Online Prediction: Rademacher Averages via Burkholder's Method

Sasha Rakhlin

UPenn

Dec 16, 2017

Joint work with D. Foster and K. Sridharan

Outline

Motivation: Online Supervised Learning

Burkholder's Method

prediction paradigm

Data $\longrightarrow$ Estimate Model $\longrightarrow$ Make Prediction

prediction paradigm

Data $\longrightarrow$ Estimate Model $\longrightarrow$ Make Prediction

movies

users

					1		
	5						
			3				
					2		

movies

users

					1		
	5						
			3				
	?						
					2		

movies

users

					1		
	5						
			3				
	4						
					2		

movies

users

					1		
	5					?	
			3				
	4						
					2		

movies

users								
						1		
		5					5	
				3				
		4						
						2		

Example 0: Bit Prediction

Suppose we want to predict a $0/1$ sequence

$$y_1, y_2, \ldots$$

Example 0: Bit Prediction

Suppose we want to predict a $0/1$ sequence

$$y_1, y_2, \ldots$$

If iid Bernoulli, then predicting majority $\mathbf{I}\{\bar{y}_t \geq .5\}$ ensures that proportion of correct predictions $\bar{c}_t$ satisfies

$$\bar{c}_t \rightsquigarrow \max\{p, 1-p\} \approx \max\{\bar{y}_t, 1-\bar{y}_t\}$$

Example 0: Bit Prediction

Suppose we want to predict a $0/1$ sequence

$$y_1, y_2, \ldots$$

If iid Bernoulli, then predicting majority $\mathbf{I}\{\bar{y}_t \geq .5\}$ ensures that proportion of correct predictions $\bar{c}_t$ satisfies

$$\bar{c}_t \;\rightsquigarrow\; \max\{p, 1-p\} \;\approx\; \max\{\bar{y}_t, 1-\bar{y}_t\}$$

More precisely:

$$\liminf_{n\to\infty} \; (\bar{c}_n - \max\{\bar{y}_n, 1-\bar{y}_n\}) \geq 0 \qquad \text{almost surely} \qquad (*)$$

Example 0: Bit Prediction

Suppose we want to predict a $0/1$ sequence

$$y_1, y_2, \ldots$$

If iid Bernoulli, then predicting majority $\mathbf{I}\{\bar{y}_t \geq .5\}$ ensures that proportion of correct predictions $\bar{c}_t$ satisfies

$$\bar{c}_t \quad \rightsquigarrow \quad \max\{p, 1-p\} \quad \approx \quad \max\{\bar{y}_t, 1-\bar{y}_t\}$$

More precisely:

$$\liminf_{n\to\infty} \; (\bar{c}_n - \max\{\bar{y}_n, 1-\bar{y}_n\}) \geq 0 \qquad \text{almost surely} \qquad (*)$$

Claim: *there is a method that ensures* $(*)$ *for an arbitrary sequence.*

Any idea how to do it? Majority will not work. Need randomized strategy.

Minimax vs Bayes prediction

By

D. Blackwell

University of California, Berkeley

Let $x = (x_1, x_2, \ldots)$ be an infinite sequence of 0s and 1s, initially unknown to you. On day $n = 1, 2, \ldots$ you observe $h_n = (x_1, \ldots, x_{n-1})$, the first $n - 1$ terms of the sequence, and must predict x_n. What is a good prediction method, and how well can you do?

A *prediction method* p is just a function that associates with each finite sequence h of 0s and 1s a prediction $p(h) = 0$ or 1, your prediction of the next x when you have observed history h. Denote by $w_n(p, x)$ the proportion of correct predictions that method p makes against sequence x in the first n days.

Looking for a p th
is an x for which
struct x sequentiall

Then choose x_2 so

etc.

But we can impro
toss a fair coin everyday, predicting 1 on heads, 0 on tails, the strong law of large numbers guarantees that, for every x, the proportion of correct predictions will approach 50% as $n \to \infty$, with probability 1. A *random prediction method* is a func-

The contrast between the minimax predictor p_0 and the Bayes predictor y is strong. The minimax predictor is not obvious, it is randomized, it satisfies (*) for every x, but the proof is not easy. The Bayes predictor is extremely obvious, it is not randomized, it satisfies (*) only for almost all x, and the proof is simple.

Online Supervised Learning

For $t = 1, \dots, n$

observe side info $x_t \in \mathcal{X}$

predict $\widehat{y}_t$

observe outcome y_t

Online Supervised Learning

For $t = 1, \ldots, n$
observe side info $x_t \in \mathcal{X}$
predict $\widehat{y}_t$
observe outcome y_t

Goal:

$$\forall (x_t, y_t)_{t=1}^n, \quad \sum_{t=1}^n |\widehat{y}_t - y_t| \quad \leq \qquad \text{small if sequence is “nice”}$$

Online Supervised Learning

For $t = 1, \ldots, n$

observe side info $x_t \in \mathcal{X}$

predict $\widehat{y}_t$

observe outcome y_t

Goal:

$$\forall (x_t, y_t)_{t=1}^n, \quad \sum_{t=1}^n |\widehat{y}_t - y_t| \quad \le \quad \phi(x_1, y_1, \ldots, x_n, y_n)$$

Online Supervised Learning

For $t = 1, \ldots, n$

observe side info $x_t \in \mathcal{X}$

predict $\widehat{y}_t$

observe outcome y_t

Goal:

$$\forall (x_t, y_t)_{t=1}^n, \quad \sum_{t=1}^n |\widehat{y}_t - y_t| \quad \le \quad \inf_{\mathbf{w} \in \mathcal{F}} \sum_{t=1}^n |\langle \mathbf{w}, x_t \rangle - y_t| + C_n(x_1, \ldots, x_n).$$

Online Supervised Learning

For $t = 1, \ldots, n$
  observe side info $x_t \in \mathcal{X}$
  predict $\widehat{y}_t$
  observe outcome y_t

Goal:

$$\forall (x_t, y_t)_{t=1}^{n}, \quad \sum_{t=1}^{n} |\widehat{y}_t - y_t| \leq \inf_{\mathbf{w} \in \mathcal{F}} \sum_{t=1}^{n} |\langle \mathbf{w}, x_t \rangle - y_t| + C_n(x_1, \ldots, x_n).$$

If $C_n(x_1, \ldots, x_n) = C_n$ is data-independent, we have full characterization.

Best possible: Empirical Rademacher

$$C_n(x_1, \ldots, x_n) \sim \mathbb{E}_\epsilon \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|$$

Example: Matrix Completion.

Motivation: Online Supervised Learning

Proper learning: $\widehat{y}_t = \langle \mathbf{w}_t, x_t \rangle$,

$$\sum_{t=1}^{n} \underbrace{|\widehat{y}_t - y_t|}_{f_t(\mathbf{w}_t)} \leq \inf_{\mathbf{w} \in \mathcal{F}} \sum_{t=1}^{n} \underbrace{|\langle \mathbf{w}, x_t \rangle - y_t|}_{f_t(\mathbf{w})} + C_n(x_1, \ldots, x_n)$$

Motivation: Online Supervised Learning

Proper learning: $\widehat{y}_t = \langle \mathbf{w}_t, x_t \rangle$,

$$\sum_{t=1}^{n} \underbrace{|\widehat{y}_t - y_t|}_{f_t(\mathbf{w}_t)} \le \inf_{\mathbf{w} \in \mathcal{F}} \sum_{t=1}^{n} \underbrace{|\langle \mathbf{w}, x_t \rangle - y_t|}_{f_t(\mathbf{w})} + C_n(x_1, \ldots, x_n)$$

Gradient Descent:

$$w_{t+1} = w_t - \eta \nabla f_t(w_t)$$

Mirror Descent: different geometry

$$\nabla R(w_{t+1}) = \nabla R(w_t) - \eta \nabla f_t(w_t)$$

(e.g. Exponential Weights Algorithm: $R(w) = \sum w(i) \log w(i)$ is strongly convex w.r.t. ℓ_1 norm on simplex)

Motivation: Online Supervised Learning

Gradient/mirror descent with adaptive step size:

$$C_n \propto \sqrt{\sum_{t=1}^{n} \|\nabla f_t(\mathbf{w}_t)\|^2} = \sqrt{\sum_{t=1}^{n} \|x_t\|^2}$$

Can be much worse than

$$\mathbb{E}_\epsilon \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\| .$$

Is this easy to fix?

Key issue: GD is not keeping the right *statistics* about the sequence

- Need additional information about geometry of functions: not just the size of gradients but also their "spread"
- Beyond usual notions of smoothness and strong convexity?

Key issue: GD is not keeping the right *statistics* about the sequence

- Need additional information about geometry of functions: not just the size of gradients but also their "spread"
- Beyond usual notions of smoothness and strong convexity?

We would be searching in the dark if not for the connections:

$$\text{Online Prediction} \quad \overset{\text{minimax}}{\longleftrightarrow} \quad \text{Martingale Inequalities} \quad \overset{\text{Burkholder}}{\longleftrightarrow} \quad \text{Geometry}$$

Key issue: GD is not keeping the right *statistics* about the sequence

- Need additional information about geometry of functions: not just the size of gradients but also their "spread"
- Beyond usual notions of smoothness and strong convexity?

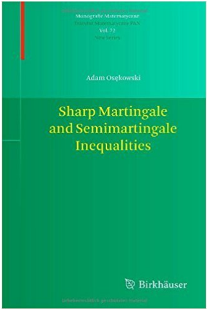
We would be searching in the dark if not for the connections:

Online Prediction $\xleftrightarrow{\text{minimax}}$ Martingale Inequalities $\xleftrightarrow{\text{Burkholder}}$ Geometry

Outline

Reference:

Adam Osekowski, *Sharp Martingale and Semimartingale Inequalities*, 2012



Next: adaptation of some of these ideas to our setting.

$\epsilon_1, \ldots, \epsilon_t, \ldots$ i.i.d. Rademacher, $\mathcal{F}_t = \sigma(\epsilon_1, \ldots, \epsilon_t)$.

$X_1, \ldots, X_t, \ldots$ martingale difference sequence w.r.t. $(\mathcal{F}_t)$.

$$\mathbb{E}[X_t | \mathcal{F}_{t-1}] = 0$$

Note: X_t can be written as $\epsilon_t \mathbf{x}_t(\epsilon_{1:t-1})$ for some function $\mathbf{x}_t$.

A martingale inequality can be generically written as

$$\mathbb{E}B(X_1, \ldots, X_n) \leq 0 \tag{1}$$

for some $B : \cup\mathcal{X}^n \to \mathbb{R}$.

A martingale inequality can be generically written as

$$\mathbb{E}B(X_1, \ldots, X_n) \le 0 \qquad (1)$$

for some $B : \cup \mathcal{X}^n \to \mathbb{R}$.

(FRS'17+): (1) holds $\forall$ martingale difference sequences (X_t) and all n

iff

there exists a function $\mathbf{U} : \cup \mathcal{X}^n \to \mathbb{R}$ satisfying three properties:

A martingale inequality can be generically written as

$$\mathbb{E}B(X_1, \ldots, X_n) \leq 0 \tag{1}$$

for some $B : \cup\mathcal{X}^n \to \mathbb{R}$.

(FRS'17+): (1) holds $\forall$ martingale difference sequences (X_t) and all n

iff

there exists a function $U : \cup\mathcal{X}^n \to \mathbb{R}$ satisfying three properties:

1°. $U(x_1, \ldots, x_t) \geq B(x_1, \ldots, x_t)$ for all t

2°. $U(\cdot) \leq 0$

3°. for all $x_1, \ldots, x_t$

$$\mathbb{E}_\epsilon U(x_1, \ldots, x_{t-1}, \epsilon x_t) \leq U(x_1, \ldots, x_{t-1})$$

proof

$(\Leftarrow)$:

$$\mathbb{E}B(X_1,\ldots,X_n) \overset{1^\circ}{\le} \mathbb{E}U(X_1,\ldots,X_n) \overset{3^\circ}{\le} \mathbb{E}U(X_1,\ldots,X_{n-1}) \ldots \le U(\cdot) \overset{2^\circ}{\le} 0.$$

proof

$(\Leftarrow)$:

$$\mathbb{E}B(X_1,\ldots,X_n) \overset{1^\circ}{\leq} \mathbb{E}U(X_1,\ldots,X_n) \overset{3^\circ}{\leq} \mathbb{E}U(X_1,\ldots,X_{n-1})\ldots \leq U(\cdot) \overset{2^\circ}{\leq} 0.$$

$(\Rightarrow)$: Define

$$U^*(x_1,\ldots,x_t) \triangleq \sup_{n\geq t,(X)_{t+1}^n} \mathbb{E}B(x_1,\ldots,x_t,X_{t+1},\ldots,X_n).$$

Claim: U^* satisfies $1^\circ - 3^\circ$.

Claim: U^* is the smallest function that satisfies $1° - 3°$.

Proof:

For any $n \geq t, (X)_{t+1}^{n}$,

$$\mathbb{E}B(x_1, \ldots, x_t, X_{t+1}, \ldots, X_n) \leq \mathbb{E}U'(x_1, \ldots, x_t, X_{t+1}, \ldots, X_n) \leq U'(x_1, \ldots, x_t)$$

Hence, $U^*(x_1, \ldots, x_t) \leq U'(x_1, \ldots, x_t)$.

Example 1: Smoothness/Strong Convexity

Let $\|\cdot\|$ be some norm on $\mathcal{X}$. Martingale inequality corresponding to smoothness:

$$\mathbb{E}\left\|\sum_{t=1}^{n} X_t\right\|^2 \leq K \cdot \mathbb{E}\sum_{t=1}^{n} \|X_t\|^2$$

Hence,

$$B(x_1, \ldots, x_n) = B\left(\sum_{t=1}^{n} x_t, \sum_{t=1}^{n} \|x_t\|^2\right) = \left\|\sum_{t=1}^{n} x_t\right\|^2 - K\sum_{t=1}^{n} \|x_t\|^2$$

Example 1: Smoothness/Strong Convexity

Let $\|\cdot\|$ be some norm on $\mathcal{X}$. Martingale inequality corresponding to smoothness:

$$\mathbb{E}\left\|\sum_{t=1}^{n} X_t\right\|^2 \leq K \cdot \mathbb{E}\sum_{t=1}^{n} \|X_t\|^2$$

Hence,

$$B(x_1, \ldots, x_n) = B\left(\sum_{t=1}^{n} x_t, \sum_{t=1}^{n} \|x_t\|^2\right) = \left\|\sum_{t=1}^{n} x_t\right\|^2 - K\sum_{t=1}^{n} \|x_t\|^2$$

Optimal U^* inherits

$$U^*(x, \alpha^2) = U^*(x, 0) - K\alpha^2.$$

Example 1: Smoothness/Strong Convexity

Let $\|\cdot\|$ be some norm on $\mathcal{X}$. Martingale inequality corresponding to smoothness:

$$\mathbb{E}\left\|\sum_{t=1}^{n} X_t\right\|^2 \leq K \cdot \mathbb{E}\sum_{t=1}^{n}\|X_t\|^2$$

Hence,

$$B(x_1,\ldots,x_n) = B\left(\sum_{t=1}^{n} x_t, \sum_{t=1}^{n}\|x_t\|^2\right) = \left\|\sum_{t=1}^{n} x_t\right\|^2 - K\sum_{t=1}^{n}\|x_t\|^2$$

Optimal U^* inherits

$$U^*(x,\alpha^2) = U^*(x,0) - K\alpha^2.$$

Restricted concavity $3°$ is

$$\mathbb{E}_\epsilon U^*(x+\epsilon y, \alpha^2 + \|y\|^2) \leq U^*(x,\alpha^2).$$

Example 1: Smoothness/Strong Convexity

Let $\|\cdot\|$ be some norm on $\mathcal{X}$. Martingale inequality corresponding to smoothness:

$$\mathbb{E}\left\|\sum_{t=1}^{n} X_t\right\|^2 \leq K \cdot \mathbb{E}\sum_{t=1}^{n}\|X_t\|^2$$

Hence,

$$B(x_1,\ldots,x_n) = B\left(\sum_{t=1}^{n} x_t, \sum_{t=1}^{n}\|x_t\|^2\right) = \left\|\sum_{t=1}^{n} x_t\right\|^2 - K\sum_{t=1}^{n}\|x_t\|^2$$

Optimal U^* inherits

$$U^*(x,\alpha^2) = U^*(x,0) - K\alpha^2.$$

Restricted concavity 3° is

$$\mathbb{E}_\epsilon U^*(x+\epsilon y, \alpha^2 + \|y\|^2) \leq U^*(x,\alpha^2).$$

Corollary: $x \mapsto U^*(x,0)$ is smooth wrt $\|\cdot\|$ (and its dual is strongly cvx).

Proof:

$$
\begin{aligned}
\Phi(x) &= U^*(x,0) \\
&\geq \mathbb{E}_\epsilon U^*(x+\epsilon y, \|y\|^2) \\
&= \frac{1}{2}(U^*(x+y,0) - K\|y\|^2) + \frac{1}{2}(U^*(x-y,0) - K\|y\|^2) \\
&= \frac{1}{2}\Phi(x+y) + \frac{1}{2}\Phi(x-y) - K\|y\|^2
\end{aligned}
$$

Example 2

Hilbert space:

$$\mathbb{E}\left\|\sum_{t=1}^{n} X_t\right\| \le 2\mathbb{E}\sqrt{\sum_{t=1}^{n} \|X_t\|^2}$$

Hence,

$$B(x, a) = \|x\| - 2y$$

and interested in

$$B\left(\sum_{t=1}^{n} x_t, \sqrt{\sum_{t=1}^{n} \|x_t\|^2}\right)$$

Example 2

Hilbert space:

$$\mathbb{E}\left\|\sum_{t=1}^{n} X_t\right\| \leq 2\mathbb{E}\sqrt{\sum_{t=1}^{n} \|X_t\|^2}$$

Hence,

$$B(x, a) = \|x\| - 2y$$

and interested in

$$B\left(\sum_{t=1}^{n} x_t, \sqrt{\sum_{t=1}^{n} \|x_t\|^2}\right)$$

Elementary algebra gives

$$U(x, a) = \begin{cases} -\sqrt{2a^2 - \|x\|^2}, & a \geq \|x\| \\ \|x\| - 2a, & a < \|x\| \end{cases}$$

satisfies concavity property

$$U(x + d, \sqrt{a^2 + d^2}) \leq U(x, a) + U_x(x, a)d$$

Example 2

Hilbert space:

$$\mathbb{E}\left\|\sum_{t=1}^{n} X_t\right\| \le 2\mathbb{E}\sqrt{\sum_{t=1}^{n}\|X_t\|^2}$$

Hence,

$$B(x,a) = \|x\| - 2y$$

and interested in

$$B\left(\sum_{t=1}^{n} x_t, \sqrt{\sum_{t=1}^{n}\|x_t\|^2}\right)$$

Elementary algebra gives

$$U(x,a) = \begin{cases} -\sqrt{2a^2 - \|x\|^2}, & a \ge \|x\| \\ \|x\| - 2a, & a < \|x\| \end{cases}$$

satisfies concavity property

$$U(x+d, \sqrt{a^2+d^2}) \le U(x,a) + \frac{x \cdot d}{\sqrt{2a^2 - \|x\|^2}}$$

Example 3: Empirical Rademacher

(FRS'16): Empirical Rademacher bound for Online Supervised Learning is possible if (and close to "iff")

$$\mathbb{E}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t \mathbf{x}_t \right\|^p \leq K \mathbb{E}_{\epsilon,\epsilon'} \left\| \sum_{t=1}^{n} \epsilon'_t \mathbf{x}_t \right\|^p$$

where $\mathbf{x}_t$ is $\mathcal{F}_{t-1}$-measurable, $p \geq 1$. *Decoupling inequality.*

Example 3: Empirical Rademacher

(FRS'16): Empirical Rademacher bound for Online Supervised Learning is possible if (and close to "iff")

$$\mathbb{E}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t \mathbf{x}_t \right\|^p \leq K \mathbb{E}_{\epsilon, \epsilon'} \left\| \sum_{t=1}^{n} \epsilon'_t \mathbf{x}_t \right\|^p$$

where $\mathbf{x}_t$ is $\mathcal{F}_{t-1}$-measurable, $p \geq 1$. *Decoupling inequality.*

Two-sided version of above is equivalent to *deterministic UMD*

$$\forall \sigma_1, \ldots, \sigma_n \in \{\pm 1\}, \quad \mathbb{E} \left\| \sum_{t=1}^{n} \sigma_t X_t \right\|^p \leq K \mathbb{E} \left\| \sum_{t=1}^{n} X_t \right\|^p$$

Example 3: Empirical Rademacher

(FRS'16): Empirical Rademacher bound for Online Supervised Learning is possible if (and close to "iff")

$$\mathbb{E}_{\epsilon} \left\| \sum_{t=1}^{n} \epsilon_t \mathbf{x}_t \right\|^p \leq K \mathbb{E}_{\epsilon,\epsilon'} \left\| \sum_{t=1}^{n} \epsilon'_t \mathbf{x}_t \right\|^p$$

where $\mathbf{x}_t$ is $\mathcal{F}_{t-1}$-measurable, $p \geq 1$. *Decoupling inequality.*

Two-sided version of above is equivalent to *deterministic UMD*

$$\forall \sigma_1, \ldots, \sigma_n \in \{\pm 1\}, \quad \mathbb{E} \left\| \sum_{t=1}^{n} \sigma_t X_t \right\|^p \leq K \mathbb{E} \left\| \sum_{t=1}^{n} X_t \right\|^p$$

This gives

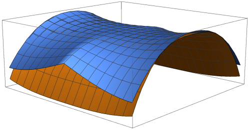
$$B(x_1, \ldots, x_n) = B\left(\sum_{t=1}^{n} x_t, \sum_{t=1}^{n} \sigma_t x_t \right) = \left\| \sum_{t=1}^{n} \sigma_t x_t \right\|^p - K \left\| \sum_{t=1}^{n} x_t \right\|^p$$

Example 3: Empirical Rademacher

Then U satisfying $1^\circ - 3^\circ$ is called Burkholder function. Property 3° reads

$$\mathbb{E}_\epsilon U(x + \epsilon z, y \pm \epsilon z) \le U(x, y)$$

which is equivalent to *zigzag concavity*.



We can now go back and use U to derive an algorithm for Online Supervised Learning with Empirical Rademacher regret bound.

Back to Online Supervised Learning

Enough to solve linearized problem

$$\sum_{t=1}^{n} \ell'_t \cdot \widehat{y}_t \leq \min_{w \in \mathcal{F}} \sum_{t=1}^{n} \ell'_t \cdot \langle w, x_t \rangle + C_n(x_1, \ldots, x_n)$$

Back to Online Supervised Learning

Enough to solve linearized problem

$$\sum_{t=1}^{n} \ell'_t \cdot \widehat{y}_t \le \min_{w\in\mathcal{F}} \sum_{t=1}^{n} \ell'_t \cdot \langle w, x_t\rangle + C_n(x_1,\ldots,x_n)$$

which can be written as

$$\sum_{t=1}^{n} \ell'_t \cdot \widehat{y}_t + \underbrace{\left\|\sum_{t=1}^{n} \ell'_t x_t\right\| - C\mathbb{E}_\epsilon \left\|\sum_{t=1}^{n} \epsilon_t \ell'_t x_t\right\|}_{\le \mathbb{E}_\epsilon U\left(\sum_{t=1}^{n} \ell'_t x_t, \sum_{t=1}^{n} \epsilon_t \ell'_t x_t\right)} \le 0$$

Last step:

$$\min_{\widehat{y}_n} \max_{\ell'_n \in [-1,1]} \left\{ \ell'_n \cdot \widehat{y}_n + \mathbb{E}_\epsilon U \left(\sum_{t=1}^{n} \ell'_t x_t, \sum_{t=1}^{n} \epsilon_t \ell'_t x_t \right) \right\}$$

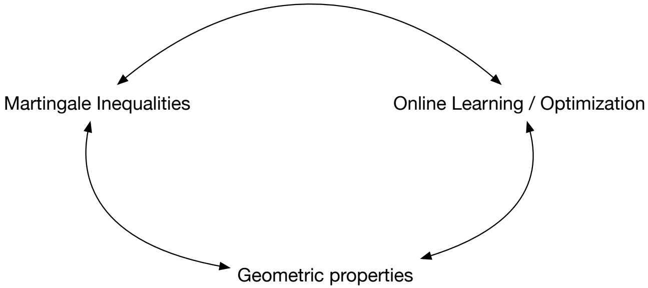
Choosing $\widehat{y}_n = -G'(0)$ for

$$G(\alpha) = \mathbb{E}_\sigma U \left(\sum_{t=1}^{n-1} \ell'_t x_t + \alpha x_t, \sum_{t=1}^{n-1} \epsilon_t \ell'_t x_t + \sigma \alpha x_t \right)$$

ensures

$$-\ell'_n \cdot G'(0) + G(\ell_n) \leq G(0)$$

by diagonal concavity and yields clean recursion.


Martingale Inequalities
Online Learning / Optimization
Geometric properties

Conclusions

- Gradient/Mirror Descent does not keep the right "statistics" about the sequence.
- Strong convexity/smoothness is not enough as a geometric primitive.
- Can find the right primitive by exploiting connections between probabilistic inequalities and geometry.