# No-Regret Linear Bandits beyond Realizability (Supplementary Material)

**Chong Liu**[1] **Ming Yin**[1] **Yu-Xiang Wang**[1]

[1] Department of Computer Science, University of California, Santa Barbara, CA 93106, USA

## 1 PROOF OF PROPOSITION 3

Equivalently, $\rho$-gap-adjusted misspecification (Definition 2) satisfies

$$|f(x) - f_0(x)| \leq \rho\,|f^* - f_0(x)|\,, \ \ \forall x \in \mathcal{X}. \tag{1}$$

*Proof of preservation of max value:* $\max_{x\in\mathcal{X}} f(x) = f^*$. Let $f_w^* := \max_{x\in\mathcal{X}} f(x)$. We first prove $f_w^* \leq f^*$ by contradiction. Suppose $f_w^* > f^*$, since $\mathcal{X}$ is compact, there exists $x_w \in \mathcal{X}$ such that $f(x_w) = f_w^* > f^*$. Then by eq. (1) this implies

$$f(x_w) - f_0(x_w) \leq \rho(f^* - f_0(x_w)) \Rightarrow f^* < f_w^* = f(x_w) \leq \rho f^* + (1-\rho) f_0(x_w) \leq f^*$$

Contraction! Therefore, $f_w^* \leq f^*$. On the other hand, choose $x_0 \in \operatorname{argmax}_{x\in\mathcal{X}} f_0(x)$, then by eq. (1) $f(x_0) = f_0(x_0) = f^*$. This implies $f_w^* \geq f^*$. Combing both results to obtain $f_w^* = f^*$. □

*Proof of preservation of maximizers:* $\operatorname{argmax}_x f(x) = \operatorname{argmax}_x f_0(x)$. Using that $f(x) \leq \rho f^* + (1-\rho) f_0(x)$ and $\max_{x\in\mathcal{X}} f(x) = f^*$, it is easy to verify $\operatorname{argmax}_x f(x) \subset \operatorname{argmax}_x f_0(x)$. On the other hand, if $x' \in \operatorname{argmax}_x f_0(x)$, then by eq. (1) $f(x') = f_0(x') = f^*$ and this means $\operatorname{argmax}_x f_0(x) \subset \operatorname{argmax}_x f(x)$. □

*Proof of self-bounding property.* This directly comes from the definition. □

## 2 PROPERTY OF WEAK $\rho$-GAP-ADJUSTED MISSPECIFICATION

First we recall Definition 4.

**Definition 1** (Restatement of Weak $\rho$-gap-adjusted misspecification)**.** *Denote $f_w^* = \max_{x\in\mathcal{X}} f(x)$. Then we say $f$ is (weak) $\rho$-gap-adjusted misspecification approximation of $f_0$ for a parameter $0 \leq \rho < 1$ if:*

$$\sup_{x\in\mathcal{X}} \left| \frac{f(x) - f_w^* + f^* - f_0(x)}{f^* - f_0(x)} \right| \leq \rho.$$

Under the weak $\rho$-gap-adjusted misspecification condition, it no longer holds $f_w^* = f^*$. However, it still preserves the maximizers.

**Proposition 2.** *Under the weak $\rho$-gap-adjusted misspecification condition, it holds*

$$\operatorname*{argmax}_x f(x) = \operatorname*{argmax}_x f_0(x).$$

*Accepted for the 39th Conference on Uncertainty in Artificial Intelligence* (UAI 2023).

*Proof.* Suppose $x' \in \operatorname{argmax}_x f(x)$, then by definition

$$|f^* - f_0(x')| = |f(x') - f_w^* + f^* - f_0(x')| \le \rho|f^* - f_0(x')| \Rightarrow (1-\rho)|f^* - f_0(x')| \le 0 \Rightarrow x' \in \operatorname*{argmax}_x f_0(x).$$

On the other hand, if $x' \in \operatorname{argmax}_x f_0(x)$, then

$$|f_w^* - f(x')| = |f(x') - f_w^* + f^* - f_0(x')| \le \rho|f^* - f_0(x')| = 0 \Rightarrow x' \in \operatorname*{argmax}_x f(x).$$

□

The next proposition shows the weak $\rho$-adjusted misspecification condition characterizes the suboptimality gap between $f$ and $f_0$.

**Proposition 3.** *Denote $g(x) := f_w^* - f(x) \ge 0$, $g_0(x) := f^* - f_0(x) \ge 0$, then the weak $\rho$-gap-adjusted misspecification condition implies:*

$$(1-\rho)g_0(x) \le g(x) \le (1+\rho)g_0(x), \quad x \in \mathcal{X}.$$

This can be proved directly by the triangular inequality. This reveals the weak $\rho$-gap-adjusted misspecification condition requires $g(x)$ to live in the band $[(1-\rho)g_0(x), (1+\rho)g_0(x)]$, and the concrete maximum values $f_w^*$ and $f^*$ can be arbitrarily different.

## 3 LINEAR BANDITS UNDER THE WEAK $\rho$-GAP-ADJUSTED MISSPECIFICATION

We need to slightly modify LinUCB [Abbasi-yadkori et al., 2011] and work with the following LinUCBw algorithm.

**Algorithm 1** LinUCBw (adapted from Abbasi-yadkori et al. [2011])

```
Input: Predefined sequence $\beta_t$ for $t = 1, 2, 3, ...$ as in eq. (2); Set $\lambda = \sigma^2/C_w^2$ and $\text{Ball}_0 = \mathcal{W}$.
for $t = 0, 1, 2, ...$ do
    Select $x_t = \operatorname{argmax}_{x\in\mathcal{X}} \max_{[w^\top, c]\in\text{Ball}_t} [w^\top, c]\begin{bmatrix} x \\ 1 \end{bmatrix}$.
    Observe $y_t = f_0(x_t) + \eta_t$.
    Update
        $\Sigma_{t+1} = \lambda I_{d+1} + \sum_{i=0}^{t} \begin{bmatrix} x_i \\ 1 \end{bmatrix} \cdot [x_i^\top, 1]$ where $\Sigma_0 = \lambda I_{d+1}$.
    Update
        $\begin{bmatrix} \hat{w}_{t+1} \\ \hat{c}_{t+1} \end{bmatrix} = \operatorname*{argmin}_{w,c} \lambda \left\| \begin{bmatrix} w \\ c \end{bmatrix} \right\|_2^2 + \sum_{i=0}^{t} (w^\top x_i + c - y_i)_2^2.$
    Update
        $\text{Ball}_{t+1} = \left\{ \begin{bmatrix} w \\ c \end{bmatrix} \,\middle|\, \left\| \begin{bmatrix} w \\ c \end{bmatrix} - \begin{bmatrix} \hat{w}_{t+1} \\ \hat{c}_{t+1} \end{bmatrix} \right\|_{\Sigma_{t+1}}^2 \le \beta_{t+1} \right\}.$
end for
```

**Theorem 4.** *Suppose Assumptions 5, 6, and 7 hold. W.l.o.g., assuming $c^* = f^* - f_w^* \le F$. Set*

$$\beta_t = 8\sigma^2 \left(1 + (d+1)\log\left(1 + \frac{tC_b^2(C_w^2 + F^2)}{d\sigma^2}\right) + 2\log\left(\frac{\pi^2 t^2}{3\delta}\right)\right). \tag{2}$$

*Then Algorithm 1 guarantees w.p.* $> 1-\delta$ *simultaneously for all* $T = 1, 2, ...$

$$R_T \leq F + c^* + \sqrt{\frac{8(T-1)\beta_{T-1}(d+1)}{(1-\rho)^2}\log\left(1+\frac{TC_b^2(C_w^2+F^2)}{d\sigma^2}\right)}.$$

**Remark 5.** *The result again shows that LinUCBw algorithm achieves* $\tilde{O}(\sqrt{T})$ *cumulative regret and thus it is also a no-regret algorithm under the weaker condition (Definition 4). Note Definition 4 is quite weak which even doesn't require the true function sits within the approximation function class.*

*Proof.* The analysis is similar to the $\rho$-gap-adjusted case but includes $c^* = f^* - f_w^*$. For instance, let $\Delta_t^w$ denote the deviation term of our linear function from the true function at $x_t$, then

$$\Delta_t^w = f_0(x_t) - w_*^\top x_t - c^*,$$

And our observation model (eq. (1)) becomes

$$y_t = f_0(x_t) + \eta_t = w_*^\top x_t + c^* + \Delta_t^w + \eta_t.$$

Then similar to Lemma 10, we have the following lemma, whose proof is nearly identical to Lemma 10.

**Lemma 6** (Bound of deviation term). $\forall t \in \{0, 1, \ldots, T-1\}$,

$$|\Delta_t| \leq \frac{\rho}{1-\rho} w_*^\top (x_* - x_t).$$

We also provide the following lemma, which is the counterpart of Lemma 13.

**Lemma 7.** *Define* $u_t = \left\|\begin{bmatrix} x_t \\ 1 \end{bmatrix}\right\|_{\Sigma_t^{-1}}$ *and assume* $\beta_t$ *is chosen such that* $w_* \in \mathrm{Ball}_t$*. Then*

$$w_*^\top (x_* - x_t) \leq 2\sqrt{\beta_t} u_t.$$

*Proof.* Let $\tilde{w}, \tilde{c}$ denote the parameter that achieves $\mathrm{argmax}_{w,c\in\mathrm{Ball}_t}\, w^\top x_t + c$, by the optimality of $x_t$,

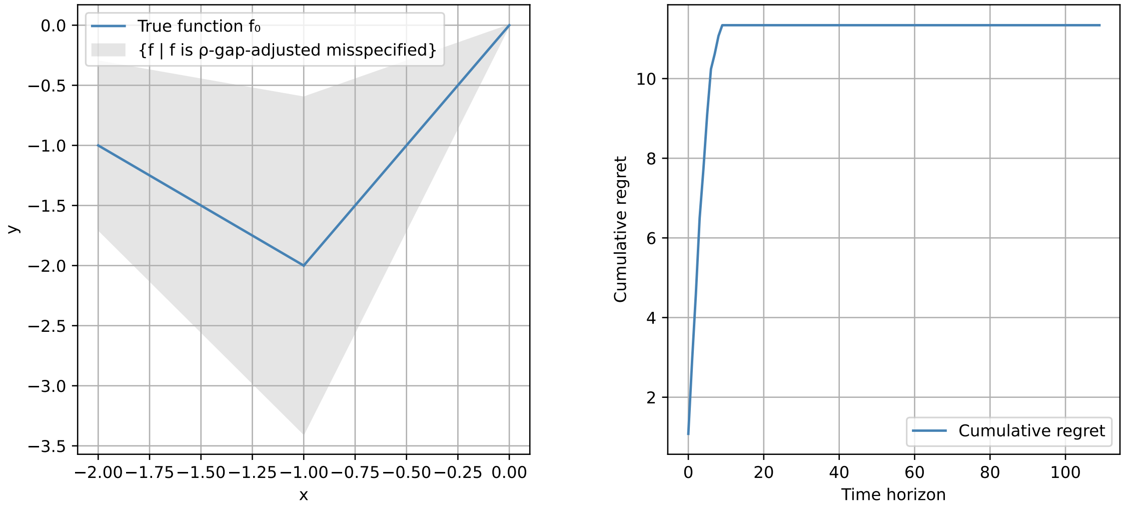
$$\begin{aligned} w_*^\top x_* - w_*^\top x_t &= \left[w_*^\top, c^*\right]\begin{bmatrix} x_* \\ 1 \end{bmatrix} - \left[w_*^\top, c^*\right]\begin{bmatrix} x_t \\ 1 \end{bmatrix} \\ &\leq \left[\tilde{w}^\top, \tilde{c}\right]\begin{bmatrix} x_t \\ 1 \end{bmatrix} - \left[w_*^\top, c^*\right]\begin{bmatrix} x_t \\ 1 \end{bmatrix} \\ &= \left(\left[\tilde{w}^\top, \tilde{c}\right] - \left[\hat{w}_t^\top, \hat{c}_t\right] + \left[\hat{w}_t^\top, \hat{c}_t\right] - \left[w_*^\top, c^*\right]\right)\begin{bmatrix} x_t \\ 1 \end{bmatrix} \\ &\leq \left\|\left[\tilde{w}^\top, \tilde{c}\right] - \left[\hat{w}_t^\top, \hat{c}_t\right]\right\|_{\Sigma_t}\left\|\begin{bmatrix} x_t \\ 1 \end{bmatrix}\right\|_{\Sigma_t^{-1}} + \left\|\left[\hat{w}_t^\top, \hat{c}_t\right] - \left[w_*^\top, c^*\right]\right\|_{\Sigma_t}\left\|\begin{bmatrix} x_t \\ 1 \end{bmatrix}\right\|_{\Sigma_t^{-1}} \\ &\leq 2\sqrt{\beta_t} u_t \end{aligned}$$

where the second inequality applies Holder's inequality; the last line uses the definition of $\mathrm{Ball}_t$ (note that both $\left[\tilde{w}^\top, \tilde{c}\right], \left[w_*^\top, c^*\right] \in \mathrm{Ball}_t$). □

The rest of the analysis follows the analysis of Theorem 8. □

## 4 SIMULATION

In this section, we run a simulation on a 1-dimensional test function shown in Figure 1(a). Here we run the first 10 iterations with uniform sampling and the remaining 100 iterations are using LinUCB algorithm. In Figure 1(b) we can see that cumulative regret is increasing with uniform sampling but it doesn't increase when running LinUCB. The reason behind it is that under the gap-adjusted misspecification, LinUCB is able to quickly find the optimal point $x_* = 0$.



(a) 1-dimensional test function.

(b) Cumulative regret

Figure 1: Simulation function and result.

## References

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.