Contextual-Bandits with Surrogate Losses:
Margin Bounds and Efficient Algorithms

Microsoft Research

Dylan Foster and Akshay Krishnamurthy

Cornell University

Contextual Bandits and Background

Contextual Bandit Protocol

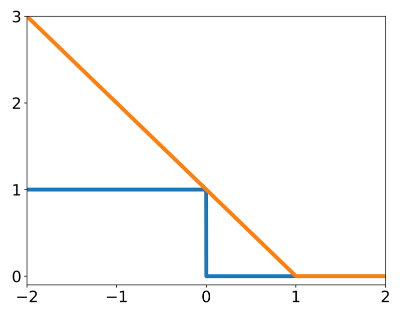
On each of T rounds

1. Observe context x_t
2. Choose action a_t
3. Observe loss $\ell_t(a_t, x_t)$.

Goal: Minimize loss!

Applications: online personalization, medical trials, etc.

Surrogate Losses in Supervised Learning



- Computational: convex/continuous relaxations of NP-hard problems.
- Statistical: sharper generalization bounds, e.g., distribution-dependent, dimension-free, etc.

Why not use surrogate losses in contextual bandits?

Our results

- **New regret bound for margin-based contextual bandits with generic function class.**
  - **Generalizes and improves several prior results.**
  - **Extends sequential complexity bounds for online learning (Rakhlin et al., 2015) to contextual bandits.**
- **A new CB algorithm for parametric/convex classes with $O(\sqrt{dT})$ regret.**
  - **First efficient bandit-multiclass algorithm with $O(\sqrt{dT})$ regret against hinge loss.**
- **A new analysis of (smoothed) Follow-the-Leader with large non-parametric classes.**

Prior Results

- Parametric methods: Simple, efficient, but rely on realizability. Can we get guarantees without realizability?
- Agnostic methods: Few assumptions, but computationally inefficient in general. Can we gain tractability?
- Bandit Multiclass: Surrogate losses common, but loss functions do not generalize to cost-sensitive.
- Statistical/Online Learning: Surrogate losses ubiquitous. Can we extend to partial information?

Surrogate Loss Functions

Setting

- Adversarial contextual bandits with K actions: $x_t \in \mathcal{X}$, $\ell_t \in [0,1]^K$ chosen by adaptive adversary.
- Bandit feedback: On each round, choose an action a_t, incur loss $\ell_t(a)$. Only loss of chosen action is observed.
- Standard goal: Compete with policy class $\Pi : \mathcal{X} \to [K]$, measured via regret

$$\mathrm{Regret}(T,\Pi) \triangleq \sum_{t=1}^{T} \mathbb{E}\left[\ell_t(a_t)\right] - \inf_{\pi\in\Pi} \sum_{t=1}^{T} \mathbb{E}\left[\ell_t(\pi(x_t))\right].$$

- Regressors: We derive Π from a class $\mathcal{F} : \mathcal{X} \to \mathbb{R}^K_{=0}$ of functions ($\mathbb{R}^K_{=0} = \{s \in \mathbb{R}^K : \sum_a s_a = 0\}$).
- Surrogates: Ramp $\phi^\gamma(s) \triangleq \min(\max(1+s/\gamma, 0), 1)$ and hinge $\psi^\gamma(s) \triangleq \max(1+s/\gamma, 0)$. (Pires et al. 2013)

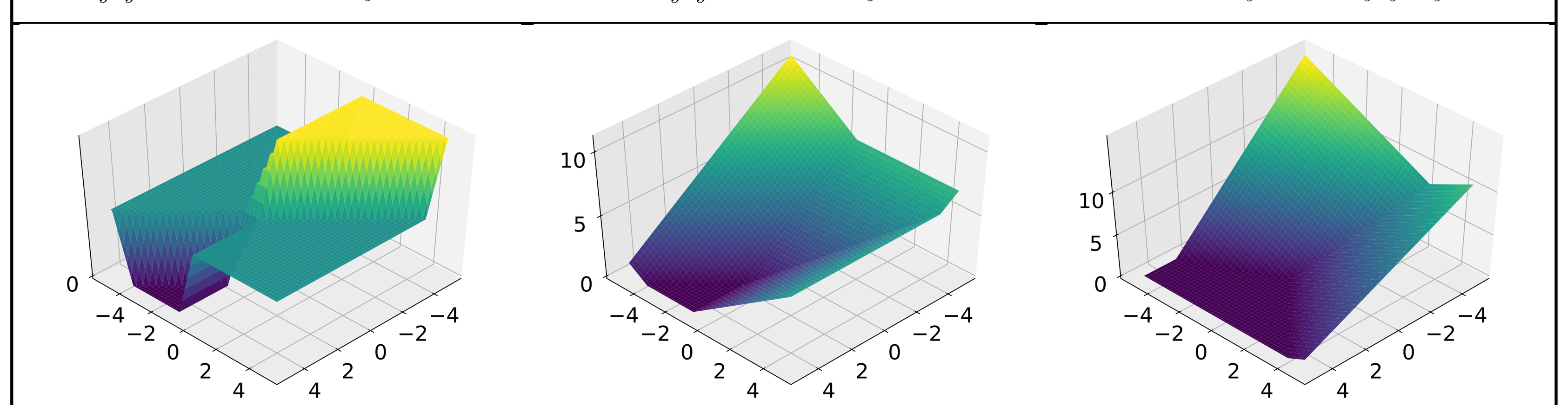
Key observation: Surrogate losses induce randomized policies.

Lemma 1. *For $s \in \mathbb{R}^K_{=0}$, define $a^\star = \mathrm{argmax}_a s_a$ and $\pi_{ramp}(s), \pi_{hinge}(s) \in \Delta([K])$ by $\pi_{ramp}(s)_a \propto \phi^\gamma(s_a)$ and $\pi_{hinge}(s)_a \propto \psi^\gamma(s_a)$. For any $\ell \in \mathbb{R}^K_+$ we have*

$$\ell(a^\star) \le \langle \pi_{ramp}(s), \ell\rangle \le \langle \ell, \phi^\gamma(s)\rangle \le \sum_a \ell(a)\mathbf{1}\{s_a \ge -\gamma\} \quad \textit{and} \quad \ell(a^\star) \le \langle \pi_{hinge}(s), \ell\rangle \le K^{-1}\langle \ell, \psi^\gamma(s)\rangle.$$

- $\langle \ell, \phi^\gamma(f(x))\rangle$ or $\langle \ell, \psi^\gamma(f(x))\rangle$ serve as *surrogate losses* for f.
- For ramp, also obtain *margin regret*: $L_T^\gamma(f) \triangleq \sum_{t=1}^T \sum_a \ell_t(a)\mathbf{1}\{f(x_t)_a \ge -\gamma\}$.

CC-Ramp	CC-Hinge	MC-Hinge
$\langle \ell, \min(\max(1+s,0),1)\rangle$	$\langle \ell, \max(1+s,0)\rangle$	—
$\sum_{y\ne y^\star} \min(\max(1+s_y,0),1)$	$\sum_{y\ne y^\star} \max(1+s_y,0)$	$\max(1-(s_{y^\star} - \max_{y\ne y^\star} s_y), 0)$



Achievable Regret Bounds

Theorem 2. *For any constants $\beta > \alpha > 0$, smoothing parameter $\mu \in (0,1)$ and margin parameter $\gamma > 0$ there exists an adversarial CB strategy with expected loss bounded as:*

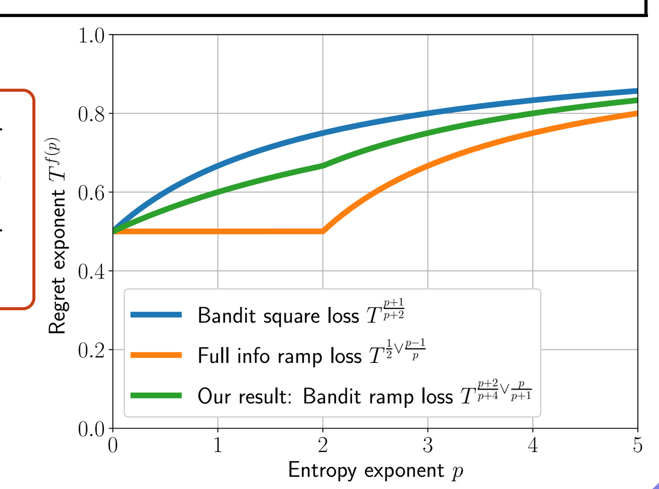
$$\mathbb{E}\left[\sum_{t=1}^T \ell_t(a_t)\right] \le \inf_{f\in\mathcal{F}} \mathbb{E}\left[L_T^\gamma(f)\right] + 4\sqrt{2K^2T\log\mathcal{N}_{\infty,\infty}(\beta/2,\mathcal{F},T)} + \mu KT$$
$$+ \frac{8}{\mu}\log\mathcal{N}_{\infty,\infty}(\beta/2,\mathcal{F},T) + \frac{1}{\gamma}\left(3e^2\alpha KT + 24e\sqrt{\frac{KT}{\mu}}\int_\alpha^\beta \sqrt{\log\mathcal{N}_{\infty,\infty}(\varepsilon,\mathcal{F},T)}d\varepsilon\right)$$

where $\mathcal{N}_{\infty,\infty}(\varepsilon,\mathcal{F},T)$ is the L_∞/ℓ_∞-sequential covering number for $\mathcal{F}$.

- Also yields a policy regret bound, against policy class derived from $\mathcal{F}$.
- Requires knowledge of margin parameter γ, unlike uniform guarantees for statistical learning.

Class	Rate	Notes
Finite classes	$K\sqrt{T\log\lvert\mathcal{F}\rvert}$	Can get optimal $O(\sqrt{KT\log\lvert\Pi\rvert})$ policy regret with our proof.
Parametric	$K\sqrt{Td\log(KT/\gamma)}$	$\log\mathcal{N}_{\infty,\infty}(\varepsilon,\mathcal{F},T) \propto d\log(1/\varepsilon)$, as in the LinUCB setting.
Rademacher	$K(\mathfrak{R}(\mathcal{F},T)/\gamma)^{2/3}T^{1/3}$	Involves Rademacher complexity of scalar restrictions of benchmark. For full information, rate is $\Theta(\max_a \mathfrak{R}(\mathcal{F}\vert_a, T))$.
Linear classes	$K(T/\gamma)^{2/3}$	Generalizes Banditron to smooth Banach spaces.
Nonparametric	$(KT)^{\frac{p+2}{p+4}}\gamma^{-\frac{2p}{p+4}}$	$\log\mathcal{N}_{\infty,\infty}(\varepsilon,\mathcal{F},T) \propto \varepsilon^{-p}, p \in (0,2]$
Nonparametric	$(KT)^{\frac{p}{p+1}}\gamma^{-\frac{p}{p+1}}$	$\log\mathcal{N}_{\infty,\infty}(\varepsilon,\mathcal{F},T) \propto \varepsilon^{-p}, p \ge 2$

Lipschitz CB. For $\mathcal{X} = [0,1]^p$, the class $\mathcal{F}$ of Lipschitz functions has sequential entropy growth ε^{-p}. We obtain $O(T^{\frac{p+2}{p+4}\vee\frac{p}{p+1}})$ margin/policy regret, improving the $O(T^{\frac{p+1}{p+2}})$ bound of Cesa-Bianchi et al. (2017).



Right: exponent on T vs. entropy exponent. "Square loss" denotes optimal rate under square-loss realizability (Slivkins, 2011).

Proof Ideas

Full information bound

If full info bound involves *local norms*, can obtain bandit bound via importance weighting. E.g., Exp4

$$\mathrm{Regret}(T,\Pi) \le \frac{\eta}{2}\sum_{t=1}^T \mathbb{E}_{\pi\sim p_t}\langle \pi(x_t), \hat{\ell}_t\rangle^2 + \frac{\log(\lvert\Pi\rvert)}{\eta}$$

We show existence of full-info algorithm with regret scaling with (1) local norms and (2) sequential covering. Uses *adaptive minimax* technique of Foster et al. (2015). We show for $\mathcal{G} : \mathcal{X} \to \mathcal{S}$

$$\mathcal{V} \triangleq \left\langle\!\!\left\langle \sup_{x_t\in\mathcal{X}} \inf_{p_t\in\Delta(\mathcal{S})} \sup_{\ell_t} \mathbb{E}_{s\sim p_t} \right\rangle\!\!\right\rangle_{t=1}^T \left[\sum_{t=1}^T \langle s_t, \ell_t\rangle - \inf_{g\in\mathcal{G}}\sum_{t=1}^T \langle g(x_t), \ell_t\rangle - B(p_{1:T}, \ell_{1:T})\right] \le C,$$

where $B(p_{1:T}, \ell_{1:T}) = \sum_{t=1}^T \eta_1 \lVert\ell_t\rVert_1 + \eta_2\lVert\ell_t\rVert_1^2 + 2\eta_3\mathbb{E}_{s\sim p_t}\langle s, \ell_t\rangle^2$ and C depends only on $\eta_{1:3}$ and $\mathcal{N}_{\infty,\infty}(\mathcal{G})$.

- Yields benign dependence on loss range and Dudley-type integral with sequential metric entropy.
- To give main theorem, use $\mathcal{G} = \phi^\gamma \circ \mathcal{F}$.

Bandit Reduction

- Use full info algorithm with class $\mathcal{G} = \phi^\gamma \circ \mathcal{F}$, to obtain $p_t \in \Delta(\mathcal{S})$
- Define $P_t(a) = \mathbb{E}_{s\sim p_t}\frac{s(a)}{\sum_{a'} s(a')}$ sample $a_t \sim P_t^\mu \triangleq (1-K\mu)P_t + \mu)$
- Feed *importance weighted loss* $\hat{\ell}_t(a) = \ell_t(a_t)\mathbf{1}\{a_t = a\}/p_t(a)$ to full-info algorithm.
- **Challenge: Variance control for surrogate losses. Solution: Randomized policies.**

Lemma 3 (Variance control for randomized policies)**.** *With $\sup_{x,f}\lVert f(x)\rVert_\infty \le B$ we have*

$$\mathbb{E}_{a_t\sim P_t^\mu}\left[\mathbb{E}_{s_t\sim p_t}\langle s_t, \hat{\ell}_t\rangle^2\right] \le \begin{cases} K, & \textit{for } \mathcal{S} \subset \Delta(\mathcal{A}). \\ K^2, & \textit{for } \mathcal{S} = \phi^\gamma \circ \mathcal{F}. \\ \left(1+\frac{B}{\gamma}\right)^2 K^2, & \textit{for } \mathcal{S} = \psi^\gamma \circ \mathcal{F}. \end{cases}$$

Hinge-LMC

Key Insights

- Stationary distribution of LMC Markov chain is Exponential weights distribution.
- With hinge surrogate and convexity, sampling problem is log-concave ⇒ efficient algorithm!
- Sampler uses randomized smoothing and ℓ_2 regularization for strong convexity.
- Also use geometric resampling to estimate importance weight.

Algorithm 1 Hinge-LMC

Input: Class Θ, learning rate η, rounds T, margin γ.
Define $w_0(\theta) = 1$ for all $\theta \in \Theta$.
for $t = 1, \ldots, T$ **do**
  $\theta_t \leftarrow \mathrm{LMC}(\eta w_{t-1})$.
  Set $p_t(\cdot;\theta_t) \propto \psi^\gamma(f(x_t;\theta_t))$, $p_t^\mu(\cdot;\theta_t) = (1-K\mu)p_t + \mu$.
  Receive x_t, play $a_t \sim p_t^\mu(\cdot;\theta_t)$, observe $\ell_t(a_t)$.
  for $m = 1, \ldots, M$ **do**
    $\tilde{\theta}_t \leftarrow \mathrm{LMC}(\eta w_{t-1})$. `// Geometric resampling.`
    Sample $\tilde{a}_t \sim p_t^\mu(\cdot;\tilde{\theta}_t)$, if $\tilde{a}_t = a_t$, break
  end for
  Set $m_t = m$, and $\tilde{\ell}_t(a) = \ell_t(a_t)\cdot m_t\mathbf{1}\{a_t = a\}$
  Update $w_t(\theta) \leftarrow w_{t-1}(\theta) + \langle \tilde{\ell}_t, \psi^\gamma(f(x_t;\theta))\rangle$
end for

Algorithm 2 Langevin Monte Carlo (LMC)

Input: Function F, parameters m, u, λ, N, α.
Set $\tilde{\theta}_0 \leftarrow 0 \in \mathbb{R}^d$
for $k = 1, \ldots, N$ **do**
  Draw $z_1, \ldots, z_m \overset{iid}{\sim} \mathcal{N}(0, u^2 I_d)$ and define
  $$\tilde{F}_k(\theta) = \tfrac{1}{m}\textstyle\sum_{i=1}^m F(\theta + z_i) + \tfrac{\lambda}{2}\lVert\theta\rVert_2^2$$
  Draw $\xi_k \sim \mathcal{N}(0, I_d)$ and update
  $$\tilde{\theta}_k \leftarrow \mathcal{P}_\Theta\left(\tilde{\theta}_{k-1} - \frac{\alpha}{2}\nabla\tilde{F}_k(\tilde{\theta}_{k-1}) + \sqrt{\alpha}\xi_k\right).$$
end for
Return $\tilde{\theta}_N$.

Theorem 4. *Assume $\mathcal{F}$ is parametrized by a compact convex set $\Theta \subset \mathbb{R}^d$, $f(x;\theta)$ is convex and L-Lipschitz in θ, and $\sup_{x,\theta}\lVert f(x;\theta)\rVert_\infty \le B$. For any γ, Hinge-LMC guarantees*

$$\mathbb{E}\left[\sum_{t=1}^T \ell_t(a_t)\right] - \min_{\theta\in\Theta}\mathbb{E}\left[\frac{1}{K}\sum_{t=1}^T \langle \ell_t, \psi^\gamma(f(x_t;\theta))\rangle\right] \le \tilde{O}\left(\frac{B}{\gamma}\sqrt{dT}\right)$$

Moreover the running time is $\tilde{O}\left(\frac{d^{14}T^{10}}{K^2\gamma^2}\right)$.

- **Bandit Multiclass:** First efficient $\sqrt{dT}$ algorithm against a loss without curvature!
- **Realizability:** If $\theta^\star$ has $f(x;\theta^\star)_a = K\gamma\mathbf{1}\{\ell(a) \le \min_{a'}\ell(a')\} - \gamma$ then obtain $\frac{B}{\gamma}\sqrt{dT}$ policy regret.
- **Practical Aspects:** Likely can significantly improve runtime and extend to non-convex classes.

Smooth-FTL and Lipschitz CB

Setting: Stochastic contextual bandits, $(x_t, \ell_t) \sim \mathcal{D}$ iid on each round.
Algorithm: Epoch based, with epoch m lasts for $n_m = 2^m$ rounds.
To begin m^{th} epoch, compute empirical importance-weighted hinge-loss minimizer:

$$\hat{f}_{m-1} = \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{\tau=n_{m-1}}^{n_m-1} \left\langle \hat{\ell}_\tau, \psi^\gamma(f(x_\tau))\right\rangle.$$

Note, uses only data from previous epoch.
For all rounds in m^{th} epoch, play as $(1-K\mu)\pi_{\mathrm{hinge}}(\hat{f}_{m-1}(x_t)) + \mu$. (Essentially ϵ-greedy.)

Theorem 5. *Suppose that $\mathcal{F}$ satisfies $\log\mathcal{N}_{\infty,\infty}(\varepsilon,\mathcal{F},T) \propto \varepsilon^{-p}$ for $p \ge 2$. Then for stochastic CB,* SmoothFTL *guarantees*

$$\mathbb{E}\left[\sum_{t=1}^T \ell_t(a_t)\right] - \min_{f\in\mathcal{F}}\frac{T}{K}\mathbb{E}\left[\langle \ell, \psi^\gamma(f(x))\rangle\right] \le \tilde{O}\left((T/\gamma)^{\frac{p}{p+1}}\right)$$

- **Oracle Efficient:** Makes $\log(T)$ calls to hinge-loss minimization oracle.
- **Lipschitz CB:** Also yields $T^{\frac{p}{p+1}}$ algorithm for Lipschitz CB with p-dimensional context space and finite action space. Yields best known guarantee for Lipschitz CB.
- **(Sub)optimality?** Matches information-theoretic results, but ϵ-greedy typically suboptimal.

References

1. Nicolo Cesa-Bianchi, Pierre Gaillard, Claudio Gentile, and Sebastien Gerchinovitz. Algorithmic chaining and the role of partial feedback in online nonparametric learning. In Conference on Learning Theory, 2017.
2. Dylan J. Foster, Alexander Rakhlin, and Karthik Sridharan. Adaptive online learning. In Advances in Neural Information Processing Systems, 2015.
3. Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Efficient bandit algorithms for online multiclass prediction. In International Conference on Machine learning, 2008.
4. Bernardo Avila Pires, Csaba Szepesvari, and Mohammad Ghavamzadeh. Cost-sensitive multiclass classifica- tion risk bounds. In International Conference on Machine Learning, 2013.
5. Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via Sequential Complexities. In Journal of Machine Learning Research, 2015.
6. Aleksandrs Slivkins. Contextual bandits with similarity information. In Conference on Learning Theory, 2011.

Learn more at https://arxiv.org/abs/1806.10745