

CSE 541: Interactive Learning

Lecture 5: Online convex optimization and multi-armed bandits

Kevin Jamieson

Paul G. Allen School of Computer Science & Engineering
University of Washington

Wednesday, October 14, 2026



Last time: squared loss, batch and online

Batch (IID), finite $\mathcal{F}$

- Hoeffding alone: $\sqrt{\log |\mathcal{F}|/n}$.
- Loss difference $W^f = (f - Y)^2 - (f^* - Y)^2$ has $\text{Var } W^f \leq 4 \mathbb{E} W^f$ when $f^* = \mathbb{E}[Y \mid X] \in \mathcal{F}$; Bernstein $\Rightarrow$
 $L(\hat{f}) - L(f^*) \leq \frac{28 \log(2|\mathcal{F}|/\delta)}{3n}$.
- Convexity, not realizability: for convex $\mathcal{F}$ the projection inequality gives the same variance control against $f_{\mathcal{F}}^*$.
- Improper learning: ERM over $\text{conv}(\mathcal{F})$ gets the fast rate with no realizability.

Online (adversarial), finite $\mathcal{F}$

- Predict the p_t -average $\hat{y}_t = \mathbb{E}_{f \sim p_t} f(x_t)$; weights $w_{t+1}(f) = w_t(f) e^{-\eta(f(x_t) - y_t)^2}$.
- $g(z) = e^{-\eta z^2}$ is concave on $[-1, 1]$ for $\eta \leq \frac{1}{2}$, so Jensen replaces Hoeffding's lemma: no slack.
- $\text{Regret}_T \leq 2 \ln |\mathcal{F}|$: no T at all, deterministic, improper.

Today

Classes too large to enumerate: convexity and gradients (OCO). Then a new kind of feedback: you only see the reward of the arm you pulled (bandits).

Online convex optimization

Trading the 0/1 loss for tractability

Back to classification: we have no practical algorithm

- **Adversarial:** exponential weights keeps a weight for every $h \in \mathcal{H}$. For $\mathcal{H} =$ linear classifiers in $\mathbb{R}^d$, $|\mathcal{H}| = \infty$; even discretized it is $e^{\Theta(d)}$.
- **IID:** ERM for the 0/1 loss, $\arg \min_h \sum_s \mathbf{1}\{h(x_s) \neq y_s\}$, is NP-hard for hyperplanes.
- **Portfolios:** the comparator was a continuum Δ_d and the loss was not linear.

Back to classification: we have no practical algorithm

- **Adversarial:** exponential weights keeps a weight for every $h \in \mathcal{H}$. For $\mathcal{H} =$ linear classifiers in $\mathbb{R}^d$, $|\mathcal{H}| = \infty$; even discretized it is $e^{\Theta(d)}$.
- **IID:** ERM for the 0/1 loss, $\arg \min_h \sum_s \mathbf{1}\{h(x_s) \neq y_s\}$, is NP-hard for hyperplanes.
- **Portfolios:** the comparator was a continuum Δ_d and the loss was not linear.

Solution: relax the objective

Parameterize $h_w(x) = \text{sign}(\langle w, x \rangle)$ with $w \in \mathcal{K} \subseteq \mathbb{R}^d$ convex, and replace $\mathbf{1}\{h_w(x) \neq y\}$ by a **convex surrogate** in w , e.g. the logistic loss

$$\ell(w, (x, y)) = \log(1 + \exp(-y\langle w, x \rangle)) \quad \text{or the hinge loss } \max\{0, 1 - y\langle w, x \rangle\}.$$

Measure regret in the surrogate: $\text{Regret}_T = \sum_{t=1}^T \ell_t(w_t) - \min_{w \in \mathcal{K}} \sum_{t=1}^T \ell_t(w).$

Back to classification: we have no practical algorithm

- **Adversarial:** exponential weights keeps a weight for every $h \in \mathcal{H}$. For $\mathcal{H} =$ linear classifiers in $\mathbb{R}^d$, $|\mathcal{H}| = \infty$; even discretized it is $e^{\Theta(d)}$.
- **IID:** ERM for the 0/1 loss, $\arg \min_h \sum_s \mathbf{1}\{h(x_s) \neq y_s\}$, is NP-hard for hyperplanes.
- **Portfolios:** the comparator was a continuum Δ_d and the loss was not linear.

Solution: relax the objective

Parameterize $h_w(x) = \text{sign}(\langle w, x \rangle)$ with $w \in \mathcal{K} \subseteq \mathbb{R}^d$ convex, and replace $\mathbf{1}\{h_w(x) \neq y\}$ by a **convex surrogate** in w , e.g. the logistic loss

$$\ell(w, (x, y)) = \log(1 + \exp(-y\langle w, x \rangle)) \quad \text{or the hinge loss } \max\{0, 1 - y\langle w, x \rangle\}.$$

Measure regret in the surrogate: $\text{Regret}_T = \sum_{t=1}^T \ell_t(w_t) - \min_{w \in \mathcal{K}} \sum_{t=1}^T \ell_t(w).$

Both the portfolio problem and this one are instances of one protocol.

Online convex optimization (OCO)

Input: convex set $\mathcal{K} \subseteq \mathbb{R}^d$

for $t = 1, 2, \dots, T$:

Player picks $a_t \in \mathcal{K}$

Adversary reveals a convex $\ell_t : \mathcal{K} \rightarrow \mathbb{R}$

Player suffers $\ell_t(a_t)$

Special cases. Experts: $\mathcal{K} = \Delta_d$, loss

$p \mapsto \langle p, \ell_t \rangle$. Portfolios: $\mathcal{K} = \Delta_d$,

$\ell_t(p) = -\log \langle p, r_t \rangle$. Classification:

$\mathcal{K} = \{\|w\|_2 \leq B\}$, logistic loss.

Online convex optimization and online gradient descent

Online convex optimization (OCO)

Input: convex set $\mathcal{K} \subseteq \mathbb{R}^d$

for $t = 1, 2, \dots, T$:

Player picks $a_t \in \mathcal{K}$

Adversary reveals a convex $\ell_t : \mathcal{K} \rightarrow \mathbb{R}$

Player suffers $\ell_t(a_t)$

Special cases. Experts: $\mathcal{K} = \Delta_d$, loss $p \mapsto \langle p, \ell_t \rangle$. Portfolios: $\mathcal{K} = \Delta_d$, $\ell_t(p) = -\log \langle p, r_t \rangle$. Classification: $\mathcal{K} = \{\|w\|_2 \leq B\}$, logistic loss.

Online gradient descent (OGD)

Input: $\mathcal{K}$, step size $\eta > 0$; any $a_1 \in \mathcal{K}$

for $t = 1, 2, \dots, T$:

Play a_t ; observe ℓ_t ; let $g_t \in \partial \ell_t(a_t)$

$a_{t+1} = \Pi_{\mathcal{K}}(a_t - \eta g_t)$

where $\Pi_{\mathcal{K}}(v) = \arg \min_{a \in \mathcal{K}} \|a - v\|_2$

Why gradients suffice. Convexity gives a global linear lower bound from local information:

$$\ell_t(a) \geq \ell_t(a_t) + \langle g_t, a - a_t \rangle \quad \forall a \in \mathcal{K}.$$

One gradient and one projection per round, regardless of how “many” hypotheses there are.

Theorem (OGD)

Let $\mathcal{K}$ be convex with diameter $D = \sup_{a,b \in \mathcal{K}} \|a - b\|_2$, and suppose $\|g_t\|_2 \leq G$ for all t . For any $a^* \in \mathcal{K}$,

$$\sum_{t=1}^T \ell_t(a_t) - \ell_t(a^*) \leq \frac{D^2}{2\eta} + \frac{\eta G^2 T}{2} \stackrel{\eta = D/(G\sqrt{T})}{=} DG\sqrt{T}.$$

Theorem (OGD)

Let $\mathcal{K}$ be convex with diameter $D = \sup_{a,b \in \mathcal{K}} \|a - b\|_2$, and suppose $\|g_t\|_2 \leq G$ for all t . For any $a^* \in \mathcal{K}$,

$$\sum_{t=1}^T \ell_t(a_t) - \ell_t(a^*) \leq \frac{D^2}{2\eta} + \frac{\eta G^2 T}{2} \stackrel{\eta=D/(G\sqrt{T})}{=} DG\sqrt{T}.$$

Proof. Convexity: $\ell_t(a_t) - \ell_t(a^*) \leq \langle g_t, a_t - a^* \rangle$. Non-expansiveness of $\Pi_{\mathcal{K}}$:

$$\|a_{t+1} - a^*\|^2 \leq \|a_t - \eta g_t - a^*\|^2 = \|a_t - a^*\|^2 - 2\eta \langle g_t, a_t - a^* \rangle + \eta^2 \|g_t\|^2.$$

Rearrange, sum over t , and telescope the distances:

$$\sum_{t=1}^T \langle g_t, a_t - a^* \rangle \leq \frac{\|a_1 - a^*\|^2 - \|a_{T+1} - a^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|g_t\|^2 \leq \frac{D^2}{2\eta} + \frac{\eta G^2 T}{2}. \quad \blacksquare$$

Online gradient descent: regret bound

Theorem (OGD)

Let $\mathcal{K}$ be convex with diameter $D = \sup_{a,b \in \mathcal{K}} \|a - b\|_2$, and suppose $\|g_t\|_2 \leq G$ for all t . For any $a^* \in \mathcal{K}$,

$$\sum_{t=1}^T \ell_t(a_t) - \ell_t(a^*) \leq \frac{D^2}{2\eta} + \frac{\eta G^2 T}{2} \stackrel{\eta=D/(G\sqrt{T})}{=} DG\sqrt{T}.$$

Proof. Convexity: $\ell_t(a_t) - \ell_t(a^*) \leq \langle g_t, a_t - a^* \rangle$. Non-expansiveness of $\Pi_{\mathcal{K}}$:

$$\|a_{t+1} - a^*\|^2 \leq \|a_t - \eta g_t - a^*\|^2 = \|a_t - a^*\|^2 - 2\eta \langle g_t, a_t - a^* \rangle + \eta^2 \|g_t\|^2.$$

Rearrange, sum over t , and telescope the distances:

$$\sum_{t=1}^T \langle g_t, a_t - a^* \rangle \leq \frac{\|a_1 - a^*\|^2 - \|a_{T+1} - a^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|g_t\|^2 \leq \frac{D^2}{2\eta} + \frac{\eta G^2 T}{2}. \quad \blacksquare$$

Example (online logistic regression). $\|x_t\|_2 \leq 1$, $\mathcal{K} = \{w : \|w\|_2 \leq B\}$: $G \leq 1$, $D = 2B$, so regret $\leq 2B\sqrt{T}$ against an adversary, with $O(d)$ work per round.

Portfolios with OGD: stay away from the boundary

$\ell_t(p) = -\log\langle p, r_t \rangle$ is convex on Δ_d with $\nabla\ell_t(p) = -r_t/\langle p, r_t \rangle$, which blows up as $\langle p, r_t \rangle \rightarrow 0$. OGD needs $\|\nabla\ell_t\|_2 \leq G$, so restrict to

$$\mathcal{K}_\alpha = \{p \in \Delta_d : p_i \geq \alpha/d \ \forall i\}, \quad \text{on which} \quad \langle p, r_t \rangle \geq \frac{\alpha}{d} \sum_i r_t(i) \geq \frac{\alpha}{d} \|r_t\|_\infty,$$

$$\text{so} \quad \|\nabla\ell_t(p)\|_2 = \frac{\|r_t\|_2}{\langle p, r_t \rangle} \leq \frac{\sqrt{d} \|r_t\|_\infty}{(\alpha/d) \|r_t\|_\infty} = \frac{d^{3/2}}{\alpha}.$$

Portfolios with OGD: stay away from the boundary

$\ell_t(p) = -\log\langle p, r_t \rangle$ is convex on Δ_d with $\nabla\ell_t(p) = -r_t/\langle p, r_t \rangle$, which blows up as $\langle p, r_t \rangle \rightarrow 0$. OGD needs $\|\nabla\ell_t\|_2 \leq G$, so restrict to

$$\mathcal{K}_\alpha = \{p \in \Delta_d : p_i \geq \alpha/d \ \forall i\}, \quad \text{on which} \quad \langle p, r_t \rangle \geq \frac{\alpha}{d} \sum_i r_t(i) \geq \frac{\alpha}{d} \|r_t\|_\infty,$$

$$\text{so} \quad \|\nabla\ell_t(p)\|_2 = \frac{\|r_t\|_2}{\langle p, r_t \rangle} \leq \frac{\sqrt{d} \|r_t\|_\infty}{(\alpha/d) \|r_t\|_\infty} = \frac{d^{3/2}}{\alpha}.$$

Regret on $\mathcal{K}_\alpha$ ($D \leq \sqrt{2}$): OGD gives $\leq \sqrt{2} d^{3/2} \sqrt{T}/\alpha$.

Cost of shrinking the comparator: for $p^* \in \Delta_d$, $\tilde{p} = (1 - \alpha)p^* + \alpha \mathbf{1}/d \in \mathcal{K}_\alpha$ and $\log\langle \tilde{p}, r_t \rangle \geq \log\langle p^*, r_t \rangle + \log(1 - \alpha) \geq \log\langle p^*, r_t \rangle - 2\alpha$ for $\alpha \leq \frac{1}{2}$.

Total: $\sqrt{2} d^{3/2} \sqrt{T}/\alpha + 2\alpha T$. With $\alpha = d^{3/4} T^{-1/4}$, regret = $O(d^{3/4} T^{3/4})$.

	Regret	Per round
Universal portfolio	$d \log T$	integral on Δ_d
OGD on $\mathcal{K}_\alpha$	$d^{3/4} T^{3/4}$	$O(d)$

Tractability at a steep price in regret, because only convexity was used. Exploiting exp-concavity (Online Newton Step, Hazan–Agarwal–Kale 2007) gives $O(d \log T)$ in $\text{poly}(d)$ time per round.

Five lectures in one table

Setting	Assumption	Algorithm	Guarantee
Batch, realizable	IID	ERM	$R(\hat{h}) \leq \log(\mathcal{H} /\delta)/n$
Batch, agnostic	IID	ERM	$R(\hat{h}) - R(h_*) \leq \sqrt{2 \log(2 \mathcal{H} /\delta)/n}$
Online, realizable	IID	ERM each round	$O(\log \mathcal{H} \log T)$ mistakes
Online, realizable	none	Halving	$\log_2 \mathcal{H} $ mistakes
Online, agnostic	IID	ERM each round	$O(\sqrt{T \log \mathcal{H} })$ regret
Online, agnostic	none	Exponential weights	$\sqrt{T \log(\mathcal{H})/2}$ regret
Portfolio vs. best stock	none	Exp. weights ($\eta=1$)	$\log d$ log-wealth regret
Portfolio vs. best CRP	none	CEW ($\eta=1$, Cover)	$d \log T + 2$
Portfolio vs. best CRP	none	OGD on $\mathcal{K}_\alpha$	$O(d^{3/4} T^{3/4})$
Batch, squared loss	IID, convex $\mathcal{F}$	ERM	$L(\hat{f}) - L(f_{\mathcal{F}}^*) \leq \frac{28}{3} \log(2 \mathcal{F} /\delta)/n$
Online, squared loss	none	Averaged exp. weights	$2 \ln \mathcal{F} $ regret
Online convex	none	OGD	$DG\sqrt{T}$

Five lectures in one table

Setting	Assumption	Algorithm	Guarantee
Batch, realizable	IID	ERM	$R(\hat{h}) \leq \log(\mathcal{H} /\delta)/n$
Batch, agnostic	IID	ERM	$R(\hat{h}) - R(h_*) \leq \sqrt{2 \log(2 \mathcal{H} /\delta)/n}$
Online, realizable	IID	ERM each round	$O(\log \mathcal{H} \log T)$ mistakes
Online, realizable	none	Halving	$\log_2 \mathcal{H} $ mistakes
Online, agnostic	IID	ERM each round	$O(\sqrt{T \log \mathcal{H} })$ regret
Online, agnostic	none	Exponential weights	$\sqrt{T \log(\mathcal{H})/2}$ regret
Portfolio vs. best stock	none	Exp. weights ($\eta=1$)	$\log d$ log-wealth regret
Portfolio vs. best CRP	none	CEW ($\eta=1$, Cover)	$d \log T + 2$
Portfolio vs. best CRP	none	OGD on $\mathcal{K}_\alpha$	$O(d^{3/4} T^{3/4})$
Batch, squared loss	IID, convex $\mathcal{F}$	ERM	$L(\hat{f}) - L(f_{\mathcal{F}}^*) \leq \frac{28}{3} \log(2 \mathcal{F} /\delta)/n$
Online, squared loss	none	Averaged exp. weights	$2 \ln \mathcal{F} $ regret
Online convex	none	OGD	$DG\sqrt{T}$

The message

Dropping IID cost nothing in the rates. What you pay for is the **comparator** and **computation**.

Multi-armed bandits

Partial feedback: you only learn about the arm you pulled

Stochastic multi-armed bandit (notes, §10.1)

n arms with unknown distributions $P_1, \dots, P_n$ and means $\theta_i^* = \mathbb{E}_{X \sim P_i}[X]$

for $t = 1, 2, \dots$:

algorithm chooses an arm $I_t \in [n]$

observes a reward $X_{I_t, t} \sim P_{I_t}$, and nothing else

Rewards are 1-sub-Gaussian, $\mathbb{E} e^{\lambda(X - \theta_i^*)} \leq e^{\lambda^2/2}$
(Gaussian, or bounded in $[-1, 1]$). Only the means matter; the **gaps** are $\Delta_i = \max_j \theta_j^* - \theta_i^*$.

What changed

Experts: each round you saw every action's loss. Bandits: one number, for the arm you chose; learning about an arm costs a pull of it. Data collection and inference happen in one closed loop. (Also stochastic, not adversarial; EXP3 comes later.)

Stochastic multi-armed bandit (notes, §10.1)

n arms with unknown distributions $P_1, \dots, P_n$ and means $\theta_i^* = \mathbb{E}_{X \sim P_i}[X]$

for $t = 1, 2, \dots$:

algorithm chooses an arm $I_t \in [n]$

observes a reward $X_{I_t, t} \sim P_{I_t}$, and nothing else

Rewards are 1-sub-Gaussian, $\mathbb{E} e^{\lambda(X - \theta_i^*)} \leq e^{\lambda^2/2}$ (Gaussian, or bounded in $[-1, 1]$). Only the means matter; the **gaps** are $\Delta_i = \max_j \theta_j^* - \theta_i^*$.

Two goals: the two paradigms from Lecture 1.

- **Regret minimization:** $R_T = \max_j \theta_j^* T - \mathbb{E}[\sum_{t=1}^T X_{I_t, t}]$; want $R_T = o(T)$.
- **Best-arm identification:** output $\arg \max_i \theta_i^*$ with probability $\geq 1 - \delta$, using as few pulls as possible.

What changed

Experts: each round you saw every action's loss. Bandits: one number, for the arm you chose; learning about an arm costs a pull of it. Data collection and inference happen in one closed loop. (Also stochastic, not adversarial; EXP3 comes later.)

Regret counts pulls of bad arms

Let $T_i = \sum_{t \leq T} \mathbf{1}\{I_t = i\}$ be the number of pulls of arm i . Since I_t is chosen before $X_{I_t,t}$ is drawn, $\mathbb{E}[X_{i,t} \mathbf{1}\{I_t = i\}] = \theta_i^* \mathbb{P}(I_t = i)$, and therefore

$$R_T = \max_j \theta_j^* T - \mathbb{E}\left[\sum_{t=1}^T X_{I_t,t}\right] =$$

Regret counts pulls of bad arms

Let $T_i = \sum_{t \leq T} \mathbf{1}\{I_t = i\}$ be the number of pulls of arm i . Since I_t is chosen before $X_{I_t, t}$ is drawn, $\mathbb{E}[X_{i, t} \mathbf{1}\{I_t = i\}] = \theta_i^* \mathbb{P}(I_t = i)$, and therefore

$$R_T = \max_j \theta_j^* T - \mathbb{E}\left[\sum_{t=1}^T \sum_{i=1}^n X_{i, t} \mathbf{1}\{I_t = i\}\right] = \max_j \theta_j^* T - \sum_{i=1}^n \theta_i^* \mathbb{E}[T_i] = \sum_{i=1}^n \Delta_i \mathbb{E}[T_i],$$

using $\sum_i T_i = T$. Minimizing regret means pulling each suboptimal arm rarely, and arms with large gaps very rarely.

Regret counts pulls of bad arms

Let $T_i = \sum_{t \leq T} \mathbf{1}\{I_t = i\}$ be the number of pulls of arm i . Since I_t is chosen before $X_{I_t, t}$ is drawn, $\mathbb{E}[X_{i, t} \mathbf{1}\{I_t = i\}] = \theta_i^* \mathbb{P}(I_t = i)$, and therefore

$$R_T = \max_j \theta_j^* T - \mathbb{E}\left[\sum_{t=1}^T \sum_{i=1}^n X_{i, t} \mathbf{1}\{I_t = i\}\right] = \max_j \theta_j^* T - \sum_{i=1}^n \theta_i^* \mathbb{E}[T_i] = \sum_{i=1}^n \Delta_i \mathbb{E}[T_i],$$

using $\sum_i T_i = T$. Minimizing regret means pulling each suboptimal arm rarely, and arms with large gaps very rarely.

The two goals pull in different directions

Regret wants to abandon an arm as soon as it looks bad. Identification wants to keep pulling the top contenders until it is *sure*, and does not care how much reward is lost along the way.

Warm-up: A/B testing ($n = 2$)

Say $\theta_1^* > \theta_2^*$ and $\Delta = \theta_1^* - \theta_2^*$. The simplest possible algorithm:

Explore then commit

Input: $\tau \in \mathbb{N}$. Pull each arm exactly τ times; compute empirical means $\hat{\theta}_1, \hat{\theta}_2$.

For all $t > 2\tau$ play $\arg \max_i \hat{\theta}_i$.

How large must τ be so that $\hat{\theta}_1 > \hat{\theta}_2$ with probability $\geq 1 - \delta$? The CLT says $\hat{\theta}_i - \theta_i^* \approx \mathcal{N}(0, 1/\tau)$, but we want a non-asymptotic statement: the **Chernoff method** from Lecture 1, with sub-Gaussianity in place of Hoeffding's lemma,

$$\mathbb{P}(\hat{\theta}_i - \theta_i^* > \epsilon) \leq \inf_{\lambda > 0} e^{-\lambda \tau \epsilon} \mathbb{E} \left[e^{\lambda \sum_{s=1}^{\tau} (X_{i,s} - \theta_i^*)} \right] \leq \inf_{\lambda > 0} e^{-\lambda \tau \epsilon + \tau \lambda^2 / 2} = e^{-\tau \epsilon^2 / 2}.$$

Warm-up: A/B testing ($n = 2$)

Say $\theta_1^* > \theta_2^*$ and $\Delta = \theta_1^* - \theta_2^*$. The simplest possible algorithm:

Explore then commit

Input: $\tau \in \mathbb{N}$. Pull each arm exactly τ times; compute empirical means $\hat{\theta}_1, \hat{\theta}_2$.

For all $t > 2\tau$ play $\arg \max_i \hat{\theta}_i$.

How large must τ be so that $\hat{\theta}_1 > \hat{\theta}_2$ with probability $\geq 1 - \delta$? The CLT says $\hat{\theta}_i - \theta_i^* \approx \mathcal{N}(0, 1/\tau)$, but we want a non-asymptotic statement: the **Chernoff method** from Lecture 1, with sub-Gaussianity in place of Hoeffding's lemma,

$$\mathbb{P}(\hat{\theta}_i - \theta_i^* > \epsilon) \leq \inf_{\lambda > 0} e^{-\lambda \tau \epsilon} \mathbb{E} \left[e^{\lambda \sum_{s=1}^{\tau} (X_{i,s} - \theta_i^*)} \right] \leq \inf_{\lambda > 0} e^{-\lambda \tau \epsilon + \tau \lambda^2 / 2} = e^{-\tau \epsilon^2 / 2}.$$

So $E_i = \{|\hat{\theta}_i - \theta_i^*| \leq \sqrt{2 \log(4/\delta)/\tau}\}$ fails with probability $\leq \delta/2$, and on $E_1 \cap E_2$ with $\tau = \lceil 8\Delta^{-2} \log(4/\delta) \rceil$ (so that $\sqrt{2 \log(4/\delta)/\tau} \leq \Delta/2$),

$$\hat{\theta}_1 > \theta_1^* - \frac{\Delta}{2} = \theta_2^* + \frac{\Delta}{2} \geq \hat{\theta}_2 - \frac{\Delta}{2} + \frac{\Delta}{2} = \hat{\theta}_2.$$

With probability $\geq 1 - \delta$ we identify the best arm after $2\tau = O(\Delta^{-2} \log(1/\delta))$ pulls, then play it forever.