u_T - p_1 \cdot u_1 + \sum_{t=2}^{T} (u_{t-1} - u_t) \cdot p_t.
\end{aligned}$$

We know that $\|p\| \leq 1$ and for any 2 probability vectors p and q we have that $0 \leq p \cdot q \leq 1$. Thus we can write

$$\sum_{t=1}^{T} \langle \eta \nabla g_t(p_t), p_t - u_t \rangle \leq \sum_{t=1}^{T} \frac{\eta^2}{2} \|\nabla g_t(p_t)\|^2 + \frac{1}{2} + 1 + \sum_{t=2}^{T} (u_{t-1} - u_t) \cdot p_t.$$

Note that

$$\|\nabla g_t(p_t)\|^2 = \sum_{k=1}^{|\mathcal{S}|} |f(x_t, s_k)|^2 \leq |\mathcal{S}| G^2,$$

This, along with the fact that $\sum_{t=2}^{T} (u_{t-1} - u_t) \cdot p_t \leq \sum_{t=2}^{T} \|u_{t-1} - u_t\| \|p_t\| \leq \sum_{t=2}^{T} \|u_{t-1} - u_t\|$, allows us to write

$$\sum_{t=1}^{T} \langle \nabla g_t(p_t), p_t - u_t \rangle \leq \frac{\eta T}{2} |\mathcal{S}| G^2 + \frac{3}{2\eta} + \frac{1}{\eta} \sum_{t=2}^{T} \|u_{t-1} - u_t\|.$$

By assumption, $\sum_{t=2}^{T} \|u_{t-1} - u_t\| \leq h'(T)$ with probability at least $1-\delta$ for some function $h'(\cdot)$. Then we have

$$\sum_{t=1}^{T} \langle \nabla g_t(p_t), p_t - u_t \rangle \leq \frac{\eta T}{2} |\mathcal{S}| G^2 + \frac{3}{2\eta} + \frac{h'(T)}{\eta},$$

with probability at least $1-\delta$. Choosing the optimal $\eta = \sqrt{\frac{3+2h'(T)}{TG^2|\mathcal{S}|}}$, we can write our result as

$$\sum_{t=1}^{T} \langle \nabla g_t(p_t), p_t - u_t \rangle \leq 2\sqrt{\frac{T}{2}|\mathcal{S}|G^2 \cdot (\frac{3}{2} + h'(T))}.$$

Then we can write

$$\sum_{t=1}^{T} \langle \nabla g_t(p_t), p_t - u_t \rangle \leq G|\mathcal{S}|^{\frac{1}{2}} \sqrt{3T + 2Th'(T)}.$$

Since $g_t(p) = -\mathbb{E}_{s\sim p}\left[f(x_t, s)\right]$ is linear in p for all $t = 1, ..., T$, it follows

$$\begin{aligned}
&\sum_{t=1}^{T} \left(\mathbb{E}_{s\sim u_t}\left[f(x_t, s)\right] - \mathbb{E}_{s\sim p_t}\left[f(x_t, s)\right]\right) \\
&= \sum_{t=1}^{T} (g_t(p_t) - g_t(u_t)) = \sum_{t=1}^{T} \langle \nabla g_t(p_t), p_t - u_t \rangle \\
&\leq G|\mathcal{S}|^{\frac{1}{2}} \sqrt{3T + 2Th'(T)}.
\end{aligned}$$

Now we choose in each round $t = 1, ..., T$ the worst-case $u_t := \arg\max_{p\in\mathcal{P}_t} \mathbb{E}_{s\sim p}\left[f(x_t, s)\right] \in \mathcal{P}_t$ and recall $x_t = \arg\min_{x\in\mathcal{X}} \mathbb{E}_{s\sim p_t}\left[f(x, s)\right]$ to obtain

$$\sum_{t=1}^{T} \left(\mathbb{E}_{s\sim u_t}\left[f(x_t, s)\right] - \mathbb{E}_{s\sim p_t}\left[f(x_t, s)\right]\right) = \sum_{t=1}^{T} \left(\max_{p\in\mathcal{P}_t} \mathbb{E}_{s\sim p}\left[f(x_t, s)\right] - \min_{x\in\mathcal{X}} \mathbb{E}_{s\sim p_t}\left[f(x, s)\right]\right).$$

Since $p_t \in \mathcal{P}_{t-1}$, we know that $\min_{x\in\mathcal{X}} \mathbb{E}_{s\sim p_t}\left[f(x, s)\right] \leq \min_{x\in\mathcal{X}} \max_{p\in\mathcal{P}_{t-1}} \mathbb{E}_{s\sim p}\left[f(x, s)\right]$ for all $t = 1, ..., T$ and thus we can conclude

$$\sum_{t=1}^{T} \left(\max_{p\in\mathcal{P}_t} \mathbb{E}_{s\sim p}\left[f(x_t, s)\right] - \min_{x\in\mathcal{X}} \max_{p\in\mathcal{P}_{t-1}} \mathbb{E}_{s\sim p}\left[f(x, s)\right]\right) \leq G|\mathcal{S}|^{\frac{1}{2}} \sqrt{3T + 2Th'(T)},$$

with a probability of at least $1-\delta$. We add and subtract $\min_{x\in\mathcal{X}} \max_{p\in\mathcal{P}_t} \mathbb{E}_{s\sim p}\left[f(x, s)\right]$ on the LHS. Rearranging the terms this allows us to write the LHS as

$$\begin{aligned}
&\sum_{t=1}^{T} \left(\max_{p\in\mathcal{P}_t} \mathbb{E}_{s\sim p}\left[f(x_t, s)\right] - \min_{x\in\mathcal{X}} \max_{p\in\mathcal{P}_t} \mathbb{E}_{s\sim p}\left[f(x, s)\right]\right) \\
&+ \sum_{t=1}^{T} \min_{x\in\mathcal{X}} \max_{p\in\mathcal{P}_t} \mathbb{E}_{s\sim p}\left[f(x, s)\right] - \min_{x\in\mathcal{X}} \max_{p\in\mathcal{P}_{t-1}} \mathbb{E}_{s\sim p}\left[f(x, s)\right].
\end{aligned}$$

Observing that the last two terms telescope, bringing them to the RHS and using the upper bound G on $|f(x, s)|$ we can conclude

$$\sum_{t=1}^{T} \left(\max_{p\in\mathcal{P}_t} \mathbb{E}_{s\sim p}\left[f(x_t, s)\right] - \min_{x\in\mathcal{X}} \max_{p\in\mathcal{P}_t} \mathbb{E}_{s\sim p}\left[f(x, s)\right]\right) \leq G|\mathcal{S}|^{\frac{1}{2}} \sqrt{3T + 2Th'(T)} + 2G \qquad \text{w.p. } 1-\delta.$$

Dividing by T on both sides completes the proof. □

Theorem 5 *Given ambiguity sets of the form specified in Section* **??**, *we have*

$$\frac{1}{2}\sum_{t=1}^{T}\|p_t - q_t\|^2 \leq h(T) \text{ and } \sum_{t=2}^{T}\|p_t - q_t\| \leq h'(T) \text{ for all } p_t \in \mathcal{P}_{t-1}, q_t \in \mathcal{P}_t,$$

with probability at least $1-\delta$. *The functions* $h(T)$ *and* $h'(T)$ *for different categories of ambiguity sets are as given below:*

1. *Confidence Intervals:*

$$h(T) = 8|\mathcal{S}|\log(\pi T)(2+\log T)$$

$$h'(T) = 8\sqrt{|\mathcal{S}|T\log(\pi T)}$$

2. *Kernel based ambiguity sets, where* λ *denotes the smallest eigenvalue of the kernel matrix* M*:*

$$h(T) = \frac{1}{2}\left(2+\frac{4\sqrt{C}}{\lambda}\right)^2 + \frac{32C}{\lambda^2}\log(\pi T/\sqrt{6\delta})(1+\log T)$$

$$h'(T) = \frac{16\sqrt{C}}{\lambda}\sqrt{T\log(\pi T/\sqrt{6\delta})}$$

3. ℓ_2*-norm ambiguity sets:*

$$h(T) = 8|\mathcal{S}|\log\frac{\pi T}{\sqrt{3\delta}}(2+\log T)$$

$$h'(T) = 8\sqrt{|\mathcal{S}|T\log\frac{\pi T}{\sqrt{3\delta}}}$$

Confidence Intervals. Calculating the function $h(T)$, we have

$$\begin{aligned}\frac{1}{2}\sum_{t=1}^{T}\|p_t - q_t\|^2 &\leq \frac{1}{2}16|\mathcal{S}|\log\pi + \frac{1}{2}\sum_{t=2}^{T}16|\mathcal{S}|\frac{\log(\pi(t-1))}{t-1}\\ &\leq 8|\mathcal{S}|\log\pi + 8|\mathcal{S}|\log(\pi(T-1))\sum_{t=1}^{T-1}\frac{1}{t}\\ &\leq 8|\mathcal{S}|\log\pi + 8|\mathcal{S}|\log(\pi(T-1))(1+\log(T-1))\\ &\leq 8|\mathcal{S}|\log(\pi T)(2+\log T).\end{aligned}$$

Here, the first inequality arises from Lemma 8. The second and third inequalities are from bounding t and from observing that $\sum_{t=1}^{T-1}(1/t) \leq 1+\log(T-1)$.
Now for the function $h'(T)$, we can calculate

$$\begin{aligned}\sum_{t=2}^{T}\|p_t - q_t\| &\leq \sum_{t=2}^{T}4\frac{\sqrt{|\mathcal{S}|\log(\pi(t-1))}}{\sqrt{t-1}}\\ &\leq 4|\mathcal{S}|^{\frac{1}{2}}\sqrt{\log(\pi T)}\sum_{t=2}^{T}\frac{1}{\sqrt{t-1}}\\ &\leq 8|\mathcal{S}|^{\frac{1}{2}}\sqrt{\log(\pi T)}\sqrt{T}.\end{aligned}$$

Here, the first inequality arises from Lemma 8. The second and third inequalities are from bounding t and from observing that $\sum_{t=2}^{T}(1/\sqrt{t-1}) \leq 2\sqrt{T-1} \leq 2\sqrt{T}$.

Kernel based ambiguity sets. Calculating the function $h(T)$, we have

$$\begin{aligned}\frac{1}{2}\sum_{t=1}^{T}\|p_t-q_t\|_2^2 &\leq \frac{1}{2}\left(2+\frac{4\sqrt{C}}{\lambda}\right)^2+\sum_{t=2}^{T}\frac{32C}{\lambda^2(t-1)}\log(\pi(t-1)/\sqrt{6\delta})\\ &\leq \frac{1}{2}\left(2+\frac{4\sqrt{C}}{\lambda}\right)^2+\frac{32C}{\lambda^2}\log(\pi T/\sqrt{6\delta})\sum_{t=2}^{T}\frac{1}{t-1}\\ &\leq \frac{1}{2}\left(2+\frac{4\sqrt{C}}{\lambda}\right)^2+\frac{32C}{\lambda^2}\log(\pi T/\sqrt{6\delta})(1+\log T).\end{aligned}$$

Here, the first inequality arises from Lemma 11. The second and third inequalities are from bounding t and from observing that $\sum_{t=1}^{T-1}(1/t)\leq 1+\log(T-1)\leq 1+\log T$.

Now, for the function $h'(T)$, we have

$$\begin{aligned}\sum_{t=2}^{T}\|p_t-q_t\|_2 &\leq \sum_{t=2}^{T}\frac{8}{\lambda}\frac{\sqrt{C}}{\sqrt{t-1}}\sqrt{\log(\pi(t-1)/\sqrt{6\delta})}\\ &\leq \frac{8\sqrt{C}}{\lambda}\sqrt{\log(\pi T/\sqrt{6\delta})}\sum_{t=2}^{T}\frac{1}{\sqrt{t-1}}\\ &\leq \frac{16\sqrt{C}}{\lambda}\sqrt{T\log(\pi T/\sqrt{6\delta})}.\end{aligned}$$

Here, the first inequality arises from Lemma 11. The second and third inequalities are from bounding t and from observing that $\sum_{t=1}^{T-1}(1/\sqrt{t-1})\leq 2\sqrt{T-1}\leq 2\sqrt{T}$.

ℓ_2-norm ambiguity sets. Calculating the function $h(T)$, we have

$$\begin{aligned}\frac{1}{2}\sum_{t=1}^{T}\|p_t-q_t\|^2 &\leq \frac{1}{2}\left(4\sqrt{|\mathcal{S}|\log(\pi/\sqrt{3\delta})}\right)^2+\frac{1}{2}\sum_{t=2}^{T}16|\mathcal{S}|\frac{\log(\pi(t-1)/\sqrt{3\delta})}{t-1}\\ &\leq 8|\mathcal{S}|\log(\pi/\sqrt{3\delta})+8|\mathcal{S}|\log(\pi T/\sqrt{3\delta})\sum_{t=2}^{T}\frac{1}{t-1}\\ &\leq 8|\mathcal{S}|\log(\pi T/\sqrt{3\delta})(2+\log T).\end{aligned}$$

Here, the first inequality arises from Lemma 10. The second and third inequalities are proven similar to the case of $h(T)$ for the interval sets.

Now for the function $h'(T)$, we can calculate

$$\begin{aligned}\sum_{t=2}^{T}\|p_t-q_t\| &\leq \sum_{t=2}^{T}4\sqrt{\frac{|\mathcal{S}|\log(\pi(t-1)/\sqrt{3\delta})}{t-1}}\\ &\leq 4\sqrt{|\mathcal{S}|\log(\pi T/\sqrt{3\delta})}\sum_{t=2}^{T}\sqrt{\frac{1}{t-1}}\\ &\leq 8\sqrt{|\mathcal{S}|T\log(\pi T/\sqrt{3\delta})}.\end{aligned}$$

Here, the first inequality arises from Lemma 10. The second and third inequalities are proven similar to the case of $h'(T)$ for the interval sets. □

A.3 Numerical Experiments

In this section, we provide the details for the numerical experiments conducted in Section 4. We also provide an additional set of experiments on an optimal routing problem to further illustrate our algorithms.

A.3.1 Benchmark Instances

All mixed-integer linear optimization problems used in our numerical experiments can be found in Table 5. The entries show the number of variables and constraints for each problem. The same holds for the quadratic problems listed in Table 6.

Name	Variables				Constraints
	All	Bin.	Int.	Cont.	
blend2	353	239	25	89	274
flugpl	18	0	11	7	19
gr4x6	48	24	0	24	34
neos-1430701	312	156	0	156	668
noswot	128	75	25	28	182
prod1	250	149	0	101	208
prod2	301	200	0	101	211
ran13x13	338	169	0	169	195
supportcase14	304	304	0	0	234
supportcase16	319	319	0	0	130
beavma	390	195	0	195	372
k16x240b	480	240	0	240	256
neos-3610040	430	85	0	345	335
neos-3611689	421	88	0	333	323
timtab1CUTS	397	77	94	226	371

Table 5: Overview of MIP instances

Name	Variables			Constraints	
	All	Bin.	Cont.	All	Quadr.
7579	300	100	200	203	1
10001	485	426	59	296	1
10002	485	426	59	296	1
10004	1058	999	59	867	1
10010	269	262	7	147	1
10003	1058	999	59	867	1
10008	845	713	132	416	1
10009	605	473	132	246	1
10011	1390	1258	132	873	1
10012	967	835	132	538	1

Table 6: Overview of MIQP instances