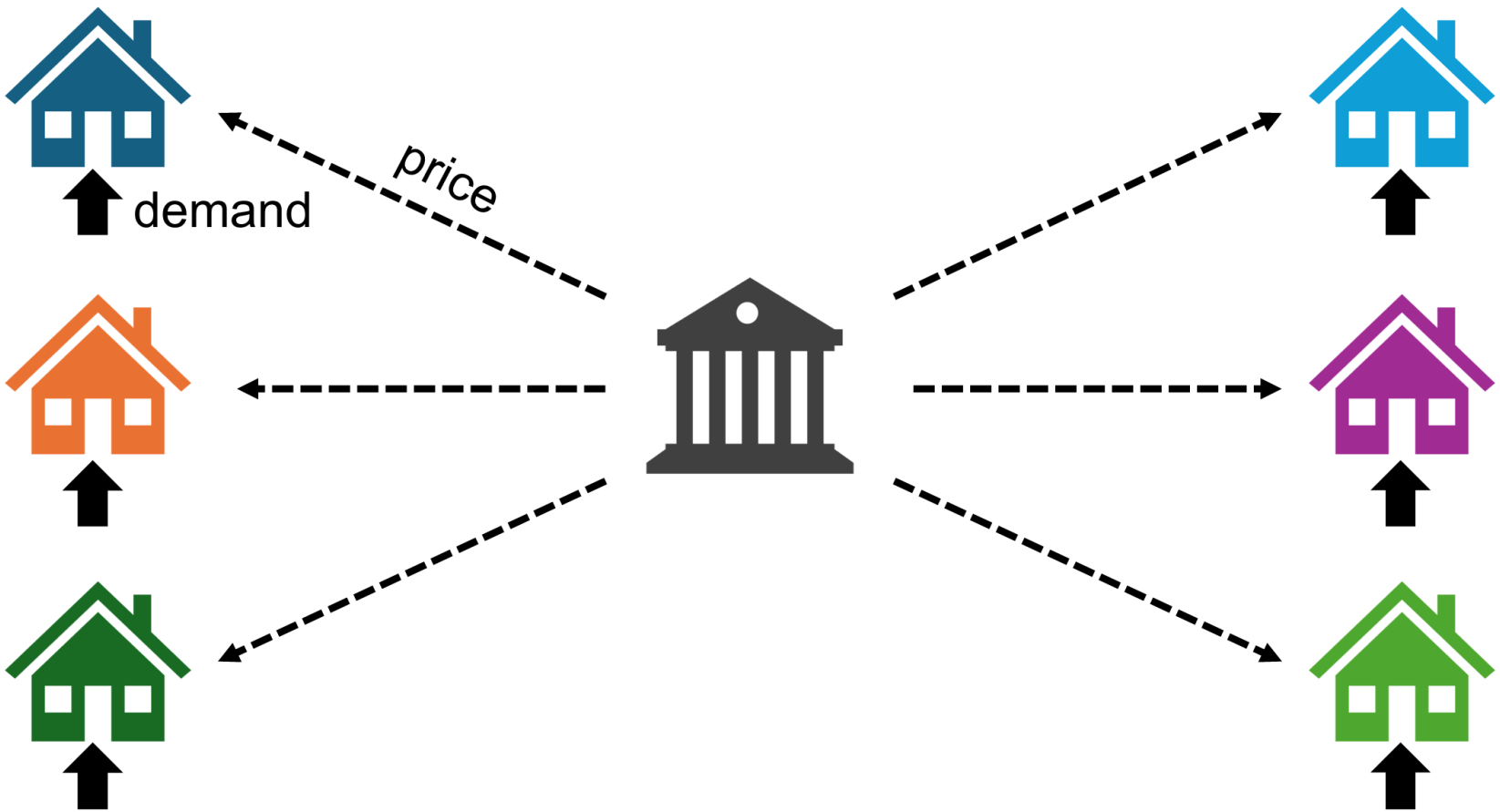
Safe Methods for Bandit and Online Optimization

Spencer Hutchinson

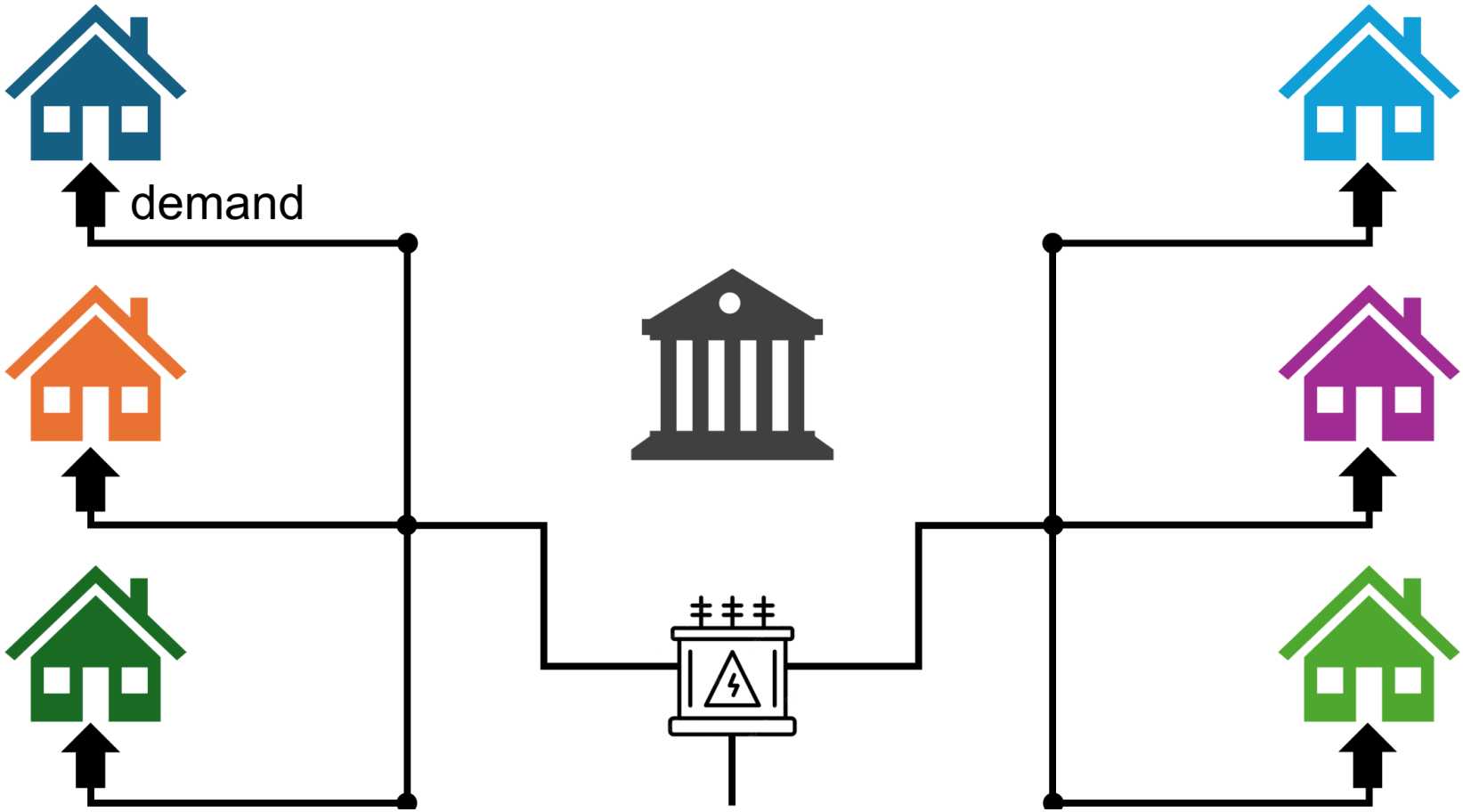
University of California, Santa Barbara
Department of Electrical and Computer Engineering

November 13, 2024

Pricing in Societal-scale Systems

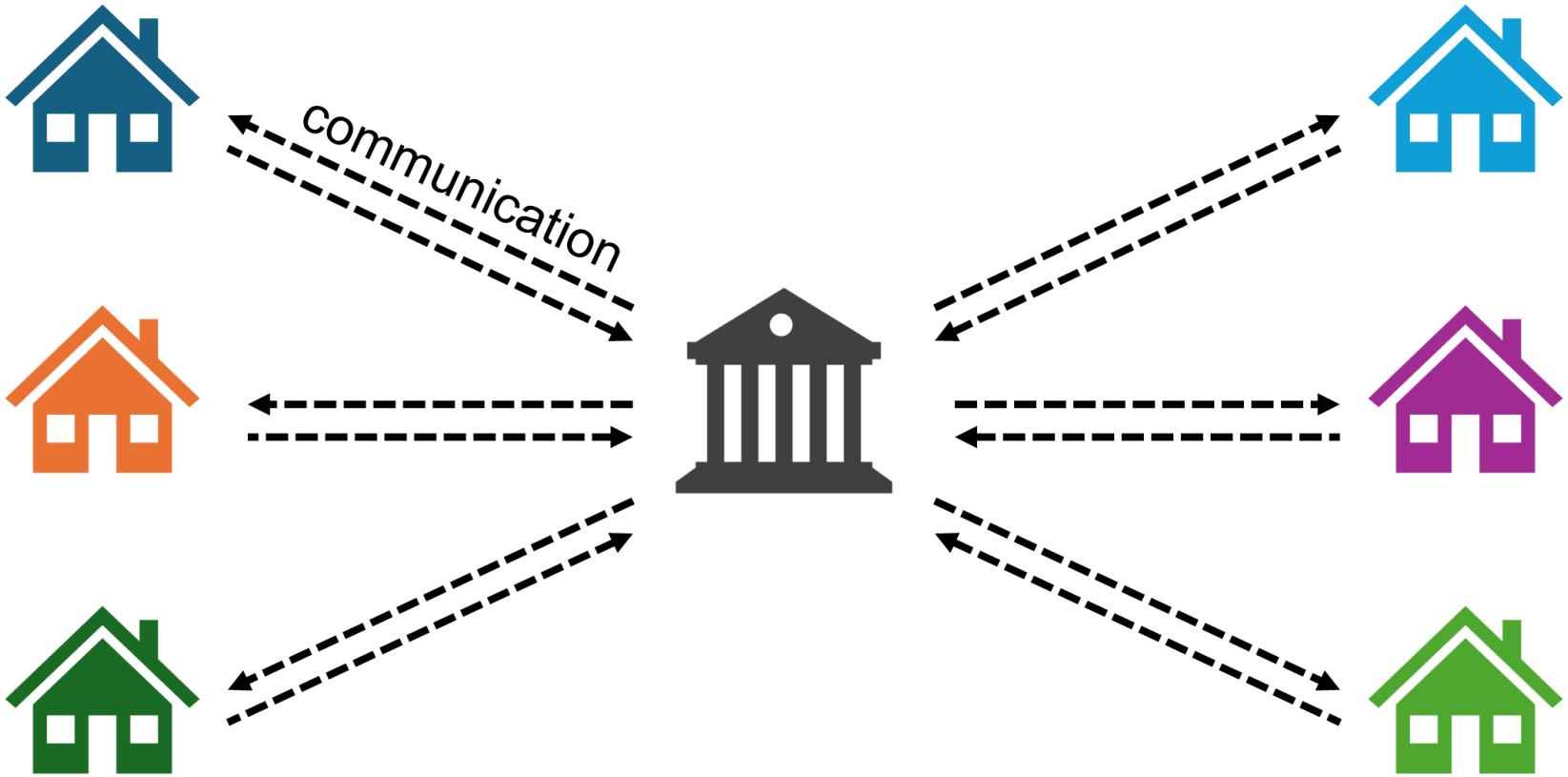


Pricing in Societal-scale Systems



Existing Approaches

Existing approaches typically use two-way communication to optimize prices *before* they are posted.[Samadi et al., 2010; Li et al., 2011]



- Requires two-way communication infrastructure.
- Requires automated demand management system for users.

Our Approach

Can we optimize prices without requiring two-way communications before posting?

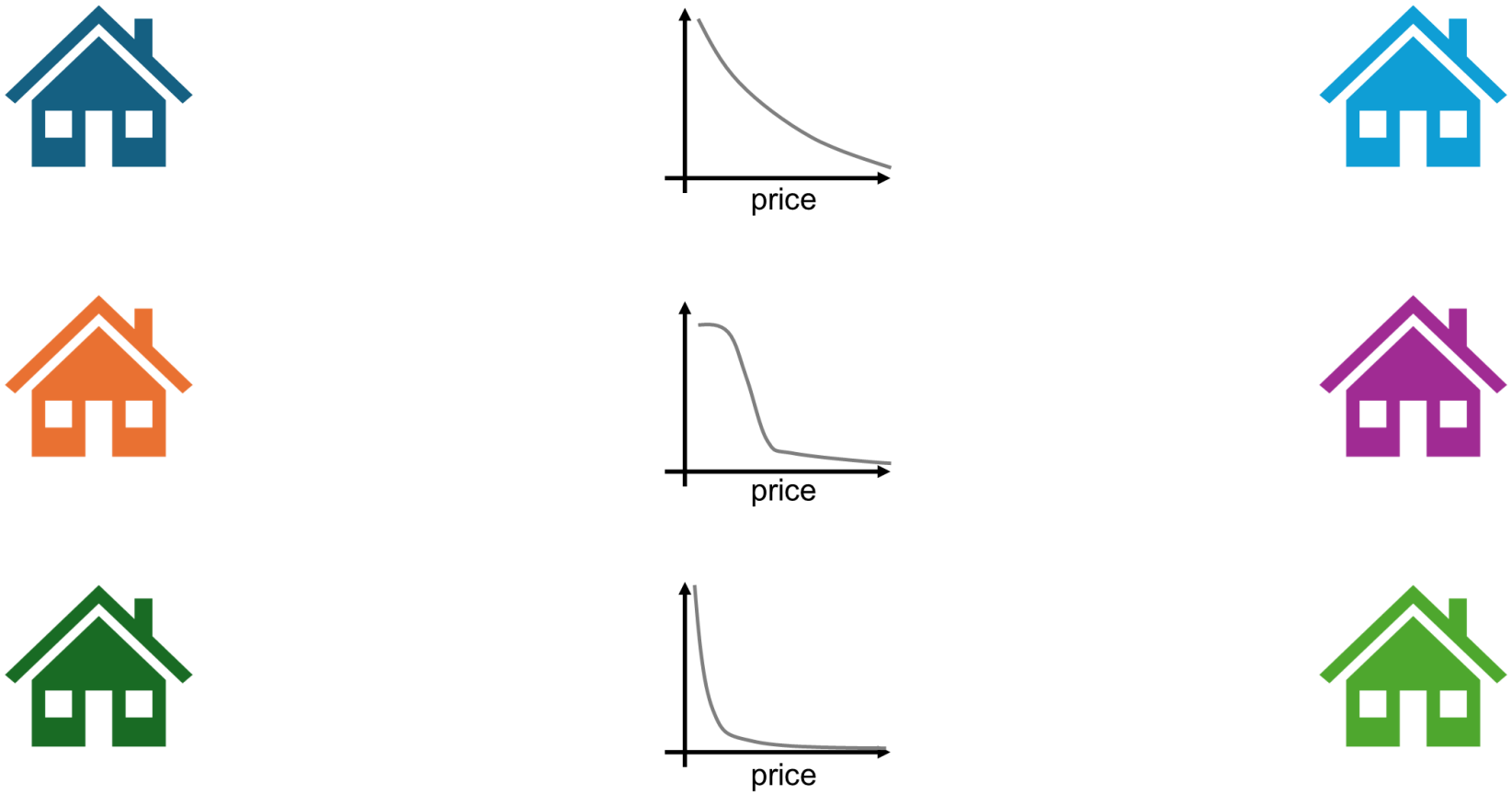
Key challenges:

- **Unknown price response:** User's price response is unknown and needs to be *learned* through interactions.
- **Limited feedback:** Can only observe user's demand *after* comitting to each price.
- **Constraint satisfaction:** Need to ensure system constraints are *always* satisfied.

Price response models:

- **Utility maximization:** Users choose utility-maximizing demand.[Turan, Hutchinson and Alizadeh, TCNS 2024]
- **Parametrically-linear:****Parametrically-linear:** Expected demand is parametrically-linear in price.[Hutchinson, Turan and Alizadeh, TCNS 2024]

Parametrically-linear Price Response



Demand feedback model: $z_t = Ax_t + \epsilon_t$.

Safe Learning with Unknown Constraints

We approach this problem via *safe learning with unknown constraints*.

In each round $t \in [T]$:

- Choose action $x_t \in \mathcal{X} \subseteq \mathbb{R}^d$.
- Observe (stochastic) constraint feedback $z_t = Ax_t + \epsilon_t \in \mathbb{R}^n$, where A is *unknown*.
- Incur cost $f_t(x_t)$ and observe cost feedbackIncur cost $f_t(x_t)$ and observe cost feedback.

Goals:

- Ensure expected constraint value is in safe set:

$$Ax_t \in \mathcal{G} \qquad \forall t \in [T]$$

- Minimize cumulative cost $\sum_{t=1}^{T} f_t(x_t)$.

Learning Paradigms

Stochastic Linear Bandit	**Online Convex Optimization**
Stochastic linear reward: $f_t(x) = \theta^\top x + \eta_t$	Adversarial convex cost f_t.
Bandit reward feedback: $y_t := \theta^\top x_t + \eta_t$	Full cost feedback f_t.
Minimize psuedo-regret: $R_T = \sum_{t=1}^{T} \theta^\top (x^\star - x_t)$ $x^\star = \arg\max_{x \in \mathcal{X}: Ax \in \mathcal{G}} \theta^\top x$	Minimize regret: $R_T = \sum_{t=1}^{T} (f_t(x_t) - f_t(x^\star))$ $x^\star = \arg\min_{x \in \mathcal{X}: Ax \in \mathcal{G}} \sum_{t=1}^{T} f_t(x)$

Learning Paradigms

Stochastic Linear Bandit	**Online Convex Optimization**
Stochastic linear reward: $f_t(x) = \theta^\top x + \eta_t$	Adversarial convex cost f_t.
Bandit reward feedback: $y_t := \theta^\top x_t + \eta_t$	Full cost feedback f_t.
Minimize psuedo-regret: $R_T = \sum_{t=1}^{T} \theta^\top (x^\star - x_t)$ $x^\star = \arg\max_{x \in \mathcal{X}: Ax \in \mathcal{G}} \theta^\top x$	Minimize regret: $R_T = \sum_{t=1}^{T} (f_t(x_t) - f_t(x^\star))$ $x^\star = \arg\min_{x \in \mathcal{X}: Ax \in \mathcal{G}} \sum_{t=1}^{T} f_t(x)$

Stochastic Linear Bandits under General Constraints

Interaction Model

In each round $t \in [T]$:

- Choose action $x_t \in \mathcal{X}$.
- Observe reward $y_t = \theta^\top x_t + \eta_t$ (with unknown θ).
- Observe constraint value $z_t = Ax_t + \epsilon_t$ (with unknown A).

Learning Goals

- Ensure constraint satisfaction for all rounds: $Ax_t \in \mathcal{G} \quad \forall t \in [T]$.
- Minimize pseudo-regret w.r.t. to best safe action,

$$R_T = \sum_{t=1}^{T} \max_{x \in \mathcal{X} : Ax \in \mathcal{G}} \theta^\top x - \sum_{t=1}^{T} \theta^\top x_t$$

Related Work

- **Amani et al., 2019:**
  constraint $\theta^\top B x_t \leqslant b$ (known B)
  explore-exploit approach
  $\tilde{\mathcal{O}}(dT^{2/3})$ regret, $\tilde{\mathcal{O}}(d\sqrt{T})$ regret when $\theta^\top B x^\star < b$
- **Khezeli and Bitar, 2020; Moradipari et al., 2020:**
  constraint $\theta^\top x_t \geqslant b$ ($\theta^\top x^\star > b$)
  explore-exploit approach
  $\tilde{\mathcal{O}}(d\sqrt{T})$ regret
- **Moradipari et al., 2021; Pacchiano et al. 2021:**
  constraint $Ax_t \leqslant b$ (i.e. $\mathcal{G} = b\mathbb{R}_-$)
  expanded confidence set approach
  $\tilde{\mathcal{O}}(d\sqrt{T})$ regret
- **Our work:**
  constraint $Ax_t \in \mathcal{G}$
  directional optimism approach
  $\tilde{\mathcal{O}}(d\sqrt{T})$ regret

Optimism in the Face of Uncertainty

When making decisions in uncertain environments, behave as though unknown quantities are as favorable as reasonably possible.[1]

Broadly popular approach, e.g. in medical trials, ad placement and RL.

[1] Lai and Robbins, 1985; Auer et al., 2002

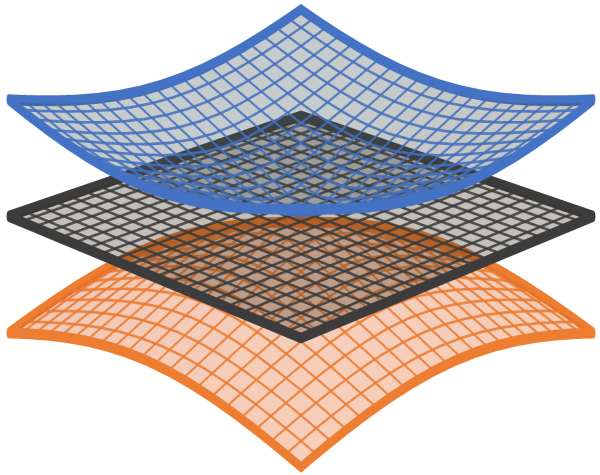
Least-squares Confidence Bounds

When making decisions in uncertain environments, behave as though unknown quantities are as favorable as reasonably possible.reasonably possible.

For least-squares estimators $\hat{\theta}_t, \hat{A}_t$ and $V_t = \sum_{t=1}^{T} x_t x_t^\top + \lambda I$, it holds w.h.p. that,[2]

$$\theta^\top x \in \hat{\theta}_t^\top x + \beta_t \|x\|_{V_t^{-1}} [-1, 1]$$

$$Ax \in \hat{A}_t x + \beta_t \|x\|_{V_t^{-1}} \mathbb{B}_\infty$$



Classical Algorithm: *Upper Confidence Bound (UCB)*[3]

$$\underset{x \in \mathcal{X} : Ax \in \mathcal{G}}{\arg\max} \ \hat{\theta}_t^\top x + \beta_t \|x\|_{V_t^{-1}}$$

The feasible set is unknown in our setting!

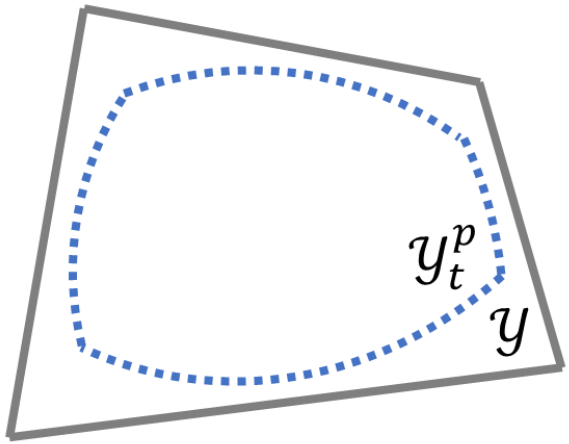
[2] Abbasi-Yadkori et al., 2011

[3] Auer et al., 2002; Dani et al., 2008

Pessimistic Feasible Set

A "pessimistic set" can be constructed with the confidence bounds:

$$\mathcal{Y}_t^p := \{x \in \mathcal{X} : (\hat{A}_t x + \beta_t \|x\|_{V_t^{-1}} \mathbb{B}_\infty) \subseteq \mathcal{G}\}$$
$$\subseteq \{x \in \mathcal{X} : Ax \in \mathcal{G}\} =: \mathcal{Y}$$



Playing in the pessimistic set ensures constraint satisfaction:

$$x_t \in \mathcal{Y}_t^p \quad \forall t \in [T] \underset{\text{w.h.p.}}{\implies} Ax_t \in \mathcal{G} \quad \forall t \in [T]$$

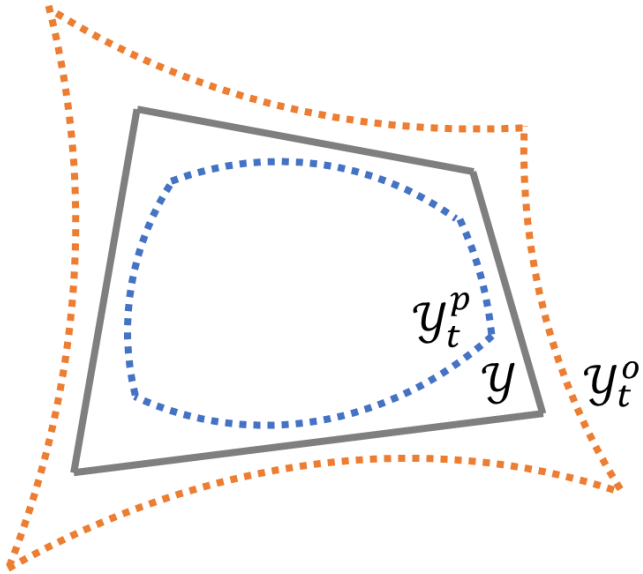
Prior work applies "expanded" UCB to the pessimistic set (for the special case when $\mathcal{G} = b\mathbb{R}_-$).[Moradipari et al., 2021; Pacchiano et al., 2021.]

$$\arg\max_{x \in \mathcal{Y}_t^p} \left(\hat{\theta}_t^\top x + \left(1 + \frac{2S}{r}\right) \beta_t \|x\|_{V_t^{-1}} \right)$$

Optimistic Sets

Our approach: We additionally use "optimistic" sets that contain the feasible set.

$$\mathcal{Y}_t^o := \{x \in \mathcal{X} : \left(\hat{A}_t x + \beta_t \|x\|_{V_t^{-1}} \mathbb{B}_\infty\right) \cap \mathcal{G} \neq \varnothing\}$$
$$\supseteq \{x \in \mathcal{X} : Ax \in \mathcal{G}\} =: \mathcal{Y}$$



We consider *optimistic actions*:

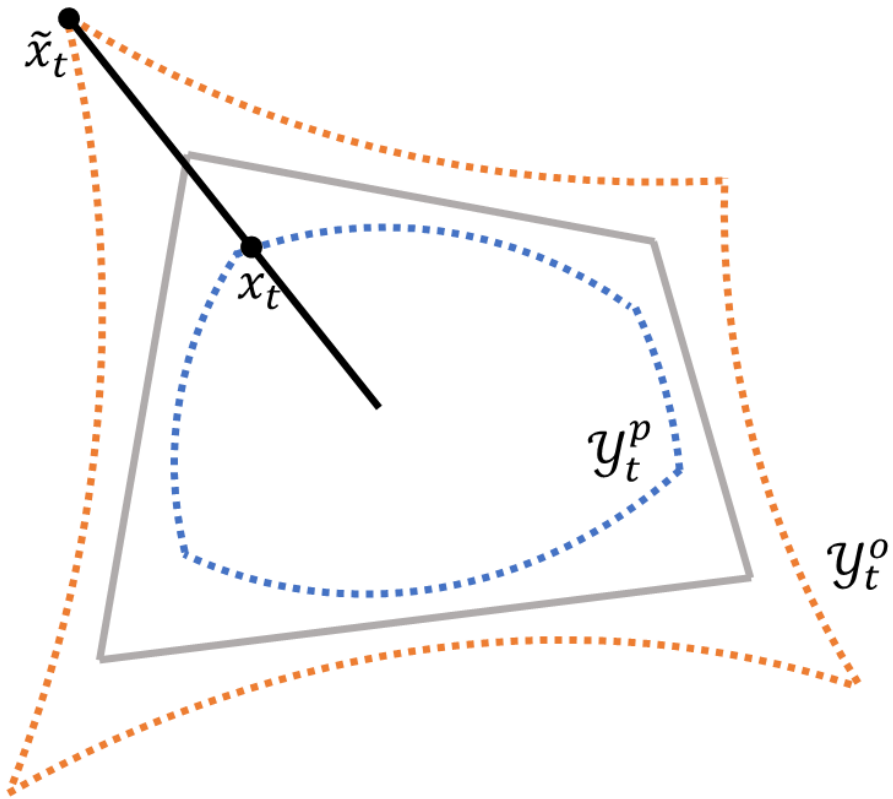
$$\tilde{x}_t \in \arg\max_{x \in \mathcal{Y}_t^o} \hat{\theta}_t^\top x + \beta_t \|x\|_{V_t^{-1}}$$

We ensure constraint satisfaction by *scaling* $\tilde{x}_t$ in to the pessimistic set.

Our Algorithm (ROFUL)

For $t \in [T]$:

1. Construct $\mathcal{Y}_t^o$ and $\mathcal{Y}_t^p$.
2. $\tilde{x}_t \in \arg\max_{x \in \mathcal{Y}_t^o} \hat{\theta}_t^\top x + \beta_t \|x\|_{V_t^{-1}}$.
3. $\gamma_t = \max\left\{\mu \in [0,1] : \mu \tilde{x}_t \in \left(\mathcal{Y}_t^p \cup \nu \mathbb{B}\right)\right\}$.
4. Play $x_t = \gamma_t \tilde{x}_t$



Bounding γ_t

Assumption ($\mathcal{G}$ is Union of Convex Sets)

There exists a set-valued mapping $\mathcal{D} : \mathcal{I} \rightrightarrows \mathbb{R}^n$ such that $\mathcal{G} = \bigcup_{i \in \mathcal{I}} \mathcal{D}(i)$, where $\mathcal{D}(i)$ is convex and $r\mathbb{B}_\infty \subseteq \mathcal{D}(i)$ for all $i \in \mathcal{I}$.

Application Example: Smart grid

- Power flow constraints are non-convex.
- Convex restrictions are sometimes used to approximate these constraints while ensuring feasibility.[4]
- Can reduce approximation error by taking the union over multiple convex restrictions.

[4] Dvijotham et al. (2017), Dongchan et al. (2019)

Bounding γ_t

Assumption ($\mathcal{G}$ is Union of Convex Sets)

There exists a set-valued mapping $\mathcal{D} : \mathcal{I} \rightrightarrows \mathbb{R}^n$ such that $\mathcal{G} = \bigcup_{i \in \mathcal{I}} \mathcal{D}(i)$, where $\mathcal{D}(i)$ is convex and $r\mathbb{B}_\infty \subseteq \mathcal{D}(i)$ for all $i \in \mathcal{I}$.

Fact

For any $\alpha \in [0, 1]$ and $z \in \mathcal{G}$, it holds that $\alpha z + (1 - \alpha) r\mathbb{B}_\infty \subseteq \mathcal{G}$.
Pf: *There exists $i \in \mathcal{I}$ such that $z \in \mathcal{D}(i)$. It follows that,*

$$\alpha z + (1-\alpha) r\mathbb{B} \subseteq \alpha \mathcal{D}(i) \oplus (1-\alpha) r\mathbb{B} \subseteq \alpha \mathcal{D}(i) \oplus (1-\alpha) \mathcal{D}(i) \underset{\text{convexity}}{=} \mathcal{D}(i) \subseteq \mathcal{G}$$

Assumption (Action Set is Star Convex)

For every $x \in \mathcal{X}$ and $\alpha \in [0, 1]$, it holds that $\alpha x \in \mathcal{X}$.

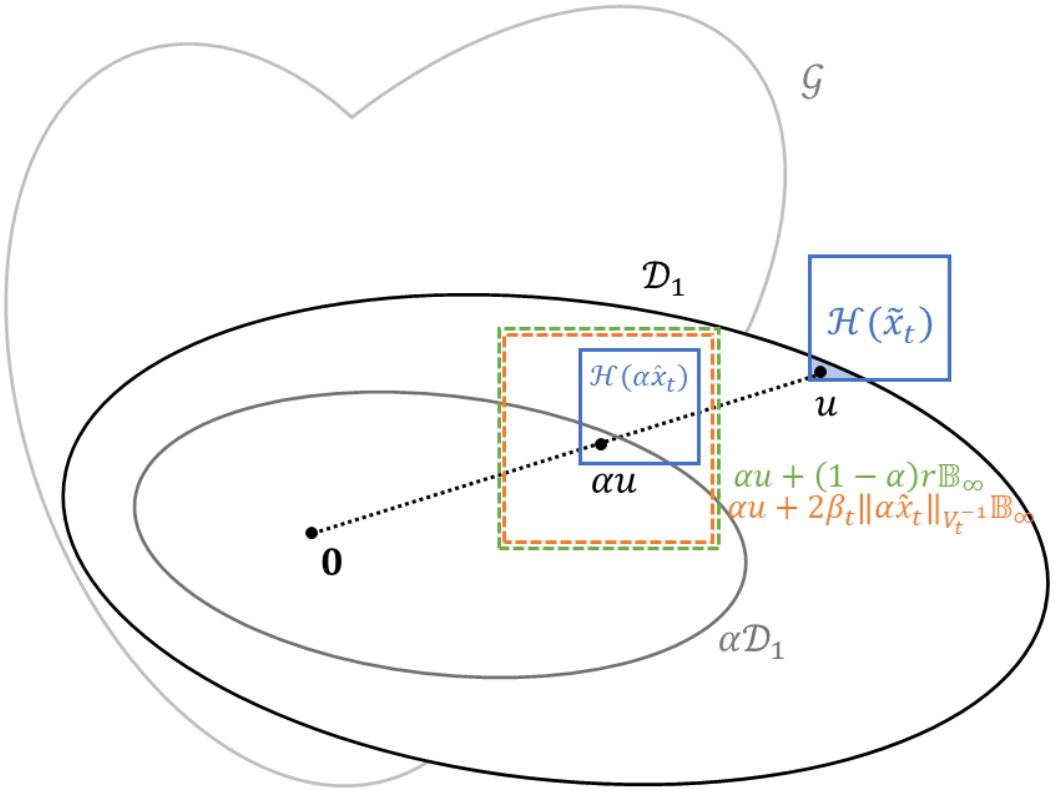
Lemma

It holds for all t that $\gamma_t \geqslant 1 - \frac{2}{r} \beta_t \|x_t\|_{V_t^{-1}}$.

Bounding γ_t: Proof Sketch

Lemma

It holds for all t that $\gamma_t \geqslant 1 - \frac{2}{r}\beta_t \|x_t\|_{V_t^{-1}} =: \alpha.$



$\mathcal{H}(x) := \hat{A}_t x + \beta_t \|x\|_{V_t^{-1}} \mathbb{B}_\infty$ (uncertainty region at x)

Regret Analysis

Theorem

Assume that $\|\theta\| \leqslant S_\theta$, $\|a_i\| \leqslant S_A \ \forall i \in [n]$ *(where* $a_i^\top$ *is* i*th row of* A*), and* $\|x\| \leqslant 1 \ \forall x \in \mathcal{X}$. *Let* $S = \max(S_\theta, S_A)$. *Then, it holds w.h.p. that,*

$$R_T = \tilde{\mathcal{O}}\left(\frac{Sd}{r}\sqrt{T}\right), \qquad \text{and,} \qquad Ax_t \in \mathcal{G} \ \ \forall t \in [T].$$

$$R_T = \sum_{t=1}^{T} \theta^\top x_* - \sum_{t=1}^{T} \theta^\top x_t = \underbrace{\sum_{t=1}^{T} \theta^\top x_* - \sum_{t=1}^{T} \theta^\top \tilde{x}_t}_{\text{Optimistic Regret}} + \underbrace{\sum_{t=1}^{T} \theta^\top \tilde{x}_t - \sum_{t=1}^{T} \theta^\top x_t}_{\text{Cost of Safety}}$$

Optimistic Regret: Regret of optimistic actions $\tilde{x}_t$.

Cost of Safety: Cost incurred by scaling into pessimistic set.

$$\leqslant C \sum_t (1 - \gamma_t)$$

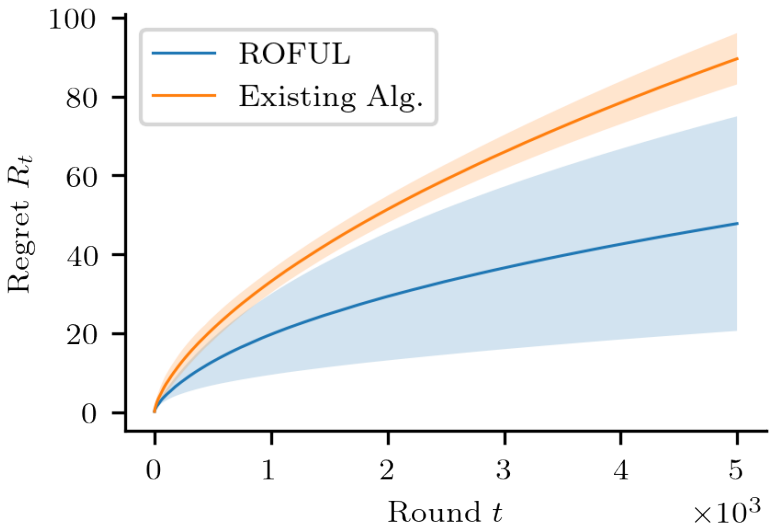
Comparison with Existing Approaches

Refined analysis shows that our algorithm (ROFUL) has smaller regret bound than "expanded UCB" algorithm[5] when,

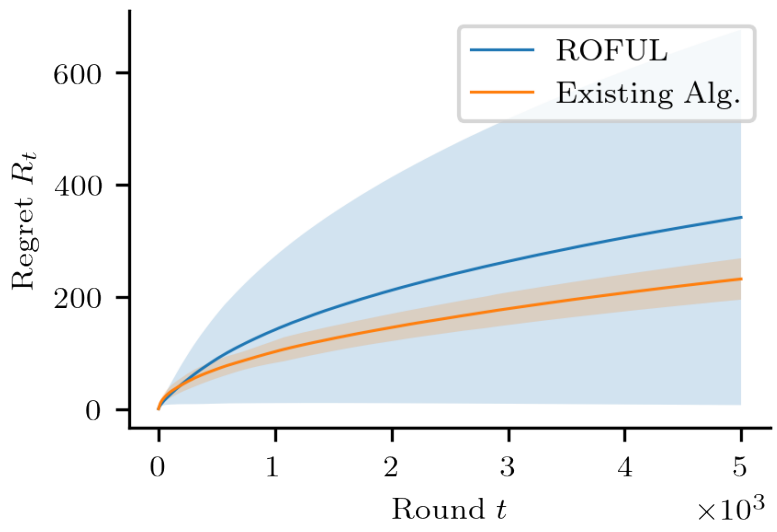
$$S_\theta - \|\theta\| > S_A - r$$

where S_θ is known bound on $\|\theta\|$ and S_A is known bound on $\|a_i\|$ where a_i is ith row of A.

$S_\theta - \|\theta\| = 0.75$
$S_A - r = 0.25$



$S_\theta - \|\theta\| = 0.25$
$S_A - r = 0.75$



[5] Moradipari et al., 2021; Pacchiano et al., 2021

Learning Paradigms

Stochastic Linear Bandit	**Online Convex Optimization**
Stochastic linear reward: $f_t(x) = \theta^\top x + \eta_t$	Adversarial convex cost f_t.
Bandit reward feedback: $y_t := \theta^\top x_t + \eta_t$	Full cost feedback f_t.
Minimize psuedo-regret: $R_T = \sum_{t=1}^{T} \theta^\top (x^\star - x_t)$ $x^\star = \arg\max_{x \in \mathcal{X}: Ax \in \mathcal{G}} \theta^\top x$	Minimize regret: $R_T = \sum_{t=1}^{T} (f_t(x_t) - f_t(x^\star))$ $x^\star = \arg\min_{x \in \mathcal{X}: Ax \in \mathcal{G}} \sum_{t=1}^{T} f_t(x)$

Online Convex Optimization with Unknown Constraints

Interaction Model

In each round $t \in [T]$:

- Choose action x_t from convex action set $\mathcal{X}$.
- Observe adversarial convex cost function $f_t : \mathcal{X} \to \mathbb{R}$.
- Observe constraint value $z_t = Ax_t + \epsilon_t$ (with unknown A).

Learning Goals

- Ensure constraint satisfaction for all rounds: $Ax_t \leqslant b \quad \forall t \in [T]$. (i.e. $\mathcal{G} = b\mathbb{R}_-$)
- Minimize regret w.r.t. to best safe action

$$R_T = \sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{X} : Ax \leqslant b} \sum_{t=1}^{T} f_t(x)$$

Prior Work

Prior work has studied the same problem:[6]

- *Regret:* $\tilde{\mathcal{O}}(T^{2/3})$
- *Constraint guarantee:* $Ax_t \leqslant b \; \forall t$ w.h.p.
- *Approach:* explore-exploit
  $t \in [1, T^{2/3}]: \; x_{t+1} \sim \mathcal{U}$
  $t \in [T^{2/3}, T]: \; x_{t+1} = \Pi_{\mathcal{Y}_t^p}(x_t - \eta \nabla f_t(x_t))$

Our results:

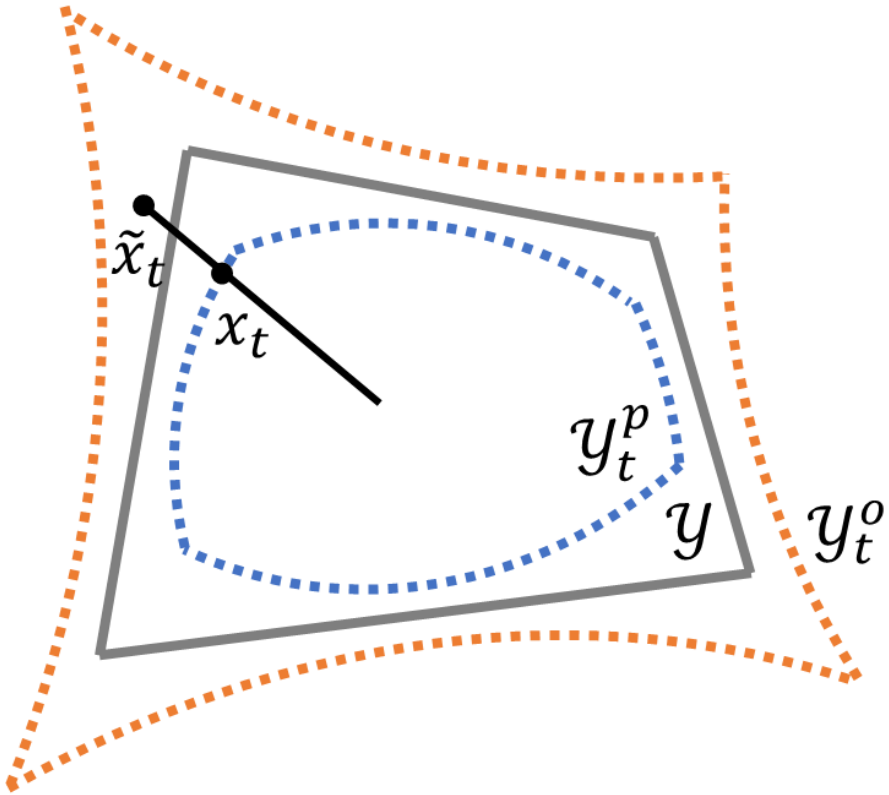
- *Regret:* $\tilde{\mathcal{O}}(\sqrt{T})$
- *Constraint guarantee:* $Ax_t \leqslant b \; \forall t$ w.h.p.
- *Approach:* optimistic

[6] Chaudhary and Kalathil, 2022; Chang et al., 2023

Design Approach

Use same general approach as stochastic linear bandit algorithm:

1. Construct $\mathcal{Y}_t^o$ and $\mathcal{Y}_t^p$.
2. Regret minimizer chooses $\tilde{x}_t \in \mathcal{Y}_t^o$.
3. $\gamma_t = \max\left\{\mu \in [0,1] : \mu\tilde{x}_t \in \mathcal{Y}_t^p\right\}$.
4. Play $x_t = \gamma_t \tilde{x}_t$



Regret Minimization on Optimistic Set

Main challenges: Optimistic set is non-convex and time-varying.
$\implies$ Standard OCO algorithms are not suitable.

Key ideas:

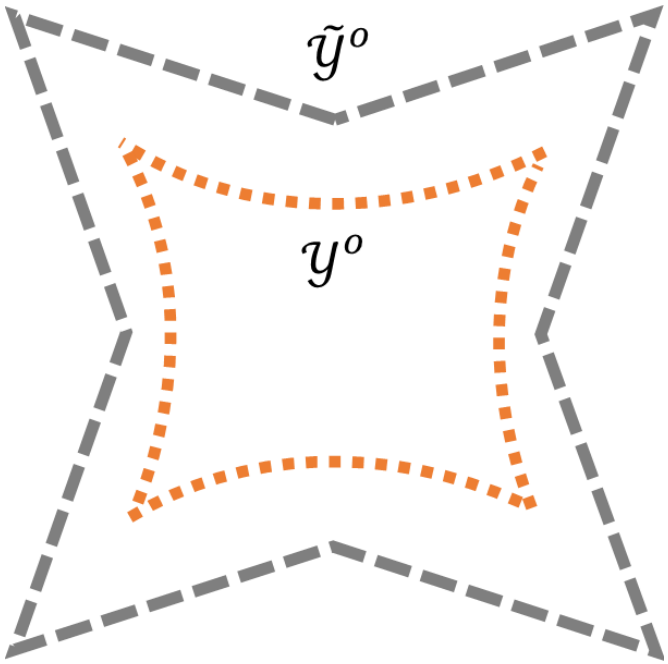
- **Phased Updates:**[7] Fix optimistic set at beginning of phase.
- **Finite Union of Convex Sets:** Represent optimistic set as finite union of convex sets.
- **Online Optimization over Union of Convex Sets:** Run gradient descent in each set and adaptively select the set to play in each round (HedgeDescent).

[7] Abbasi-Yadkori et al., 2011

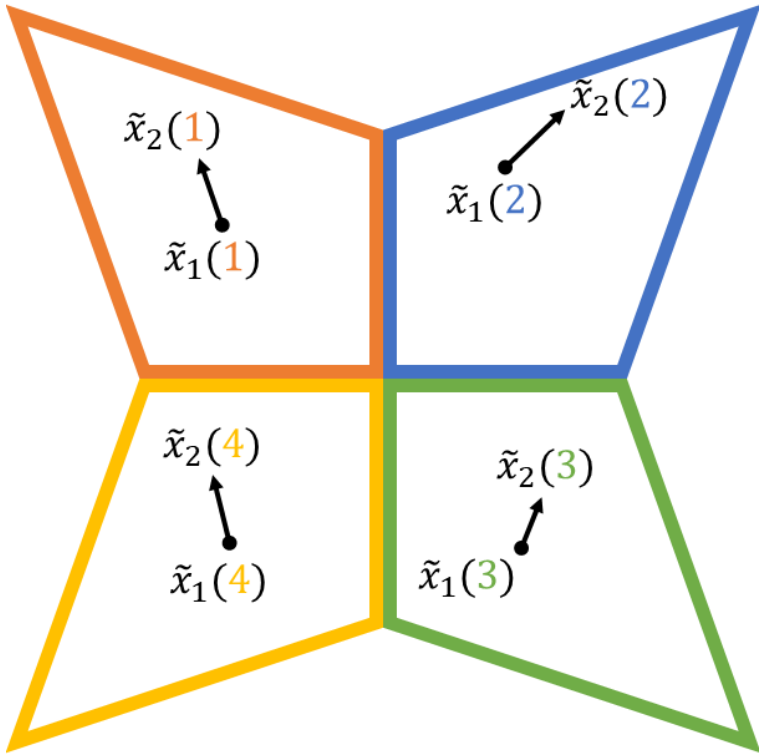
Optimistic Set as Union of Convex Sets

Lemma

$$\mathcal{Y}_t^o \subseteq \tilde{\mathcal{Y}}_t^o := \{x \in \mathcal{X} : \hat{A}_t x - \sqrt{d}\beta_t \|V_t^{-1/2} x\|_\infty \mathbf{1} \leqslant b\}$$
$$= \bigcup_{k \in [d], \xi \in \{-1,1\}} \underbrace{\{x \in \mathcal{X} : \hat{A}_t x - \xi\sqrt{d}\beta_t [V_t^{-1/2}]_k x \mathbf{1} \leqslant b\}}_{\tilde{\mathcal{Y}}_t^o(k,\xi)}.$$



HedgeDescent



HedgeDescent

Algorithm 2: HedgeDescent

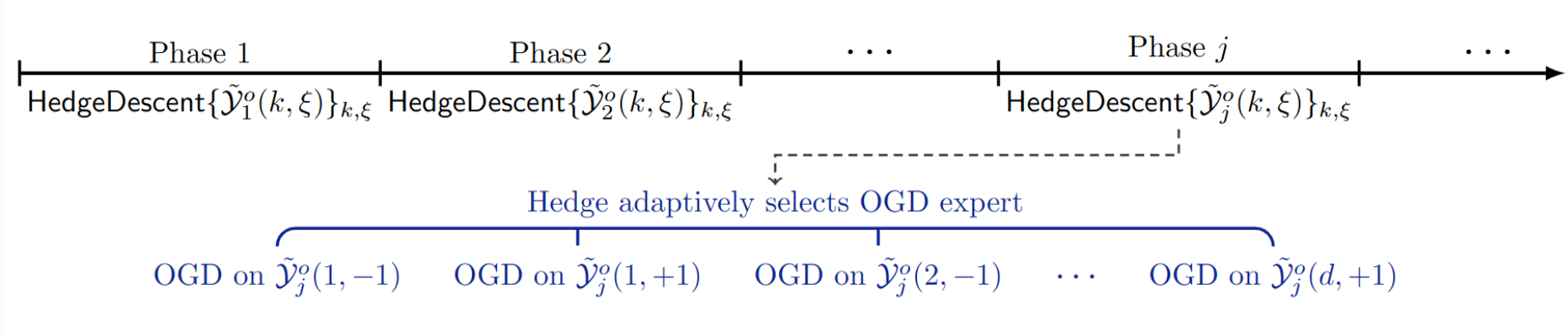
Input: $\tilde{\mathcal{Y}}^o(\cdot,\cdot)$.
Initialize: $p_1(m) = 1/(2d), \tilde{x}_1(m) = \mathbf{0}$ for all $m \in [2d]$.
for $\tau = 1, 2, \ldots$ **do**
  Sample Hedge: $m_\tau \sim p_\tau$.
  Play $\tilde{x}_\tau(m_\tau)$ and observe f_τ.
  Update Descent: $\tilde{x}_{\tau+1}(m) = \Pi_{\tilde{\mathcal{Y}}^o(m_\tau)}(\tilde{x}_\tau(m_\tau) - \eta_t \nabla f_\tau(x_{\tau,m}))$.
  Update Hedge: $p_{\tau+1}(m) \propto p_\tau(m) \exp(-\zeta_t f_\tau(\tilde{x}_\tau(m)))$.
end

Proposition

Assume $\|x\| \leqslant D$ *and* $\|\nabla f_t(x)\| \leqslant G$ *for all* $x \in \mathcal{X}$. *Choosing* $\zeta_t = \sqrt{4\log(2d)}/Gd\sqrt{t}$ *and* $\eta_t = D/G\sqrt{t}$ *ensures that, for all* $N \in \mathbb{N}$,

$$\sum_{\tau\in[N]} \mathbb{E}_{m_\tau\sim p_\tau} f_t(\tilde{x}_\tau(m_\tau)) - \min_{x\in\tilde{\mathcal{Y}}^o} \sum_{\tau\in[N]} f_t(x) \leqslant DG\sqrt{T\log(2d)} + 3DG\sqrt{T}$$

Meta-Algorithm



Meta-Algorithm

Algorithm 2: OSOCO

Initialize: $V_t = \lambda I$.

while $t \leqslant T$ **do**

  Update optimistic sets:

  $\tilde{\mathcal{Y}}_j^o(k,\xi) = \{x \in \mathcal{X} : \hat{A}_t x - \xi\sqrt{d}\beta_t[V^{-1/2}]_k x\mathbf{1} \leqslant b\}$.

  Update pessimistic set $\mathcal{Y}_j^p$.

  Initialize HedgeDescent with $\{\tilde{\mathcal{Y}}_j^o(k,\xi)\}_{k,\xi}$.

  $\bar{V}_j = V_t$.

  while $\det(V_t) \leqslant 2\det(\bar{V}_j)$ ***and*** $t \leqslant T$ **do**

    Receive $\tilde{x}_t$ from HedgeDescent.

    $\gamma_t = \max\left\{\mu \in [0,1] : \mu\tilde{x}_t \in \mathcal{Y}_j^p\right\}$.

    Play $x_t = \gamma_t\tilde{x}_t$ and observe f_t, z_t.

    Send f_t to HedgeDescent.

    $V_{t+1} = V_t + x_t x_t^\top$

  end

end

Regret Analysis

Theorem

Assume $\|x\| \leqslant D$ *and* $\|\nabla f_t(x)\| \leqslant G$ *for all* $x \in \mathcal{X}$*, and that* $\|a_i\| \leqslant S \; \forall i \in [n]$ *and* $b_{\min} := \min_{i \in [n]} b_i > 0$*. Then, w.h.p. the regret of OSOCO satisfies,*

$$R_T = \tilde{\mathcal{O}}\left(d^{3/2}\sqrt{T}\right),$$

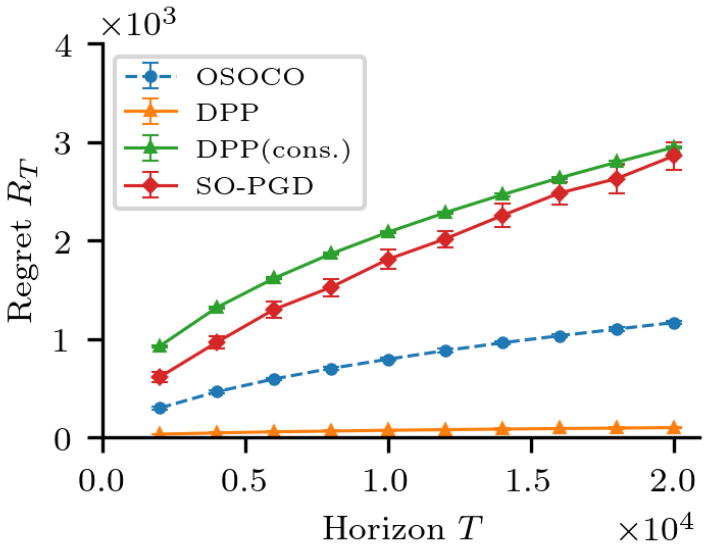
and $Ax_t \leqslant b$ *for all* $t \in [T]$*.*

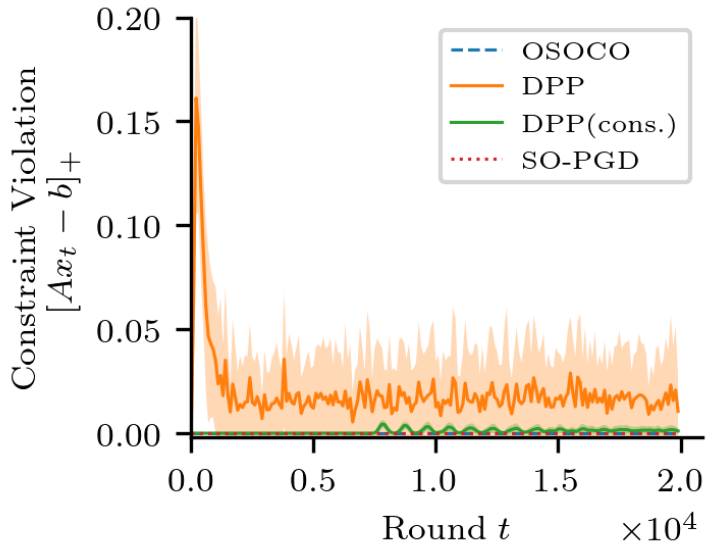
Regret decomposition:

$$R_T = \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x^\star) = \underbrace{\sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(\tilde{x}_t)}_{\text{Cost of Safety}} + \underbrace{\sum_{t=1}^{T} f_t(\tilde{x}_t) - \sum_{t=1}^{T} f_t(x^\star)}_{\text{Optimistic Regret}}$$

Numerical Experiments

Empirical comparison of our algorithm (OSOCO) with existing algorithms.





Existing algorithms are SO-PGD [Chaudhary and Kalathil, 2021] and DPP [Yu et al., 2017].

Conclusion

Introduced a general design approach for online and bandit learning with unknown constraints.

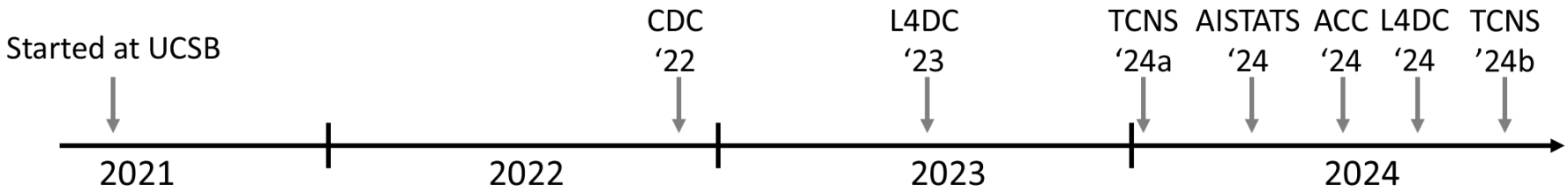
Stochastic Linear Bandits:

- Novel algorithmic approach of combining optimistic and pessimistic estimates of feasible set.
- Analysis with general constraint sets (beyond linear).
- Smaller regret bound and empirical improvements in some settings.

Online Convex Optimization:

- Improved state-of-the-art regret bound from $\tilde{\mathcal{O}}(T^{2/3})$ to $\tilde{\mathcal{O}}(\sqrt{T})$.
- Novel online optimization algorithm for action sets that are union of convex sets.

Timeline



- **CDC '22** *Hutchinson*, Turan and Alizadeh. A Safe Pricing Mechanism for Distributed Resource Allocation with Bandit Feedback.
- **L4DC '23** *Hutchinson*, Turan and Alizadeh. The Impact of the Geometric Properties of the Constraint Set in Safe Optimization with Bandit Feedback.
- **TCNS '24a** *Hutchinson*, Turan and Alizadeh. Safe Pricing Mechanisms for Distributed Resource Allocation with Bandit Feedback.
- **AISTATS '24** *Hutchinson*, Turan and Alizadeh. Directional Optimism for Safe Linear Bandits.
- **ACC '24** *Hutchinson* and Alizadeh. Safe Online Convex Optimization with First-Order Feedback.
- **L4DC '24a** *Hutchinson* and Alizadeh. Safe Online Convex Optimization with Multi-Point Feedback.
- **L4DC '24b** Turan, *Hutchinson*, Alizadeh. Safe Dynamic Pricing for Nonstationary Network Resource Allocation

Future Directions

- Efficient implementations for stochastic bandit setting.
- Demand response in the smart grid.
- Safe learning in related settings, e.g., online control, distributed online optimization.
- Safe learning with nonlinear constraints.
- Projection-free OCO.

Extra slides for stochastic linear bandit setting.

Bounding γ_t

It was shown that $\alpha\tilde{x}_t \in \mathcal{Y}_t^p$ when,

$$(1-\alpha)r = 2\beta_t \|\alpha\tilde{x}_t\|_{V_t^{-1}}.$$

Rearranging shows that,

$$\alpha = \frac{r}{2\beta_t \|\tilde{x}_t\|_{V_t^{-1}} + r}.$$

Then, since $\gamma_t \geqslant \alpha$ and $x_t = \gamma_t \tilde{x}_t$,

$$\begin{aligned}
\gamma_t &\geqslant \frac{r}{2\beta_t \|\tilde{x}_t\|_{V_t^{-1}} + r} \\
\iff \quad & 2\beta_t \|\gamma_t \tilde{x}_t\|_{V_t^{-1}} + \gamma_t r \geqslant r \\
\iff \quad & 2\beta_t \|x_t\|_{V_t^{-1}} + \gamma_t r \geqslant r \\
\iff \quad & \gamma_t \geqslant 1 - \frac{2}{r}\beta_t \|x_t\|_{V_t^{-1}}
\end{aligned}$$

Optimistic Regret

$$
\begin{aligned}
\text{Optimistic Regret} &= \sum_{t\in[T]} \left(\theta^\top x_* - \theta^\top \tilde{x}_t\right) \\
&\leqslant \sum_{t\in[T]} \left(\hat{\theta}_t^\top \tilde{x}_t + \beta_t \|\tilde{x}_t\|_{V_t^{-1}} - \theta^\top \tilde{x}_t\right) && \text{(optimism)} \\
&\leqslant \sum_{t\in[T]} 2\beta_t \|\tilde{x}_t\|_{V_t^{-1}} && \left(\theta^\top \tilde{x}_t \geqslant \hat{\theta}_t^\top \tilde{x}_t - \beta_t \|\tilde{x}_t\|_{V_t^{-1}}\right) \\
&= \sum_{t\in[T]} 2\beta_t \|x_t/\gamma_t\|_{V_t^{-1}} \\
&\leqslant \frac{2S}{r} \sum_{t\in[T]} \beta_t \|x_t\|_{V_t^{-1}} && \left(\gamma_t \geqslant \frac{\nu}{\|\tilde{x}_t\|} \geqslant \nu = \frac{r}{S}\right) \\
&\leqslant \frac{2S}{r} \beta_T \sqrt{2dT \log\left(1 + \frac{T}{\lambda d}\right)}
\end{aligned}
$$

Cost of Safety

$$
\begin{aligned}
\text{Cost of Safety} &= \sum_{t=1}^{T} \left(\theta^\top \tilde{x}_t - \theta^\top x_t\right) \\
&= \sum_{t=1}^{T} \theta^\top \tilde{x}_t \left(1 - \gamma_t\right) \\
&= S \sum_{t=1}^{T} \left(1 - \gamma_t\right) && \left(\theta^\top \tilde{x}_t \leqslant \|\tilde{x}_t\| \|\theta\| \leqslant S\right) \\
&\leqslant \frac{2S}{r} \sum_{t=1}^{T} \beta_t \|x_t\|_{V_t^{-1}} && \left(\gamma_t \geqslant 1 - \frac{2}{r} \beta_t \|x_t\|_{V_t^{-1}}\right) \\
&= \frac{2S}{r} \beta_T \sqrt{2dT \log\left(1 + \frac{T}{\lambda d}\right)}
\end{aligned}
$$

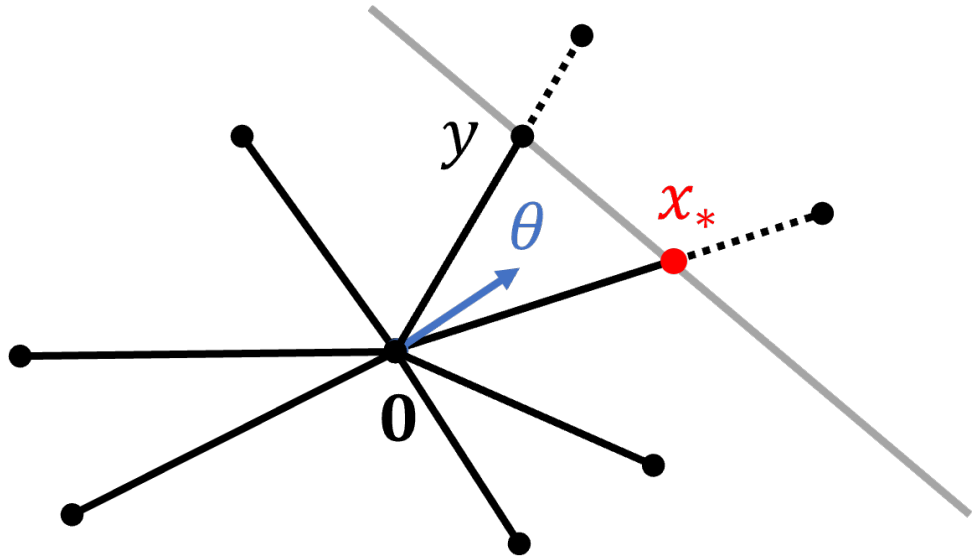
Problem-Dependent Analysis

Definition

The *reward gap* of a problem instance is defined as the gap in (expected) reward between the best and second-best directions,

$$\Delta := \inf_{x \in \mathcal{Y}:\ x \neq \alpha x_* \ \forall \alpha > 0} \theta^\top (x_* - x), \tag{1}$$

where $x_* = \arg\max_{x \in \mathcal{Y}} \theta^\top x$ and $\mathcal{Y} = \{x \in \mathcal{X} : Ax \in \mathcal{G}\}$.



$$\Delta = \theta^\top (x_* - y)$$

Problem-Dependent Analysis

Theorem

When $\Delta > 0$, the number of wrong directions chosen by ROFUL satisfies w.h.p.

$$B_T := \sum_{t=1}^{T} \mathbb{1}\{\nexists\, \alpha > 0 : x_t = \alpha x_*\} \leqslant \mathcal{O}\left(\frac{d^2}{\Delta^2} \log^2(T)\right)$$

PD-ROFUL Algorithm:

1. Play ROFUL until any single direction is played at least $\bar{B}$ times.
2. Play largest safe scaling in this direction for the remaining rounds.

Corollary

PD-ROFUL with $\bar{B} = \mathcal{O}(\frac{d^2}{\Delta^2} \log^2(T))$ (when Δ is known) results in $\tilde{\mathcal{O}}(\frac{d^2}{\Delta} + \sqrt{T})$ regret.

Problem-Dependent Analysis

Study regret due to choice of direction:

$$\tilde{R}_T := \sum_{t=1}^{T} \theta^\top (x_* - \alpha_t x_t) \qquad \alpha_t := \max\{\alpha \geqslant 0 : \alpha x_t \in \mathcal{Y}\}$$

$$\geqslant \sum_{t=1}^{T} \Delta \mathbb{1}\{\nexists\, \alpha > 0 : x_t = \alpha x_*\} = \Delta B_T$$

Lemma

It holds w.h.p. that

$$\tilde{r}_t := \theta^\top (x_* - \alpha_t x_t) \leqslant \frac{4S}{r} \beta_t \|x_t\|_{V_t^{-1}}$$

Then, because $\tilde{r}_t \leqslant (\tilde{r}_t)^2/\Delta$, it holds that

$$B_T \leqslant \frac{\tilde{R}_T}{\Delta} = \frac{1}{\Delta} \sum_{t=1}^{T} \tilde{r}_t \leqslant \frac{1}{\Delta^2} \sum_{t=1}^{T} (\tilde{r}_t)^2 \leqslant \frac{4S^2}{r^2 \Delta^2} \beta_t^2 \|x_t\|^2_{V_t^{-1}}$$

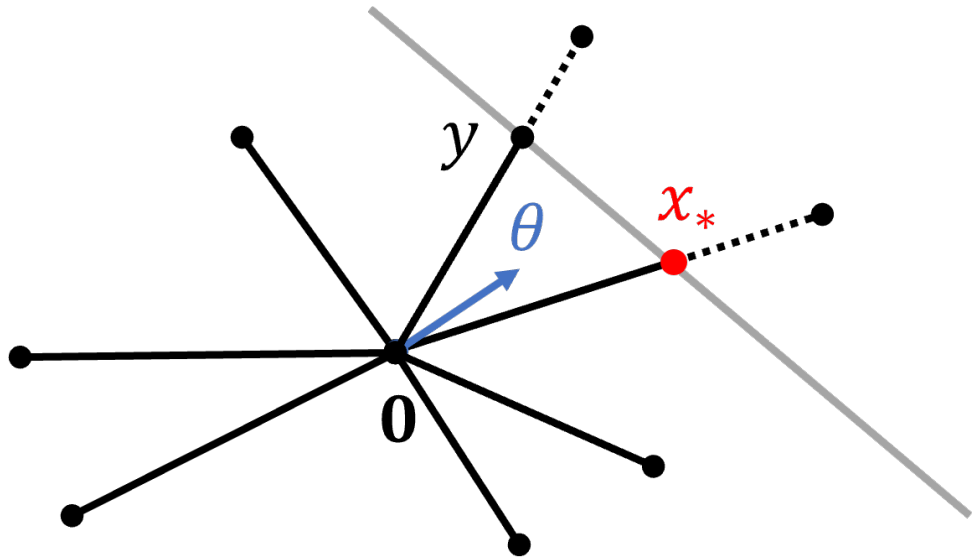
Problem-Dependent Analysis

Definition

The *reward gap* of a problem instance is defined as the gap in (expected) reward between the best and second-best directions,

$$\Delta := \inf_{x \in \mathcal{Y}:\ x \neq \alpha x_* \ \forall \alpha > 0} \theta^\top (x_* - x), \tag{2}$$

where $x_* = \arg\max_{x \in \mathcal{Y}} \theta^\top x$.



$$\Delta = \theta^\top (x_* - y)$$

Extra slides for online convex optimization setting.

Representation of Pessimistic Set

$$\mathcal{Y}_t^p = \{x \in \mathcal{X} : (\hat{A}^\top x + \beta_t \|x\|_{V_t^{-1}} \mathbb{B}_\infty) \subseteq \mathcal{G}\}$$

When $\mathcal{G} = b\mathbb{R}_-$, a given $x \in \mathcal{X}$ is in $\mathcal{Y}_t^p$ iff the following is non-positive.

$$\begin{aligned}
&\max_{u \in \mathbb{B}_\infty} \max_{i \in [n]} \left(\hat{a}_i^\top x - b_i + \beta_t \|x\|_{V_t^{-1}} u_i \right) \\
&= \max_{\substack{u_i \in [-1,1] \\ \forall i}} \max_{i \in [n]} \left(\hat{a}_i^\top x - b_i + \beta_t \|x\|_{V_t^{-1}} u_i \right) \\
&= \max_{i \in [n]} \left(\hat{a}_i^\top x - b_i + \beta_t \|x\|_{V_t^{-1}} \right).
\end{aligned}$$

Therefore, we can write,

$$\mathcal{Y}_t^p = \{x \in \mathcal{X} : \hat{A}^\top x + \beta_t \|x\|_{V_t^{-1}} \mathbf{1} \leqslant b\}.$$

Representation of Optimistic Set

$$\mathcal{Y}_t^o = \{x \in \mathcal{X} : (\hat{A}^\top x + \beta_t \|x\|_{V_t^{-1}} \mathbb{B}_\infty) \cap \mathcal{G} \neq \emptyset\}$$

When $\mathcal{G} = b\mathbb{R}_-$, a given $x \in \mathcal{X}$ is in $\mathcal{Y}_t^o$ iff the following is non-positive.

$$\begin{aligned}
&\min_{u \in \mathbb{B}_\infty} \max_{i \in [n]} \left(\hat{a}_i^\top x - b_i + \beta_t \|x\|_{V_t^{-1}} u_i \right) \\
&= \min_{\substack{u_i \in [-1,1] \\ \forall i \neq j}} \min_{u_j \in [-1,1]} \underbrace{\max_{i \in [n]} \left(\hat{a}_i^\top x - b_i + \beta_t \|x\|_{V_t^{-1}} u_i \right)}_{\text{non-decreasing in } u_j} \\
&= \min_{\substack{u_i \in [-1,1] \ \forall i \neq j \\ u_j = -1}} \max_{i \in [n]} \left(\hat{a}_i^\top x - b_i + \beta_t \|x\|_{V_t^{-1}} u_i \right) \\
&\ \ \vdots \\
&= \max_{i \in [n]} \left(\hat{a}_i^\top x - b_i - \beta_t \|x\|_{V_t^{-1}} \right)
\end{aligned}$$

Relaxed Optimistic Set

It holds that,

$$\mathcal{Y}_t^o \subseteq \tilde{\mathcal{Y}}_t^o := \{x \in \mathcal{X} : \hat{A}_t x - \sqrt{d}\beta_t \| V_t^{-1/2} x \|_\infty \mathbf{1} \leqslant b\}$$

This is because, for all $x \in \mathcal{Y}_t^o$, it holds that,

$$\begin{aligned} b &\geqslant \hat{A}_t x - \beta_t \| x \|_{V_t^{-1}} \mathbf{1} \\ &= \hat{A}_t x - \beta_t \| V_t^{-1/2} x \|_2 \mathbf{1} \\ &\geqslant \hat{A}_t x - \sqrt{d}\beta_t \| V_t^{-1/2} x \|_\infty \mathbf{1}, \end{aligned}$$

and therefore $x \in \tilde{\mathcal{Y}}_t^o$.

Relaxed Optimistic Set as Union of Convex Set

A given $x \in \mathcal{X}$ is in $\tilde{\mathcal{Y}}_t^o$ iff the following is non-positive.

$$\begin{aligned}
&\max_{i\in[n]} \left(\hat{a}_i^\top x - b_i - \sqrt{d}\beta_t \| V_t^{-1/2} x \|_\infty \right) \\
&= \max_{i\in[n]} \left(\hat{a}_i^\top x - b_i - \sqrt{d}\beta_t \max_{k\in[d],\xi\in\{-1,1\}} \xi [V_t^{-1/2}]_k x \right) \\
&= \max_{i\in[n]} \left(\hat{a}_i^\top x - b_i \right) + \min_{k\in[d],\xi\in\{-1,1\}} \left(-\sqrt{d}\beta_t \xi [V_t^{-1/2}]_k x \right) \\
&= \min_{k\in[d],\xi\in\{-1,1\}} \left(\max_{i\in[n]} \left(\hat{a}_i^\top x - b_i - \sqrt{d}\beta_t \xi [V_t^{-1/2}]_k x \right) \right)
\end{aligned}$$

Therefore, we can write,

$$\tilde{\mathcal{Y}}_t^o = \bigcup_{k\in[d],\xi\in\{-1,1\}} \{x \in \mathcal{X} : \hat{A}_t x - \xi\sqrt{d}\beta_t [V_t^{-1/2}]_k x \mathbf{1} \leqslant b\}.$$

Cost of Safety

$$
\begin{aligned}
\text{Cost of Safety} &= \sum_{t\in[T]} (f_t(x_t) - f_t(\tilde{x}_t)) \\
&= \sum_{j\in[N]} \sum_{t\in\mathcal{T}_j} (f_t(x_t) - f_t(\tilde{x}_t)) && (j = \text{phase index}) \\
&\leqslant GD \sum_{j\in[N]} \sum_{t\in\mathcal{T}_j} (1-\gamma_t) \\
&\leqslant \frac{2GD\sqrt{d}}{b_{\min}} \beta_T \sum_{j\in[N]} \sum_{t\in\mathcal{T}_j} \|x_t\|_{\bar{V}_{j_t}^{-1}} \\
&\leqslant \frac{4GD\sqrt{d}}{b_{\min}} \beta_T \sum_{t\in[T]} \|x_t\|_{V_t^{-1}} && \left(\det(V_t) \leqslant 2\det(\bar{V}_{j_t})\right)
\end{aligned}
$$

Optimistic Regret

$$
\begin{aligned}
\text{Optimistic Regret} &= \sum_{t\in[T]} (f_t(\tilde{x}_t) - f_t(x^\star)) \\
&= \sum_{t\in[T]} (f_t(\tilde{x}_t) - \mathbb{E}_t[f_t(\tilde{x}_t)]) + \sum_{t\in[T]} (\mathbb{E}_t[f_t(\tilde{x}_t)] - f_t(x^\star)) \\
&\leqslant \sum_{t\in[T]} (\mathbb{E}_t[f_t(\tilde{x}_t)] - f_t(x^\star)) + \tilde{\mathcal{O}}(\sqrt{T}) \qquad \text{(Azuma's)} \\
&= \sum_{j\in[N]} \sum_{t\in\mathcal{T}_j} (\mathbb{E}_t[f_t(\tilde{x}_t)] - f_t(x^\star)) + \tilde{\mathcal{O}}(\sqrt{T}) \\
&\leqslant C \sum_{j\in[N]} \sqrt{T_j} + \tilde{\mathcal{O}}(\sqrt{T}) \qquad \text{(HedgeDescent)} \\
&\leqslant C\sqrt{NT} + \tilde{\mathcal{O}}(\sqrt{T}) \qquad \text{(Cauchy-Schwarz)} \\
&\leqslant \tilde{\mathcal{O}}(\sqrt{dT}) \qquad (N = \mathcal{O}(d\log(T)))
\end{aligned}
$$