Multi-armed bandits for adaptive learning: A technical implementation guide

Thompson Sampling and contextual bandits have emerged as the dominant paradigm for personalized learning path optimization, outperforming both heuristic approaches and traditional A/B testing in production educational systems. Duolingo, ALEKS, and Knewton have demonstrated that bandit-based content selection can improve learner engagement by **2-7%** while reducing dropout rates by up to **50%** when combined with deep knowledge tracing. The key insight driving adoption: educational content selection is fundamentally an explore-exploit problem where the cost of suboptimal choices—wasted learner time and pedagogical regret—demands principled uncertainty quantification that bandits uniquely provide.

This report provides the mathematical foundations, production architectures, and implementation patterns needed to build bandit-based adaptive learning systems at scale.

Thompson Sampling provides the mathematical backbone for content selection

Thompson Sampling (TS) has become the algorithm of choice for educational content recommendation due to its natural handling of uncertainty and robustness to delayed feedback— (stanford)both critical in learning contexts where outcomes take days or weeks to observe.

The Beta-Bernoulli model forms the foundation for binary mastery outcomes. For each content item *k*, we maintain a Beta(α_k, β_k) posterior over its success probability θ_k. The algorithm samples θ̃_k from each posterior, selects the content with highest sampled value, observes the outcome, and updates via the conjugate prior: **α' = α + successes, β' = β + failures**. (stanford) This closed-form update requires only O(1) computation per observation and O(K) storage for K content items.

The mathematical elegance lies in the exploration-exploitation balance emerging naturally from posterior uncertainty. Unlike ε-greedy which wastes exploration budget on clearly suboptimal content, TS directs exploration toward genuinely uncertain items. (stanford) Unlike UCB which requires careful tuning of confidence parameters, TS needs no hyperparameters for the exploration-exploitation tradeoff. Empirical studies consistently show TS achieves regret competitive with the theoretical lower bound of Lai and Robbins, (ACM Digital Library) with problem-dependent regret bounded by $O(\Sigma_i \ln(T)/KL(\mu_i, \mu))$* where KL is the Kullback-Leibler divergence between suboptimal and optimal arm distributions.

Cold-start handling leverages informative priors derived from content features. For new learning activities, historical data from similar content (same difficulty level, topic, format) provides prior parameters. (stanford) A new algebra exercise similar to past items with 70% success rate might use Beta(7, 3), encoding prior belief equivalent to 10 observations. This dramatically reduces exploration costs for new content.

Non-stationarity from learner improvement presents a fundamental challenge: as students learn, success probabilities change. Three adaptation strategies address this. First, sliding-window posteriors use only the most recent τ observations, preventing the posterior from over-concentrating on outdated estimates. Second, discounted Thompson Sampling applies exponential decay: $(\alpha, \beta) \leftarrow ((1-\gamma)\alpha + \gamma\bar{\alpha}, (1-\gamma)\beta + \gamma\bar{\beta})$ with decay

parameter γ controlling memory. Third, change-point detection monitors for abrupt skill jumps and resets posteriors accordingly. (stanford) Research establishes regret bounds of **O(√LKT log T)** for non-stationary bandits with L change points. (MDPI)

Contextual bandits incorporate learner features for personalization

Pure multi-armed bandits ignore the rich context available about learners. Contextual bandits extend the framework by modeling expected reward as a function of learner features, enabling true personalization. (Wikipedia)

LinUCB assumes reward is linear in context: r_t,a = θ_*a^T · x_t,a + ε_t, where x_t,a combines learner and content features. The algorithm maintains per-arm design matrices A_a and response vectors b_a, computing θ̂_a = A_a^(-1) b_a via ridge regression. The UCB formula* ***UCB_a(x_t) = θ̂_a^T · x_t + α · √(x_t^T · A_a^(-1) · x_t)*** *balances exploitation (mean estimate) with exploration (uncertainty bonus scaled by α). Selection is deterministic: choose the arm maximizing UCB. The disjoint model learns independent parameters per arm, while hybrid LinUCB shares parameters across arms via z_t,a^T · β* + x_t,a^T · θ_a, improving generalization when content shares structural similarities.

The learner context vector x_t,a should capture multiple dimensions of the learning state:

- **Prior knowledge**: BKT mastery probabilities P(L_skill), DKT hidden states, prerequisite completion rates, diagnostic assessment scores
- **Engagement signals**: average time-on-task, session frequency, hint request rates, content completion percentages, click patterns
- **Temporal features**: time-of-day, day-of-week, time since last session, fatigue indicators from within-session performance degradation
- **Historical performance**: accuracy on similar content types, performance trajectory (improving/declining), error type frequency distinguishing careless mistakes from knowledge gaps
- **Cross-features**: learner ability minus content difficulty, cosine similarity between learner and content embeddings, mastery level on content's prerequisite knowledge components

Neural contextual bandits address LinUCB's limitation to linear reward functions. NeuralUCB (Zhou et al., 2020) uses deep neural networks for reward prediction while constructing exploration bonuses from neural tangent features. The architecture extracts learned representations φ(x) = f_neural(x; W) from raw context, then applies linear bandit methods on the extracted features: r(a) = θ_a^T · φ(x). This neural-linear approach combines the representation power of deep networks with tractable uncertainty estimation from the linear layer.

Neural Thompson Sampling (Zhang et al., 2021) achieved the first provable **O(√T) regret bounds** for neural bandits by building posteriors on neural tangent features. Practical implementations use dropout-based uncertainty estimation: keeping dropout active at inference and computing variance across multiple forward

passes provides exploration signals. The PluggableTS approach (SDM 2023) simplifies this further, requiring only univariate Gaussian sampling during serving with no changes to network training.

The exploration-exploitation tradeoff carries unique pedagogical consequences

In educational settings, "regret" isn't merely a mathematical abstraction—it represents real learning time lost to suboptimal content. This creates distinctive constraints on exploration strategies.

Pedagogical regret manifests in several forms: presenting content too easy wastes time without learning gains; content too difficult causes frustration and disengagement; content misaligned with learning style reduces retention. Unlike advertising where a suboptimal recommendation merely reduces click probability, educational missteps can compound: a confused learner falls behind on prerequisites, creating cascading failures.

Delayed reward structures complicate credit assignment. Quiz results arrive hours or days after content exposure. Mastery assessments occur at topic completion. Retention tests happen weeks later. Research demonstrates Thompson Sampling maintains optimality under delayed feedback in $\omega(\log T)$ batches, and Wu & Wager (2022) establish the first regret bounds for TS with arbitrary delay distributions including unbounded expectation. The practical solution at Duolingo uses proxy rewards—counting lesson completion within 2 hours of notification as success—to minimize organic activity contamination while capturing actionable signals.

Safety constraints bound exploration in ways unnecessary for typical recommendation. Content must respect prerequisite dependencies—you cannot explore whether calculus works for a student who hasn't learned algebra. Difficulty must remain within the learner's zone of proximal development. Minimum educational quality standards eliminate clearly harmful exploration. These constraints create a constrained exploration problem where the feasible action set depends on learner state.

Exploration budget management varies by deployment philosophy. Yahoo and Twitter concentrate exploration on 1-5% of traffic to minimize negative impact while gathering learning signal. Netflix spreads small exploration across all users to improve diversity. (Eugene Yan) (eugeneyan) Duolingo uses softmax with temperature $\tau = 0.0025$, achieving approximately 17% exploration rate (top arm not selected) plus a 5% pure exploration holdout for monitoring. The optimal strategy depends on the cost of suboptimal recommendations relative to the value of exploration data.

Production systems reveal proven architectural patterns

Duolingo's Birdbrain system processes approximately **1 billion exercises daily** using a two-generation architecture. Birdbrain V1 modeled learner ability as a simple scalar updated via stochastic gradient descent—essentially a generalized Elo rating. Birdbrain V2 upgraded to an LSTM that compresses learner history into a 40-dimensional vector representation, enabling richer modeling of learning dynamics.

The session generator algorithm dynamically selects exercises using Birdbrain predictions, targeting the engagement zone where content difficulty matches learner ability. Exercises are tagged with pedagogical features: part of speech, sentence structure, tense, vocabulary items. A "Blame Algorithm" assigns errors to specific knowledge components for targeted remediation.

Duolingo's push notification system (KDD 2020) introduced the **Recovering Difference Softmax Algorithm (RDSA)** addressing two novel bandit problems. The sleeping arms problem handles conditional eligibility—some notifications require streak prerequisites. The recovering arms problem models novelty decay for repeated notifications via exponential penalty: $s^*_{a,t} = \hat{s}_a - \gamma \times 0.5^{(d_{a,t}/h)}$ with $\gamma = 0.017$ base penalty and $h = 15$ days half-life. Results showed **+0.5% daily active users** and **+2.2% new user retention**.

ALEKS (McGraw-Hill) builds on Knowledge Space Theory (KST), a mathematical framework modeling prerequisite relationships between concepts. The system investigates trillions of potential knowledge states to create individual knowledge maps, visualized as the "ALEKS Pie" showing known versus unknown topics. The mastery learning heuristic uses a Tug-of-War (TOW) scoring system: +i points for correct responses, -j points for incorrect, requiring N points for mastery. Research proves this is optimal for certain BKT variants. Recent work integrates RNNs to improve KST-based stopping algorithms and models retention through forgetting curves.

Knewton (now Wiley) operates an adaptive ontology knowledge graph connecting modules (content pieces), concepts (abstract ideas), and relationships. The recommendation engine uses Item Response Theory for measuring student knowledge combined with probabilistic graphical modeling for knowledge inference. The API-based integration pattern enabled partnerships with major publishers, supporting tens of millions of students before acquisition.

Khan Academy implements mastery-based learning with a proficiency progression: Not Started → Attempted (70-99% correct) → Familiar → Proficient (100% correct) → Mastered. The adaptive mechanism adjusts difficulty based on performance—correct answers trigger harder questions, incorrect trigger easier ones. Mixed-skill assessments can promote or demote proficiency levels, implementing spaced retrieval practice.

Recent research advances combine deep learning with principled exploration

NeuralUCB (ICML 2020) achieved the first near-optimal **$\tilde{O}(\sqrt{T})$ regret guarantee** for neural contextual bandits. The algorithm uses neural network-based random feature mapping for UCB construction without requiring assumptions on the reward function form. This enables modeling the complex, non-linear relationships between learner features and content effectiveness that linear methods miss.

Deep Knowledge Tracing integration represents the frontier of learner modeling combined with bandit selection. RL-DKT (Scientific Reports 2025) combines Deep Knowledge Tracing with reinforcement learning, achieving **7.6% AUC improvement** over standard DKT, **12.5% reduction in task completion time**, and **50% reduction in dropout rate** through real-time difficulty adjustment. A dual-stream neural network architecture (Nature Scientific Reports 2025) simultaneously models knowledge state and cognitive load via bidirectional Transformer with graph attention for knowledge component relationships, reaching **87.5% prediction accuracy** with **4.4/5 path quality rating**.

Hierarchical Multi-Armed Bandits (arXiv 2024) addresses curriculum structure through separate MAB agents for concept selection and problem selection within concepts. The architecture determines problem difficulties, assesses latent memory decay, and uses BKT for mastery estimation— (arXiv) capturing the natural hierarchy of courses, modules, lessons, and problems.

ZPDES (Zone of Proximal Development through Exploration and Scaffolding) from the Journal of Educational Data Mining combines intrinsically motivated learning with MAB exploration, requiring minimal knowledge about exercise difficulty. (Educationaldatamining) The RiARiT variant leverages difficulty information when available. Both demonstrate that principled exploration can work even with limited pedagogical metadata.

Non-stationary bandit research (Entropy 2025) provides theoretical foundations for modeling learner improvement. Discounted TS and sliding-window TS achieve **O(√T·B_T) regret** where B_T bounds the total variation in reward distributions. Change detection algorithms (SIGIR 2018) enable adaptive user preference modeling by identifying when reward distributions shift.

Implementation patterns enable practical deployment

Vowpal Wabbit serves as the production workhorse, powering Azure Personalizer and numerous internal systems. Key commands include (--cb_explore_adf) for action-dependent features with dynamic action sets and (--cb_type dr) for doubly robust policy evaluation. The ADF format naturally encodes educational content selection:

```
shared |user student_id:123 grade_level:10 mastery_algebra:0.7
0:0.8:0.25 |content content_id:456 difficulty:medium topic:algebra
|content content_id:789 difficulty:easy topic:geometry
```

The (-q UA) flag creates user-action feature interactions automatically. Exploration strategies span epsilon-greedy (uniform random), bagging (ensemble of policies), softmax (probability proportional to exp(λ×score)), cover (explicitly diverse policies), and RND (randomized LinUCB approximation). (github)

For Python implementations, MABWiser from Fidelity provides scikit-learn-style interfaces supporting Thompson Sampling, LinUCB, LinTS, and neighborhood policies. (PyPI) The contextualbandits library offers streaming support and built-in off-policy evaluation via doubly robust estimators. (GitHub)

Feature store integration (Feast/Tecton patterns) separates feature computation from model serving. Learner features—session duration, completion rates, mastery estimates—compute in batch pipelines and serve online with sub-millisecond latency. Content features—difficulty, topic embeddings, engagement signals—update less frequently. Contextual features like time-since-last-session compute at request time.

Offline policy evaluation enables safe deployment of new policies. Inverse Propensity Scoring (IPS) reweights logged data by propensity: if the new policy would have chosen the logged action, weight by reward/propensity. Doubly Robust estimation combines IPS with a direct method reward model, reducing variance while maintaining unbiasedness. (ope-rec) Vowpal Wabbit's (--cb_type dr) implements this efficiently.

The **critical logging requirement** for counterfactual evaluation: every decision must record the probability that action was selected. Without propensity scores, offline evaluation becomes impossible. Log entries must capture event_id, timestamp, context features, available actions, chosen action, **action probability**, and eventual reward.

Cold-start mitigation follows hierarchical patterns. Doordash's approach—regional priors → subregional adjustments → user-level posteriors—transfers naturally to education: population priors → cohort adjustments → individual posteriors. Content cold-start uses pessimistic initialization (Beta(1, 99) outperforms naive Beta(1,1) per Deezer findings) combined with forced exploration for minimum impressions.

Delayed reward handling in production uses Thompson Sampling over UCB—stochastic policies naturally continue exploring during delay periods. Event-based architectures track pending rewards: store (context, action, timestamp) at selection time, update model asynchronously when rewards arrive. Proxy rewards (immediate engagement signals) supplement delayed true rewards (learning outcomes).

Engineering tradeoffs shape system design decisions

Online versus batch learning presents a fundamental architectural choice. Online learning (Duolingo's real-time LSTM updates) provides immediate adaptation but requires streaming infrastructure and risks volatility. Batch updates (daily arm score computation) offer stability and simpler infrastructure but delay adaptation. The hybrid pattern—offline feature engineering combined with online bandit parameter updates—balances these concerns. YouTube's system uses offline user clustering and item pre-selection with online per-cluster bandit models. ACM Digital Library

Exploration budget allocation depends on deployment context. Concentrated exploration (1-5% of users) minimizes negative impact on the majority while still gathering sufficient exploration data. Eugene Yan Distributed exploration (all users, small epsilon) improves content diversity and avoids filter bubbles. Decaying epsilon starts aggressive and reduces as uncertainty decreases. Pre-filtered exploration (Spotify pattern) limits exploration to already-promising candidates. Eugene Yan

Model serving latency requirements typically target sub-50ms for real-time recommendations. Strategies include caching frequently-used arm parameters, pre-computing UCB scores for popular contexts, and using diagonal approximations (Diag-LinUCB) that avoid full matrix inversions. arXiv Udemy achieved latency improvements from 750ms to 14ms through Scala rewrites and Redis caching.

Safety and guardrails constrain the optimization. Minimum exposure constraints ensure all content receives baseline impressions for fair evaluation. Diversity regularization prevents filter bubbles that narrow learner exposure. Prerequisite constraints enforce pedagogical sequencing. Fairness monitoring tracks disparate impact across learner demographics.

Conclusion: Building the next generation of adaptive learning systems

The convergence of deep knowledge tracing, neural contextual bandits, and production-scale feature engineering creates unprecedented opportunities for personalized education. The key architectural insight: treat content selection as a sequential decision problem under uncertainty where exploration has real pedagogical cost.

For teams building AI-first learning systems, the recommended stack begins with Vowpal Wabbit or MABWiser for bandit algorithms, integrates knowledge tracing outputs (BKT mastery, DKT hidden states) as context

features, implements doubly robust offline evaluation before any policy deployment, and logs propensity scores religiously. The most successful production systems—Duolingo's Birdbrain, ALEKS's Knowledge Space engine—combine rigorous psychometric foundations with modern ML infrastructure and continuous experimentation culture.

The frontier lies in end-to-end learning: jointly optimizing knowledge tracing, content selection, and long-term outcomes in a single framework. Recent RL-DKT results demonstrating 50% dropout reduction through integrated approaches suggest this integration will define the next generation of adaptive learning technology.