Second Order Methods for Bandit Optimization and Control

Jennifer Sun

Praneeth Netrapalli

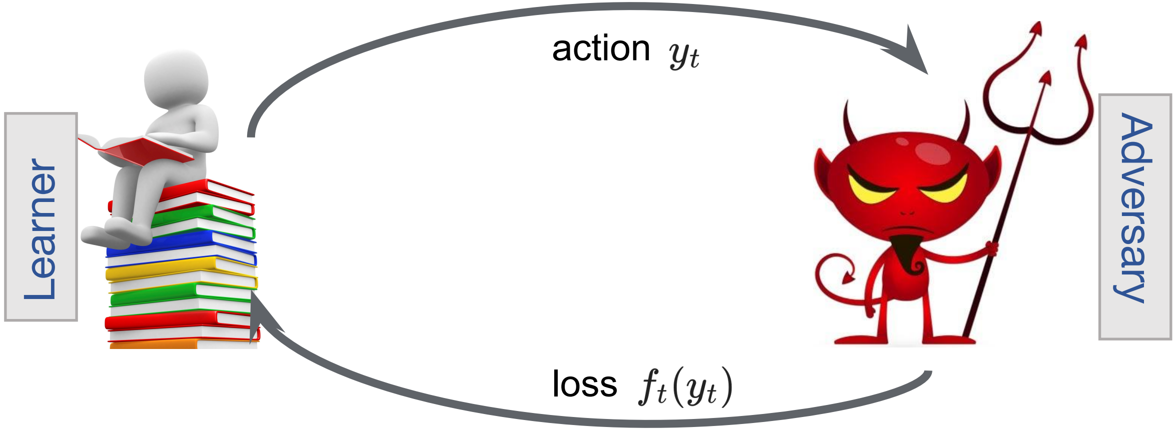
Elad Hazan

Based on work published at COLT 2021, 2024

Outline

Online Learning with Bandit Feedback

Repeated game between learner and an adversary



Goal: minimize regret $\sum_{t=1}^{T} f_t(y_t) - \min_{x \in X} f_t(x)$

Applications

- Two player zero sum games: $\min_{x \in X} \max_{y \in Y} g(x, y)$
  - example: constrained optimization, robust ML
- Online advertising systems
  - An advertiser submits a bid and only observes the reward if they won the auction
- Hyperparameter optimization in ML
  - Tuning learning rate, regularization strength etc..
- Non-stochastic control

Background

- **Gradient based methods**
  - estimate the gradient at point x as $(d/\delta) f(x + \delta u) u$
    - u is a random vector sampled from unit sphere
    - this is an unbiased estimate of gradient of $\mathbb{E}_v[f(x + \delta v)]$
  - perform stochastic gradient descent on the smoothed function

Stoke's theorem

$$\nabla \int_{\delta \mathbb{B}} f(x+v) dv = \int_{\delta \mathbb{S}} f(x+u) \frac{u}{\|u\|} du.$$

Methods	Setting	Regret
[Flaxman et. al' 04]	bounded convex	$O(T^{5/6})$
[Abernethy et. al' 09]	**linear**	$O(T^{1/2})$
[Saha-Tewari ' 11]	smooth convex	$O(T^{2/3})$
[Hazan-Levy' 14]	**strongly convex, smooth**	$O(T^{1/2})$

Background

- **General convex losses:** Regret optimal (in T) algorithms were developed by *Bubeck et al. (2017), Lattimore (2020)*

- *Lattimore (2020)* only show the existence of a strategy, but do not provide any constructive algorithm

- *Bubeck et al. (2017)* develop an extension of *exponential weights algorithm*
    - but the runtime of the algorithm is large, and is not implementable in practice

This work: Develop regret optimal, computationally efficient algorithms for a broad class of loss functions.

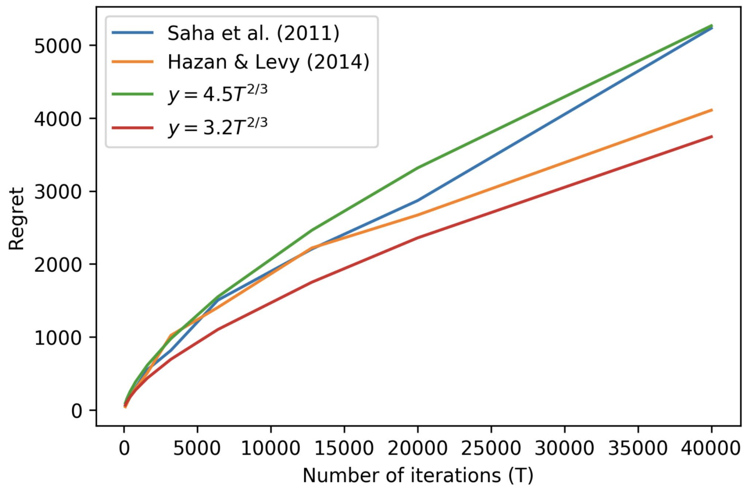
Key Observation

- **Challenge**: achieve the right exploration exploitation trade-off

- Most existing works only estimate gradients of the loss function
  - They ignore curvature information and perform poor exploration

- An ideal algorithm performs:
  - less exploration along high curvature directions
  - more exploration along low curvature directions

This work: estimate higher order information (**Hessian**) of the loss function for better exploration



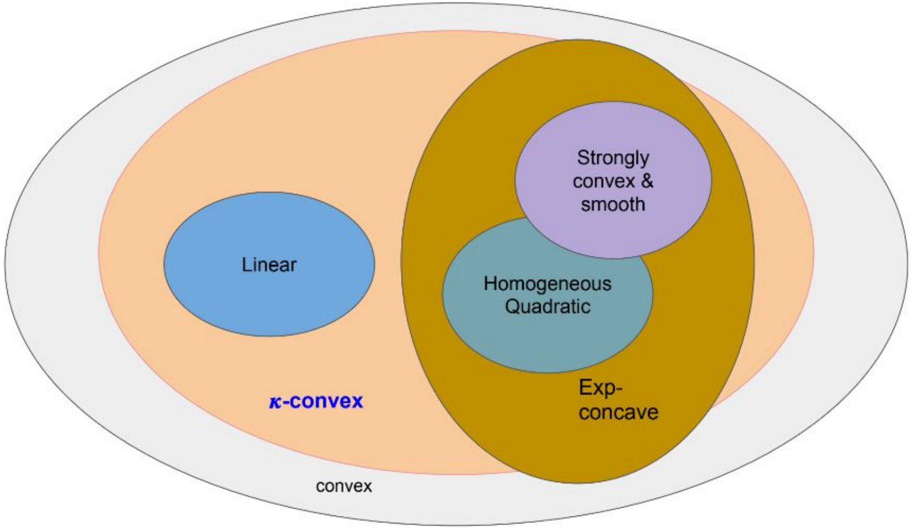
regret of various gradient based techniques on convex quadratic loss function

κ-convex loss functions

Definition: f is called κ-convex if there exist constants c_1, c_2 and a PSD matrix H such that

$$\forall x,\ c_1 H \preceq \nabla^2 f(x) \preceq c_2 H$$

where $\dfrac{c_2}{c_1} \leq \kappa, 0 \preceq H \preceq I$



Examples

- Linear, Quadratic
- Generalized Linear Models: logistic regression
- Strongly convex and Smooth

Main Result

Suppose f_t's are κ-convex and generated by an oblivious adversary

Then there exists an algorithm that achieves $O(d^{5/2}\kappa \min(d^{1/2}, \kappa)T^{1/2})$ regret in expectation

Remarks

- The algorithm is an improper algorithm that plays iterates outside the constraint set
- For bandit logistic regression this gives a regret of $O(e^D\sqrt{T})$
  - where D is the diameter of the parameter space
  - $O(\cdot)$ hides all parameters other than D, T

Online Logistic Regression

Paper	Feedback	Advers.	Proper	Regret	Comp.	Note
Hazan et al. (2007)	full	✓	✓	$\tilde{O}(e^D)$	$O(d^2)$	
Hazan et al. (2014)	full	×	✓	$\tilde{\Omega}(e^D \vee \sqrt{DT})$	–	
Foster et al. (2018)	full	✓	×	$\tilde{O}(1)$	$\mathrm{poly}(d, T)$	
Hazan and Kale (2011)	semi-bandit	✓	✓	$\tilde{O}(e^D \wedge DT^{2/3})$	$O(d^2)$	
Foster et al. (2018)	semi-bandit	✓	×	$\tilde{O}(e^D \wedge \sqrt{T})$	$\mathrm{poly}(d, T)$	
Dong et al. (2019)	bandit	×	✓	$\tilde{O}(\sqrt{T})$	$\mathrm{poly}(d)$	Bayesian
Faury et al. (2022)	bandit	×	✓	$\tilde{O}(e^D \vee \sqrt{T})$	$O(d^2)$	frequentist
Corollary 8	bandit	✓	×	$O(e^{2D}\sqrt{T})$	$O(d^2)$	

Table 4: Comparison with relevant prior works for online logistic regression. $\tilde{O}, \tilde{\Omega}$ in the regret column hide all parameters other than D, T and logarithmic factors in D, T.

Outline

Bandit Newton Step (BNS)

- Randomly sample $v_{t,1}, v_{t,2}$ from surface of a unit sphere
- Compute $y_t = x_t + \tilde{A}_{t-1}^{-1/2}(v_{t,1} + v_{t,2})$
  - $\tilde{A}_t$ is the cumulative Hessian estimate

- Estimate gradients, Hessians from single point feedback

Gradient:	$\tilde{g}_t = 2d f_t(y_t)\tilde{A}_{t-1}^{\frac{1}{2}} v_{t,1}$
Hessian:	$\tilde{H}_t = 2d^2 f_t(y_t)\tilde{A}_{t-1}^{\frac{1}{2}}(v_{t,1}v_{t,2}^\top + v_{t,2}v_{t,1}^\top)\tilde{A}_{t-1}^{\frac{1}{2}}$

- $\tilde{A}_t$ which dictates the exploration, relies on curvature of cumulative loss

$$\tilde{A}_t = I + \frac{\eta}{\kappa'} \sum_{s=1}^{t} \tilde{H}_t$$

Bandit Newton Step (BNS)

- Perform Newton step using the estimated gradients, Hessians

$$x_{t+1} = \prod_{X}^{\tilde{A}_t} \left[x_t - \eta \tilde{A}_t^{-1} \tilde{g}_t \right]$$

  - where $\prod_{X}^{\tilde{A}_t}$ is the projection onto X w.r.t $\|\cdot\|_{\tilde{A}_t}$

Key steps for deriving regret bound

- $\tilde{A}_t$ is a good approximation of the true cumulative Hessian

$$0.5A_t \preceq \tilde{A}_t \preceq 1.5A_t$$

- Expected regret can be decomposed as

$$\mathbb{E}\left[\sum_t f_t(y_t) - f_t(x)\right] = \mathbb{E}\left[\sum_t \tilde{f}_t(y_t) - \tilde{f}_t(x)\right] + \mathbb{E}\left[\sum_t (f_t(y_t) - \tilde{f}_t(y_t)) - (f_t(x) - \tilde{f}_t(x))\right]$$

where $\tilde{f}_t$ is the following smoothing of f_t

$$\tilde{f}_t(x) = \mathbb{E}_{u,v}\left[f_t\left(x + \tilde{A}_{t-1}^{-1/2}(u+v)\right)\right]$$

Key steps for deriving regret bound

$$\mathbb{E}\left[\sum_t f_t(y_t) - f_t(x)\right] = \mathbb{E}\left[\sum_t \tilde{f}_t(y_t) - \tilde{f}_t(x)\right] + \mathbb{E}\left[\sum_t (f_t(y_t) - \tilde{f}_t(y_t)) - (f_t(x) - \tilde{f}_t(x))\right]$$

- **First term:** BNS performs stochastic newton step on $\tilde{f}_t$
    - Reduction to stochastic online Newton method
    - Stochastic Online Newton Step has good regret bounds in **expectation**
- **Second term:** $f_t, \tilde{f}_t$ are close to each other over the entire domain, because of κ-convexity

Question: This only gives us regret bounds in expectation. What about high probability regret guarantees?

Bandit Newton Step : h.p. regret bounds

- Our local quadratic approximation of f_t has high variance at points far away from x_t
  - ***Quadratic approximation***: $f(y_t) + \langle \tilde{g}_t, x - x_t \rangle + (x - x_t)^T \tilde{H}_t (x - x_t)$
  - **variance** scales with $\|x - x_t\|_{A_t}$, where A_t is the cumulative Hessian

- **Focus Region:** Restrict the learner to low variance regions

$$F_t = F_{t-1} \cap \{x : \|x - x_t\|_{A_t} \leq \gamma\}$$

where γ is a constant, $F_0 = X$

Bandit Newton Step : h.p. regret bounds

- Focus region can only guarantee low regret within F_t
  - we want low regret over the entire domain X

- **Restart Condition:** At every iteration, check if the minimizer of the cumulative loss is in the focus region
  - we design a computationally efficient way to test this
  - If the test fails, restart the algorithm (regret so far is negative)!.

$$\sum_{s=1}^{t} \widehat{f}_s(x_s) - \min_{x \in F_t} \sum_{s=1}^{t} \widehat{f}_s(x) \geq -\frac{c}{\eta_1}$$

Main Result

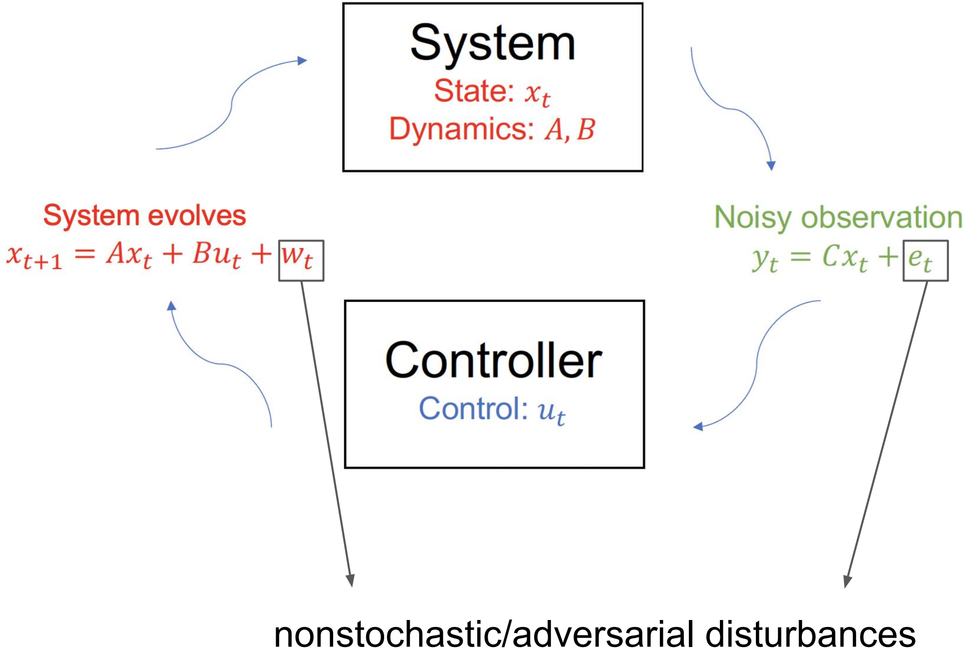
Suppose the sequence of losses are convex, quadratic.

Then BNS with focus region and restart condition gets $O(d^{16}T^{1/2})$ regret with h.p

- The result holds with high probability against adaptive adversaries
- The technique was improved by [Fokkema et al. 2024]

Outline

Application to Bandit Nonstochastic Control



Controller interacts with system and receives **bandit** feedback of convex cost $c_t(y_t, u_t)$ at time t

Goal: minimize regret

$$\sum_{t=1}^{T} c_t(y_t, u_t) - \min_{\pi \in \Pi} \sum_{t=1}^{T} c_t(y_t^\pi, u_t^\pi)$$

Application to Bandit Nonstochastic Control

Assumptions

- **stability:** system is stable
- **oblivious adversary:** perturbations are generated by an oblivious adversary and are bounded
- **quadratic costs:** the cost functions are strongly convex and smooth

- **Past works:** either considered ***full information*** *setting* or placed more ***restrictive*** assumptions on the perturbations

- **This work:** can we derive optimal algorithms for LQ problem under bandit feedback and adversarial perturbations?

DRC Policy Class

Standard technique in nonstochastic control

- Convex comparator class **Disturbance Response Controller (DRC)**
  - Play controls that are linear combinations of past noises/signals.

$$c_t(y_t, u_t)$$

linear in past m controls

$$= K y_t + \sum_{j=0}^{m-1} M^{[j]} y_{t-j}^K$$

signal depending on $\{w_s, e_s\}_{s=1}^{t-j}$

stabilizing controller

$M = M^{[0:m-1]}$ DRC matrix

Q: Identify optimal M

Reduction to Bandit Optimization with Memory

Since our system is stable

- effect of past controls decay over time
- we can reduce the control problem to bandit optimization with **memory** of the following form
  - Loss at time t depends on the past actions m: $f_t(z_t, z_{t-1}, \ldots z_{t-m})$

Goal: minimize regret: $\sum_t f_t(z_t, z_{t-1} \ldots z_{t-m}) - \min_z f_t(z, \ldots z)$

Challenge in bandit nonstochastic control of LQ problems

Cost functions in LQ are strongly convex and smooth:

Can we obtain the optimal $\sqrt{T}$ regret?

Challenge: with adversarial noises, the control cost induced loss function is not guaranteed to be strongly convex

$$c_t(y_t, u_t)$$

$$= K y_t + \sum_{j=0}^{m-1} M^{[j]} y_{t-j}^K$$

adversarial

However, the loss function is always κ-convex with known matrices

Main result in bandit control of LQ problems

Suppose a LDS is stabilizable, c_t's are quadratic (smooth, and strongly-convex) and generated by an oblivious adversary. Suppose the sequence of perturbations $\{w_t, e_t\}$ are bounded and given by an oblivious adversary.

Then there exists an algorithm based on Bandit Newton Step that achieves $\tilde{O}(\sqrt{T})$ regret in expectation w.r.t. the class of DRC controllers.

Outline

Conclusion

- Right exploration is crucial for achieving optimal regret in bandit optimization
- We proposed a Bandit Newton method for optimization of κ-convex losses
  - The algorithm is simple and relies on single point estimate of Hessian. But high probability guarantees still require focus regions and restart conditions
  - **Key insight**: κ-convexity helps us get an optimistic hessian estimate for the entire action space
- Results for bandit LQ control
  - Challenge: adversarial noises can break the strong convexity of loss induced by control costs.
  - However, the induced loss function satisfies κ-convexity with known matrix parameters due to the memory structure of nonstochastic control problems.

Future Work

- ❑ BCO
  - *"Proper"* learning algorithm for bandit optimization of κ-convex losses
  - Extensions to general convex losses
- ❑ Online non-stochastic control
  - Extension to general cost functions

Questions?

Thanks! You can reach out to arunss@google.com