On adaptive regret bounds for non-stochastic bandits

Gergely Neu

INRIA Lille, SequeL team

→ Universitat Pompeu Fabra, Barcelona

Outline

- Online learning and bandits
- Adaptive bounds in online learning
- Adaptive bounds for bandits
  - What we already have
  - What's new: First-order bounds
  - What may be possible
  - What seems impossible*

*Opinion alert!

Online learning and non-stochastic bandits

For each round $t = 1,2, \ldots, T$

- Learner chooses action $I_t \in \{1,2, \ldots, N\}$
- Environment chooses losses $\ell_{t,i} \in [0,1]$ for all i
- Learner suffers loss ℓ_{t,I_t}
- Learner observes losses $\ell_{t,i}$ for all i

Online learning and non-stochastic bandits

For each round $t = 1,2,\ldots,T$

- Learner chooses action $I_t \in \{1,2,\ldots,N\}$
- Environment chooses losses $\ell_{t,i} \in [0,1]$ for all i
- Learner suffers loss ℓ_{t,I_t}
- Learner observes losses $\ell_{t,i}$ for all i

For each round $t = 1,2,\ldots,T$

- Learner chooses action $I_t \in \{1,2,\ldots,N\}$
- Environment chooses losses $\ell_{t,i} \in [0,1]$ for all
- Learner suffers loss ℓ_{t,I_t}
- Learner observes its own loss ℓ_{t,I_t}

Online learning and non-stochastic bandits

For each round $t = 1,2,\dots,T$

- Learner chooses action $I_t \in \{1,2,\dots,N\}$
- Environment chooses losses $\ell_{t,i} \in [0,1]$ for all i
- Learner suffers loss ℓ_{t,I_t}
- Learner observes losses $\ell_{t,i}$ for all i

Need to explore!

For each round $t = 1,2,\dots,T$

- Learner chooses action $I_t \in \{1,2,\dots,N\}$
- Environment chooses losses $\ell_{t,i} \in [0,1]$ for all
- Learner suffers loss ℓ_{t,I_t}
- Learner observes its own loss ℓ_{t,I_t}

Minimax regret

- Define (expected) regret against action i as

$$R_{T,i} = \mathbf{E}\left[\sum_{t=1}^{T} \ell_{t,I_t} - \sum_{t=1}^{T} \ell_{t,i}\right]$$

- Goal: minimize regret against the best action i^*

$$\boxed{R_T = R_{T,i^*} = \max_i R_{T,i}}$$

Minimax regret

- Define (expected) regret against action i as

$$R_{T,i} = \mathbf{E}\left[\sum_{t=1}^{T} \ell_{t,I_t} - \sum_{t=1}^{T} \ell_{t,i}\right]$$

- Goal: minimize regret against the best action i^*

$$\boxed{R_T = R_{T,i^*} = \max_i R_{T,i}}$$

Full information

$$R_T = \Theta\left(\sqrt{T \log N}\right)$$

Bandit

$$R_T = \Theta\left(\sqrt{NT}\right)$$

Beyond minimax: i.i.d. losses

Full information

$$R_T = \Theta\left(\sqrt{T \log N}\right)$$

$$\Theta(\log N)$$

Bandit

$$R_T = \Theta\left(\sqrt{NT}\right)$$

$$\Theta(N \log T)$$

Beyond minimax: "Higher-order" bounds

	Full information	Bandit
minimax	$R_T = O\left(\sqrt{T \log N}\right)$	$R_T = O\left(\sqrt{NT}\right)$
first-order $L_{T,i} = \sum_t \ell_{t,i}$	$R_T = O\left(\sqrt{L_{T,i^*} \log N}\right)$	
second-order $S_{T,i} = \sum_t \ell_{t,i}^2$	$R_T = O\left(\sqrt{S_{t,i^*} \log N}\right)$ ★ Cesa-Bianchi, Mansour, Stoltz (2005)	
variance $V_{T,i} = \sum_t \left(\ell_{t,i} - m\right)^2$	$R_T = O\left(\sqrt{V_{T,i^*} \log N}\right)$ ★ Hazan and Kale (2010)	

★ with a little cheating

Beyond minimax: "Higher-order" bounds

	Full information	Bandit
minimax	$R_T = O\left(\sqrt{T \log N}\right)$	$R_T = O\left(\sqrt{NT}\right)$
first-order $L_{T,i} = \sum_t \ell_{t,i}$	$R_T = O\left(\sqrt{L_{T,i^*} \log N}\right)$	
second-order $S_{T,i} = \sum_t \ell_{t,i}^2$	$R_T = O\left(\sqrt{S_{t,i^*} \log N}\right)$ ★ Cesa-Bianchi, Mansour, Stoltz (2005)	$R_T = \tilde{O}\left(\sqrt{\sum_i S_{t,i}}\right)$ Auer et al. (2002) + some hacking
variance $V_{T,i} = \sum_t \left(\ell_{t,i} - m\right)^2$	$R_T = O\left(\sqrt{V_{T,i^*} \log N}\right)$ ★ Hazan and Kale (2010)	$R_T = \tilde{O}\left(N^2 \sqrt{\sum_i V_{t,i}}\right)$ Hazan and Kale (2011)

★ with a little cheating

Beyond minimax: "Higher-order" bounds

	Full information	Bandit
minimax	$R_T = O\left(\sqrt{T \log N}\right)$	$R_T = O\left(\sqrt{NT}\right)$
first-order $L_{T,i} = \sum_t \ell_{t,i}$	$R_T = O\left(\sqrt{L_{T,i^*} \log N}\right)$	(it's complicated)
second-order $S_{T,i} = \sum_t \ell_{t,i}^2$	$R_T = O\left(\sqrt{S_{t,i^*} \log N}\right)$ ★ Cesa-Bianchi, Mansour, Stoltz (2005)	$R_T = \tilde{O}\left(\sqrt{\sum_i S_{t,i}}\right)$ Auer et al. (2002) + some hacking
variance $V_{T,i} = \sum_t \left(\ell_{t,i} - m\right)^2$	$R_T = O\left(\sqrt{V_{T,i^*} \log N}\right)$ ★ Hazan and Kale (2010)	$R_T = \tilde{O}\left(N^2 \sqrt{\sum_i V_{t,i}}\right)$ Hazan and Kale (2011)

★ with a little cheating

First-order bounds for bandits

(it's complicated)

- "Small-gain" bounds:
  - Consider the gain game with $g_{t,i} = 1 - \ell_{t,i}$
  - Auer, Cesa-Bianchi, Freund and Schapire (2002):

$$R_T = O\left(\sqrt{N G_{T,i^*} \log N}\right)$$

$G_{T,i} = \sum_t g_{t,i}$

First-order bounds for bandits

(it's complicated)

- "Small-gain" bounds:
  - Consider the gain game with $g_{t,i} = 1 - \ell_{t,i}$
  - Auer, Cesa-Bianchi, Freund and Schapire (2002):

$$R_T = O\left(\sqrt{N G_{T,i^*} \log N}\right)$$

$G_{T,i} = \sum_t g_{t,i}$

Problem:

only good if best expert is bad!

First-order bounds for bandits

(it's complicated)

- "Small-gain" bounds: $R_T = O\left(\sqrt{N G_{T,i^*} \log N}\right)$
- A little trickier analysis gives

$$R_T = O\left(\sqrt{\sum_t \sum_i g_{t,i} \log N}\right)$$

or

$$R_T = O\left(\sqrt{\sum_t \sum_i \ell_{t,i} \log N}\right)$$

First-order bounds for bandits

(it's complicated)

- "Small-gain" bounds: $R_T = O\left(\sqrt{N G_{T,i^*} \log N}\right)$
- A little trickier analysis gives

$$R_T = O\left(\sqrt{\sum_t \sum_i g_{t,i} \log N}\right)$$

or

$$R_T = O\left(\sqrt{\sum_t \sum_i \ell_{t,i} \log N}\right)$$

Problem:

one misbehaving action ruins the bound!

First-order bounds for bandits

(it's complicated)

- "Small-gain" bounds: $R_T = O\left(\sqrt{N G_{T,i^*} \log N}\right)$
- A little trickier analysis gives $R_T = O\left(\sqrt{\sum_t \sum_i \ell_{t,i} \log N}\right)$
- Some obscure actual first-order bounds:
  - Stoltz (2005): $N\sqrt{L_T^*}$
  - Allenberg, Auer, Györfi and Ottucsák (2006): $\sqrt{N L_T^*}$
  - Rakhlin and Sridharan (2013): $N^{3/2}\sqrt{L_T^*}$

First-order bounds for bandits

(it's complicated)

- "Small-gain" bounds: $R_T = O\left(\sqrt{N G_{T,i^*} \log N}\right)$
- A little trickier analysis gives $R_T = O\left(\sqrt{\sum_t \sum_i \ell_{t,i} \log N}\right)$
- Some obscure actual first-order bounds:
  - $\sqrt{N L_T^*}$

Problem:
no real insight from analyses!

First-order bounds for non-stochastic bandits

A typical bandit algorithm

For every round $t = 1,2, \dots, T$

- Choose arm $I_t = i$ with probability $p_{t,i}$
- Compute unbiased loss estimate

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i}} \mathbf{1}_{\{I_t = i\}}$$

- Use $\hat{\ell}_{t,i}$ in a black-box online learning algorithm to compute $\boldsymbol{p}_{t+1}$

A typical regret bound

$$R_T \leq \boxed{\frac{\log N}{\eta} + \eta \mathbf{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} p_{t,i}\left(\hat{\ell}_{t,i}\right)^2\right]}$$

(η: “learning rate”)

A typical regret bound

$$R_T \leq \boxed{\frac{\log N}{\eta} + \eta \mathbf{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} p_{t,i}\left(\hat{\ell}_{t,i}\right)^2\right]}$$

$$\leq \boxed{\frac{\log N}{\eta} + \eta \mathbf{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} \hat{\ell}_{t,i}\right]}$$

(η: "learning rate")

A typical regret bound

$$R_T \le \frac{\log N}{\eta} + \eta \mathbf{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} p_{t,i}\left(\hat{\ell}_{t,i}\right)^2\right]$$

$$\le \frac{\log N}{\eta} + \eta \mathbf{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} \hat{\ell}_{t,i}\right]$$

$$= \frac{\log N}{\eta} + \eta \sum_{i=1}^{N} L_{T,i}$$

(η: "learning rate")

A typical regret bound

$$R_T \le \boxed{\frac{\log N}{\eta} + \eta \mathbf{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} p_{t,i}\left(\hat{\ell}_{t,i}\right)^2\right]}$$

$$\le \boxed{\frac{\log N}{\eta} + \eta \mathbf{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} \hat{\ell}_{t,i}\right]}$$

$$= \boxed{\frac{\log N}{\eta} + \eta \sum_{i=1}^{N} L_{T,i}} = \tilde{O}\left(\sqrt{\textstyle\sum_{i=1}^{N} L_{T,i}}\right)$$

(η: "learning rate")

(for appropriate η)

A typical regret bound

$$R_T = \tilde{O}\left(\sqrt{\sum_{i=1}^{N} L_{T,i}}\right)$$

A typical regret bound

$$R_T = \tilde{O}\left(\sqrt{\sum_{i=1}^{N} L_{T,i}}\right)$$

It's all because
$\mathbf{E}\left[\hat{L}_{T,i}\right] = L_{T,i}$!!!

A typical regret bound

$$R_T = \tilde{O}\left(\sqrt{\sum_{i=1}^{N} L_{T,i}}\right)$$

It's all because
$\mathbf{E}\left[\hat{L}_{T,i}\right] = L_{T,i}$!!!

Idea: try to enforce
$\mathbf{E}\left[\hat{L}_{T,i}\right] = O(L_{T,i^*})$

A typical regret bound

$$R_T = \ldots \left(\sqrt{\sum \ldots} \; \ldots_{T,i} \right)$$

Need optimistic estimates!

It's all because
$\mathbf{E}\left[\hat{L}_{T,i}\right] = L_{T,i}$!!!

Idea: try to enforce
$\mathbf{E}\left[\hat{L}_{T,i}\right] = O\left(L_{T,i^*}\right)$

A typical algorithm – fixed!

For every round $t = 1,2,\dots,T$

- Choose arm $I_t = i$ with probability $p_{t,i}$
- Compute unbiased loss estimate

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i}} \mathbf{1}_{\{I_t = i\}}$$

- Use $\hat{\ell}_{t,i}$ in a black-box online learning algorithm to compute $\boldsymbol{p}_{t+1}$

A typical algorithm – fixed!

For every round $t = 1,2, \dots, T$

- Choose arm $I_t = i$ with probability $p_{t,i}$
- Compute biased loss estimate

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i} + \gamma} \mathbf{1}_{\{I_t = i\}}$$

- Use $\hat{\ell}_{t,i}$ in a black-box online learning algorithm to compute $\boldsymbol{p}_{t+1}$

"Implicit exploration"
(Kocák, N, Valko and Munos, 2015)

Algorithm: Follow the Perturbed Leader
(Kalai and Vempala, 2005, Poland, 2005)

For every round $t = 1,2,\ldots,T$

- Draw perturbation $Z_{t,i} \sim \mathrm{Exp}(1)$ for all i
- Choose arm $I_t = \arg\min_i \left(\eta \widehat{L}_{t-1,i} - Z_{t,i}\right)$
- Compute biased loss estimate

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i} + \gamma} \mathbf{1}_{\{I_t = i\}}$$

- Update $\widehat{L}_{t,i} = \widehat{L}_{t-1,i} + \hat{\ell}_{t,i}$

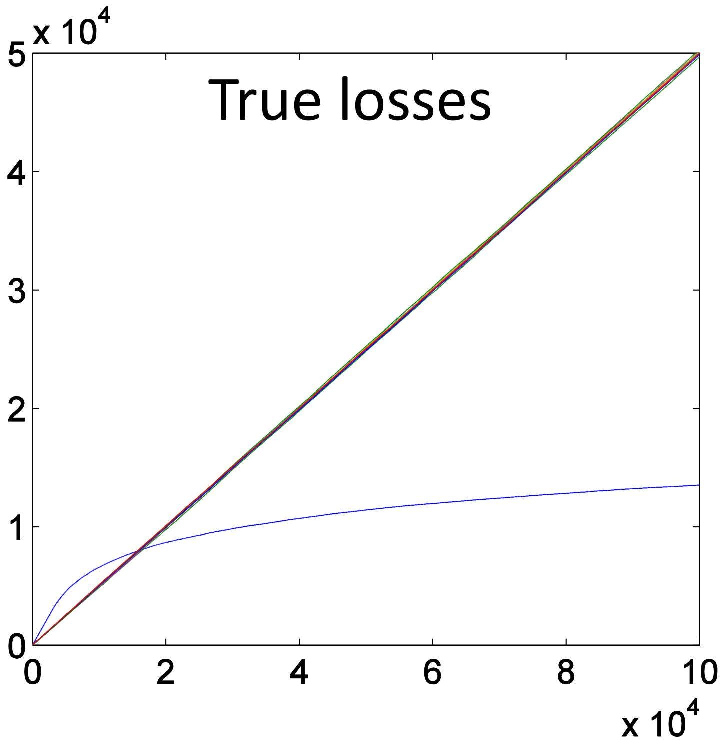
"Implicit exploration"
(Kocák, N, Valko and Munos, 2015)

Implicit exploration in action

$$\widehat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i} + \gamma} \mathbf{1}_{\{I_t = i\}}$$

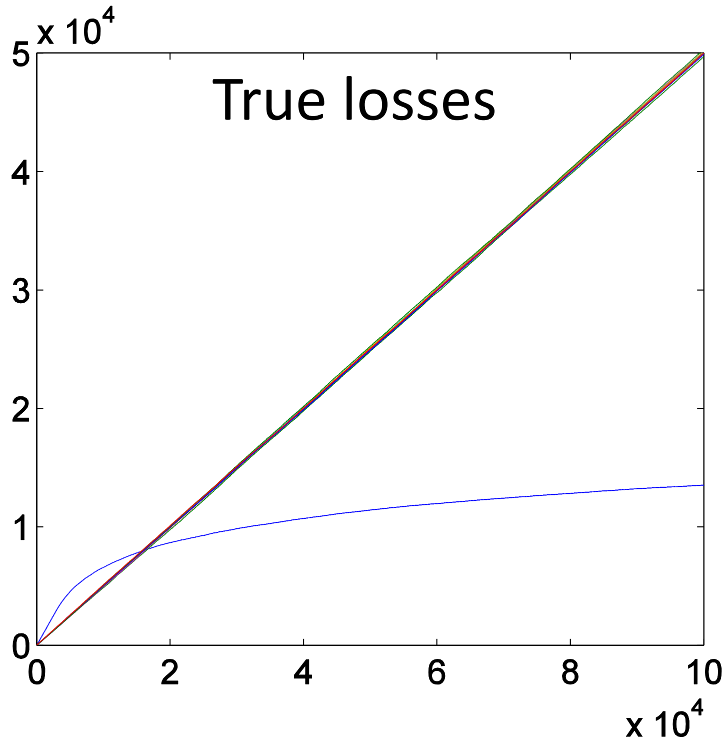
Implicit exploration in action

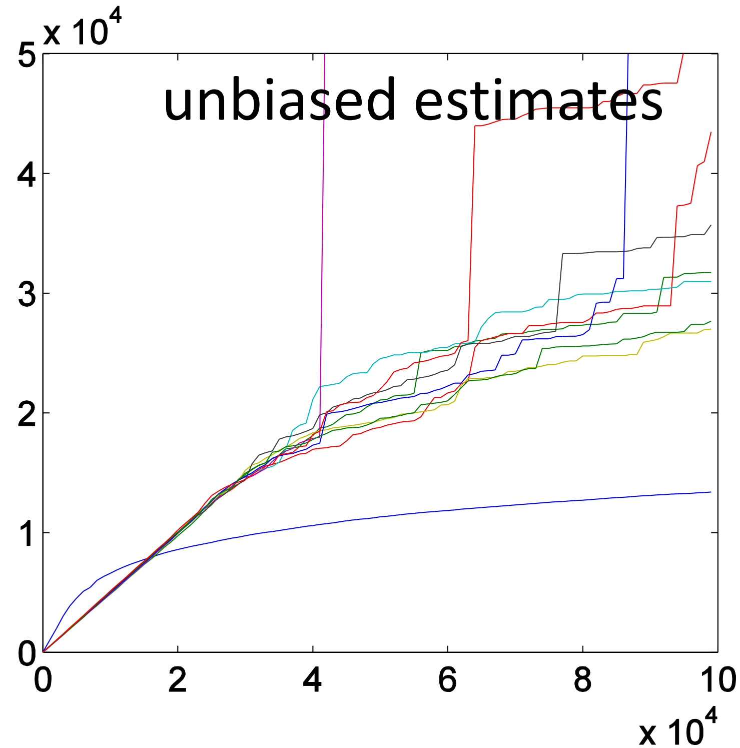
$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i} + \gamma} \mathbf{1}_{\{I_t = i\}}$$



Implicit exploration in action

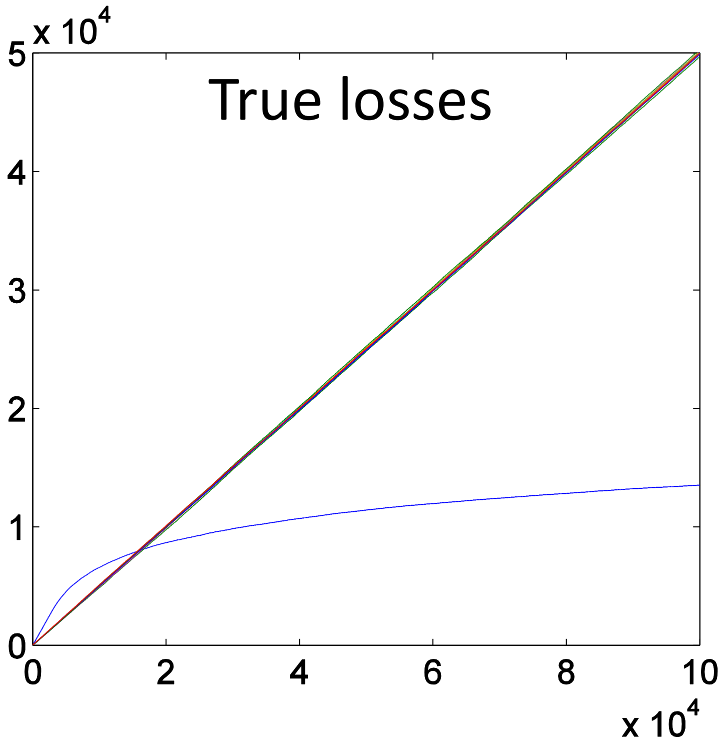
$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i} + \gamma} \mathbf{1}_{\{I_t = i\}}$$

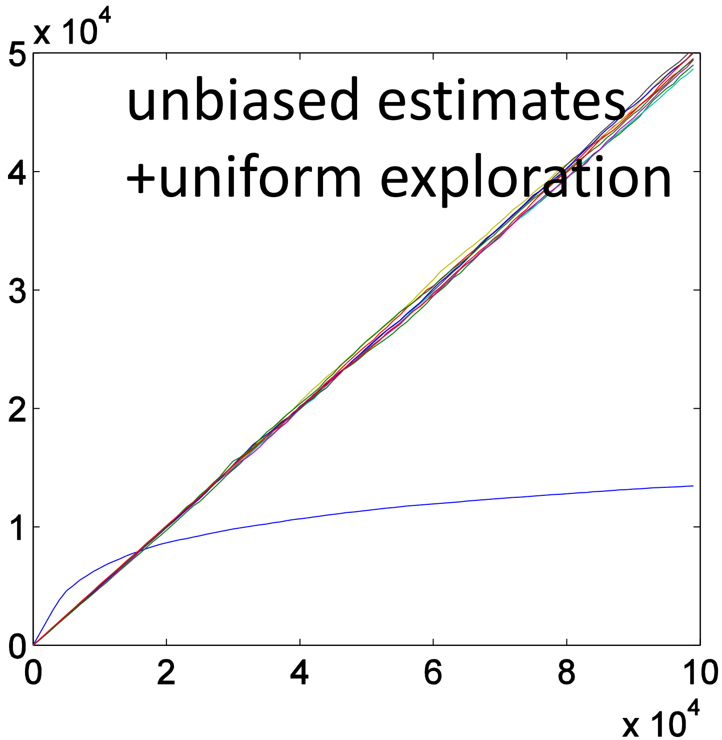




Implicit exploration in action

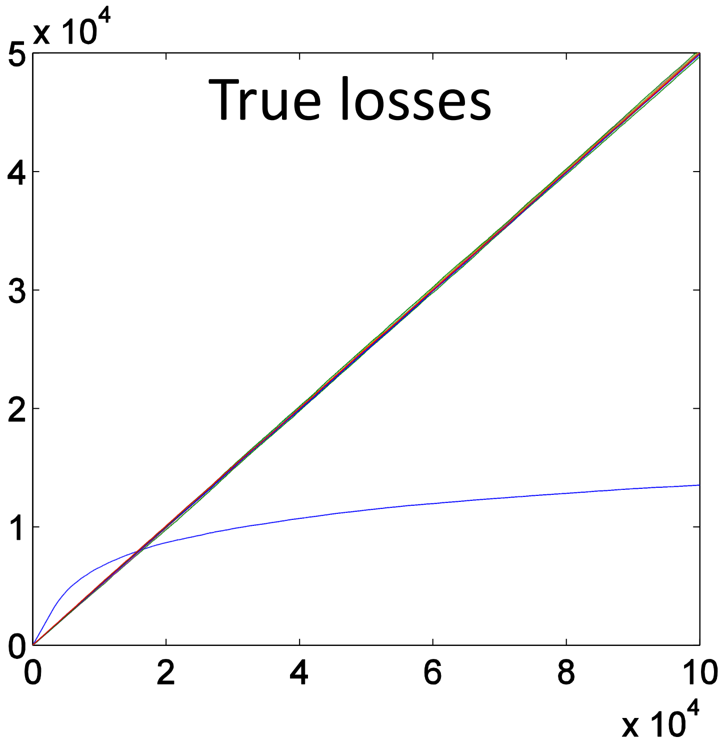
$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i} + \gamma} \mathbf{1}_{\{I_t = i\}}$$

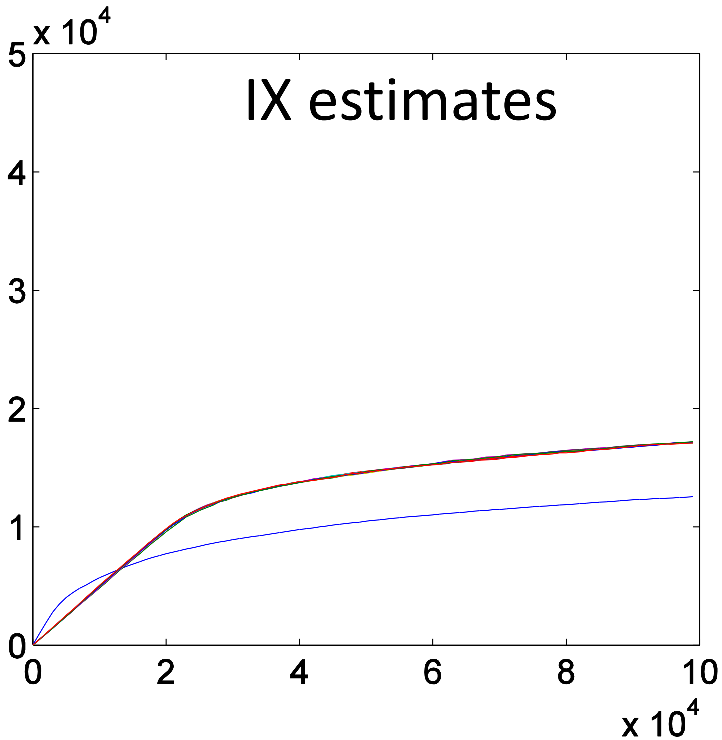




Implicit exploration in action

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i} + \gamma} \mathbf{1}_{\{I_t = i\}}$$





Optimistic estimates

Lemma (N, 2015a): Assume that $Z_{t,i} \leq B$ for all t and i. Then, for any i and j,

$$\widehat{L}_{T,i} \leq \widehat{L}_{T,j} + \frac{(\log N + B)}{\eta} + \frac{1}{\gamma}$$

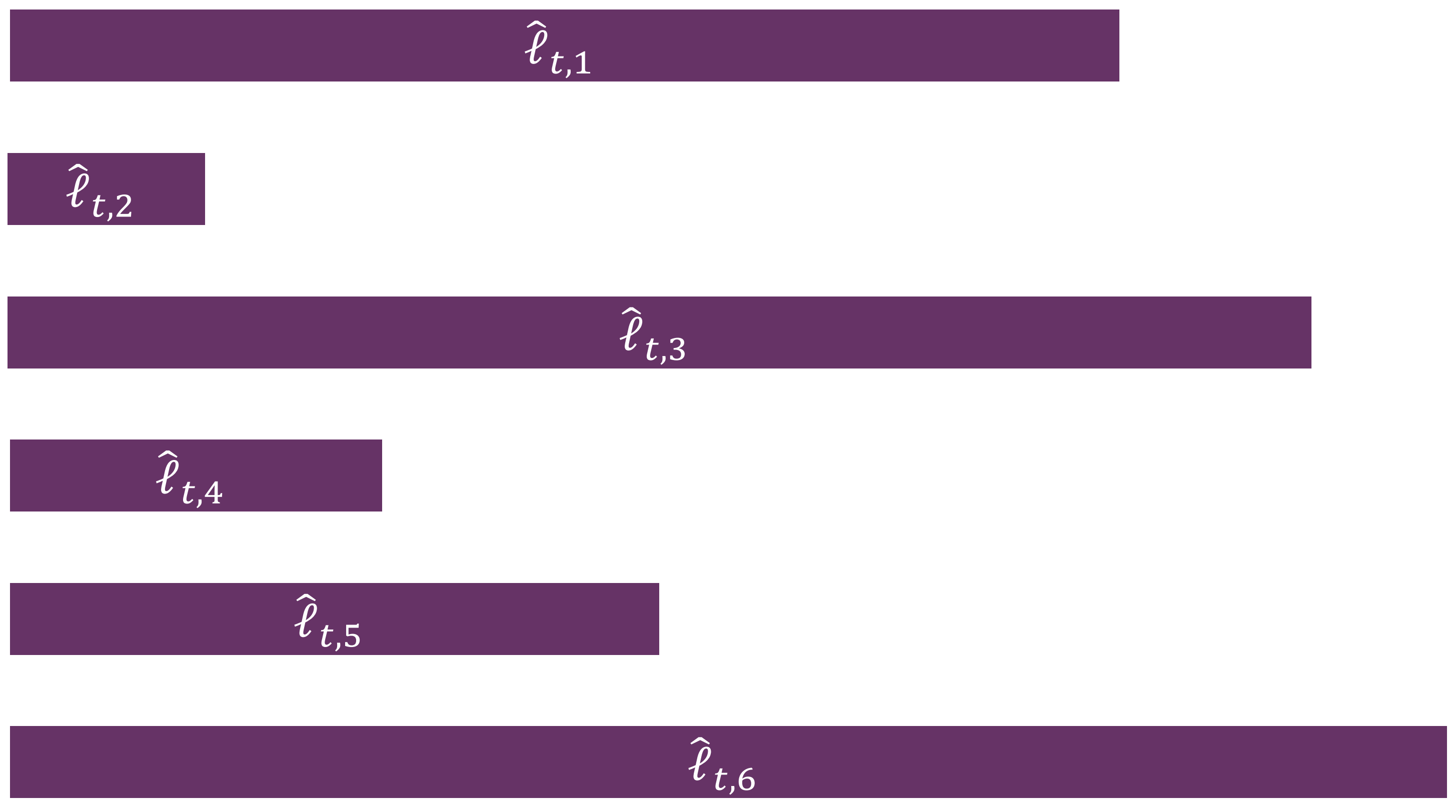
Optimistic estimates

Lemma (N, 2015a): Assume that $Z_{t,i} \leq B$ for all t and i. Then, for any i and j,

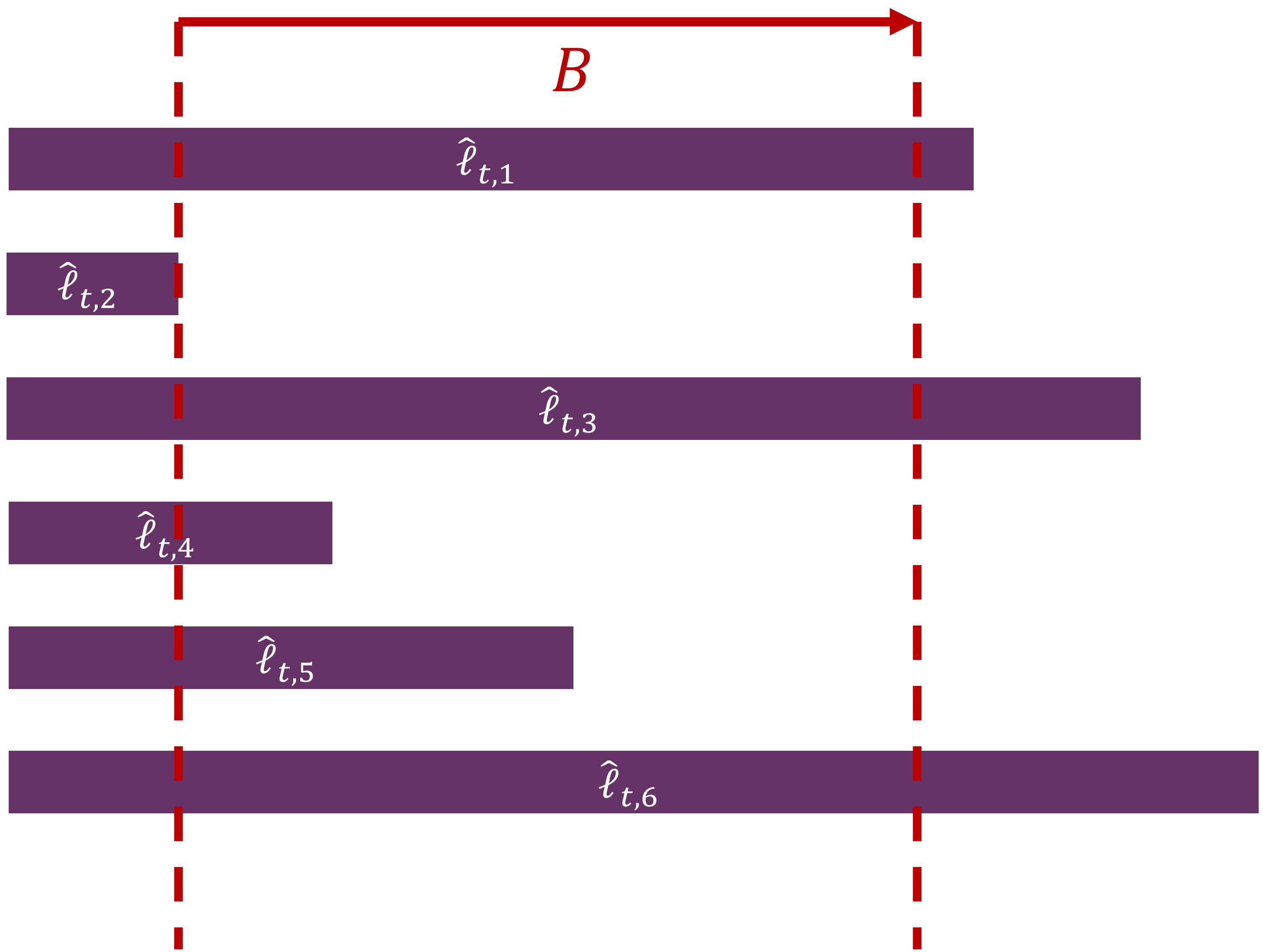
$$\hat{L}_{T,i} \leq \hat{L}_{T,j} + \frac{(\log N + B)}{\eta} + \frac{1}{\gamma}$$

All perturbations are nicely bounded with high probability → bad arms are suppressed!

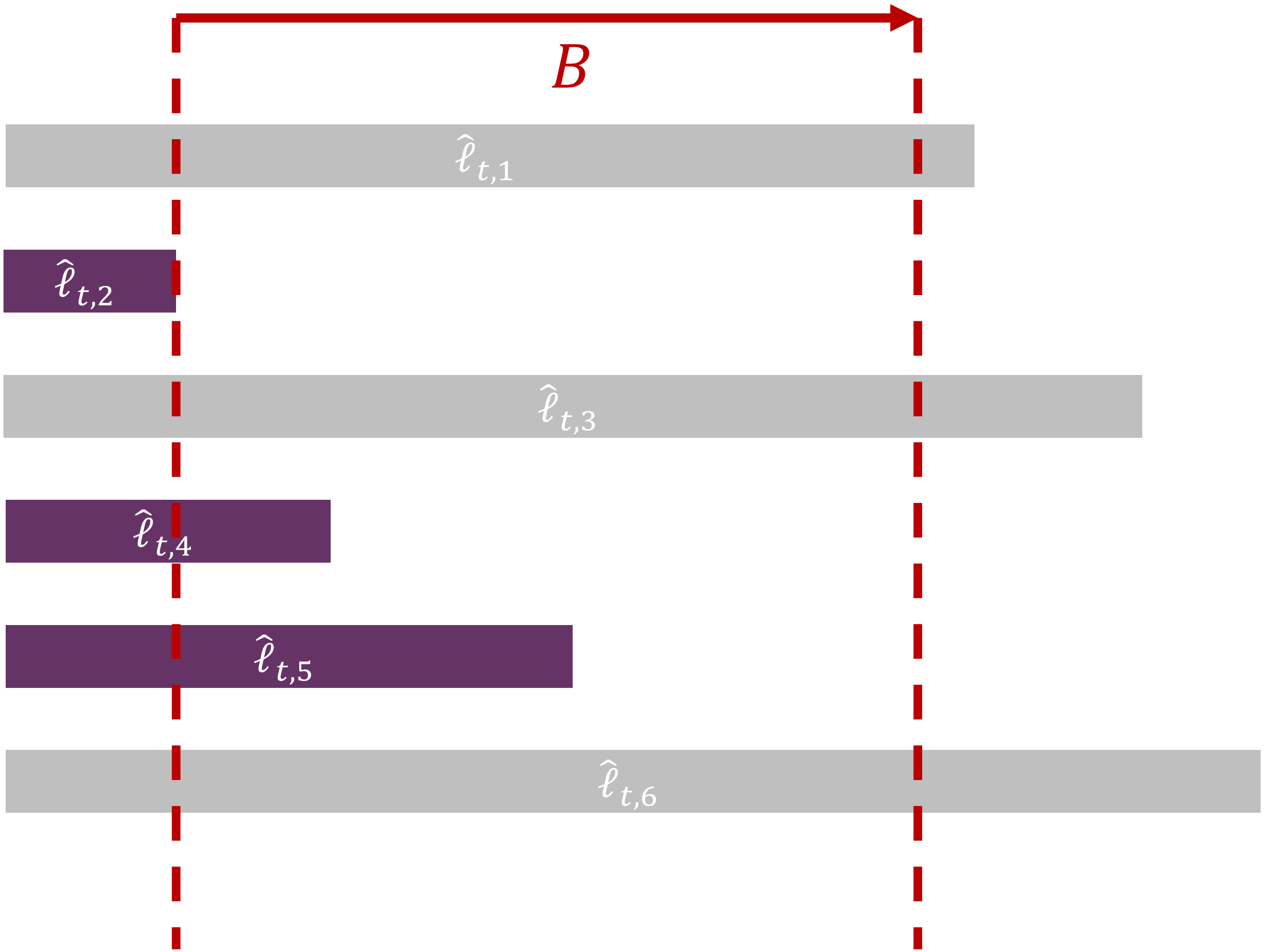
Suppressing bad arms



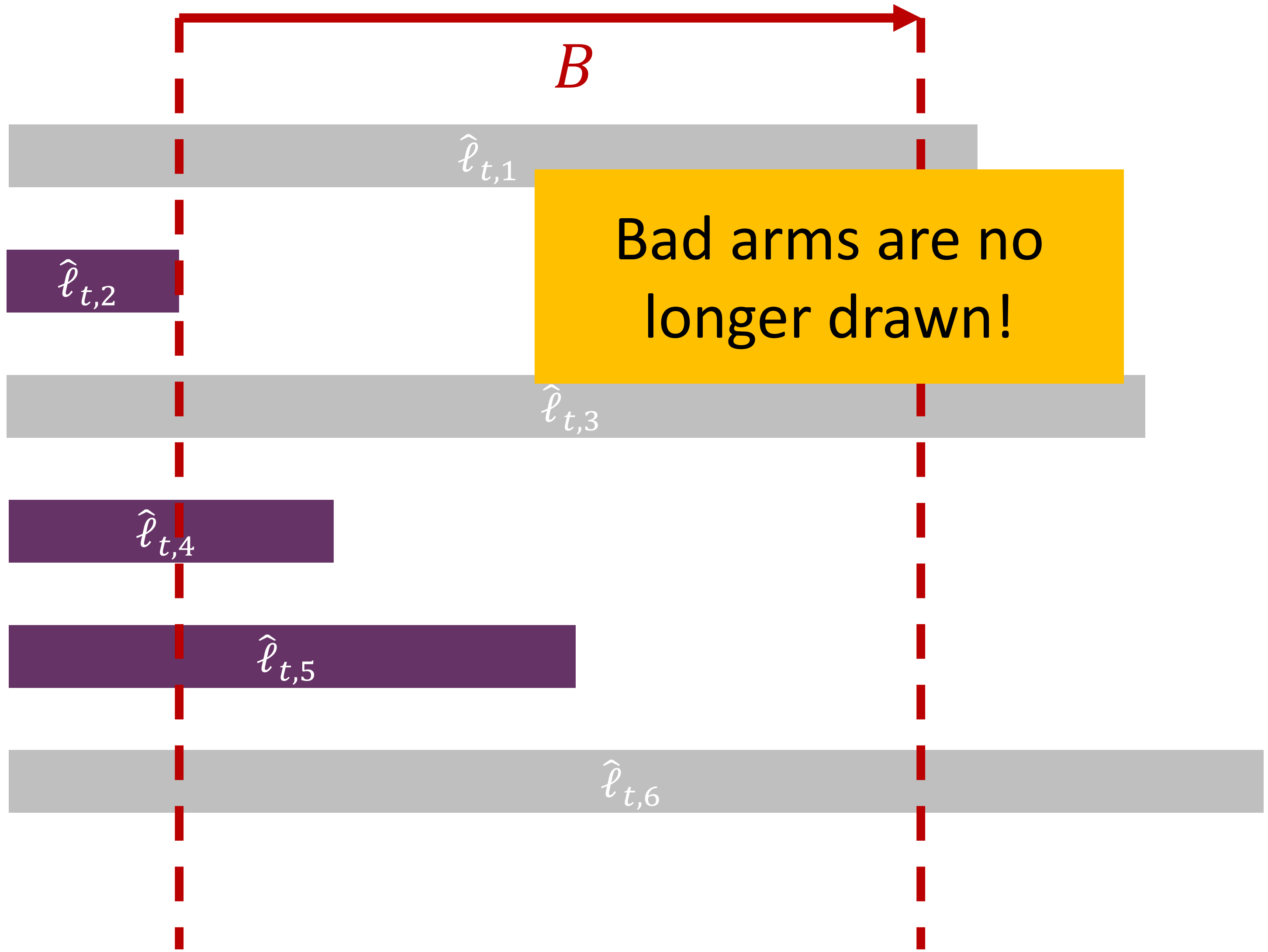
Suppressing bad arms



Suppressing bad arms

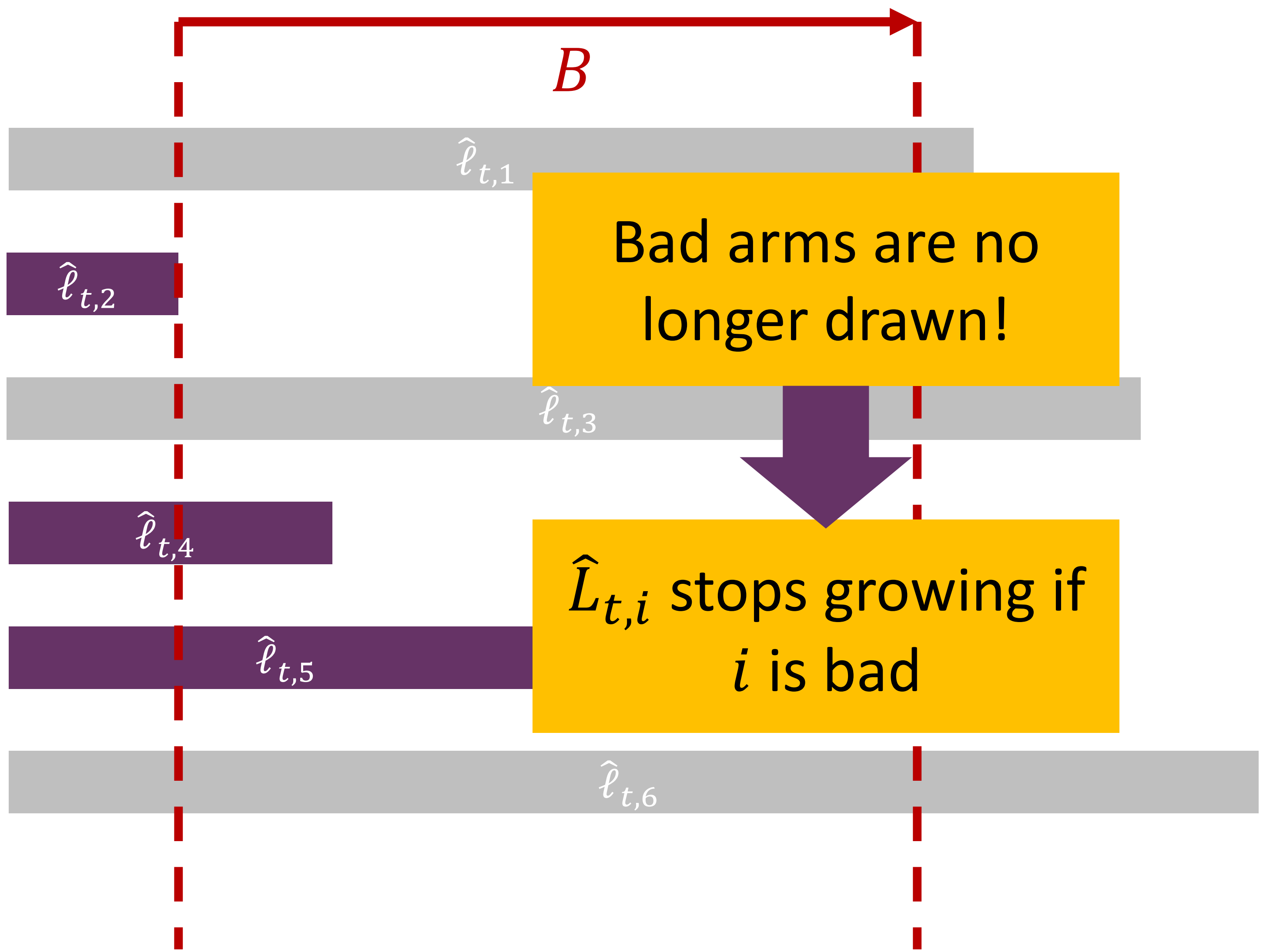


Suppressing bad arms



Bad arms are no longer drawn!

Suppressing bad arms



Finally: a first-order bound!

$$R_T \leq \boxed{\frac{\log N}{\eta} + \eta \mathbf{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} p_{t,i}\left(\hat{\ell}_{t,i}\right)^2\right]}$$

$$\leq \boxed{\frac{\log N}{\eta} + \eta \mathbf{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} \hat{\ell}_{t,i}\right]}$$

$$= \boxed{\frac{\log N}{\eta} + \eta \sum_{i=1}^{N} L_{T,i}} = \tilde{O}\left(\sqrt{\sum_{i=1}^{N} L_{T,i}}\right)$$

(for appropriate η)

Finally: a first-order bound!

$$R_T \le \frac{\log N}{\eta} + \eta \mathbf{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} p_{t,i}\left(\hat{\ell}_{t,i}\right)^2\right]$$

$$\le \frac{\log N}{\eta} + \eta \mathbf{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} \hat{\ell}_{t,i}\right]$$

Lemma

$$\le \frac{\log N}{\eta} + \eta N L_{T,i^*} + N(B + \log N) = \tilde{O}\left(\sqrt{N L_{T,i^*}}\right)$$

(for appropriate η, γ, B)

Finally: a first-order bound!

$$R_T = \tilde{O}\left(\sqrt{NL_{T,i^*}}\right)$$

Parameters:

- Set $\gamma = \eta/2$

Finally: a first-order bound!

$$R_T = \tilde{O}\left(\sqrt{N L_{T,i^*}}\right)$$

Parameters:

- Set $\gamma = \eta/2$
- If we know L_{T,i^*}: $\eta = \sqrt{\frac{(\log N + 1)}{N L_{T,i^*}}}$

Finally: a first-order bound!

$$R_T = \tilde{O}\left(\sqrt{NL_{T,i^*}}\right)$$

Parameters:

- Set $\gamma = \eta/2$
- If we know L_{T,i^*}: $\eta = \sqrt{\frac{(\log N+1)}{NL_{T,i^*}}}$
- If we don't: $\eta_t = \sqrt{\frac{(\log N+1)}{N\left(1+\sum_i \hat{L}_{t-1,i}\right)}}$

Finally: a first-order bound!

$$R_T = \tilde{O}\left(\sqrt{N L_{T,i^*}}\right)$$

Parameters:

- Set $\gamma = \eta/2$
- If we know L_{T,i^*}: $\eta = \sqrt{\frac{(\log N+1)}{N L_{T,i^*}}}$
- If we don't: $\eta_t = \sqrt{\frac{(\log N+1)}{N\left(1+\sum_i \hat{L}_{t-1,i}\right)}}$

Arguments also extend to combinatorial semi-bandits!

What’s next?

Beyond minimax: “Higher-order” bounds

	Full information	Bandit
minimax	$R_T = \Theta\left(\sqrt{T \log N}\right)$	$R_T = O\left(\sqrt{NT}\right)$
first-order $L_{T,i} = \sum_t \ell_{t,i}$	$R_T = O\left(\sqrt{L_{T,i^*} \log N}\right)$	(it’s complicated)
second-order $S_{T,i} = \sum_t \ell_{t,i}^2$	$R_T = O\left(\sqrt{S_{t,i^*} \log N}\right)$ ★ Cesa-Bianchi, Mansour, Stoltz (2005)	$R_T = \tilde{O}\left(\sqrt{\sum_i S_{t,i}}\right)$ Auer et al. (2002) + some hacking
variance $V_{T,i} = \sum_t \left(\ell_{t,i} - m\right)^2$	$R_T = O\left(\sqrt{V_{T,i^*} \log N}\right)$ ★ Hazan and Kale (2010)	$R_T = \tilde{O}\left(N^2 \sqrt{\sum_i V_{t,i}}\right)$ Hazan and Kale (2011)

★ with a little cheating

Beyond minimax: "Higher-order" bounds

	Full information	Bandit
minimax	$R_T = \Theta\left(\sqrt{T \log N}\right)$	$R_T = O\left(\sqrt{NT}\right)$
first-order $L_{T,i} = \sum_t \ell_{t,i}$	$R_T = O\left(\sqrt{L_{T,i^*} \log N}\right)$	$R_T = \tilde{O}\left(\sqrt{N L_{T,i^*}}\right)$
second-order $S_{T,i} = \sum_t \ell_{t,i}^2$	$R_T = O\left(\sqrt{S_{t,i^*} \log N}\right)$ ★ Cesa-Bianchi, Mansour, Stoltz (2005)	$R_T = \tilde{O}\left(\sqrt{\sum_i S_{t,i}}\right)$ Auer et al. (2002) + some hacking
variance $V_{T,i} = \sum_t \left(\ell_{t,i} - m\right)^2$	$R_T = O\left(\sqrt{V_{T,i^*} \log N}\right)$ ★ Hazan and Kale (2010)	$R_T = \tilde{O}\left(N^2 \sqrt{\sum_i V_{t,i}}\right)$ Hazan and Kale (2011)

★ with a little cheating

Beyond minimax: “Higher-order” bounds

	Full information	Bandit
minimax	$R_T = \Theta\left(\sqrt{T \log N}\right)$	$R_T = O\left(\sqrt{NT}\right)$
first-order $L_{T,i} = \sum_t \ell_{t,i}$	$R_T = O\left(\sqrt{L_{T,i^*} \log N}\right)$	$R_T = \tilde{O}\left(\sqrt{N L_{T,i^*}}\right)$
second-order $S_{T,i} = \sum_t \ell_{t,i}^2$	★ (5)	$R_T = \tilde{O}\left(\sqrt{\sum_i S_{t,i}}\right)$ Auer et al. (2002) + some hacking
variance $V_{T,i} = \sum_t \left(\ell_{t,i} - m\right)^2$	$R_T = O\left(\sqrt{V_{T,i^*} \log N}\right)$ ★ Hazan and Kale (2010)	$R_T = \tilde{O}\left(N^2 \sqrt{\sum_i V_{t,i}}\right)$ Hazan and Kale (2011)

What about these?

★ with a little cheating

Beyond first-order bounds?

A key tool for adaptive bounds in full-info: **PROD** (Cesa-Bianchi, Mansour and Stoltz, 2005)

$$p_{t,i} \propto \prod_{s=1}^{t-1} (1 - \eta \ell_{s,i})$$

Beyond first-order bounds?

A key tool for adaptive bounds in full-info: **PROD** (Cesa-Bianchi, Mansour and Stoltz, 2005)

$$p_{t,i} \propto \prod_{s=1}^{t-1} (1 - \eta \ell_{s,i})$$

Used for proving

- Second-order bounds (Cesa-Bianchi et al., 2005)
- Variance-dependent bounds (Hazan and Kale, 2010)
- Path-length bounds (Steinhardt and Liang, 2014)
- Quantile bounds (Koolen and Van Erven, 2015)
- Best-of-both-worlds bounds (Sani et al., 2014)
- …

Beyond first-order bounds?

A key tool for adaptive bounds in full-info: **PROD** (Cesa-Bianchi, Mansour and Stoltz, 2005)

$$p_{t,i} \propto \prod_{s=1}^{t-1} (1 - \eta \ell_{s,i})$$

But does it work for bandits?

l., 2005)

d Kale, 2010)

ng, 2014)

2015)

, 2014)

- …

Beyond first-order bounds?

A key tool for adaptive bounds in full-info: **PROD** (Cesa-Bianchi, Mansour and Stoltz, 2005)

$$p_{t,i} \propto \prod_{s=1}^{t-1} \left(1 - \eta \ell_{s,i}\right)$$

Used for proving

l., 2005)

d Kale, 2010)

ng, 2014)

2015)

bounds (Sani et al., 2014)

- …

But does it work for bandits?

NO

Why does PROD fail?

$$p_{t,i} \propto \prod_{s=1}^{t-1} \left(1 - \eta \hat{\ell}_{s,i}\right)$$

=

$$p_{t,i} \propto e^{-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{s,i}}$$

with

$$\tilde{\ell}_{t,i} = -\frac{1}{\eta} \log\left(1 - \eta \hat{\ell}_{t,i}\right)$$

Why does PROD fail?

EXP3 with a pessimistic estimate:

$$\mathbf{E}\left[\tilde{\ell}_{t,i}\right] \geq \ell_{t,i}$$

=

$$p_{t,i} \propto e^{-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{s,i}}$$

with

$$\tilde{\ell}_{t,i} = -\frac{1}{\eta}\log\left(1 - \eta\hat{\ell}_{t,i}\right)$$

Why does PROD fail?

EXP3 with a pessimistic estimate:

$$\mathbf{E}\left[\tilde{\ell}_{t,i}\right] \geq \ell_{t,i}$$

=

$$p_{t,i} \propto e^{-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{s,i}}$$

with

$$\tilde{\ell}_{t,i} = -\frac{1}{\eta}\log\left(1 - \eta\hat{\ell}_{t,i}\right)$$

Implicit exploration

$$\tilde{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i} + \gamma}\mathbf{1}_{\{I_t = i\}}$$

Why does PROD fail?

EXP3 with a
pessimistic estimate:
$$\mathbf{E}\left[\tilde{\ell}_{t,i}\right] \geq \ell_{t,i}$$

Implicit exploration

$$\tilde{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i} + \gamma} \mathbf{1}_{\{I_t = i\}}$$

$$=$$

$$\approx$$

$$p_{t,i} \propto e^{-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{s,i}}$$

with

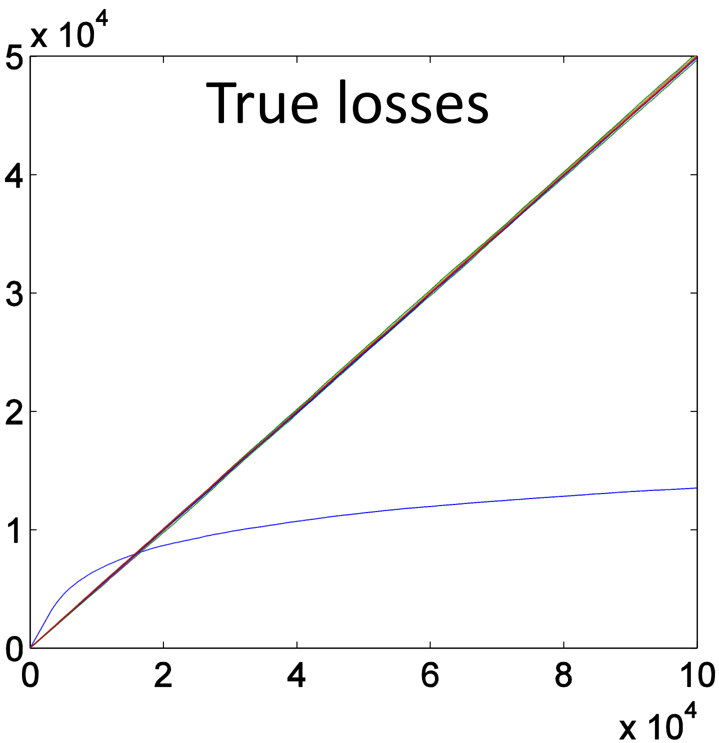
$$\tilde{\ell}_{t,i} = -\frac{1}{\eta} \log\left(1 - \eta \hat{\ell}_{t,i}\right)$$

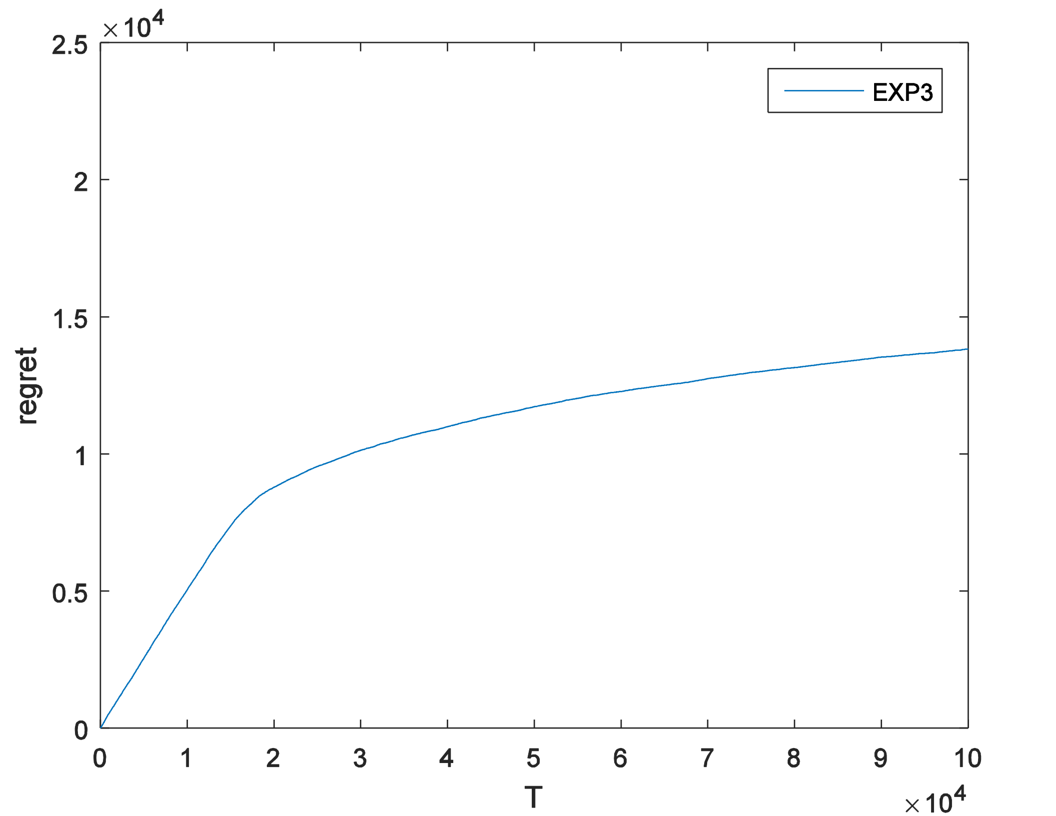
$$p_{t,i} \propto e^{-\eta \sum_{s=1}^{t-1} \tilde{\ell}_{s,i}}$$

with

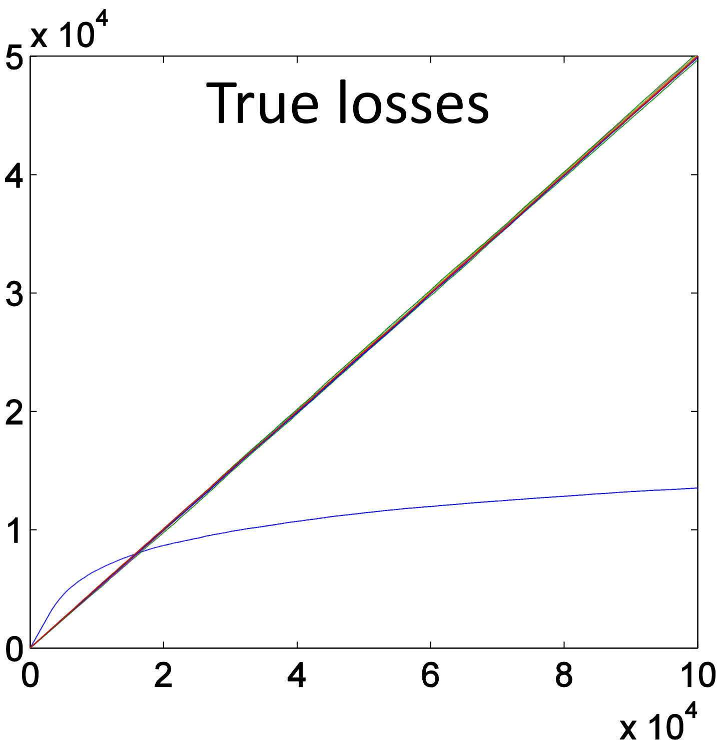
$$\tilde{\ell}_{t,i} = \frac{1}{\eta} \log\left(1 + \eta \hat{\ell}_{t,i}\right)$$

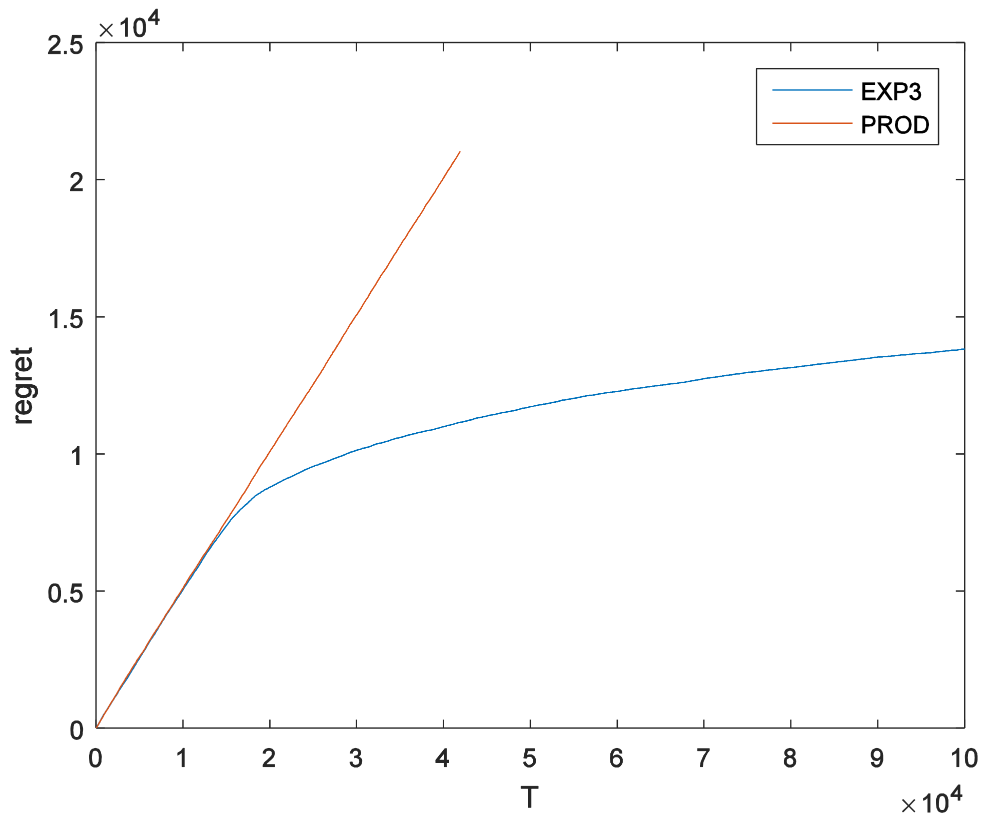
PROD for bandits





PROD for bandits





Summary

- Key for first-order bounds: implicit exploration

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i} + \gamma} \mathbf{1}_{\{I_t = i\}}$$

+ also key for high-probability bounds! (NIPS 2015)

Summary

- Key for first-order bounds: implicit exploration

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i} + \gamma} \mathbf{1}_{\{I_t = i\}}$$

+ also key for high-probability bounds! (NIPS 2015)

- Further bounds seem to be difficult to prove: smoothness conflicts with need to explore!

Summary

- Key for first-order bounds: implicit exploration

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i} + \gamma} \mathbf{1}_{\{I_t = i\}}$$

+ also key for high-probability bounds! (NIPS 2015)

- Further bounds seem to be difficult to prove: smoothness conflicts with need to explore!
- More depressing results by Lattimore (NIPS 2015)

Summary

- Key for first-order bounds: implicit exploration

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i} + \gamma} \mathbf{1}_{\{I_t = i\}}$$

+ also key for high-probability bounds! (NIPS 2015)

- Further bounds seem to be difficult to prove: smoothness conflicts with need to explore!
- More depressing results by Lattimore (NIPS 2015)
  - Second-order bounds (Cesa-Bianchi et al., 2005)
  - Variance-dependent bounds (Hazan and Kale, 2010)
  - Path-length bounds (Steinhardt and Liang, 2014)
  - Quantile bounds (Koolen and Van Erven, 2015)

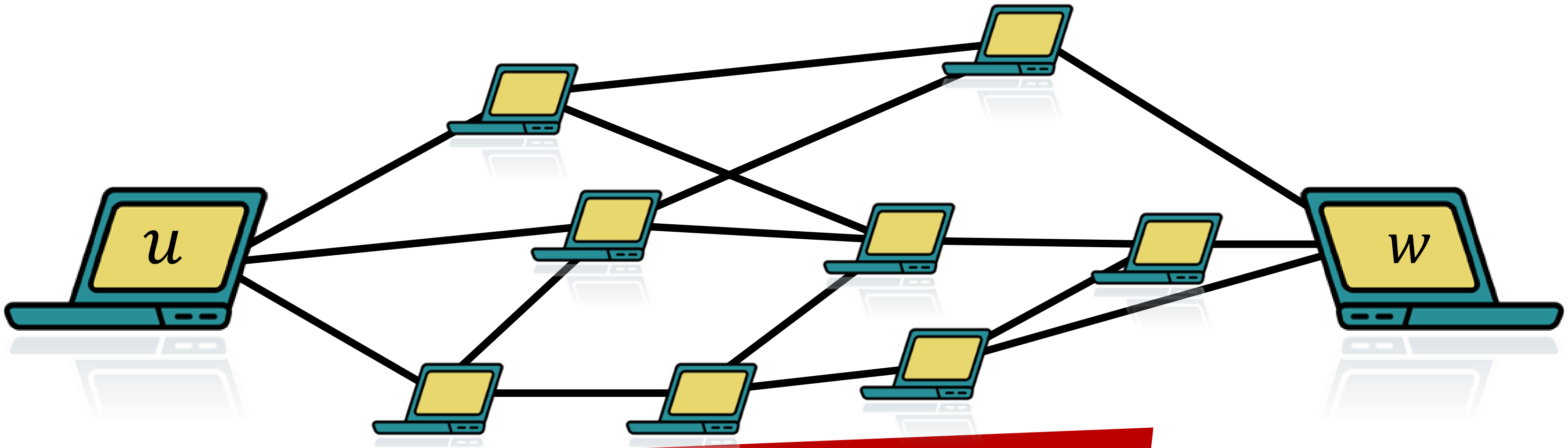
?

Thanks!

Appendix

First-order bounds for combinatorial semi-bandits

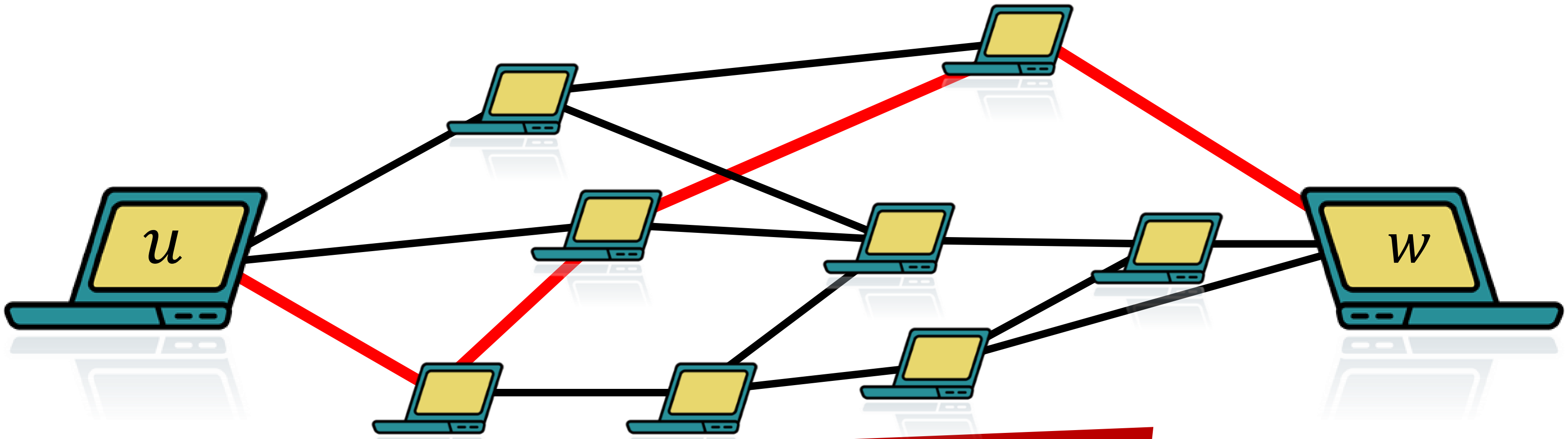
Combinatorial semi-bandits



For every round $t = 1,2, \ldots, T$

- learner picks an action $V_t \in S \subseteq \{0,1\}^d$
- Environment chooses loss vector $\ell_t \in [0,1]^d$
- Learner suffers loss $V_t^\top \ell_t$
- Learner observes losses $V_{t,i} \ell_{t,i}$

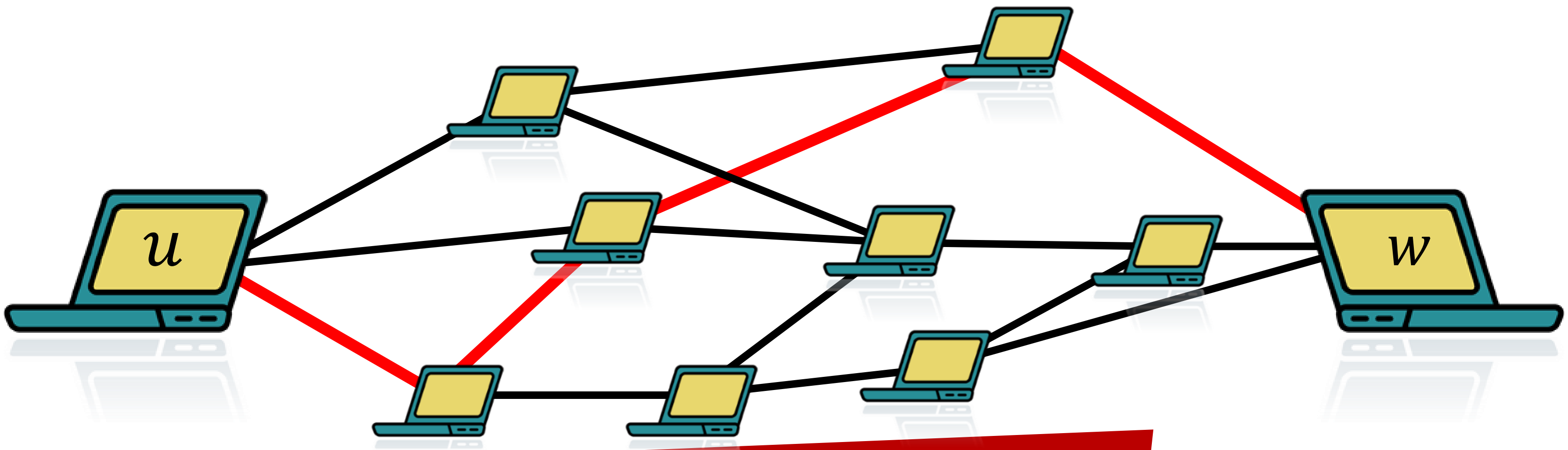
Combinatorial semi-bandits



For every round $t = 1,2,\ldots,T$

- learner picks an action $V_t \in S \subseteq \{0,1\}^d$
- Environment chooses loss vector $\ell_t \in [0,1]^d$
- Learner suffers loss $V_t^\top \ell_t$
- Learner observes losses $V_{t,i}\ell_{t,i}$

Combinatorial semi-bandits



For every round $t = 1,2,\dots,T$

- learner picks an action $V_t \in S \subseteq \{0,1\}^d$
- Environment chooses loss vector $\ell_t \in [0,1]^d$
- Learner suffers loss $V_t^\top \ell_t$
- Learner observes losses $V_{t,i}\ell_{t,i}$

Decision set:

$$S = \{v_i\}_{i=1}^N \subseteq \{0,1\}^d$$
$$\|v_i\|_1 \leq m$$

Combinatorial semi-bandits

- Goal: minimize (expected) regret
$$\hat{R}_T = \max_{v \in S} \mathbf{E}\left[\sum_{t=1}^{T} (V_t - v)^\top \ell_t\right]$$
- Minimax regret is
$$\hat{R}_T = \Theta\left(\sqrt{mdT}\right)$$
- Best efficient algorithm (FPL) gives
$$\hat{R}_T = O\left(m\sqrt{dT \log(d)}\right)$$

Combinatorial semi-bandits

- Goal: minimize (expected) regret
$$\hat{R}_T = \max_{v \in S} \mathbf{E}\left[\sum_{t=1}^{T} (V_t - v)^\top \ell_t\right]$$
- Minimax regret is
$$\hat{R}_T = \Theta\left(\sqrt{mdT}\right)$$
- Best efficient algorithm (FPL) gives
$$\hat{R}_T = O\left(m\sqrt{dT\log(d)}\right)$$
- Our bound:
$$\hat{R}_T = O\left(m\sqrt{dL_T^*\log(d)}\right)$$