Bandits in Flux: Adversarial Constraints in Dynamic Environments

Tareq Si Salem
Paris Research Center, Huawei Technologies | AISTATS 2026

Motivation and Contributions

We study adversarial multi-armed bandits with time-varying soft constraints. At each round, the learner observes only the cost and constraint of the sampled arm, while the environment may change over time.

- Introduce BCOMD, a bandit-feedback constrained online mirror descent algorithm.
- Combine primal entropic mirror descent with a dual update that tracks cumulative violation.
- Establish optimal dynamic regret under two complementary non-stationarity measures: path length P_T and temporal variation V_T.
- Achieve $\tilde{O}(\sqrt{T})$ cumulative constraint violation, improving over prior dynamic constrained bandit rates.

Learn from bandit feedback + adapt to drift + satisfy long-run constraints

Problem Setting

There are n arms, $\mathcal{A} = \{1, \dots, n\}$. At time t, the learner chooses a distribution $\mathbf{x}_t \in \triangle_n$ and samples $a_t \sim \mathbf{x}_t$. It observes only

$$f_t(a_t) \in [0,1], \qquad g_t(a_t) \in [-1,1].$$

Expected loss and constraint are linear:

$$f_t(\mathbf{x}) = \mathbf{f}_t \cdot \mathbf{x}, \qquad g_t(\mathbf{x}) = \mathbf{g}_t \cdot \mathbf{x}.$$

Soft feasibility is enforced in the long run. Slater feasibility holds: there exists $\mathbf{x} \in \triangle_n$ and $\rho > 0$ such that

$$g_t(\mathbf{x}) \le -\rho, \qquad t = 1, \dots, T.$$

The dynamic comparator is the best feasible sequence in hindsight:

$$\{\mathbf{x}_t^\star\}_{t=1}^T \in \operatorname*{arg\,min}_{\mathbf{x}_t \in \triangle_n,\ g_t(\mathbf{x}_t) \le 0} \sum_{t=1}^T f_t(\mathbf{x}_t).$$

Performance metrics:

$$\mathfrak{R}_T = \mathbb{E}\left[\sum_{t=1}^T f_t(a_t)\right] - \sum_{t=1}^T f_t(\mathbf{x}_t^\star), \qquad \mathfrak{V}_T = \mathbb{E}\left[\sum_{t=1}^T g_t(a_t)\right].$$

Measuring Non-Stationarity

Dynamic regret is impossible without controlling environmental drift. We use two incomparable measures:

$$P_T = \sum_{t=1}^{T-1} \|\mathbf{x}_{t+1}^\star - \mathbf{x}_t^\star\|_1 \quad \textbf{path length},$$

$$V_T = \sum_{t=1}^{T-1} \|\mathbf{f}_{t+1} - \mathbf{f}_t\|_\infty \quad \textbf{temporal variation}.$$

Small P_T: comparator moves slowly — incomparable — Small V_T: costs change slowly

Representative Application Domains

The model captures systems that must keep learning while respecting a long-run budget or safety envelope:

- hyperparameter selection with prediction-time constraints;
- wireless scheduling with cumulative energy constraints;
- recommendation and engagement optimization with fairness constraints;
- budget-constrained advertisement allocation;
- energy-aware client selection in federated learning.

The soft-constraint objective is not zero violations at every time, but vanishing time-average violation:

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}[g_t(a_t)] \to 0.$$

BCOMD: Bandit-feedback Constrained Online Mirror Descent

The Lagrangian at time t is

$$\Psi_t(\mathbf{x}, \lambda) = f_t(\mathbf{x}) + \lambda g_t(\mathbf{x}).$$

Under bandit feedback, use importance-weighted unbiased gradient estimates:

$$\widetilde{\mathbf{f}}_t = \frac{f_t(a_t)}{x_{t,a_t}}\mathbf{e}_{a_t}, \qquad \widetilde{\mathbf{g}}_t = \frac{g_t(a_t)}{x_{t,a_t}}\mathbf{e}_{a_t}.$$

Input: $\mathbf{x}_1 = \mathbf{1}/n$, $\lambda_1 = 0$, η, μ, γ
for $t = 1, \dots, T$ **do**
  sample $a_t \sim \mathbf{x}_t$
  observe $f_t(a_t), g_t(a_t)$
  $\widetilde{\mathbf{f}}_t = \frac{f_t(a_t)}{x_{t,a_t}}\mathbf{e}_{a_t}$
  $\widetilde{\mathbf{g}}_t = \frac{g_t(a_t)}{x_{t,a_t}}\mathbf{e}_{a_t}$
  $\widetilde{\boldsymbol{\omega}}_t = \Omega x_{t,a_t}^{-1}\mathbf{e}_{a_t}$
  $\widetilde{\mathbf{h}}_t = \widetilde{\boldsymbol{\omega}}_t + \widetilde{\mathbf{f}}_t + \lambda_t \widetilde{\mathbf{g}}_t$
  $\mathbf{y}_{t+1} = (\nabla\Phi)^{-1}(\nabla\Phi(\mathbf{x}_t) - \eta\widetilde{\mathbf{h}}_t)$
  $\mathbf{x}_{t+1} = \Pi_{\triangle_{n,\gamma}}^{\Phi}(\mathbf{y}_{t+1})$
  $\lambda_{t+1} = [\lambda_t + \mu g_t(a_t)]_+$
end for

Entropic instantiation:

$$\Phi(\mathbf{x}) = \sum_{i=1}^n x_i \log x_i, \qquad \triangle_{n,\gamma} = \triangle_n \cap [\gamma, 1]^n.$$

The update is multiplicative in the dual geometry, then Bregman-projected back to the restricted simplex.

Why Entropic Mirror Descent?

Euclidean OGD is ill-suited because the bandit estimator has unbounded second moment on the simplex:

$$\sup_{\mathbf{x} \in \triangle_n} \mathbb{E}_{a \sim \mathbf{x}}\left[\left\|\frac{f_t(a)}{x_a}\mathbf{e}_a\right\|_2^2\right] = \infty.$$

When $x_a \to 0$, the importance-weighted estimator explodes.

Key technical move. Negative entropy replaces Euclidean variance control with Bregman control:

$$\mathbb{E}\left[D_\Phi(\mathbf{x}_t, \mathbf{y}_{t+1}) \mid \mathcal{H}_{t-1}\right] \le 1.5\eta^2 n(1 + \lambda_t^2 + \Omega^2).$$

The dual variable is bounded by a carefully chosen Ω, enabling simultaneous regret and feasibility guarantees.

Euclidean OGD unbounded variance — replace geometry — Entropic OMD controlled variance

Main Theoretical Guarantees

Let $c_T = \min\{\sqrt{P_T}, V_T^{1/3}T^{1/6}\}$, $\quad \mu = \frac{1}{M\sqrt{T}}$, $\quad \eta = \frac{\max\{1, c_T\}}{M\sqrt{T}}$, with $\gamma = \Theta(T^{-1/2})$ and $M = 4\left(\frac{3n+2}{\rho} + 1\right)^2$.

Theorem. Under boundedness, Slater feasibility, and known regularity,

$$\mathfrak{R}_T(\texttt{BCOMD}) = \tilde{O}\left(\min\{\sqrt{P_T T},\ V_T^{1/3}T^{2/3}\}\right), \qquad \mathfrak{V}_T(\texttt{BCOMD}) = \tilde{O}(\sqrt{T}).$$

This regret matches the unconstrained dynamic bandit lower bound $\Omega(\min\{V_T^{1/3}T^{2/3}, \sqrt{P_T T}\})$, while keeping cumulative constraint violation sublinear.

Unknown P_T, V_T. A doubling/meta-bandit variant MBCOMD removes prior knowledge and obtains

$$\mathfrak{R}_T(\texttt{MBCOMD}) = \tilde{O}\left(\min\{\sqrt{P_T}T^{2/3},\ V_T^{1/3}T^{7/9}\}\right), \quad \mathfrak{V}_T = \tilde{O}(\sqrt{T}).$$

How the Parameters Interact

The rates separate the roles of exploration, primal adaptation, and feasibility control:

$$\gamma = \Theta(T^{-1/2}) \quad \text{keeps every arm observable,} \qquad \eta \propto \frac{\max\{1, c_T\}}{\sqrt{T}} \quad \text{tracks non-stationarity,} \qquad \mu \propto T^{-1/2} \quad \text{controls violation.}$$

The stabilizer Ω makes the pseudo-costs nonnegative and enables a uniform dual bound. In the adaptive version, a doubling schedule runs a logarithmic grid of BCOMD instances and a meta-bandit selects among them.

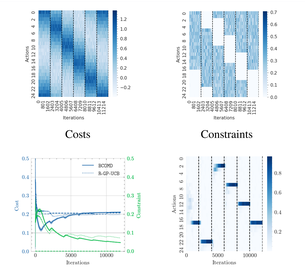
Known P_T, V_T: optimal tuning. Unknown P_T, V_T: meta-learning with only logarithmic expert overhead.

Comparison with Prior Work

Setting / work	Regularity	Regret	Violation
Static unconstrained MAB	none	$O(\sqrt{T})$	–
Dynamic unconstrained bandits	P_T or V_T	$T^{3/4}$-type to optimal	–
Deng et al. dynamic constrained	V_T	$\tilde{O}(V_T^{1/4}T^{3/4})$	$\tilde{O}(V_T^{1/4}T^{3/4})$
Cao et al. dynamic constrained	P_T	$\tilde{O}(\sqrt{P_T T})$	$\tilde{O}(P_T^{1/4}T^{3/4})$
This work: BCOMD	both	$\tilde{O}(\min\{\sqrt{P_T T}, V_T^{1/3}T^{2/3}\})$	$\tilde{O}(\sqrt{T})$
This work: adaptive	both (not known)	$\tilde{O}(\min\{\sqrt{P_T}T^{2/3}, V_T^{1/3}T^{7/9}\})$	$\tilde{O}(\sqrt{T})$

Takeaway: the constrained setting pays no extra order in regret compared with the optimal unconstrained dynamic rate, and feasibility improves to $\tilde{O}(\sqrt{T})$.

Numerical Evaluation



Synthetic non-stationary environment. Costs and constraints cyclically shift over time windows, with Gaussian noise added.

- BCOMD is compared against restarting GP-UCB for constrained non-stationary bandits.
- The proposed method attains lower cumulative cost while maintaining better constraint satisfaction.
- Learned arm distributions track feasible low-cost regions across shifts.

Proof Skeleton

1. Establish a saddle-point inequality coupling primal mirror descent and dual ascent.
2. Bound the entropic Bregman divergence induced by the bandit estimator.
3. Prove bounded dual iterates using Slater feasibility and the stabilizer Ω.
4. Telescope dynamic comparator terms, controlled either by P_T or by V_T.
5. Use the dual update to obtain $\sum_{t=1}^T \mathbb{E}[g_t(a_t)] \le \mu^{-1}\Omega = \tilde{O}(\sqrt{T})$.

Selected References

Auer et al. (2002), finite-time analysis of multi-armed bandits. Besbes et al. (2015), non-stationary stochastic optimization. Cao et al. (2018), online convex optimization with time-varying constraints and bandit feedback. Chen et al. (2025), bridging adversarial and nonstationary multi-armed bandit. Deng et al. (2022), interference constrained beam alignment for time-varying channels via kernelized bandits. Zinkevich (2003), online convex programming.

Conclusions and Future Directions

BCOMD shows that adversarial bandits with time-varying soft constraints can achieve optimal-order dynamic regret while maintaining sublinear cumulative violation.

- Entropic geometry is central: it handles the importance-weighted estimator where Euclidean OGD fails.
- Primal and dual updates decouple adaptation to changing costs from long-run feasibility.
- Future work: Tsallis entropy, combinatorial bandits, and bandit-feedback submodular maximization.