Efficient Learning with Smoothed Data

Bridging Statistical and Online Learning

Adam Block and Abhishek Shetty

COLT 2025

Key Takeaways

Smoothed data bridges efficiency of statistical learning and robustness of online learning.

Key Takeaways

Smoothed data bridges efficiency of statistical learning and robustness of online learning.

Technical tools:

Key Takeaways

Smoothed data bridges efficiency of statistical learning and robustness of online learning.

Technical tools:

(i) Surprise Lemma (compactness)

Key Takeaways

Smoothed data bridges efficiency of statistical learning and robustness of online learning.

Technical tools:

(i) Surprise Lemma (compactness)

(ii) Coupling (rejection sampling)

Tutorial Outline

Part I

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications

2. The Smoothed Model: Best of Both Worlds?

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications
2. The Smoothed Model: Best of Both Worlds?
3. **The Power of Empirical Risk Minimization**

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications
2. The Smoothed Model: Best of Both Worlds?
3. The Power of Empirical Risk Minimization

Part II

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications
2. The Smoothed Model: Best of Both Worlds?
3. The Power of Empirical Risk Minimization

Part II

1. Coupling Lemma

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications
2. The Smoothed Model: Best of Both Worlds?
3. The Power of Empirical Risk Minimization

Part II

1. Coupling Lemma
2. Handling Label Noise: The Agnostic Setting

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications
2. The Smoothed Model: Best of Both Worlds?
3. The Power of Empirical Risk Minimization

Part II

1. Coupling Lemma
2. Handling Label Noise: The Agnostic Setting
3. Oracle Efficiency: ERM Returns

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications

   (a) What constitutes success in understanding learning?

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications

   (a) What constitutes success in understanding learning?

   (b) Online Learning: What is it and why study it?

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications

   (a) What constitutes success in understanding learning?

   (b) Online Learning: What is it and why study it?

   (c) Why is Online Learning not enough?

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications

   (a) What constitutes success in understanding learning?

   (b) Online Learning: What is it and why study it?

   (c) Why is Online Learning not enough?

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$.

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$.

Mean zero

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$.

Mean zero

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$.

Mean zero

Model class can be neural networks, decision trees, etc.

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$.

Mean zero

Goal: Return $\hat{f} \in \mathcal{F}$ with small test loss $\mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T} \ell(\hat{f}(X'_s), f^\star(X'_s))\right]$.

Model class can be neural networks, decision trees, etc.

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Mean zero

Goal: Return $\hat{f} \in \mathscr{F}$ with small test loss $\mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T} \ell(\hat{f}(X'_s), f^\star(X'_s))\right]$.

Model class can be neural networks, decision trees, etc.

Think of $\ell(y, y') = (y - y')^2$.

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Mean zero

Goal: Return $\hat{f} \in \mathscr{F}$ with small test loss $\mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T} \ell(\hat{f}(X'_s), f^\star(X'_s))\right]$.

Model class can be neural networks, decision trees, etc.

Think of $\ell(y, y') = (y - y')^2$.

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: Return $\hat{f} \in \mathscr{F}$ with small **test loss** $\mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T} \ell(\hat{f}(X'_s), f^\star(X'_s))\right]$.

Algorithmic

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: Return $\hat{f} \in \mathscr{F}$ with small **test loss** $\mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T} \ell(\hat{f}(X'_s), f^\star(X'_s))\right]$.

Algorithmic

How should we choose $\hat{f}$?

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$.

Goal: Return $\hat{f} \in \mathcal{F}$ with small **test loss** $\mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T} \ell(\hat{f}(X'_s), f^\star(X'_s))\right]$.

Algorithmic

How should we choose $\hat{f}$?

What computation do we need?

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: Return $\hat{f} \in \mathscr{F}$ with small **test loss** $\mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T} \ell(\hat{f}(X'_s), f^\star(X'_s))\right]$.

Algorithmic

How should we choose $\hat{f}$?

What computation do we need?

Epistemic

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: Return $\hat{f} \in \mathscr{F}$ with small **test loss** $\mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T} \ell(\hat{f}(X'_s), f^\star(X'_s))\right]$.

Algorithmic

How should we choose $\hat{f}$?

What computation do we need?

Epistemic

How much data?

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: Return $\hat{f} \in \mathscr{F}$ with small **test loss** $\mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T} \ell(\hat{f}(X'_s), f^\star(X'_s))\right]$.

Algorithmic

How should we choose $\hat{f}$?

What computation do we need?

Epistemic

How much data?

What makes a problem hard?

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: Return $\hat{f} \in \mathscr{F}$ with small **test loss** $\mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T} \ell(\hat{f}(X'_s), f^\star(X'_s))\right]$.

Algorithmic

How should we choose $\hat{f}$?

What computation do we need?

Epistemic

How much data?

What makes a problem hard?

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$.

Goal: Return $\hat{f} \in \mathcal{F}$ with small **test loss** $\mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T} \ell(\hat{f}(X'_s), f^\star(X'_s))\right]$.

Empirical Risk Minimization

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} L_T(f)$$

$$L_T(f) = \frac{1}{T}\sum_{t=1}^{T} \ell(f(X_t), Y_t)$$

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: Return $\hat{f} \in \mathscr{F}$ with small **test loss** $\mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T} \ell(\hat{f}(X'_s), f^\star(X'_s))\right]$.

Algorithmic

How should we choose $\hat{f}$?

What computation do we need?

Epistemic

How much data?

What makes a problem hard?

Performance of ERM

Theorem [KP'00, KP'01, BBL'02]: If ℓ is Lipschitz and $\hat{f}$ is an ERM, then

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\mathscr{G}_T(\mathscr{F})}{T} \lesssim \sqrt{\frac{\mathsf{vc}(\mathscr{F})}{T}}.$$

Performance of ERM

Theorem [KP'00, KP'01, BBL'02]: If ℓ is Lipschitz and $\hat{f}$ is an ERM, then

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\mathscr{G}_T(\mathscr{F})}{T} \lesssim \sqrt{\frac{\mathrm{VC}(\mathscr{F})}{T}}.$$

Definition: Gaussian complexity $\mathscr{G}_T(\mathscr{F}) = \mathbb{E}\left[\sup_{f\in\mathscr{F}} \sum_{t=1}^{T} \xi_t \cdot f(X_t)\right]$.

Performance of ERM

Theorem [KP'00, KP'01, BBL'02]: If ℓ is Lipschitz and $\hat{f}$ is an ERM, then

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\mathscr{G}_T(\mathscr{F})}{T} \lesssim \sqrt{\frac{\mathsf{VC}(\mathscr{F})}{T}}.$$

Definition: Gaussian complexity $\mathscr{G}_T(\mathscr{F}) = \mathbb{E}\left[\sup_{f \in \mathscr{F}} \sum_{t=1}^{T} \xi_t \cdot f(X_t)\right]$.

Definition: VC dimension is size of largest set shattered by $\mathscr{F}$.

Performance of ERM (Square Loss)

Theorem [BBM'02, LRS'15]: If ℓ is square loss and $\hat{f}$ is an ERM, then

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\mathsf{vc}(\mathscr{F})}{T}.$$

Performance of ERM (Square Loss)

Theorem [BBM'02, LRS'15]: If ℓ is square loss and $\hat{f}$ is an ERM, then

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\mathrm{vc}(\mathscr{F})}{T}.$$

Theorem [KP'00, KP'01, BBL'02]: If ℓ is Lipschitz and $\hat{f}$ is an ERM, then

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\mathscr{G}_T(\mathscr{F})}{T} \lesssim \sqrt{\frac{\mathrm{vc}(\mathscr{F})}{T}}.$$

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: Return $\hat{f} \in \mathscr{F}$ with small **test loss** $\mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T} \ell(\hat{f}(X'_s), f^\star(X'_s))\right]$.

Algorithmic

How should we choose $\hat{f}$?

What computation do we need?

Epistemic

How much data?

What makes a problem hard?

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications

   (a) What constitutes success in understanding learning?

   (b) Online Learning: What is it and why study it?

   (c) Why is Online Learning not enough?

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: Return $\hat{f} \in \mathscr{F}$ with small **test loss** $\mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T} \ell(\hat{f}(X'_s), f^\star(X'_s))\right]$.

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

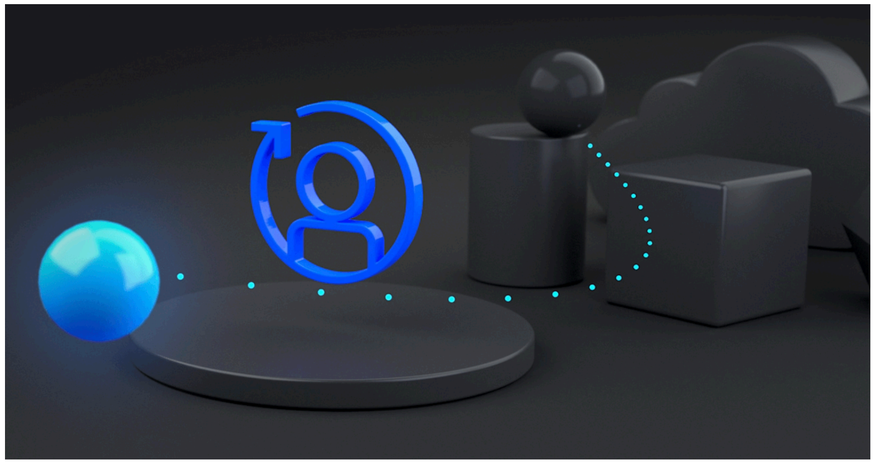
Goal: Return $\hat{f} \in \mathscr{F}$ with small **test loss** $\mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T} \ell(\hat{f}(X'_s), f^\star(X'_s))\right]$.

Is the independence assumption too strong?

Many Problems are Sequential and Adaptive

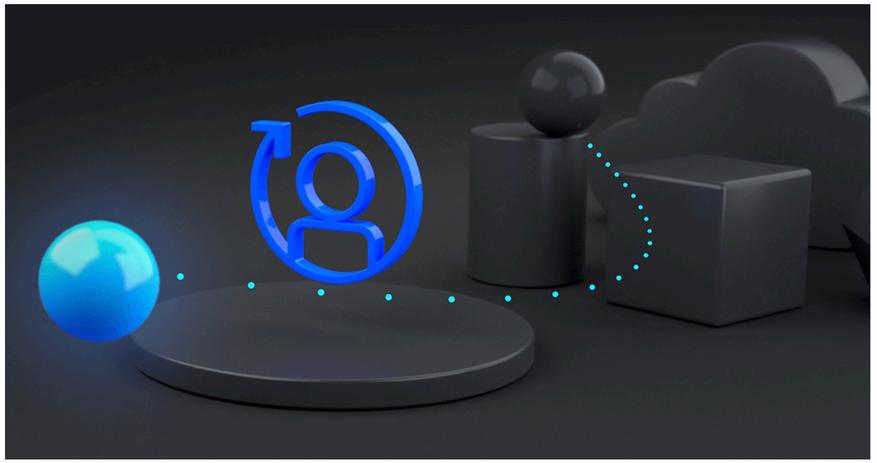
Sora, Shazeer et al ’17, ChatGPT, Chi et al ‘24

Many Problems are Sequential and Adaptive



Sora, Shazeer et al ’17, ChatGPT, Chi et al ‘24

Many Problems are Sequential and Adaptive



You

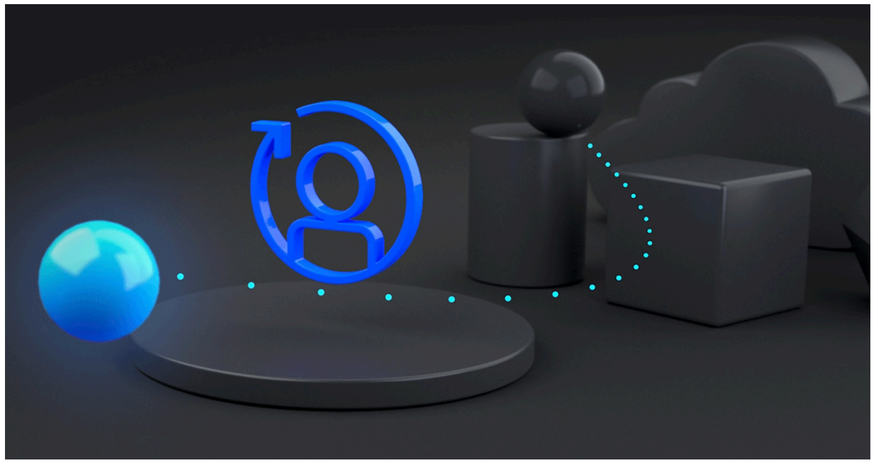
Write one short paragraph that simply explains how you generate text

ChatGPT

I generate text using a machine learning model called GPT (Generative Pre-trained Transformer), which has been trained on a vast amount of text data. Using its understanding of language patterns and semantics, GPT predicts the most likely next words based on the context it's given, producing coherent and contextually relevant text through a process of sampling from probability distributions.

Sora, Shazeer et al ’17, ChatGPT, Chi et al ‘24

Many Problems are Sequential and Adaptive

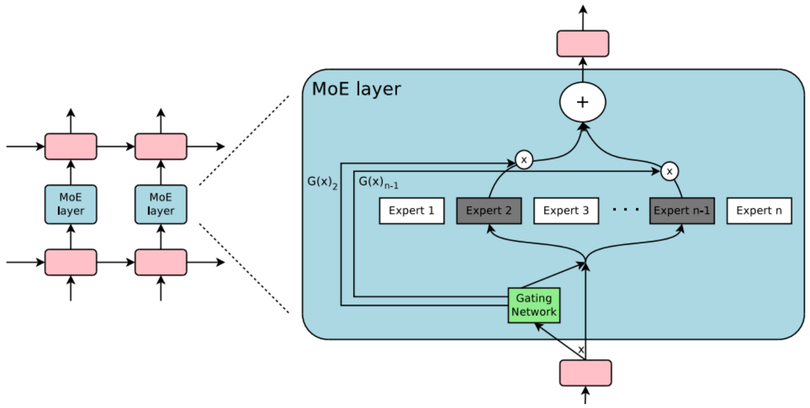


You
Write one short paragraph that simply explains how you generate text

ChatGPT
I generate text using a machine learning model called GPT (Generative Pre-trained Transformer), which has been trained on a vast amount of text data. Using its understanding of language patterns and semantics, GPT predicts the most likely next words based on the context it's given, producing coherent and contextually relevant text through a process of sampling from probability distributions.

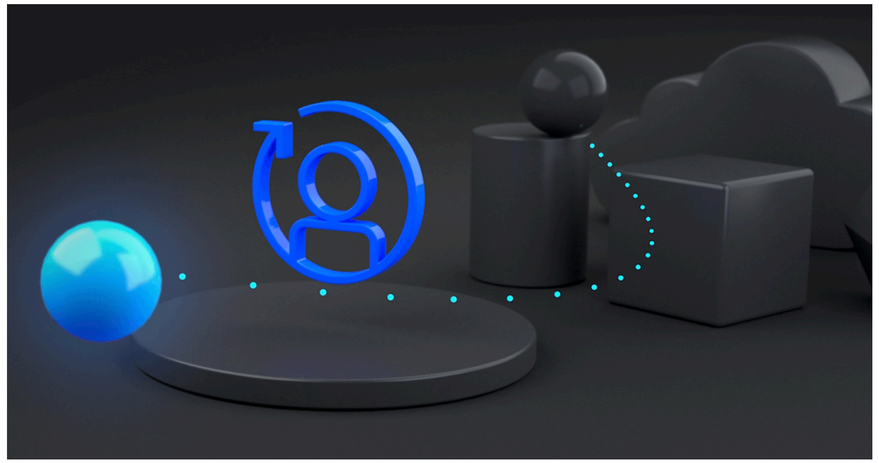
OUTRAGEOUSLY LARGE NEURAL NETWORKS:
THE SPARSELY-GATED MIXTURE-OF-EXPERTS LAYER

Noam Shazeer[1], Azalia Mirhoseini*[1], Krzysztof Maziarz*[2], Andy Davis[1], Quoc Le[1], Geoffrey Hinton[1] and Jeff Dean[1]



Sora, Shazeer et al ’17, ChatGPT, Chi et al ‘24

Many Problems are Sequential and Adaptive

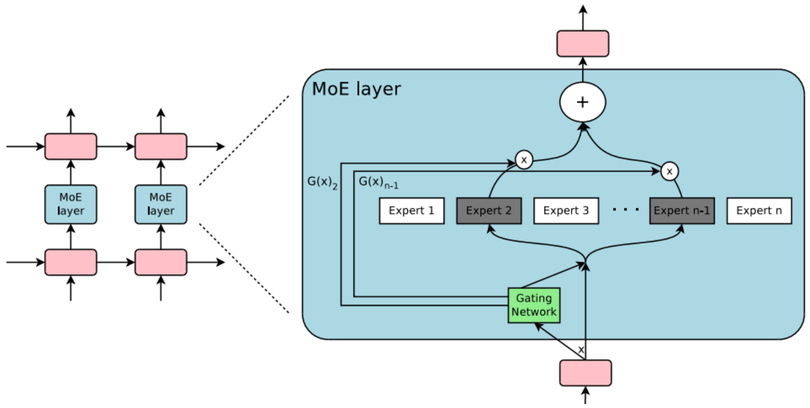


You
Write one short paragraph that simply explains how you generate text

ChatGPT
I generate text using a machine learning model called GPT (Generative Pre-trained Transformer), which has been trained on a vast amount of text data. Using its understanding of language patterns and semantics, GPT predicts the most likely next words based on the context it's given, producing coherent and contextually relevant text through a process of sampling from probability distributions.

Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer

Noam Shazeer[1], Azalia Mirhoseini*†[1], Krzysztof Maziarz*[2], Andy Davis[1], Quoc Le[1], Geoffrey Hinton[1] and Jeff Dean[1]



Sora, Shazeer et al ’17, ChatGPT, Chi et al ‘24

Online Learning

1. We get T data points X_t **generated adversarially** and $Y_t = f^\star(X_t) + \eta_t$.

Online Learning

1. We get T data points X_t **generated adversarially** and $Y_t = f^\star(X_t) + \eta_t$

For agnostic, see part II!

Online Learning

1. We get T data points X_t **generated adversarially** and $Y_t = f^\star(X_t) + \eta_t$.

Online Learning

1. We get T data points X_t **generated adversarially** and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Online Learning

1. We get T data points X_t **generated adversarially** and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: For each t, return $f_t \in \mathscr{F}$ with small error

$$\mathbb{E}\left[\mathrm{Err}_T\right] = \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \ell(f_t(X_t), f^\star(X_t))\right].$$

Online Learning

1. We get T data points X_t **generated adversarially** and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$.

Goal: For each t, return $f_t \in \mathcal{F}$ with small **error**

$$\mathbb{E}\left[\mathrm{Err}_T\right] = \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \ell(f_t(X_t), f^\star(X_t))\right].$$

Large body of work reduces sequential decision making to online learning!

Online Learning

1. We get T data points X_t **generated adversarially** and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: For each t, return $f_t \in \mathscr{F}$ with small **error**

$$\mathbb{E}\left[\mathrm{Err}_T\right] = \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \ell(f_t(X_t), f^\star(X_t))\right].$$

Algorithmic

How should we choose $\hat{f}$?

What computation do we need?

Epistemic

How much data?

What makes a problem hard?

Sample Complexity of Online Learning

Theorem [L'88,BPS'09,RST'14,RST'15**]: If ℓ is Lipschitz one can achieve**

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\mathscr{G}_T^{\mathsf{seq}}(\mathscr{F})}{T} \lesssim \sqrt{\frac{\mathsf{LDim}(\mathscr{F})}{T}}.$$

Sample Complexity of Online Learning

Theorem [L'88,BPS'09,RST'14,RST'15**]: If ℓ is Lipschitz one can achieve**

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\mathscr{G}_T^{\mathsf{seq}}(\mathscr{F})}{T} \lesssim \sqrt{\frac{\mathsf{LDim}(\mathscr{F})}{T}}.$$

We always have $\mathsf{vc}(\mathscr{F}) \leq \mathsf{LDim}(\mathscr{F})$.

Sample Complexity of Online Learning

Theorem [L'88,BPS'09,RST'14,RST'15**]: If ℓ is Lipschitz one can achieve**

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\mathscr{G}_T^{\mathsf{seq}}(\mathscr{F})}{T} \lesssim \sqrt{\frac{\mathsf{LDim}(\mathscr{F})}{T}}.$$

We always have $\mathsf{vc}(\mathscr{F}) \leq \mathsf{LDim}(\mathscr{F})$.

Typically, $\mathsf{vc}(\mathscr{F}) \ll \mathsf{LDim}(\mathscr{F}) = \infty$.

Sample Complexity of Online Learning

Theorem [L'88,BPS'09,RST'14,RST'15**]: If ℓ is Lipschitz one can achieve**

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\mathscr{G}_T^{\mathsf{seq}}(\mathscr{F})}{T} \lesssim \sqrt{\frac{\mathsf{LDim}(\mathscr{F})}{T}}.$$

We always have $\mathsf{vc}(\mathscr{F}) \leq \mathsf{LDim}(\mathscr{F})$.

Typically, $\mathsf{vc}(\mathscr{F}) \ll \mathsf{LDim}(\mathscr{F}) = \infty$.

Online learning is computationally hard even under nice oracle assumptions [HK'16**].**

Tutorial Outline

Part I

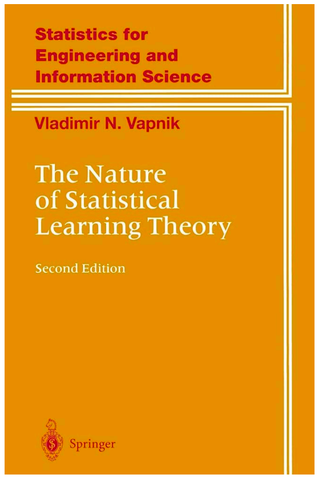
1. Statistical and Online Learning: Definitions and Applications

   (a) What constitutes success in understanding learning?

   (b) Online Learning: What is it and why study it?

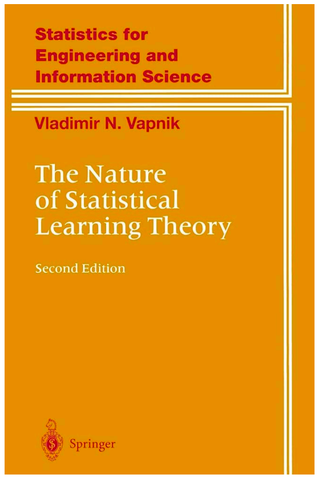
   (c) Why is Online Learning not enough?

Difficulty of Learning

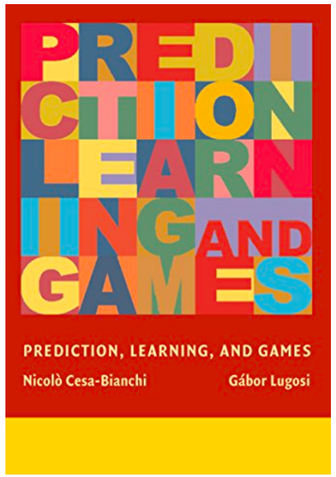


Difficulty of Learning

Statistical Learning



Difficulty of Learning



Statistical Learning — **Online Learning**

Statistical Learning

Online Learning

Statistical Learning

Pros:

Intuitive notion of learning.

Online Learning

Statistical Learning

Pros:

Intuitive notion of learning.

Learning reduces to optimization:

Learning = Data + Opt

Online Learning

Statistical Learning

Pros:

Intuitive notion of learning.

Learning reduces to optimization:

Learning = Data + Opt

Cons:

Strong modeling assumption.

Online Learning

Statistical Learning

Pros:

Intuitive notion of learning.

Learning reduces to optimization:

Learning = Data + Opt

Cons:

Strong modeling assumption.

Not robust.

Online Learning

Statistical Learning

Pros:

Intuitive notion of learning.

Learning reduces to optimization:

Learning = Data + Opt

Cons:

Strong modeling assumption.

Not robust.

Online Learning

Pros:

Minimal assumptions needed.

Statistical Learning

Pros:

Intuitive notion of learning.

Learning reduces to optimization:

Learning = Data + Opt

Cons:

Strong modeling assumption.

Not robust.

Online Learning

Pros:

Minimal assumptions needed.

Models robustness.

Statistical Learning

Pros:

Intuitive notion of learning.

Learning reduces to optimization:

Learning = Data + Opt

Cons:

Strong modeling assumption.

Not robust.

Online Learning

Pros:

Minimal assumptions needed.

Models robustness.

Cons:

Statistical hardness.

Statistical Learning

Pros:

Intuitive notion of learning.

Learning reduces to optimization:

Learning = Data + Opt

Cons:

Strong modeling assumption.

Not robust.

Online Learning

Pros:

Minimal assumptions needed.

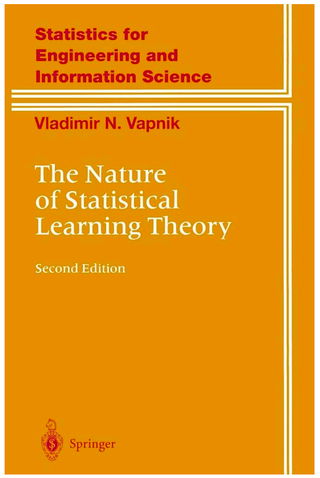
Models robustness.

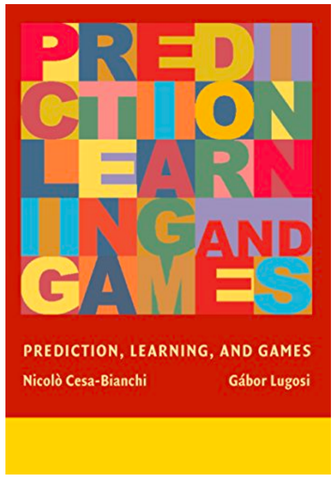
Cons:

Statistical hardness.

Computational hardness:

Learning ≠ Data + Opt





??????????

Statistical Learning

Online Learning

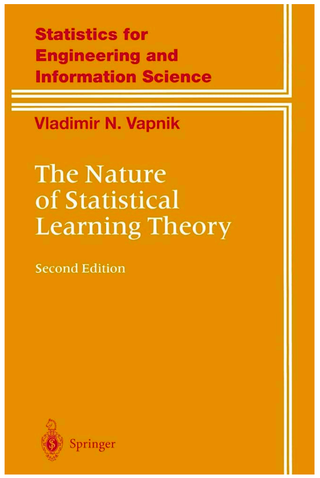
Tutorial Outline

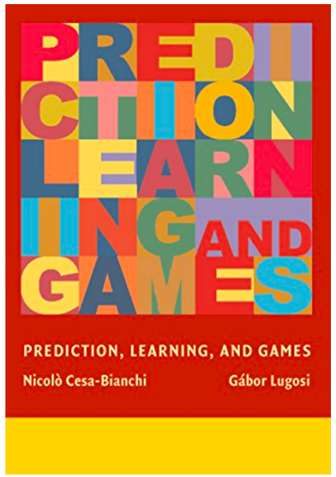
Part I

1. Statistical and Online Learning: Definitions and Applications
2. The Smoothed Model: Best of Both Worlds?
3. The Power of Empirical Risk Minimization

Part II

1. Coupling Lemma
2. Handling Label Noise: The Agnostic Setting
3. Oracle-Efficiency: ERM Returns

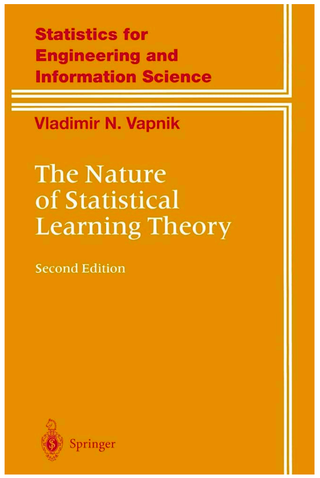


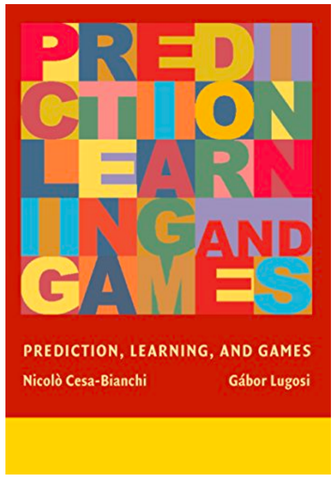


??????????

Statistical Learning

Online Learning





Smoothed data

Statistical Learning

Online Learning

Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is **σ-smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_{\infty} \leq \sigma^{-1}.$$

Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is **σ-smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_{\infty} \leq \sigma^{-1}.$$

Motivation from smoothed analysis of algorithms [ST'02].

Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is **σ-smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

Motivation from smoothed analysis of algorithms [ST'02].

Forces data to be **anti-concentrated**.

Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is **σ-smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

Motivation from smoothed analysis of algorithms [ST'02].

Forces data to be **anti-concentrated**.

Example: μ is uniform on a discrete $\mathcal{X}$.

Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is **σ-smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

Motivation from smoothed analysis of algorithms [ST'02].

Forces data to be **anti-concentrated**.

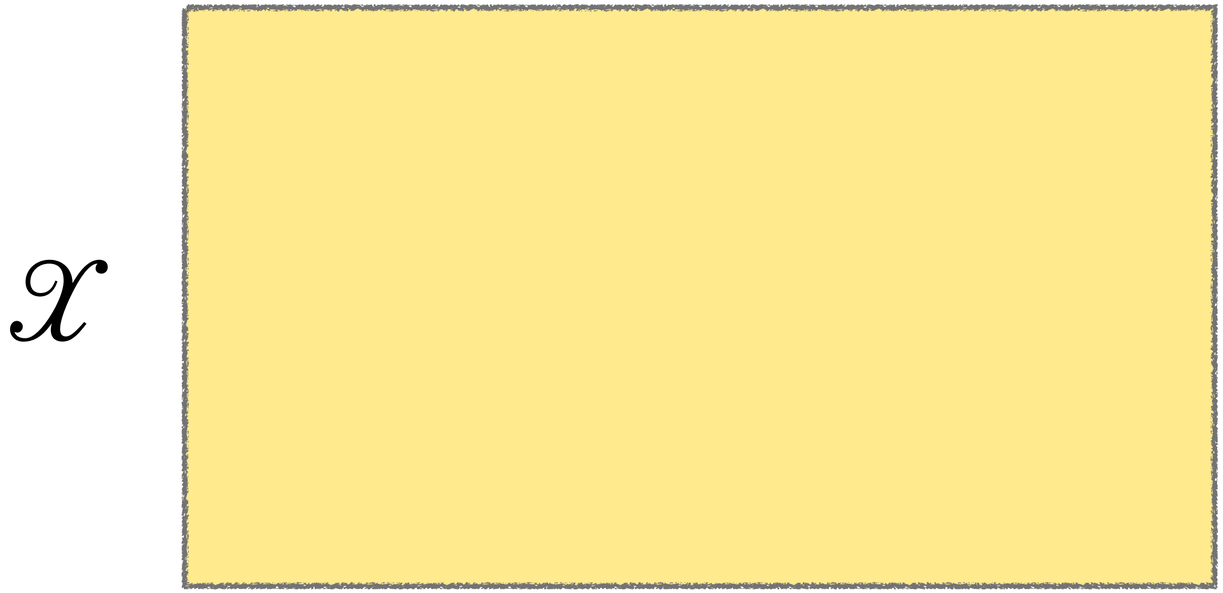
Example: μ is uniform on a discrete $\mathcal{X}$.

Example: μ is Lebesgue on $\mathbb{R}^d$ and $X = \overline{X} + \text{noise}$.

Online Learning

1. We get T data points X_t **generated adversarially** and $Y_t = f^\star(X_t) + \eta_t$.

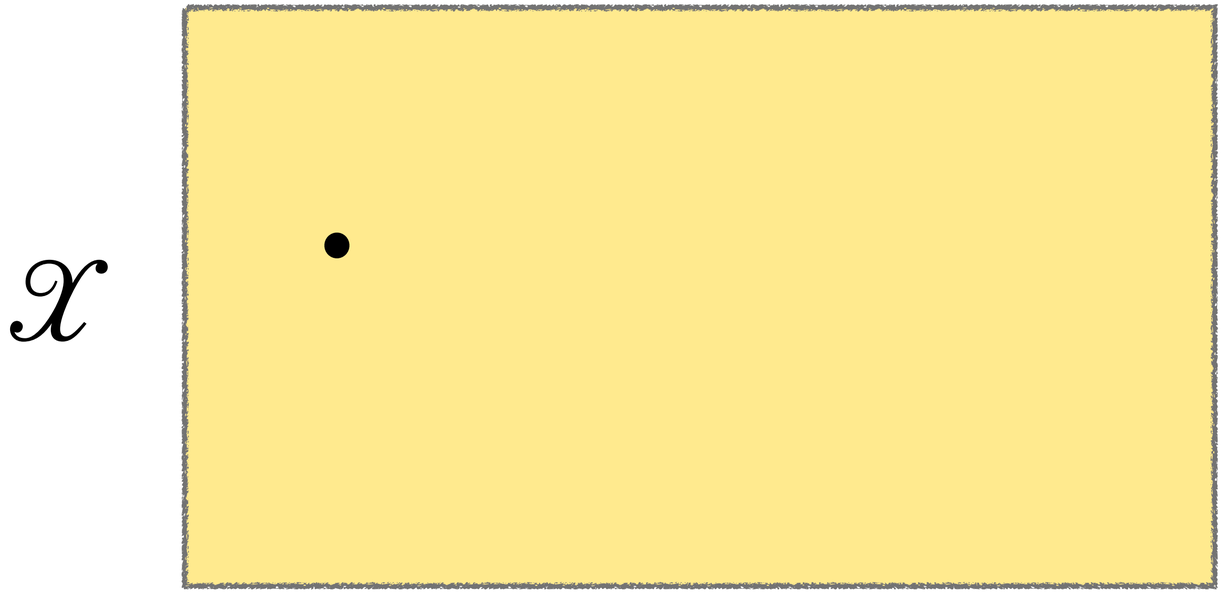
2. We have access to a **model class** $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$.



Online Learning

1. We get T data points X_t **generated adversarially** and $Y_t = f^\star(X_t) + \eta_t$.

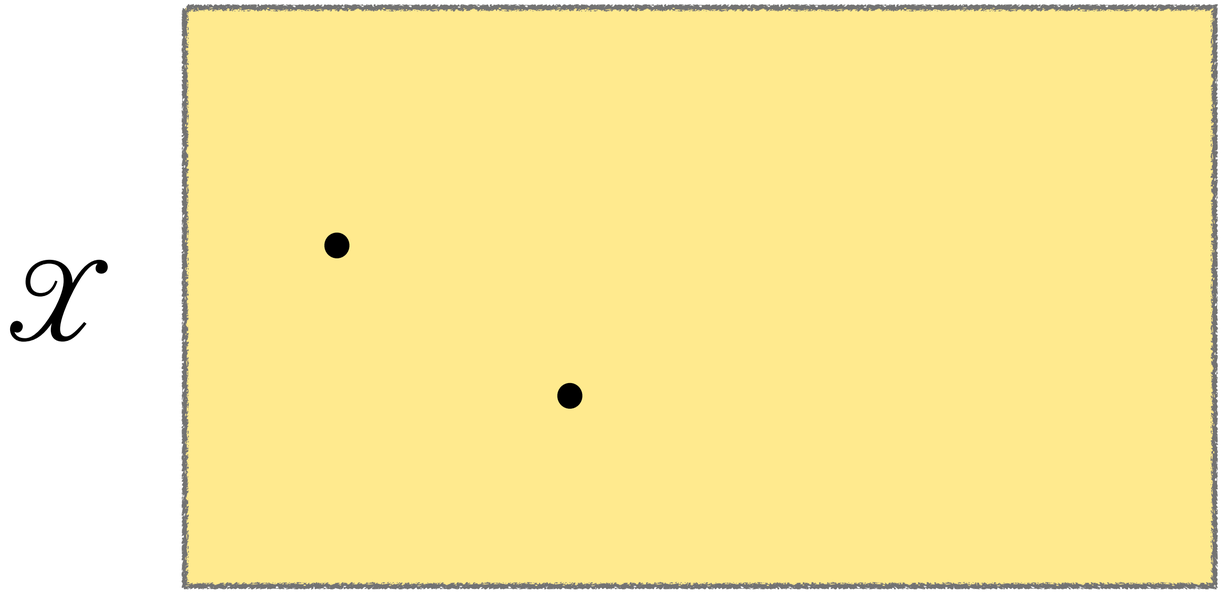
2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.



Online Learning

1. We get T data points X_t **generated adversarially** and $Y_t = f^\star(X_t) + \eta_t$.

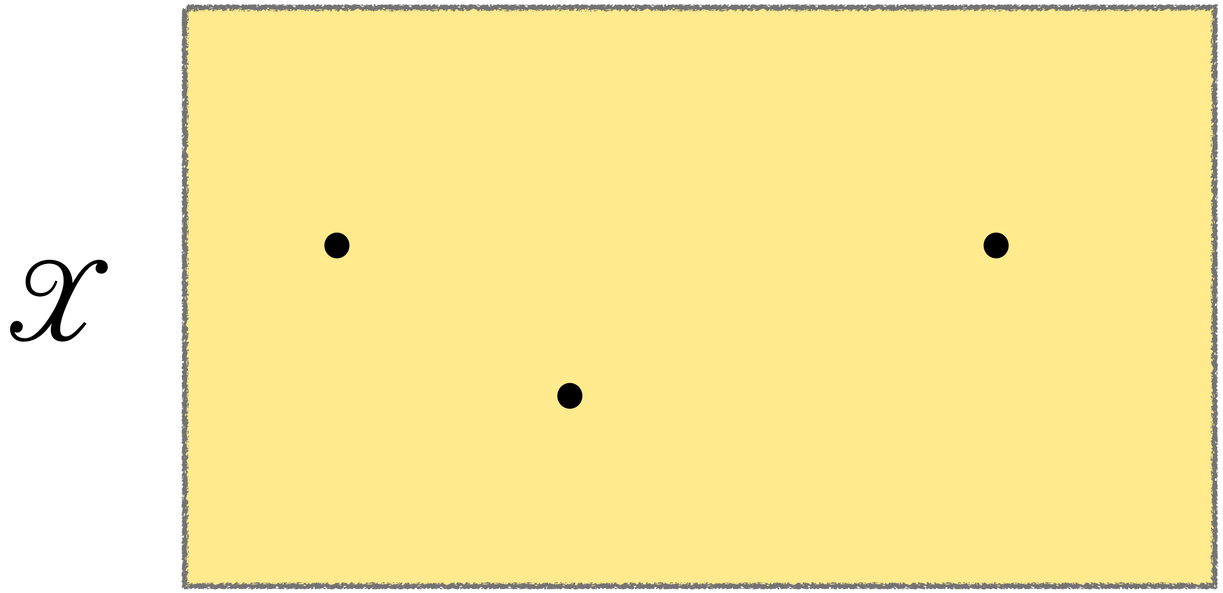
2. We have access to a **model class** $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$.



Online Learning

1. We get T data points X_t **generated adversarially** and $Y_t = f^\star(X_t) + \eta_t$.

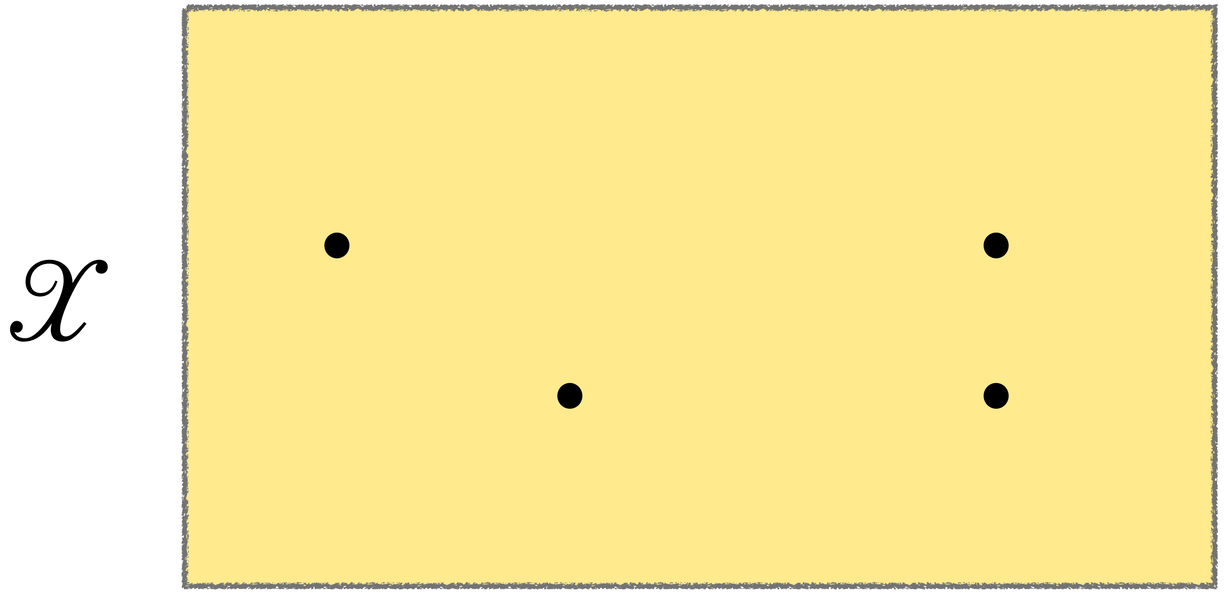
2. We have access to a **model class** $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$.



Online Learning

1. We get T data points X_t **generated adversarially** and $Y_t = f^\star(X_t) + \eta_t$.

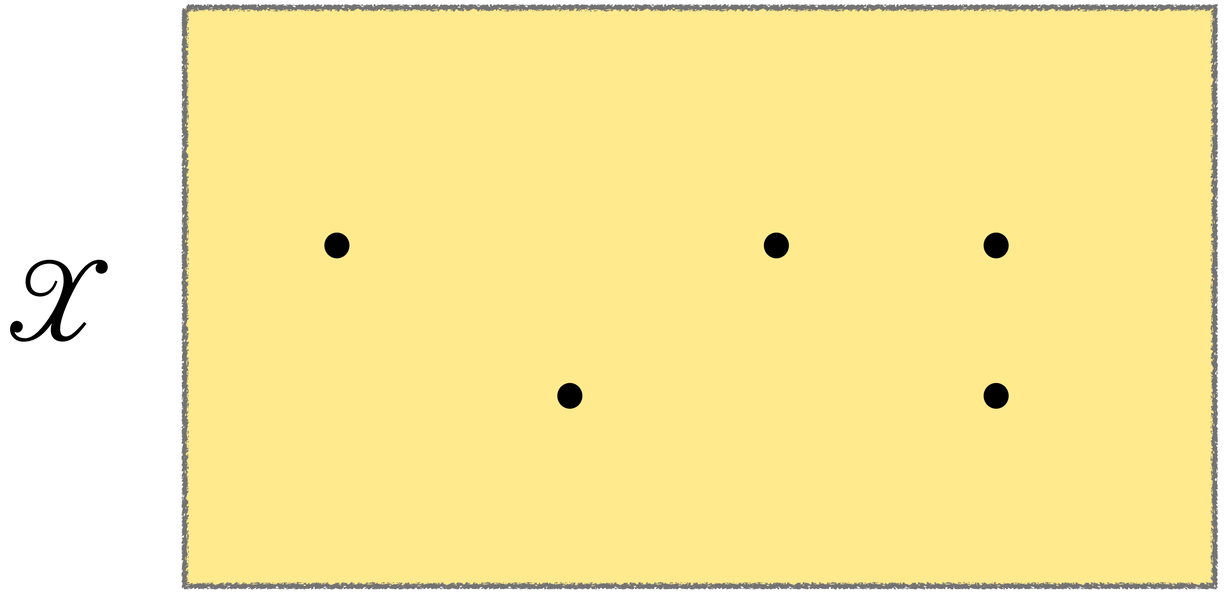
2. We have access to a **model class** $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$.



Online Learning

1. We get T data points X_t **generated adversarially** and $Y_t = f^\star(X_t) + \eta_t$.

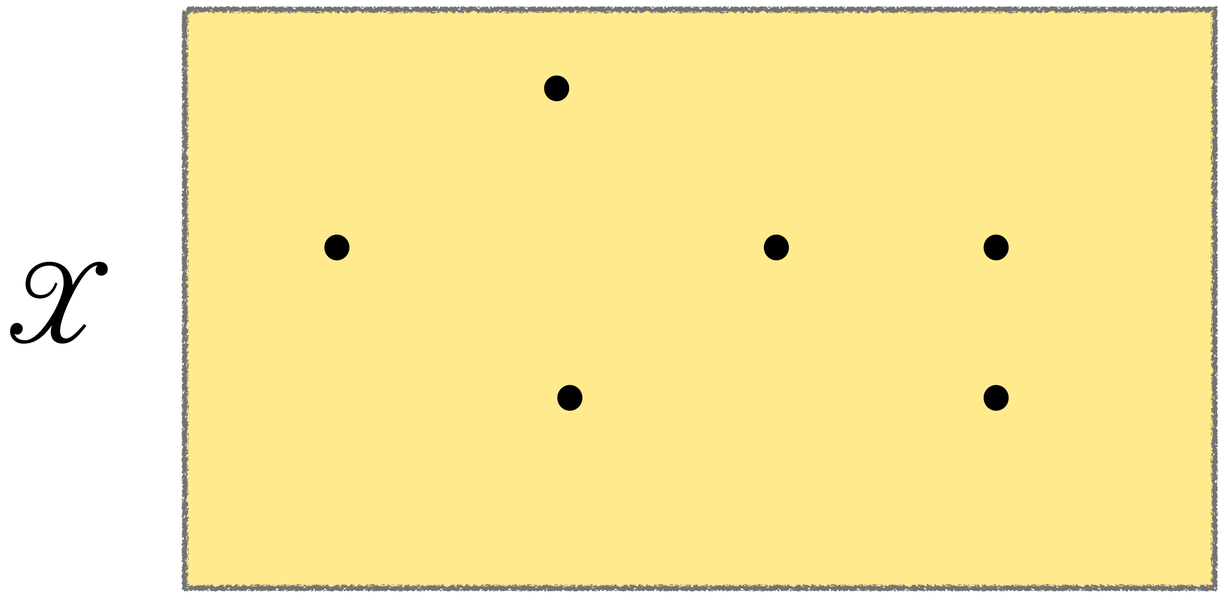
2. We have access to a **model class** $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$.



Online Learning

1. We get T data points X_t **generated adversarially** and $Y_t = f^\star(X_t) + \eta_t$.

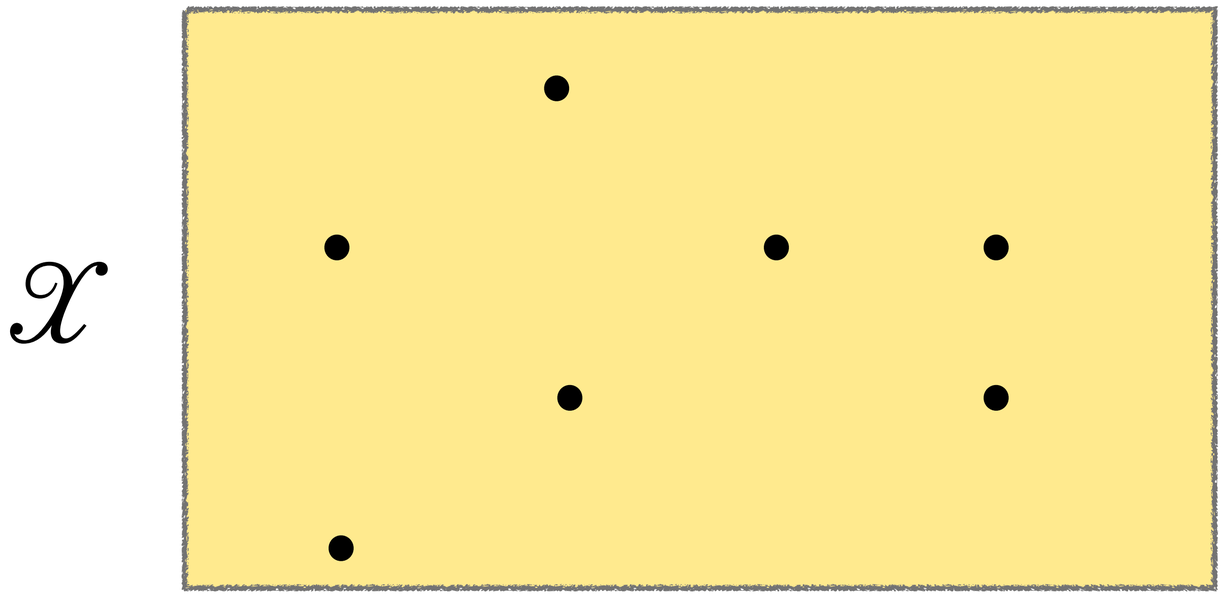
2. We have access to a **model class** $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$.



Online Learning

1. We get T data points X_t **generated adversarially** and $Y_t = f^\star(X_t) + \eta_t$.

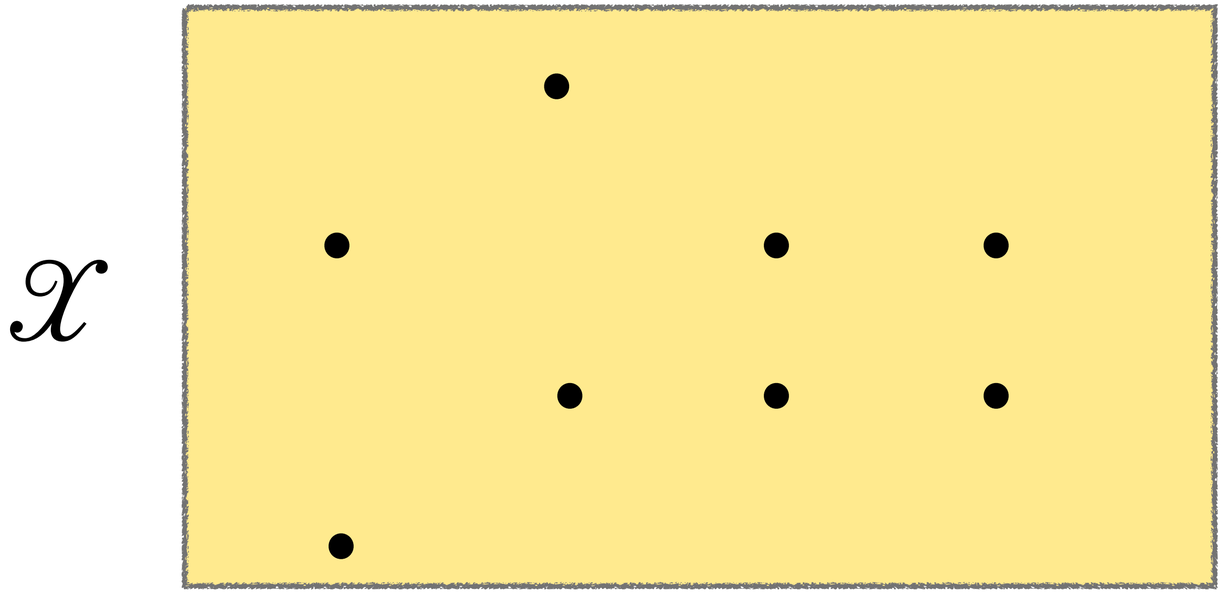
2. We have access to a **model class** $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$.



Online Learning

1. We get T data points X_t **generated adversarially** and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$.



Smoothed Online Learning

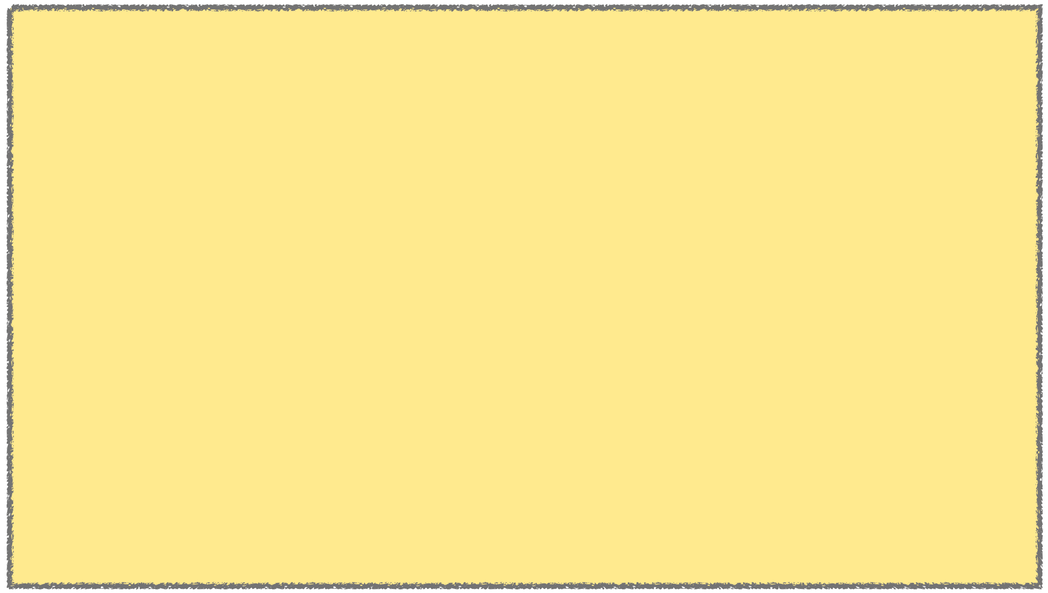
1. We get T data points X_t **smooth** w.r.t μ and $Y_t = f^\star(X_t) + \eta_t$.

Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is **σ-smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$\mu = \text{Unif}(\mathcal{X})$

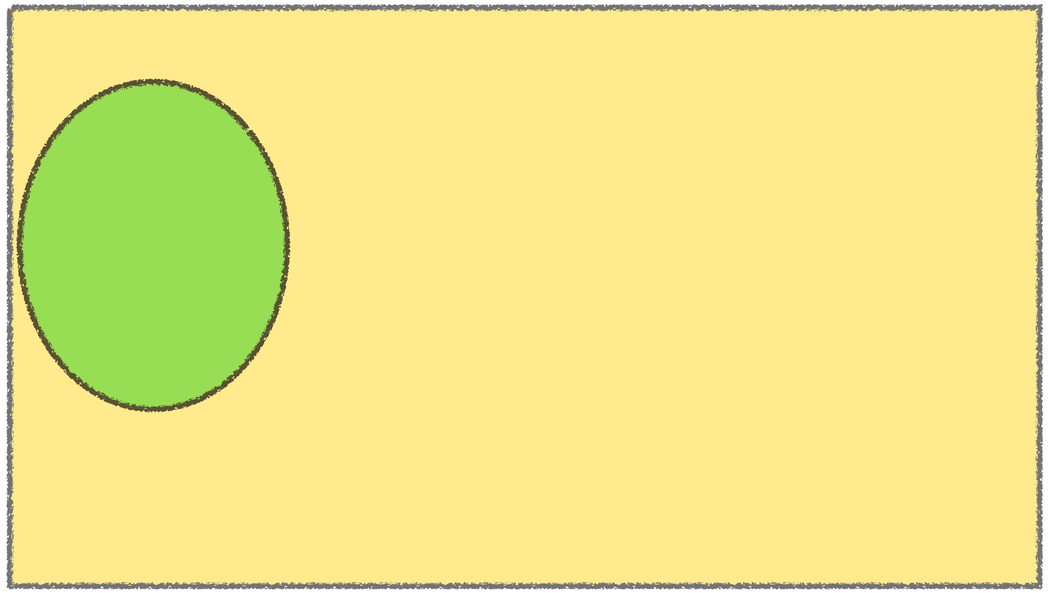


Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$$\mu = \mathrm{Unif}(\mathcal{X})$$

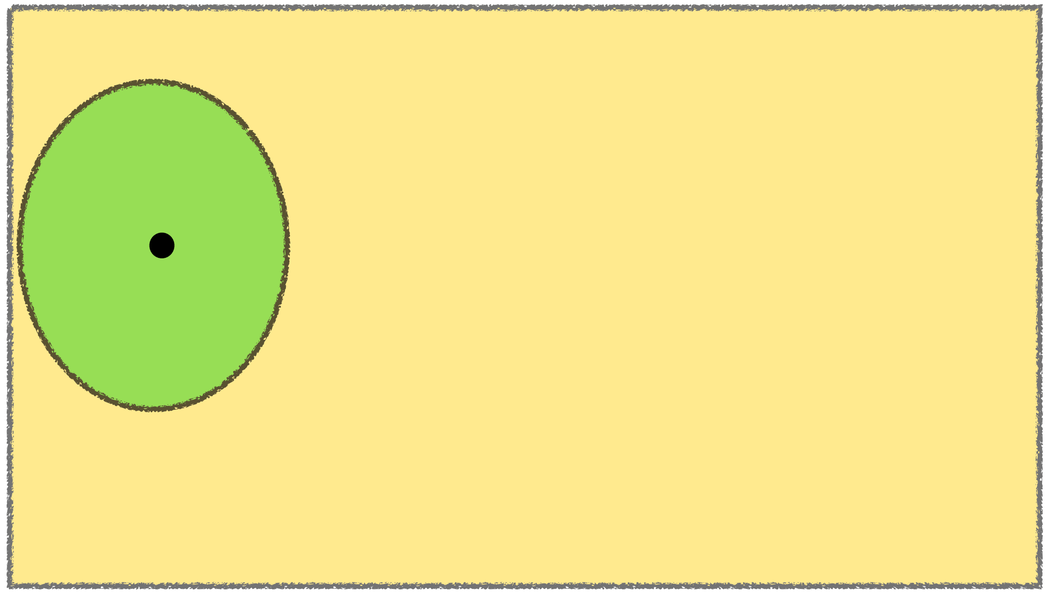


Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$\mu = \text{Unif}(\mathcal{X})$

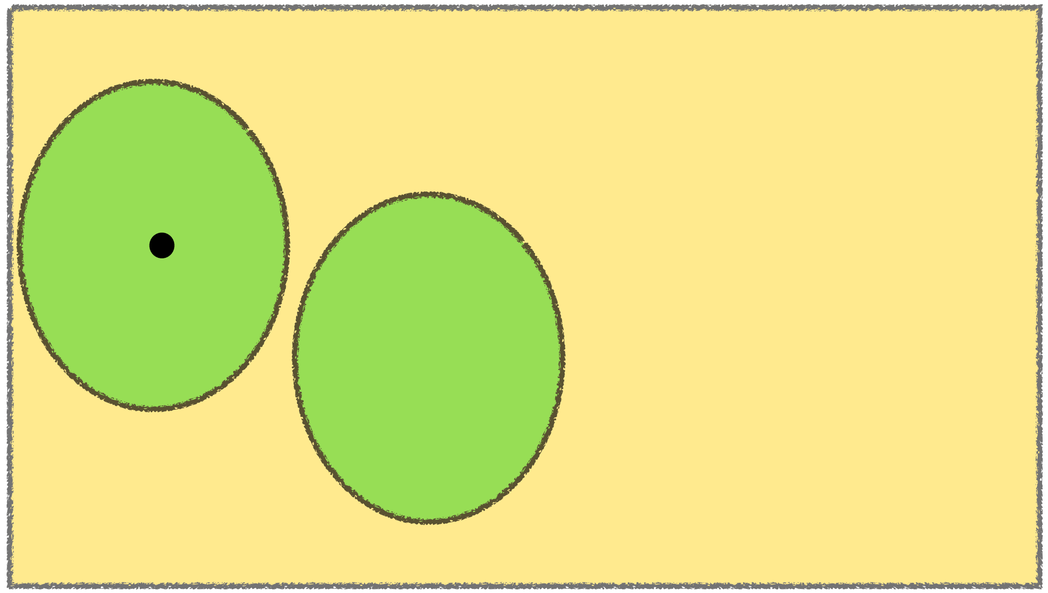


Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$$\mu = \text{Unif}(\mathcal{X})$$

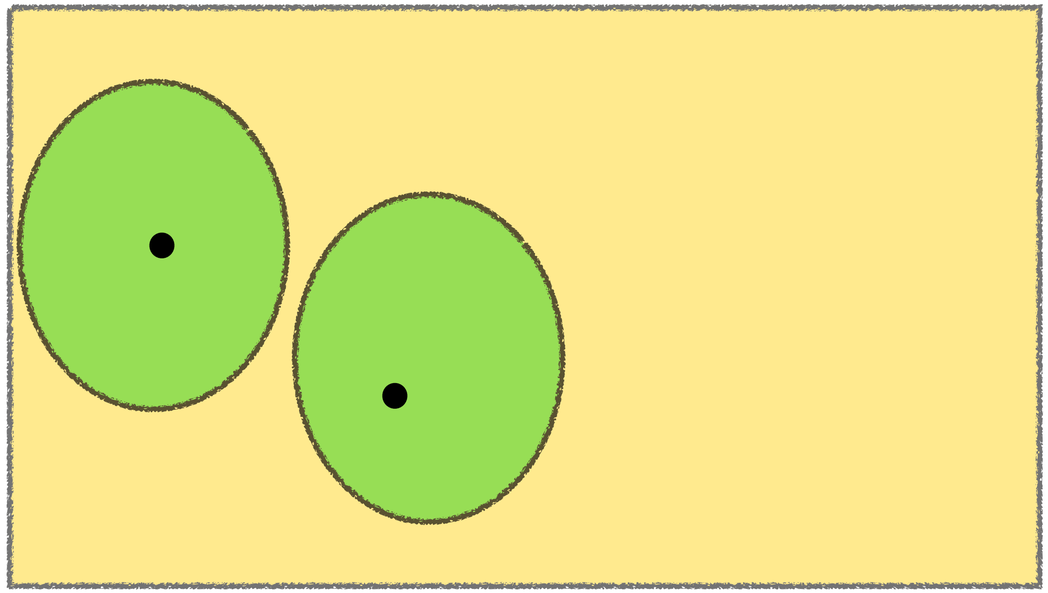


Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$$\mu = \text{Unif}(\mathcal{X})$$

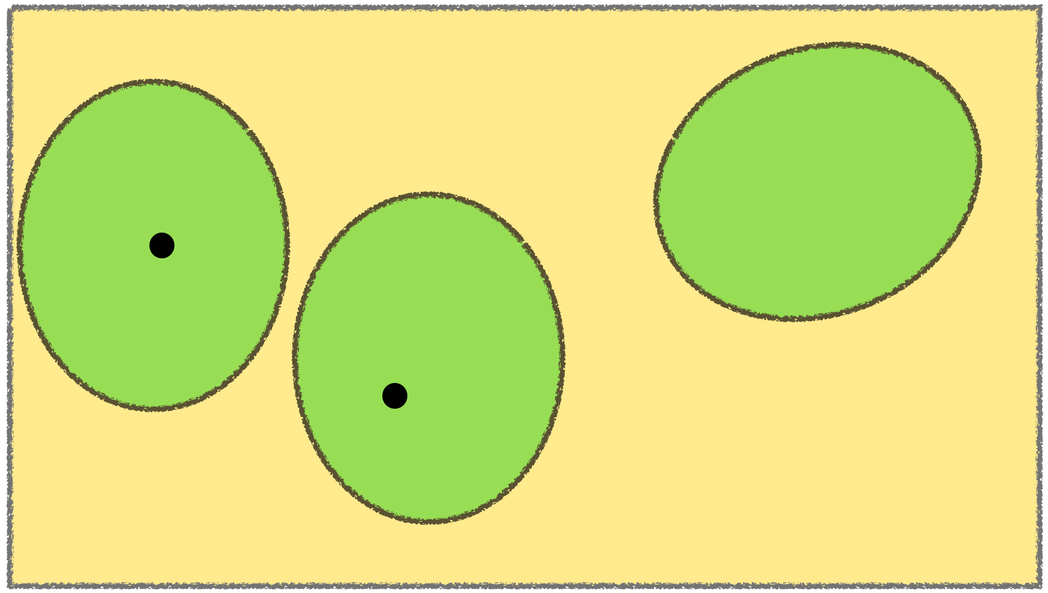


Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$$\mu = \mathrm{Unif}(\mathcal{X})$$

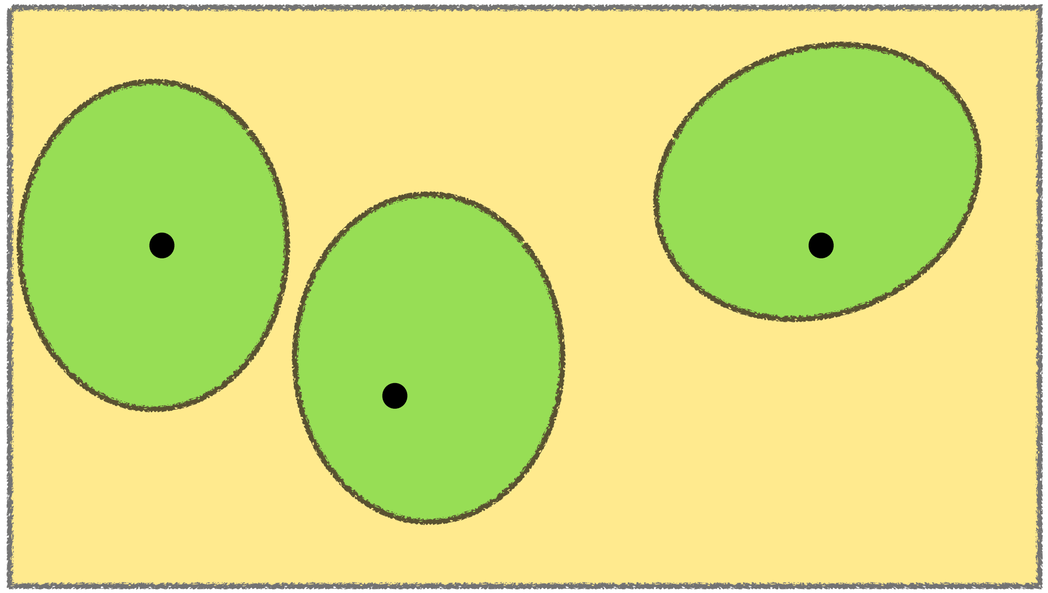


Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$$\mu = \mathrm{Unif}(\mathcal{X})$$

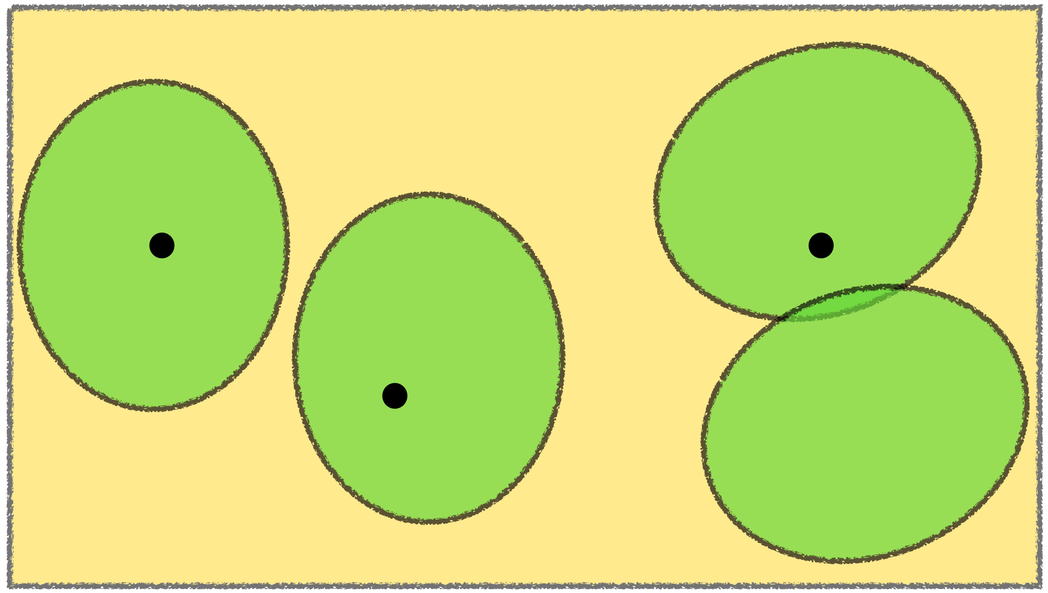


Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$$\mu = \text{Unif}(\mathcal{X})$$

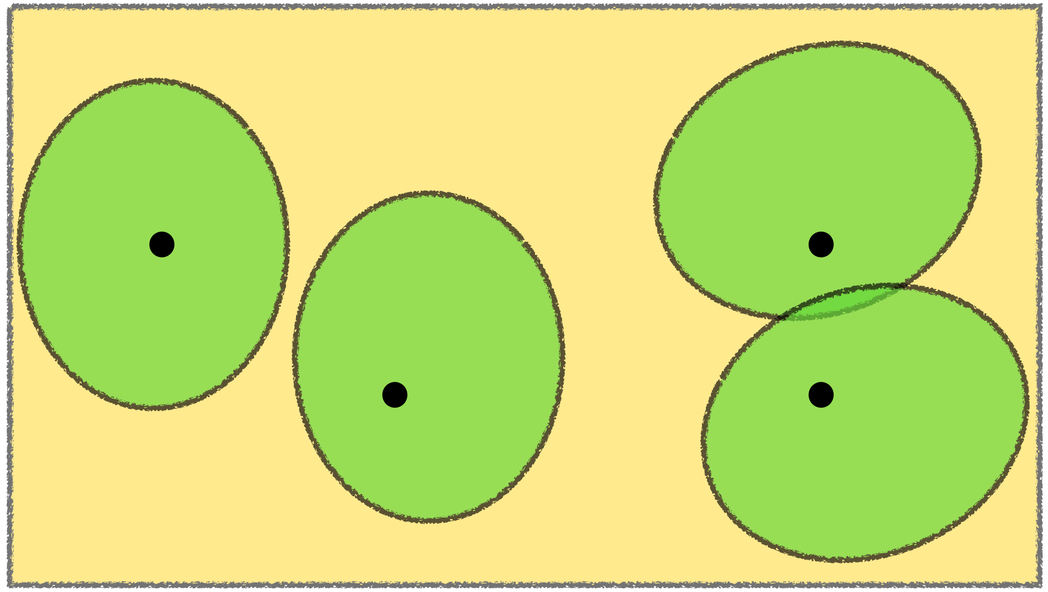


Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$$\mu = \text{Unif}(\mathcal{X})$$

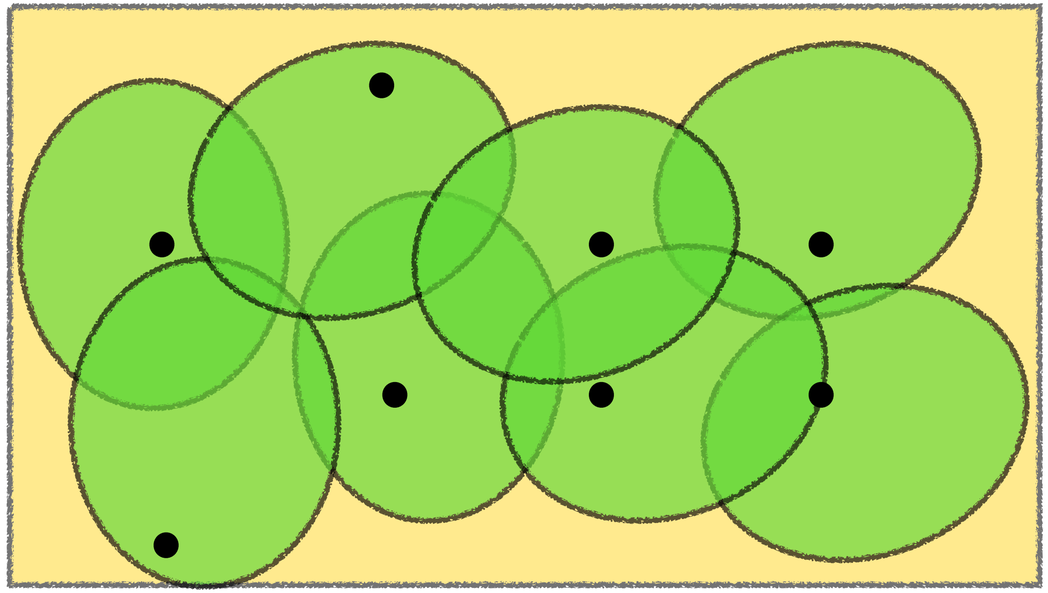


Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$$\mu = \mathrm{Unif}(\mathcal{X})$$



Smoothed Online Learning

1. We get T data points X_t **smooth** w.r.t μ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: For each t, return $f_t \in \mathscr{F}$ with small **error**

$$\mathbb{E}\left[\mathrm{Err}_T\right] = \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \ell(f_t(X_t), f^\star(X_t))\right].$$

Smoothed Online Learning

For agnostic, see part II!

1. We get T data points X_t **smooth** w.r.t μ and $Y_t = f^\star(X_t) + \eta_t$

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: For each t, return $f_t \in \mathscr{F}$ with small **error**

$$\mathbb{E}\left[\mathrm{Err}_T\right] = \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \ell(f_t(X_t), f^\star(X_t))\right].$$

Smoothed Online Learning

For agnostic, see part II!

1. We get T data points X_t **smooth** w.r.t μ and $Y_t = f^\star(X_t) + \eta_t$

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: For each t, return $f_t \in \mathscr{F}$ with small **error**

$$\mathbb{E}\left[\mathrm{Err}_T\right] = \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \ell(f_t(X_t), f^\star(X_t))\right].$$

Is μ known to the learner?

Smoothed Online Learning

For agnostic, see part II!

1. We get T data points X_t **smooth** w.r.t μ and $Y_t = f^\star(X_t) + \eta_t$

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: For each t, return $f_t \in \mathscr{F}$ with small **error**

$$\mathbb{E}\left[\mathrm{Err}_T\right] = \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \ell(f_t(X_t), f^\star(X_t))\right].$$

Is μ known to the learner?

Mostly doesn't matter in Part I.

Smoothed Online Learning

For agnostic, see part II!

1. We get T data points X_t **smooth** w.r.t μ and $Y_t = f^\star(X_t) + \eta_t$

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: For each t, return $f_t \in \mathscr{F}$ with small **error**

$$\mathbb{E}\left[\mathrm{Err}_T\right] = \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \ell(f_t(X_t), f^\star(X_t))\right].$$

Is μ known to the learner?

Mostly doesn't matter in Part I.

We will come back to this in Part II.

Smoothed Online Learning

For agnostic, see part II!

1. We get T data points X_t **smooth** w.r.t μ and $Y_t = f^\star(X_t) + \eta_t$

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: For each t, return $f_t \in \mathscr{F}$ with small **error**

$$\mathbb{E}\left[\mathrm{Err}_T\right] = \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \ell(f_t(X_t), f^\star(X_t))\right].$$

Why is smoothness helpful?

Smoothed Online Learning

For agnostic, see part II!

1. We get T data points X_t **smooth** w.r.t μ and $Y_t = f^\star(X_t) + \eta_t$

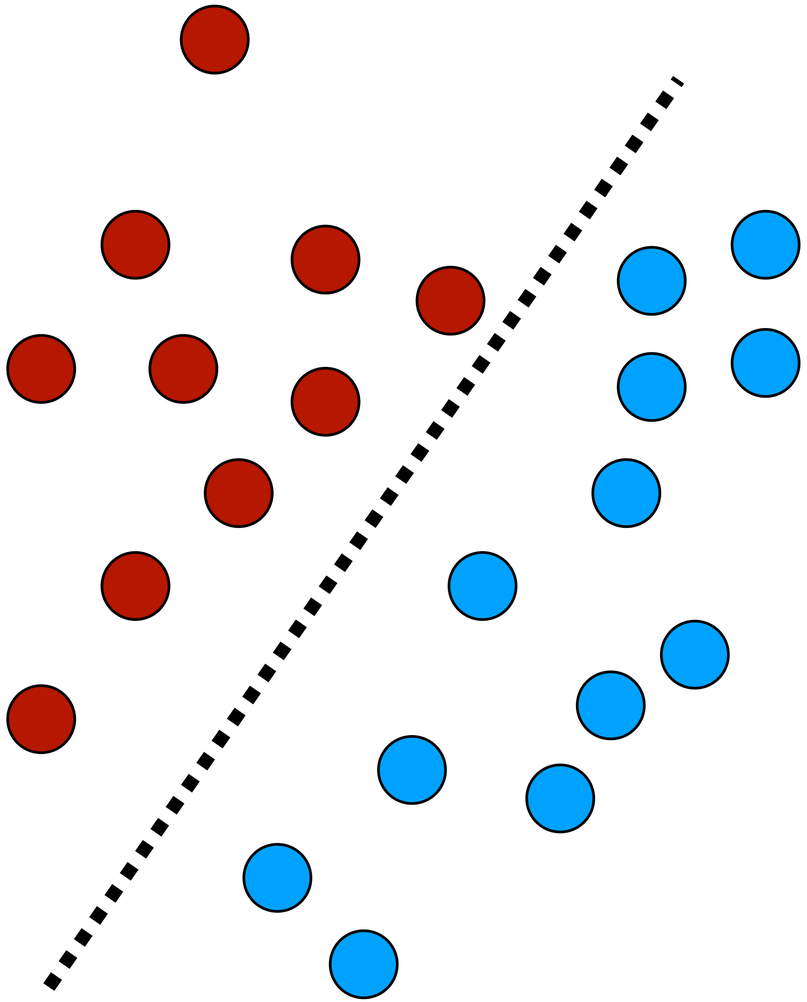
2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: For each t, return $f_t \in \mathscr{F}$ with small **error**

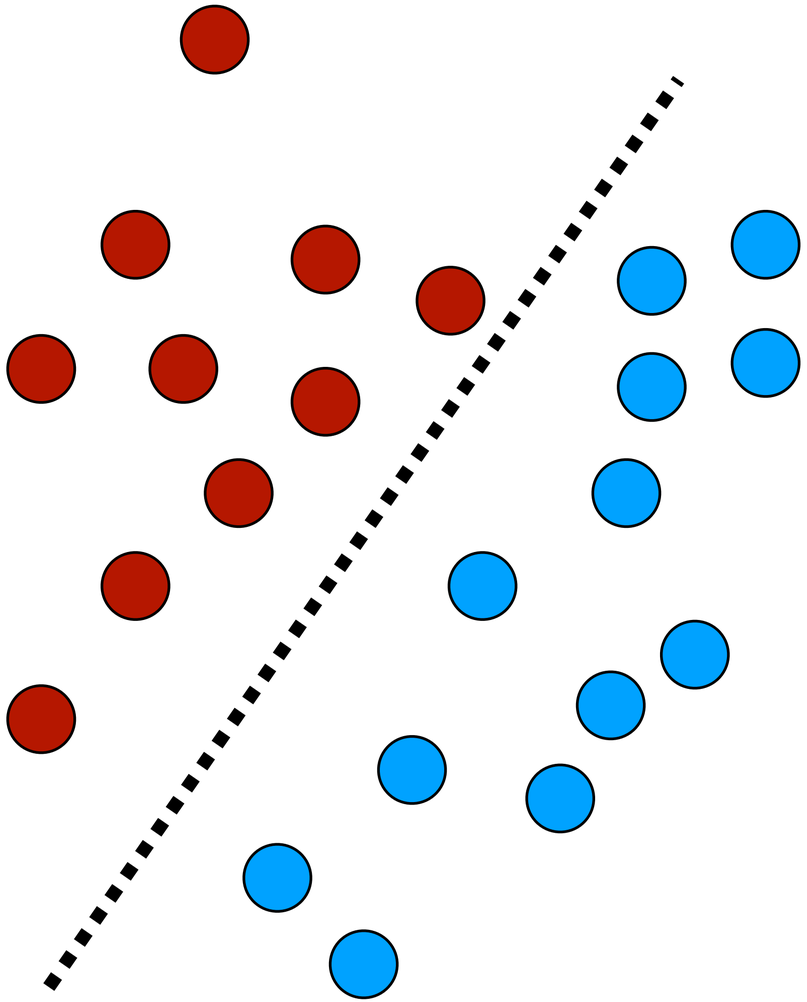
$$\mathbb{E}\left[\mathrm{Err}_T\right] = \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \ell(f_t(X_t), f^\star(X_t))\right].$$

Why is smoothness helpful?

Linear thresholds



Linear thresholds



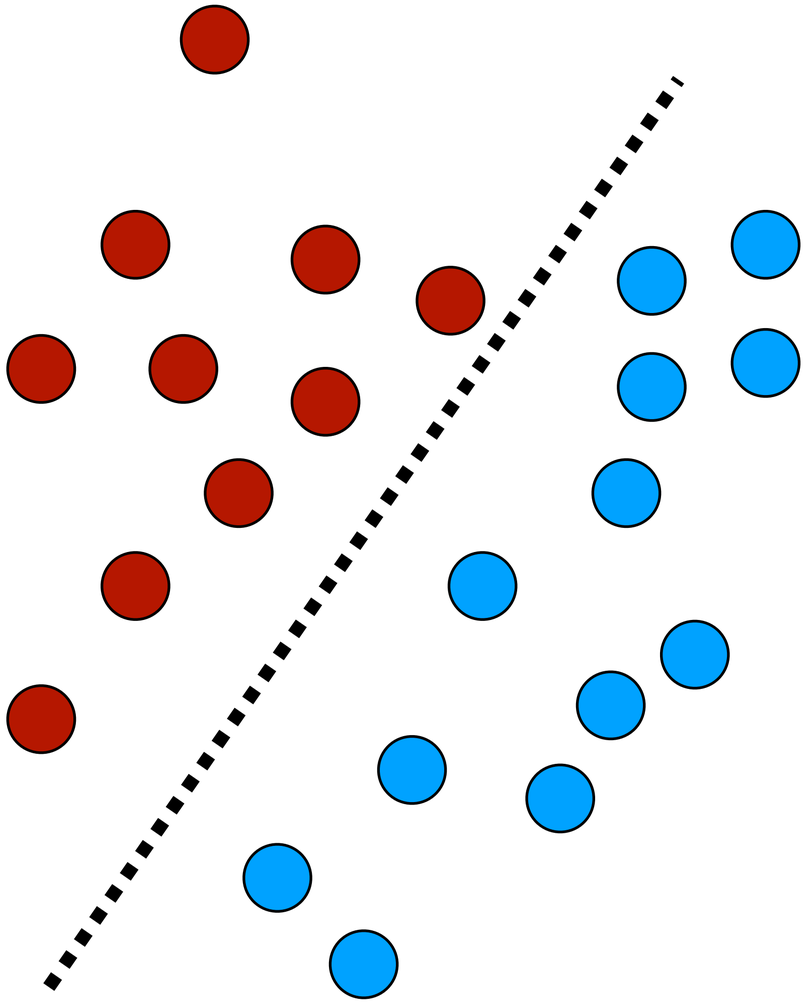
Psychological Review
Vol. 65, No. 6, 1958

THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN[1]

F. ROSENBLATT

Cornell Aeronautical Laboratory

Linear thresholds



Psychological Review
Vol. 65, No. 6, 1958

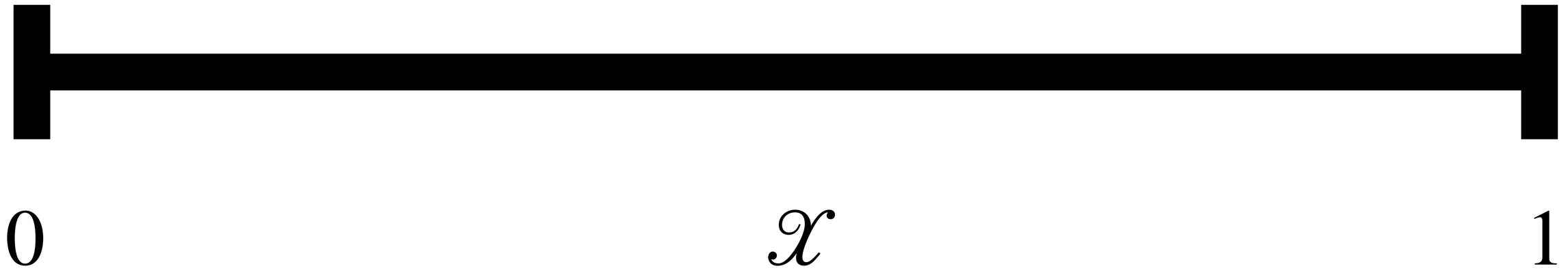
THE PERCEPTRON: A PROBABILISTIC MODEL FOR INFORMATION STORAGE AND ORGANIZATION IN THE BRAIN [1]

F. ROSENBLATT

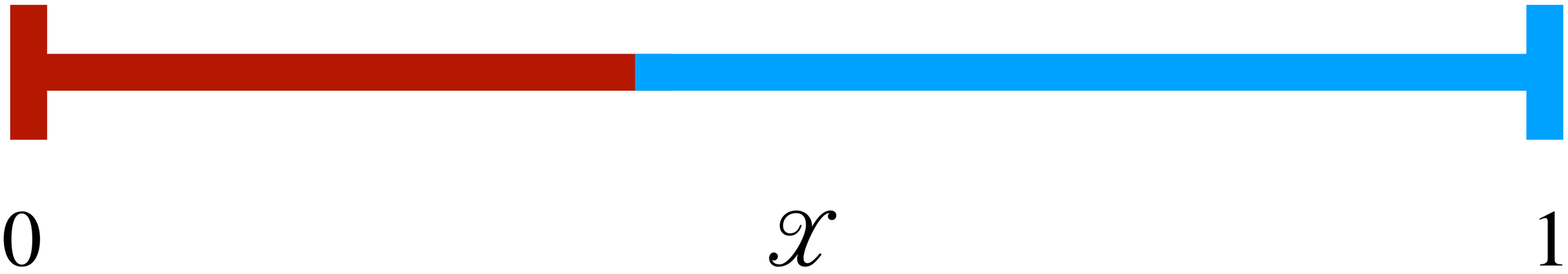
Cornell Aeronautical Laboratory

$$y = \operatorname{sign}\left(\langle \mathbf{w}, \mathbf{x} \rangle + b\right)$$

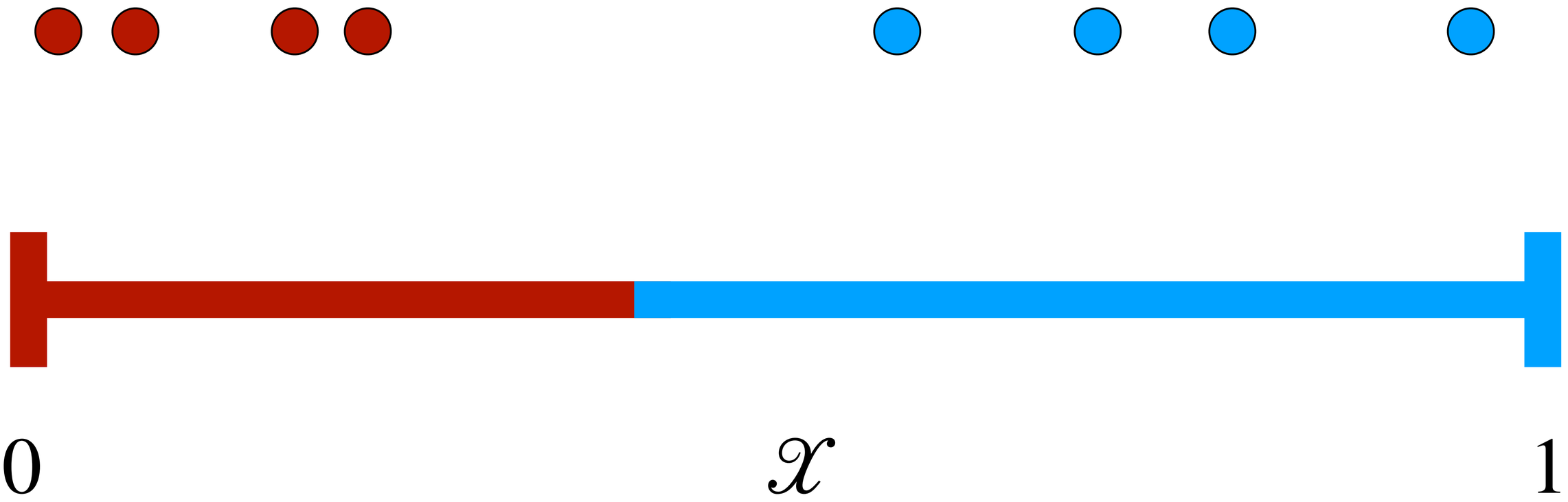
What Makes Online Learning Hard?



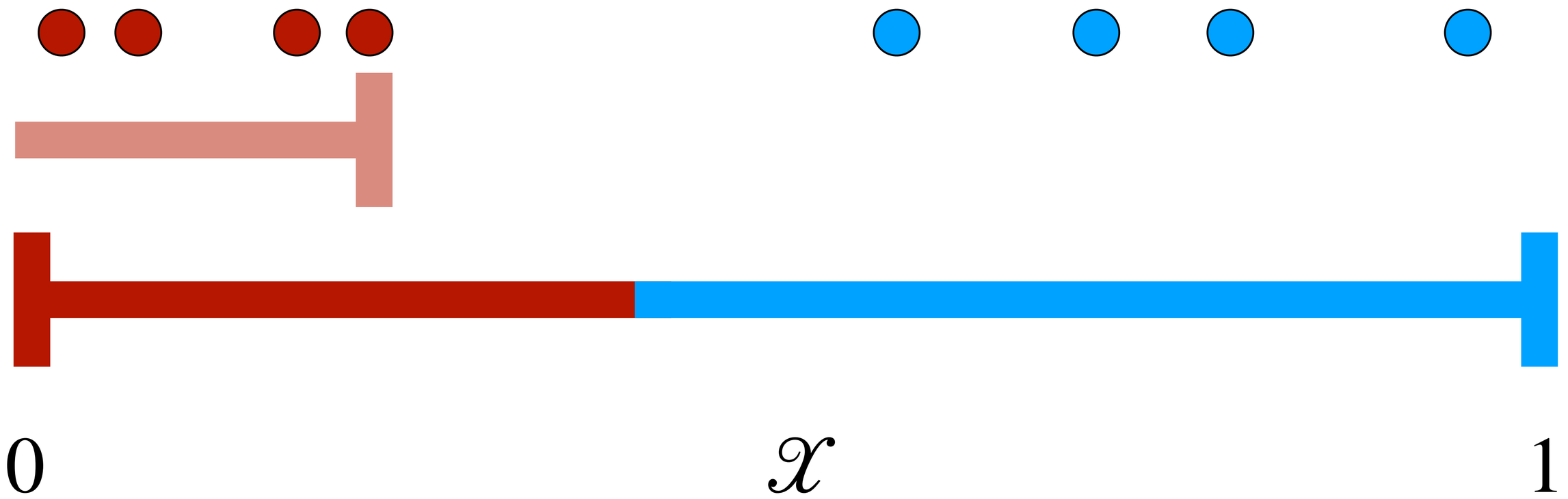
What Makes Online Learning Hard?



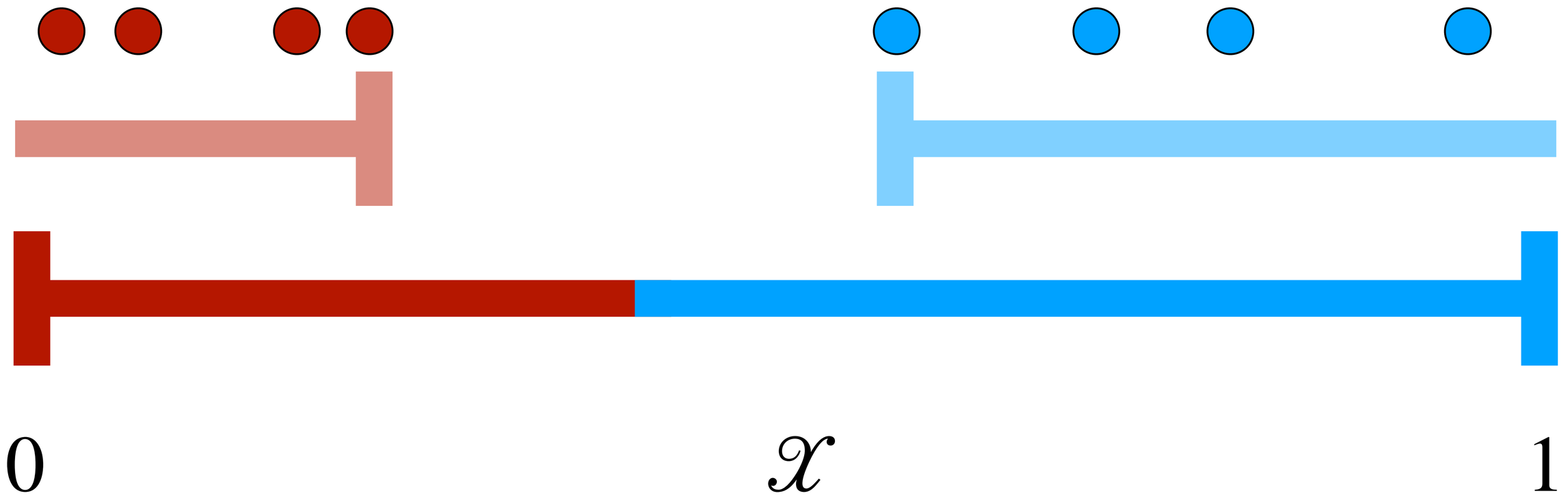
What Makes Online Learning Hard?



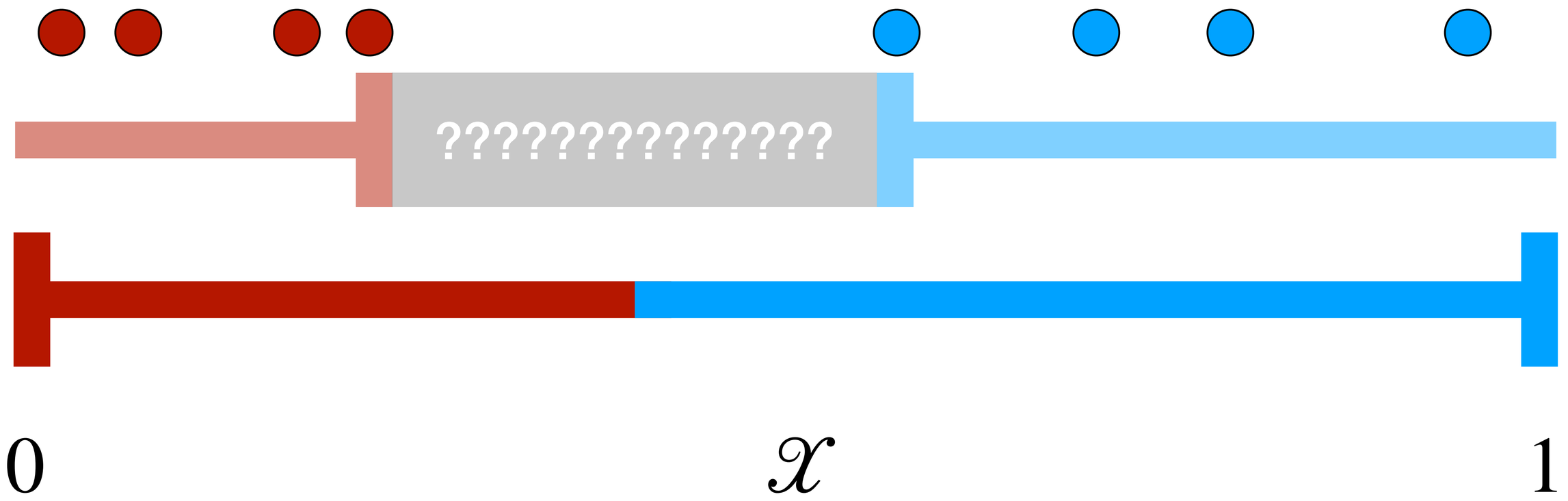
What Makes Online Learning Hard?



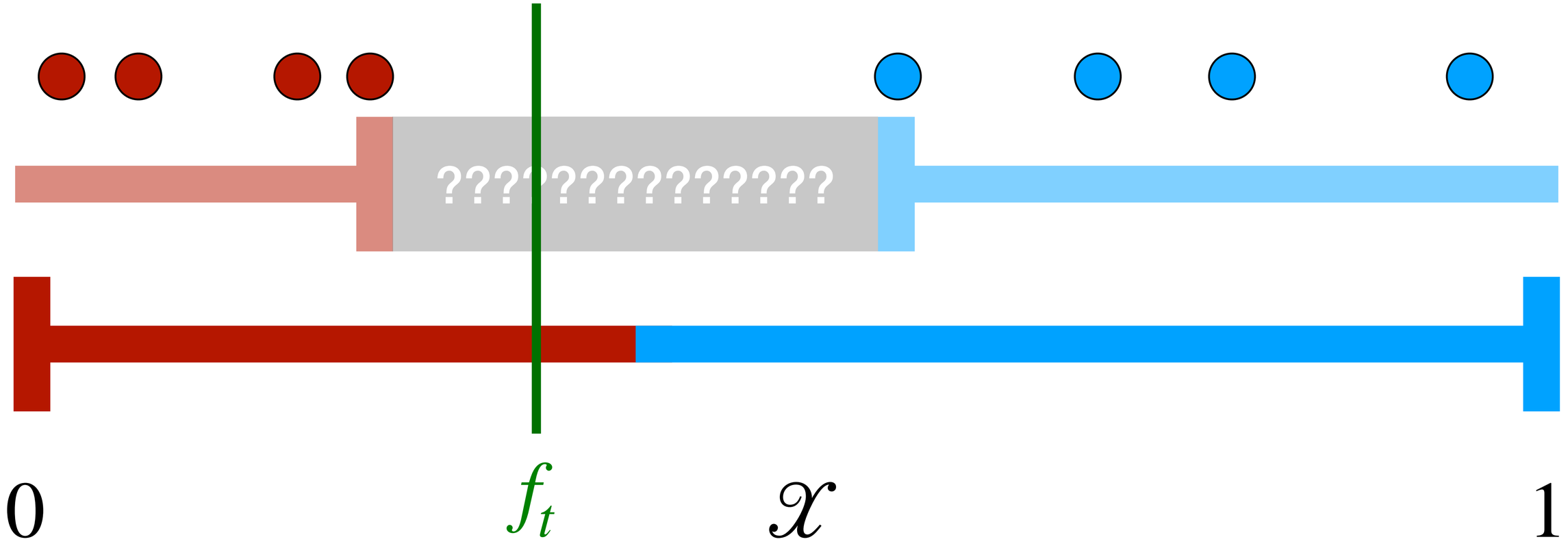
What Makes Online Learning Hard?



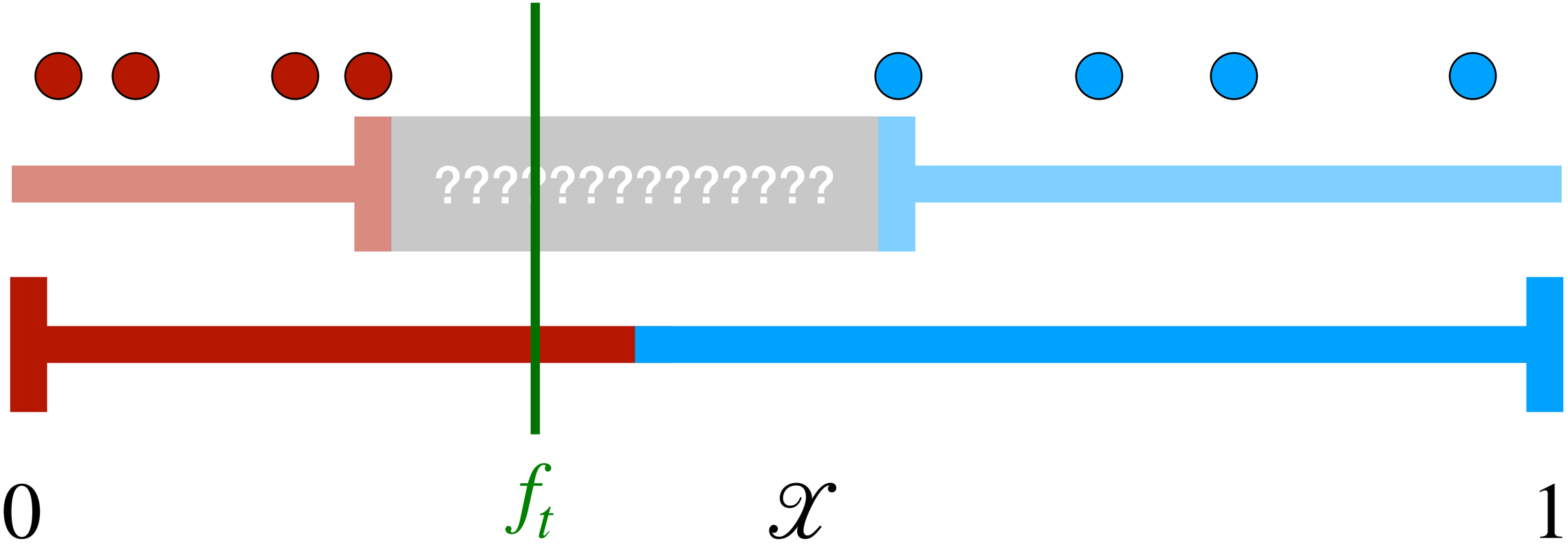
What Makes Online Learning Hard?



Why is Smoothness Helpful?



Why is Smoothness Helpful?



If gray region has length ε, w.p $\geq 1 - \varepsilon/\sigma$, new point not in gray region!

Why is Smoothness Helpful?

Theorem [BS'22]: If $\mathscr{F} = \left\{ x \mapsto \operatorname{sign}\left(\langle \theta, x \rangle\right) : \|\theta\| = 1 \right\}$ is linear thresholds, and data are **realizable** and **smooth** w.r.t. Lebesgue, then an **efficient** algorithm achieves

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{d \cdot \log\left(\frac{T}{\sigma}\right)}{T}.$$

Why is Smoothness Helpful?

Theorem [BS'22]: If $\mathscr{F} = \left\{ x \mapsto \operatorname{sign}\left(\langle \theta, x \rangle\right) : \|\theta\| = 1 \right\}$ is linear thresholds, and data are **realizable** and **smooth** w.r.t. Lebesgue, then an **efficient** algorithm achieves

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{d \cdot \log\left(\frac{T}{\sigma}\right)}{T}.$$

Algo works by choosing center of John Ellipsoid of feasible set.

Why is Smoothness Helpful?

Theorem [BS'22]: If $\mathscr{F} = \left\{ x \mapsto \operatorname{sign}\left(\langle \theta, x \rangle\right) : \|\theta\| = 1 \right\}$ is linear thresholds, and data are **realizable** and **smooth** w.r.t. Lebesgue, then an **efficient** algorithm achieves

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{d \cdot \log\left(\frac{T}{\sigma}\right)}{T}.$$

Algo works by choosing center of John Ellipsoid of feasible set.

Ensures constant fraction of wrong functions removed with each mistake.

Why is Smoothness Helpful?

Theorem [BS'22]: If $\mathscr{F} = \left\{ x \mapsto \operatorname{sign}\left(\langle \theta, x \rangle\right) : \|\theta\| = 1 \right\}$ is linear thresholds, and data are **realizable** and **smooth** w.r.t. Lebesgue, then an **efficient** algorithm achieves

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{d \cdot \log\left(\frac{T}{\sigma}\right)}{T}.$$

Algo works by choosing center of John Ellipsoid of feasible set.

Ensures constant fraction of wrong functions removed with each mistake.

Analysis uses duality between $\mathscr{F}$ and $\mathscr{X}$.

Performance of ERM (Square Loss)

Theorem [BBM'02, LRS'15]: If ℓ is square loss and $\hat{f}$ is an ERM, then

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\mathsf{vc}(\mathscr{F})}{T}.$$

Performance of ERM (Square Loss)

Theorem [BBM'02, LRS'15]: If ℓ is square loss and $\hat{f}$ is an ERM, then

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\mathsf{vc}(\mathscr{F})}{T}.$$

Statistical Learning: $\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{d}{T}$

Performance of ERM (Square Loss)

Theorem [BBM'02, LRS'15]: If ℓ is square loss and $\hat{f}$ is an ERM, then

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\mathrm{vc}(\mathscr{F})}{T}.$$

Statistical Learning: $\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{d}{T}$

Smoothed Online Learning: $\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{d \cdot \log(T/\sigma)}{T}$

Why is Smoothness Helpful?

Theorem [BS'22]: If $\mathscr{F} = \left\{ x \mapsto \operatorname{sign}\left(\langle \theta, x \rangle\right) : \|\theta\| = 1 \right\}$ is linear thresholds, and data are **realizable** and **smooth** w.r.t. Lebesgue, then an **efficient** algorithm achieves

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{d \cdot \log\left(\frac{T}{\sigma}\right)}{T}.$$

Pros:

End-to-end efficient algo.

Why is Smoothness Helpful?

Theorem [BS'22]: If $\mathscr{F} = \left\{ x \mapsto \operatorname{sign}\left(\langle \theta, x \rangle\right) : \|\theta\| = 1 \right\}$ is linear thresholds, and data are **realizable** and **smooth** w.r.t. Lebesgue, then an **efficient** algorithm achieves

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{d \cdot \log\left(\frac{T}{\sigma}\right)}{T}.$$

Pros:

- End-to-end efficient algo.
- Fast $(1/T)$ rate in error.

Why is Smoothness Helpful?

Theorem [BS'22]: If $\mathscr{F} = \left\{ x \mapsto \operatorname{sign}\left(\langle \theta, x \rangle\right) : \|\theta\| = 1 \right\}$ is linear thresholds, and data are **realizable** and **smooth** w.r.t. Lebesgue, then an **efficient** algorithm achieves

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{d \cdot \log\left(\frac{T}{\sigma}\right)}{T}.$$

Pros:

End-to-end efficient algo.

Fast $(1/T)$ rate in error.

Cons:

Requires realizable data.

Why is Smoothness Helpful?

Theorem [BS'22]: If $\mathscr{F} = \left\{ x \mapsto \operatorname{sign}\left(\langle \theta, x \rangle\right) : \|\theta\| = 1 \right\}$ is linear thresholds, and data are **realizable** and **smooth** w.r.t. Lebesgue, then an **efficient** algorithm achieves

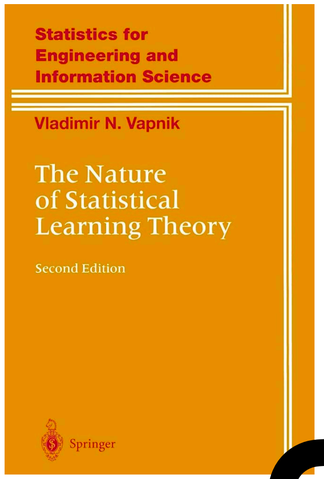
$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{d \cdot \log\left(\frac{T}{\sigma}\right)}{T}.$$

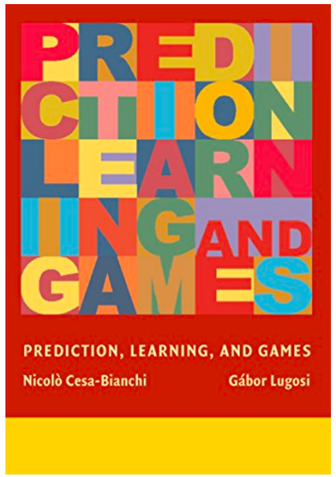
Pros:

- End-to-end efficient algo.
- Fast $(1/T)$ rate in error.

Cons:

- Requires realizable data.
- Only thresholds, Lebesgue.

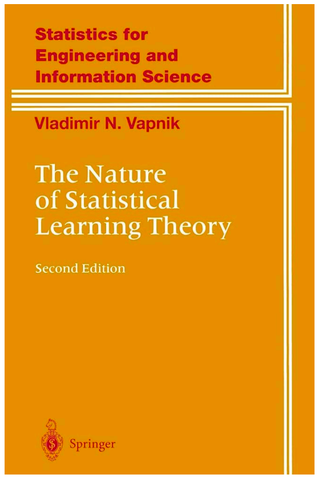




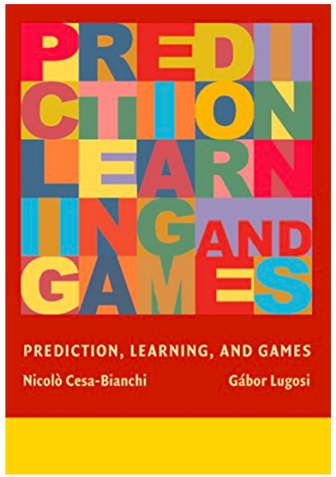
Smoothed data
(Linear Thresholds)

Statistical Learning

Online Learning



Smoothed data



??????????

Statistical Learning

Online Learning

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications
2. The Smoothed Model: Best of Both Worlds?
3. **The Power of Empirical Risk Minimization**

Part II

1. Coupling Lemma
2. Handling Label Noise: The Agnostic Setting
3. Oracle-Efficiency: ERM Returns

Tutorial Outline

Part I

3. The Power of Empirical Risk Minimization

Tutorial Outline

Part I

3. The Power of Empirical Risk Minimization

(a) Beyond Thresholds with the ERM

Tutorial Outline

Part I

3. The Power of Empirical Risk Minimization

(a) Beyond Thresholds with the ERM

(b) Key Analysis Techniques

Tutorial Outline

Part I

3. The Power of Empirical Risk Minimization

(a) Beyond Thresholds with the ERM

(b) Key Analysis Techniques

Statistical Learning

1. We get T **data points** (X_t, Y_t) such that $X_t \overset{\text{iid}}{\sim} \mu$ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: Return $\hat{f} \in \mathscr{F}$ with small **test loss** $\mathbb{E}\left[\frac{1}{T}\sum_{s=1}^{T} \ell(\hat{f}(X'_s), f^\star(X'_s))\right]$.

Empirical Risk Minimization

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathscr{F}} L_T(f)$$

$$L_T(f) = \frac{1}{T}\sum_{t=1}^{T} \ell(f(X_t), Y_t)$$

Online Empirical Risk Minimization

1. We get T data points X_t **smooth** w.r.t μ and $Y_t = f^\star(X_t) + \eta_t$.

Online Empirical Risk Minimization

1. We get T data points X_t **smooth** w.r.t μ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Online Empirical Risk Minimization

1. We get T data points X_t **smooth** w.r.t μ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

3. For each t let $f_t \in \operatorname*{argmin}_{f \in \mathscr{F}} L_{t-1}(f)$ with $L_{t-1}(f) = \dfrac{1}{t-1} \displaystyle\sum_{s=1}^{t-1} \ell(f(X_s), Y_s)$.

Online Empirical Risk Minimization

1. We get T data points X_t **smooth** w.r.t μ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

3. For each t let $f_t \in \operatorname*{argmin}_{f \in \mathscr{F}} L_{t-1}(f)$ with $L_{t-1}(f) = \dfrac{1}{t-1} \sum_{s=1}^{t-1} \ell(f(X_s), Y_s)$.

Goal: For each t, return $f_t \in \mathscr{F}$ with small **error**

$$\mathbb{E}\left[\mathrm{Err}_T\right] = \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \ell(f_t(X_t), f^\star(X_t))\right].$$

Online Empirical Risk Minimization

1. We get T data points X_t **smooth** w.r.t μ and $Y_t = f^\star(X_t) + \eta_t$.

2. We have access to a **model class** $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$.

3. For each t let $f_t \in \operatorname*{argmin}_{f \in \mathcal{F}} L_{t-1}(f)$ with $L_{t-1}(f) = \dfrac{1}{t-1} \sum_{s=1}^{t-1} \ell(f(X_s), Y_s)$.

Goal: For each t, return $f_t \in \mathcal{F}$ with small **error**

$$\mathbb{E}\left[\mathrm{Err}_T\right] = \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \ell(f_t(X_t), f^\star(X_t))\right].$$

ERM Performance

Theorem [BRS'24]: If data are σ-smooth w.r.t. μ and f_t is ERM, then

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\log(T/\sigma)}{\sigma \cdot T} + \sqrt{\frac{\mathrm{vc}(\mathscr{F}) \cdot \log(T/\sigma)}{\sigma \cdot T}}.$$

ERM Performance

Theorem [BRS'24]: If data are σ-smooth w.r.t. μ and f_t is ERM, then

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\log(T/\sigma)}{\sigma \cdot T} + \sqrt{\frac{\mathrm{vc}(\mathscr{F}) \cdot \log(T/\sigma)}{\sigma \cdot T}}.$$

Compare to IID, where $\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\mathrm{vc}(\mathscr{F})}{T}$.

ERM Performance

Theorem [BRS'24]: If data are σ-smooth w.r.t. μ and f_t is ERM, then

$$\mathbb{E}[\mathrm{Err}_T] \lesssim \frac{\log(T/\sigma)}{\sigma \cdot T} + \sqrt{\frac{\mathrm{vc}(\mathscr{F}) \cdot \log(T/\sigma)}{\sigma \cdot T}}.$$

Compare to IID, where $\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\mathrm{vc}(\mathscr{F})}{T}$.

ERM Performance

Theorem [BRS'24]: If data are σ-smooth w.r.t. μ and f_t is ERM, then

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\log(T/\sigma)}{\sigma \cdot T} + \sqrt{\frac{\mathrm{vc}(\mathscr{F}) \cdot \log(T/\sigma)}{\sigma \cdot T}}.$$

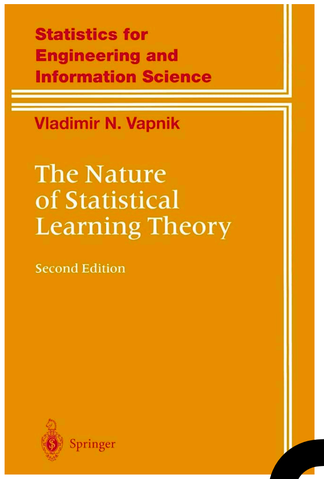
Theorem [BRS'24]: For all d there is $\mathscr{F}$ with $\mathrm{vc}(\mathscr{F}) \leq d$ and a **realizable** adversary such any algorithm (if μ is **unknown**) must pay

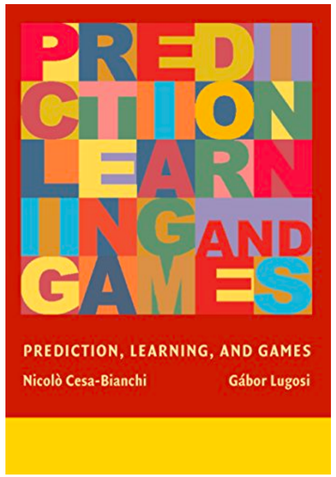
$$\mathbb{E}\left[\mathrm{Err}_T\right] \gtrsim \sqrt{\frac{d}{\sigma^{1/d} \cdot T}}.$$

ERM Performance

Smoothed Online Learning

$$\sqrt{\frac{\mathsf{vc}(\mathscr{F})}{\sigma^{1/\mathsf{vc}(\mathscr{F})} \cdot T}} \lesssim \mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \max\left(\sqrt{\frac{\mathsf{vc}(\mathscr{F}) \cdot \log(T/\sigma)}{\sigma \cdot T}}, \frac{\log(T/\sigma)}{\sigma \cdot T}\right).$$





Smoothed data

Statistical Learning

Online Learning

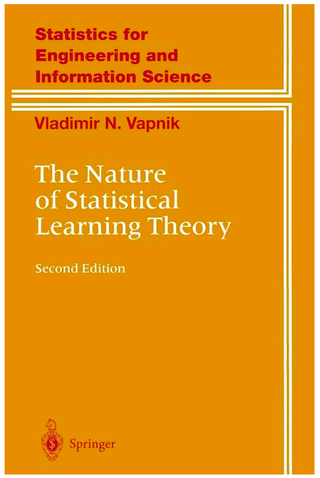
ERM Performance

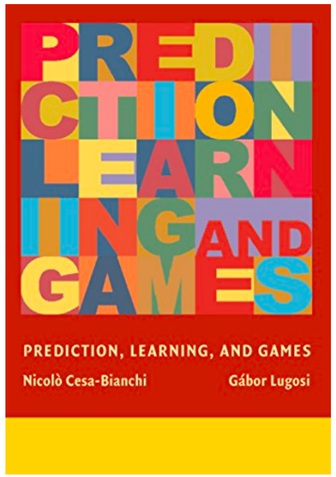
Smoothed Online Learning

$$\sqrt{\frac{\mathsf{vc}(\mathscr{F})}{\sigma^{1/\mathsf{vc}(\mathscr{F})} \cdot T}} \lesssim \mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \max\left(\sqrt{\frac{\mathsf{vc}(\mathscr{F}) \cdot \log(T/\sigma)}{\sigma \cdot T}}, \frac{\log(T/\sigma)}{\sigma \cdot T}\right).$$

Statistical Learning

$$\mathbb{E}\left[\mathrm{Err}_T\right] \asymp \frac{\mathsf{vc}(\mathscr{F})}{T}.$$





Smoothed data

Statistical Learning

Online Learning

Tutorial Outline

Part I

3. The Power of Empirical Risk Minimization

(a) Beyond Thresholds with the ERM

(b) Key Analysis Techniques

Tutorial Outline

Part I

3. The Power of Empirical Risk Minimization

(a) Beyond Thresholds with the ERM

(b) Key Analysis Techniques

(i) Overall Framework.

Tutorial Outline

Part I

3. The Power of Empirical Risk Minimization

(a) Beyond Thresholds with the ERM

(b) Key Analysis Techniques

(i) Overall Framework.

(ii) Key Technique 1: Surprise Lemma

Tutorial Outline

Part I

3. The Power of Empirical Risk Minimization

(a) Beyond Thresholds with the ERM

(b) Key Analysis Techniques

(i) Overall Framework.

(ii) Key Technique 1: Surprise Lemma

(iii) Key Technique 2: Coupling and Monotonicity

Tutorial Outline

Part I

3. The Power of Empirical Risk Minimization

(a) Beyond Thresholds with the ERM

(b) Key Analysis Techniques

(i) Overall Framework.

(ii) Key Technique 1: Surprise Lemma

(iii) Key Technique 2: Coupling and Monotonicity

ERM with Realizable, IID Data

For each t let $f_t \in \operatorname*{argmin}_{f \in \mathcal{F}} L_{t-1}(f)$ with $L_{t-1}(f) = \dfrac{1}{t-1} \sum_{s=1}^{t-1} \ell(f(X_s), f^\star(X_s))$.

ERM with Realizable, IID Data

For each t let $f_t \in \operatorname*{argmin}_{f \in \mathcal{F}} L_{t-1}(f)$ with $L_{t-1}(f) = \dfrac{1}{t-1} \sum_{s=1}^{t-1} \ell(f(X_s), f^\star(X_s))$.

If data IID, for fixed $f \in \mathcal{F}$, $\mathbb{E}\left[L_{t-1}(f)\right] = \mathbb{E}[\ell(f(X'_t), f^\star(X'_t))]$.

ERM with Realizable, IID Data

For each t let $f_t \in \operatorname*{argmin}_{f \in \mathcal{F}} L_{t-1}(f)$ with $L_{t-1}(f) = \dfrac{1}{t-1} \sum_{s=1}^{t-1} \ell(f(X_s), f^\star(X_s))$.

If data IID, for fixed $f \in \mathcal{F}$, $\mathbb{E}\left[L_{t-1}(f)\right] = \mathbb{E}[\ell(f(X'_t), f^\star(X'_t))]$.

If data IID, $\mathbb{E}\left[\sup_{g \in \mathcal{G}} \dfrac{1}{t} \sum_{s=1}^{t-1} g(X'_s)^2 - 2 \cdot g(X_s)^2\right] \lesssim \dfrac{\mathrm{comp}(\mathcal{G})}{t}$.

ERM with Realizable, IID Data

For each t let $f_t \in \operatorname*{argmin}_{f \in \mathcal{F}} L_{t-1}(f)$ with $L_{t-1}(f) = \dfrac{1}{t-1} \sum_{s=1}^{t-1} \ell(f(X_s), f^\star(X_s))$.

If data IID, for fixed $f \in \mathcal{F}$, $\mathbb{E}\left[L_{t-1}(f)\right] = \mathbb{E}[\ell(f(X'_t), f^\star(X'_t))]$.

If data IID, $\mathbb{E}\left[\sup_{g \in \mathcal{G}} \dfrac{1}{t} \sum_{s=1}^{t-1} g(X'_s)^2 - 2 \cdot g(X_s)^2\right] \lesssim \dfrac{\mathrm{comp}(\mathcal{G})}{t}$.

ERM with Realizable, IID Data

For each t let $f_t \in \operatorname*{argmin}_{f \in \mathcal{F}} L_{t-1}(f)$ with $L_{t-1}(f) = \dfrac{1}{t-1} \sum_{s=1}^{t-1} \ell(f(X_s), f^\star(X_s))$.

If data IID, for fixed $f \in \mathcal{F}$, $\mathbb{E}\left[L_{t-1}(f)\right] = \mathbb{E}[\ell(f(X'_t), f^\star(X'_t))]$.

If data IID, $\mathbb{E}\left[\sup_{g \in \mathcal{G}} \dfrac{1}{t} \sum_{s=1}^{t-1} g(X'_s)^2 - 2 \cdot g(X_s)^2\right] \lesssim \dfrac{\mathrm{comp}(\mathcal{G})}{t}$.

ERM with Realizable, IID Data

For each t let $f_t \in \operatorname*{argmin}_{f \in \mathcal{F}} L_{t-1}(f)$ with $L_{t-1}(f) = \dfrac{1}{t-1} \sum_{s=1}^{t-1} \ell(f(X_s), f^\star(X_s))$.

If data IID, for fixed $f \in \mathcal{F}$, $\mathbb{E}\left[L_{t-1}(f)\right] = \mathbb{E}[\ell(f(X'_t), f^\star(X'_t))]$.

If data IID, $\mathbb{E}\left[\sup_{g \in \mathcal{G}} \dfrac{1}{t} \sum_{s=1}^{t-1} g(X'_s)^2 - 2 \cdot g(X_s)^2\right] \lesssim \dfrac{\mathrm{comp}(\mathcal{G})}{t}$.

ERM with Realizable, IID Data

For each t let $f_t \in \operatorname*{argmin}_{f \in \mathcal{F}} L_{t-1}(f)$ with $L_{t-1}(f) = \dfrac{1}{t-1} \sum_{s=1}^{t-1} \ell(f(X_s), f^\star(X_s))$.

If data IID, for fixed $f \in \mathcal{F}$, $\mathbb{E}\left[L_{t-1}(f)\right] = \mathbb{E}[\ell(f(X'_t), f^\star(X'_t))]$.

If data IID, $\mathbb{E}\left[\sup_{f \in \mathcal{F}} \dfrac{1}{t} \sum_{s=1}^{t-1} (f - f^\star)^2(X'_s) - 2 \cdot (f - f^\star)^2(X_s)\right] \lesssim \dfrac{\mathrm{comp}(\mathcal{F})}{t}$.

ERM with Realizable, IID Data

For each t let $f_t \in \operatorname*{argmin}_{f \in \mathcal{F}} L_{t-1}(f)$ with $L_{t-1}(f) = \dfrac{1}{t-1} \sum_{s=1}^{t-1} \ell(f(X_s), f^\star(X_s))$.

If data IID, for fixed $f \in \mathcal{F}$, $\mathbb{E}\left[L_{t-1}(f)\right] = \mathbb{E}[\ell(f(X'_t), f^\star(X'_t))]$.

If data IID, $\mathbb{E}\left[\sup_{f \in \mathcal{F}} \dfrac{1}{t} \sum_{s=1}^{t-1} (f - f^\star)^2(X'_s) - 2 \cdot (f - f^\star)^2(X_s)\right] \lesssim \dfrac{\mathrm{comp}(\mathcal{F})}{t}$.

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(f_t(X'_t) - f^\star(X'_t))^2\right] \lesssim \mathbb{E}\left[\frac{2}{T}\sum_{t=1}^{T}(f_t(X_t) - f^\star(X_t))^2\right] + \frac{\mathrm{comp}(\mathcal{F}) \cdot \log(T)}{T}$$

ERM with Realizable, IID Data

For each t let $f_t \in \operatorname*{argmin}_{f \in \mathcal{F}} L_{t-1}(f)$ with $L_{t-1}(f) = \dfrac{1}{t-1} \sum_{s=1}^{t-1} \ell(f(X_s), f^\star(X_s))$.

If data IID, for fixed $f \in \mathcal{F}$, $\mathbb{E}\left[L_{t-1}(f)\right] = \mathbb{E}[\ell(f(X'_t), f^\star(X'_t))]$.

If data IID, $\mathbb{E}\left[\sup_{f \in \mathcal{F}} \dfrac{1}{t} \sum_{s=1}^{t-1} (f - f^\star)^2(X'_s) - 2 \cdot (f - f^\star)^2(X_s)\right] \lesssim \dfrac{\mathrm{comp}(\mathcal{F})}{t}$.

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(f_t(X'_t) - f^\star(X'_t))^2\right] \lesssim \mathbb{E}\left[\frac{2}{T}\sum_{t=1}^{T}(f^\star(X_t) - f^\star(X_t))^2\right] + \frac{\mathrm{comp}(\mathcal{F}) \cdot \log(T)}{T}$$

ERM with Realizable, IID Data

For each t let $f_t \in \operatorname*{argmin}_{f \in \mathcal{F}} L_{t-1}(f)$ with $L_{t-1}(f) = \dfrac{1}{t-1} \sum_{s=1}^{t-1} \ell(f(X_s), f^\star(X_s))$.

If data IID, for fixed $f \in \mathcal{F}$, $\mathbb{E}\left[L_{t-1}(f)\right] = \mathbb{E}[\ell(f(X'_t), f^\star(X'_t))]$.

If data IID, $\mathbb{E}\left[\sup_{f \in \mathcal{F}} \dfrac{1}{t} \sum_{s=1}^{t-1} (f - f^\star)^2(X'_s) - 2 \cdot (f - f^\star)^2(X_s)\right] \lesssim \dfrac{\mathrm{comp}(\mathcal{F})}{t}$.

$$\mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} (f_t(X'_s) - f^\star(X'_t))^2\right] \lesssim \frac{\mathrm{comp}(\mathcal{F}) \cdot \log(T)}{T}$$

Can we Extend to Smoothed Data?

Key facts used:

Can we Extend to Smoothed Data?

Key facts used:

If data IID, for fixed $f \in \mathscr{F}$, $\mathbb{E}\left[L_{t-1}(f)\right] = \mathbb{E}[\ell(f(X'_t), f^\star(X'_t))]$.

Can we Extend to Smoothed Data?

Key facts used:

If data IID, for fixed $f \in \mathscr{F}$, $\mathbb{E}\left[L_{t-1}(f)\right] = \mathbb{E}[\ell(f(X'_t), f^\star(X'_t))]$.

If data IID, $\mathbb{E}\left[\sup_{g \in \mathscr{G}} \frac{1}{t} \sum_{s=1}^{t-1} g(X'_s)^2 - 2 \cdot g(X_s)^2\right] \lesssim \frac{\mathrm{comp}(\mathscr{G})}{t}$.

Can we Extend to Smoothed Data?

Key facts used:

If data smooth, for fixed $f \in \mathscr{F}$, $\mathbb{E}\left[L_{t-1}(f)\right] \neq \mathbb{E}[\ell(f(X'_t), f^\star(X'_t))]$.

If data smooth, $\mathbb{E}\left[\sup_{g \in \mathscr{G}} \frac{1}{t} \sum_{s=1}^{t-1} g(X'_s)^2 - 2 \cdot g(X_s)^2\right] \lesssim \frac{\mathrm{comp}(\mathscr{G})}{t}$ **?**

Tutorial Outline

Part I

3. The Power of Empirical Risk Minimization

(a) Beyond Thresholds with the ERM

(b) Key Analysis Techniques

(i) Overall Framework.

(ii) Key Technique 1: Surprise Lemma

(iii) Key Technique 2: Coupling and Monotonicity

Can we Extend to Smoothed Data?

Key facts used:

If data smooth, for fixed $f \in \mathscr{F}$, $\mathbb{E}\left[L_{t-1}(f)\right] \neq \mathbb{E}[\ell(f(X'_t), f^\star(X'_t))]$.

If data smooth, $\mathbb{E}\left[\sup_{g \in \mathscr{G}} \frac{1}{t} \sum_{s=1}^{t-1} g(X'_s)^2 - 2 \cdot g(X_s)^2\right] \lesssim \frac{\mathrm{comp}(\mathscr{G})}{t}$?

Can we Extend to Smoothed Data?

Key facts used:

If data smooth, for fixed $f \in \mathscr{F}$, $\mathbb{E}\left[L_{t-1}(f)\right] \neq \mathbb{E}[\ell(f(X_t'), f^\star(X_t'))]$.

$$\hat{f}_t \in \operatorname*{argmin}_{f \in \mathscr{F}} L_{t-1}(f) \qquad L_{t-1}(f) = \frac{1}{t-1} \sum_{s=1}^{t-1} \ell(f(X_s), f^\star(X_s))$$

Can we Extend to Smoothed Data?

Key facts used:

If data smooth, for fixed $f \in \mathscr{F}$, $\mathbb{E}\left[L_{t-1}(f)\right] \neq \mathbb{E}[\ell(f(X'_t), f^\star(X'_t))]$.

$$\hat{f}_t \in \operatorname*{argmin}_{f \in \mathscr{F}} L_{t-1}(f) \qquad L_{t-1}(f) = \frac{1}{t-1}\sum_{s=1}^{t-1} \ell(f(X_s), f^\star(X_s))$$

$$\mathbb{E}\left[\mathrm{Err}_T\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(f_t(X_t) - f^\star(X_t))^2\right]$$

Can we Extend to Smoothed Data?

Key facts used:

If data smooth, for fixed $f \in \mathscr{F}$, $\mathbb{E}\left[L_{t-1}(f)\right] \neq \mathbb{E}[\ell(f(X'_t), f^\star(X'_t))]$.

$$\hat{f}_t \in \operatorname*{argmin}_{f \in \mathscr{F}} L_{t-1}(f) \qquad L_{t-1}(f) = \frac{1}{t-1} \sum_{t=1}^{t-1} \ell(f(X_s), f^\star(X_s))$$

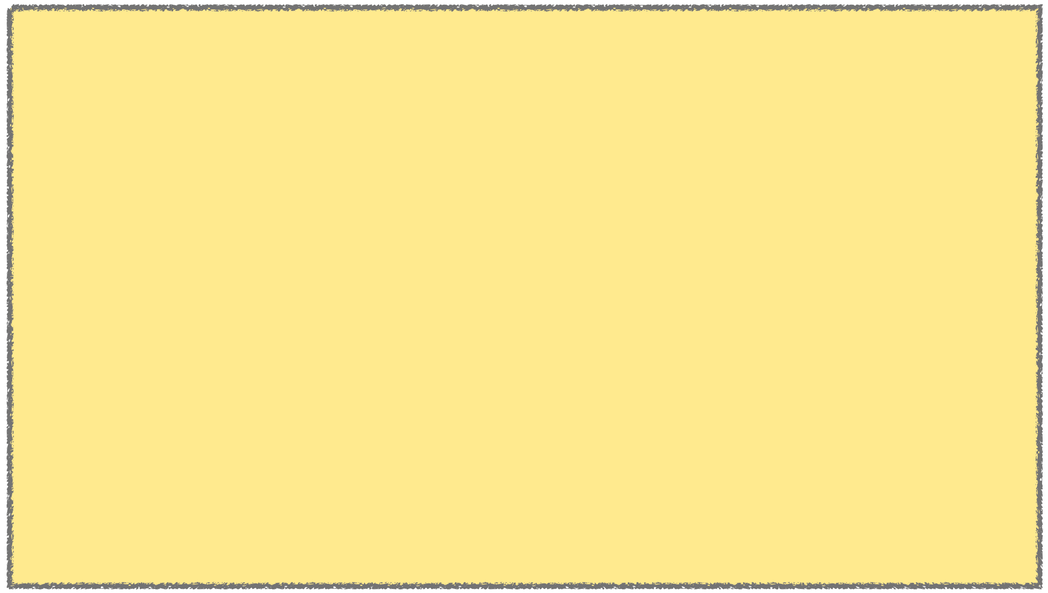
$$\mathbb{E}\left[\mathrm{Err}_T\right] = \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} (f_t(X_t) - f^\star(X_t))^2\right]$$

Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is **σ-smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$\mu = \text{Unif}(\mathcal{X})$

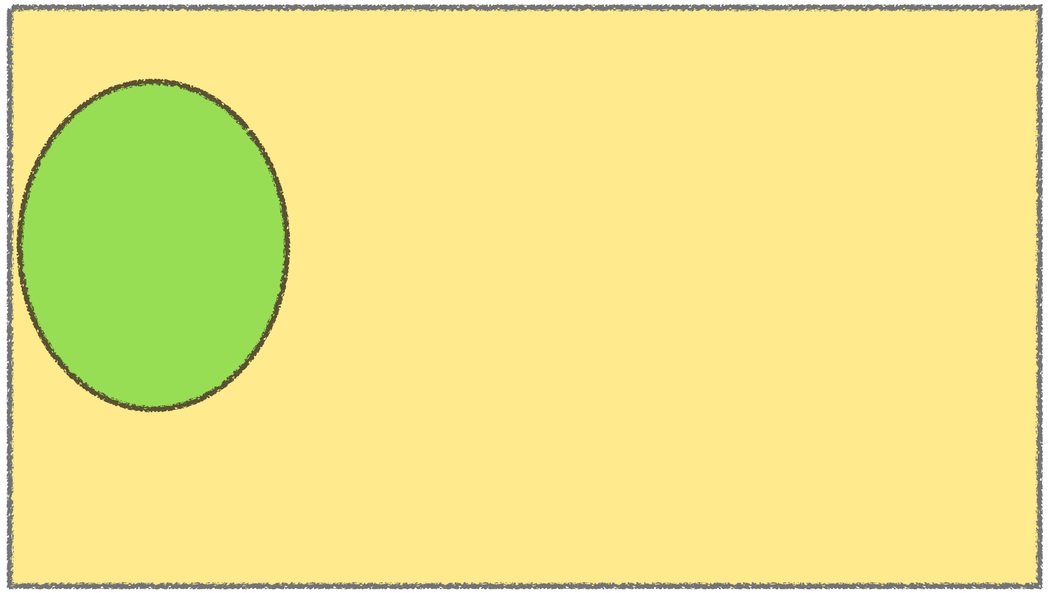


Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$$\mu = \mathrm{Unif}(\mathcal{X})$$

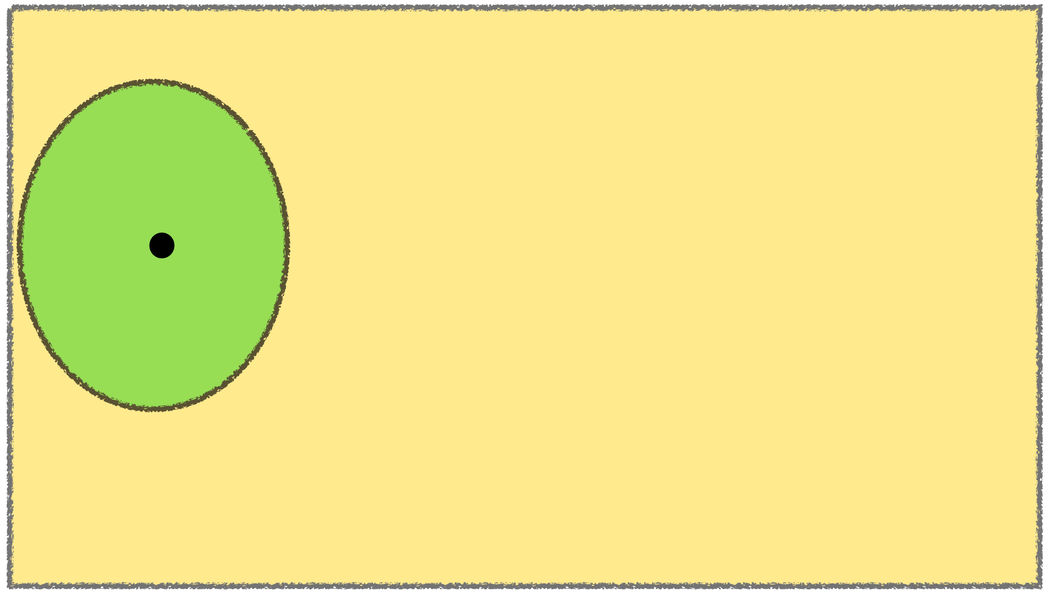


Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$\mu = \text{Unif}(\mathcal{X})$

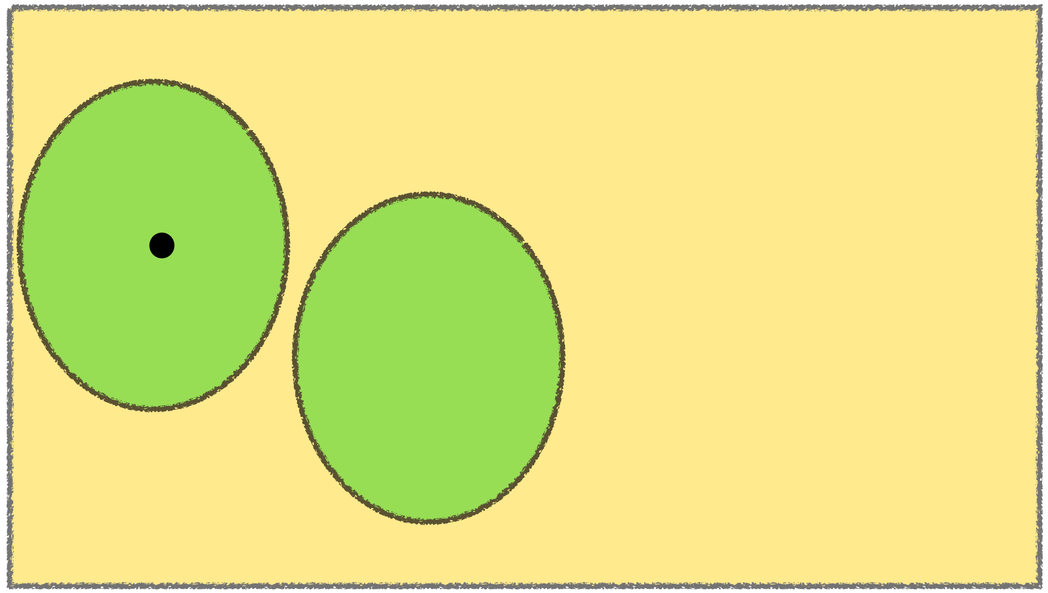


Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$$\mu = \text{Unif}(\mathcal{X})$$

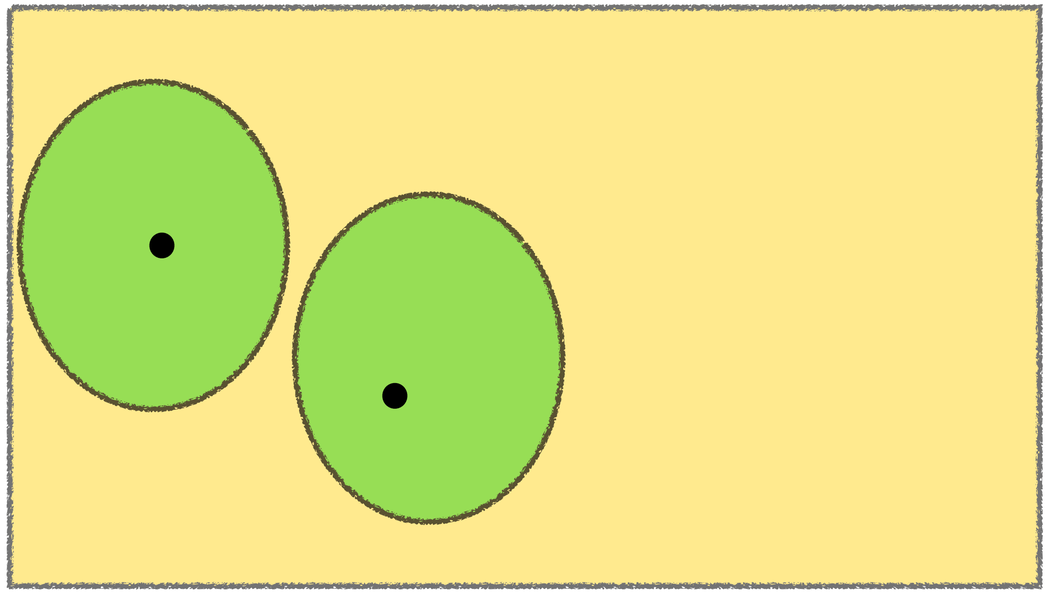


Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$$\mu = \text{Unif}(\mathcal{X})$$

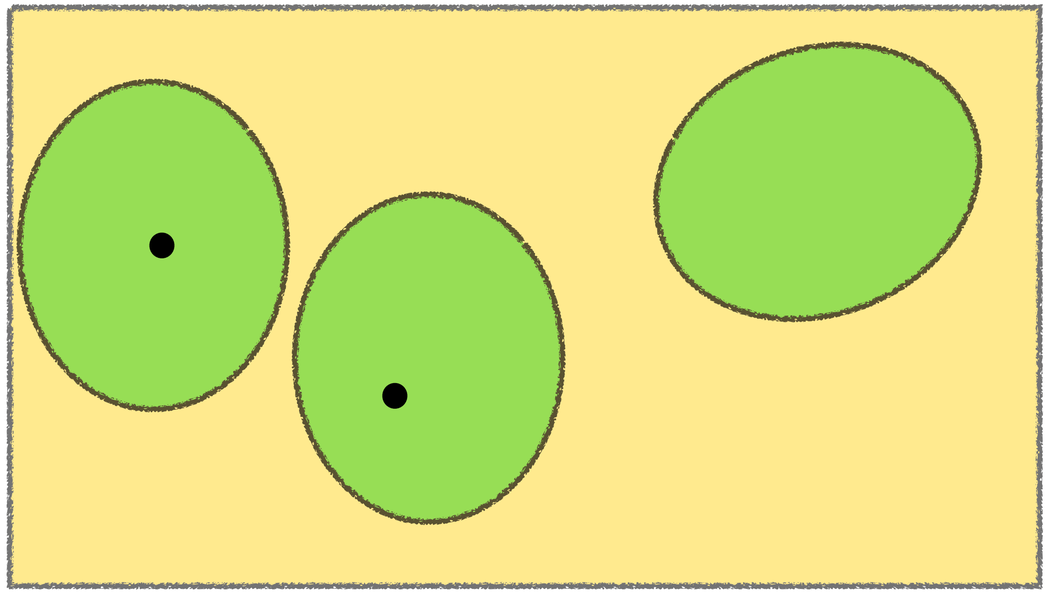


Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$$\mu = \mathrm{Unif}(\mathcal{X})$$

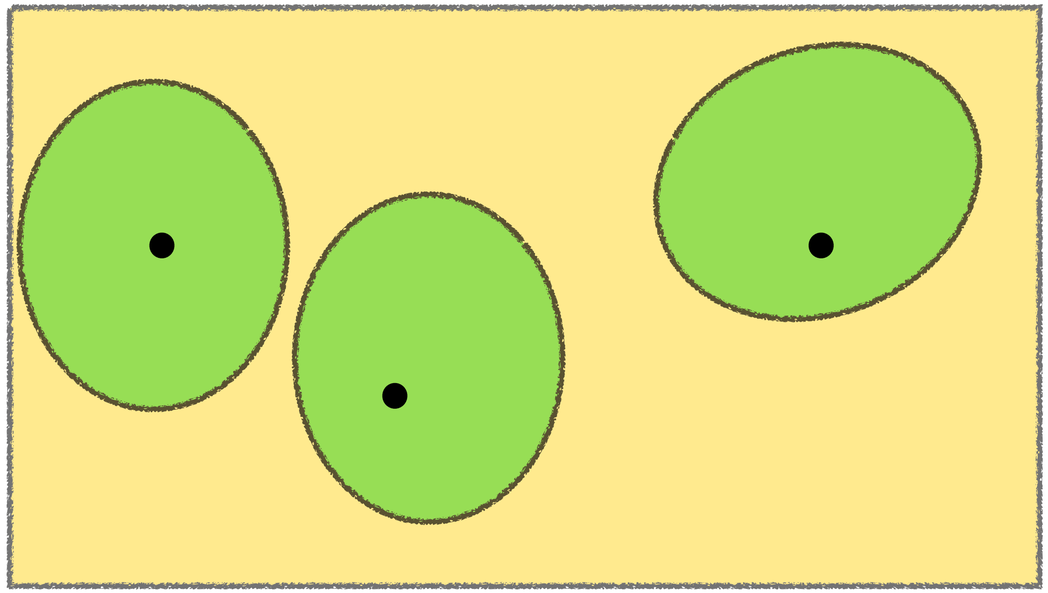


Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$$\mu = \mathrm{Unif}(\mathcal{X})$$

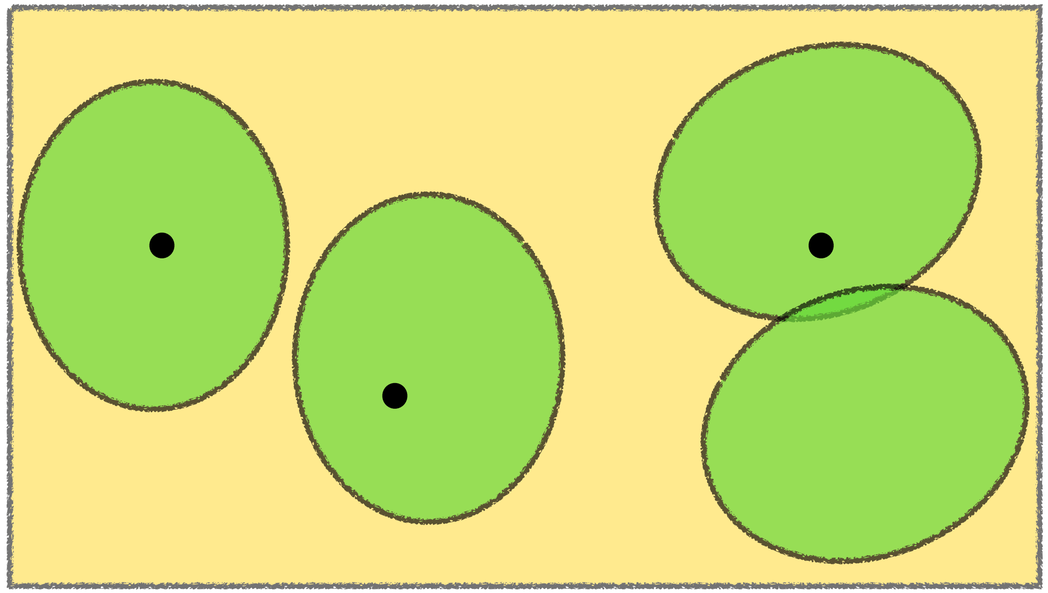


Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$$\mu = \text{Unif}(\mathcal{X})$$

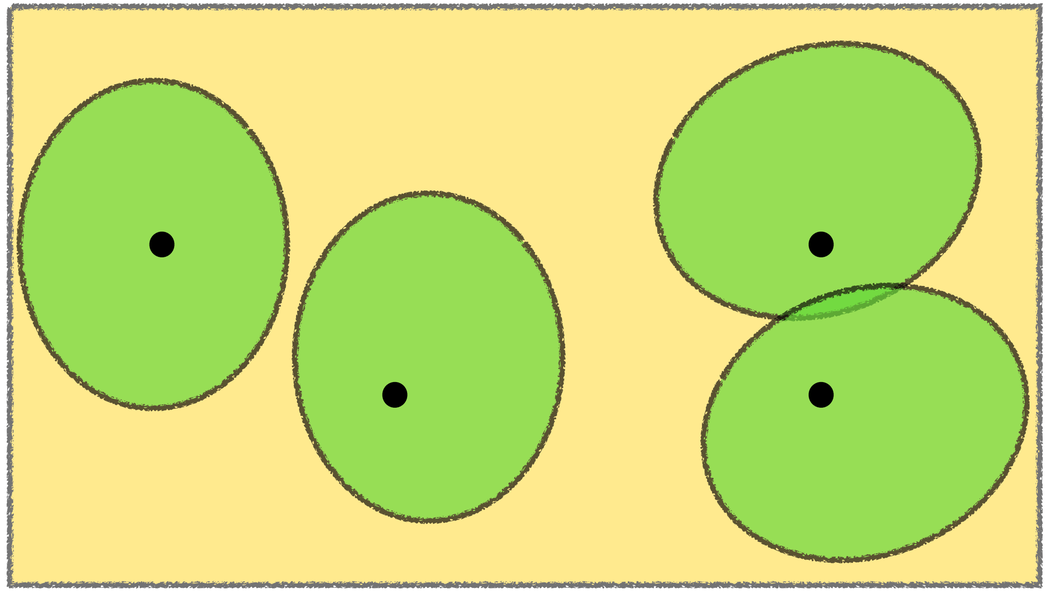


Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$$\mu = \text{Unif}(\mathcal{X})$$

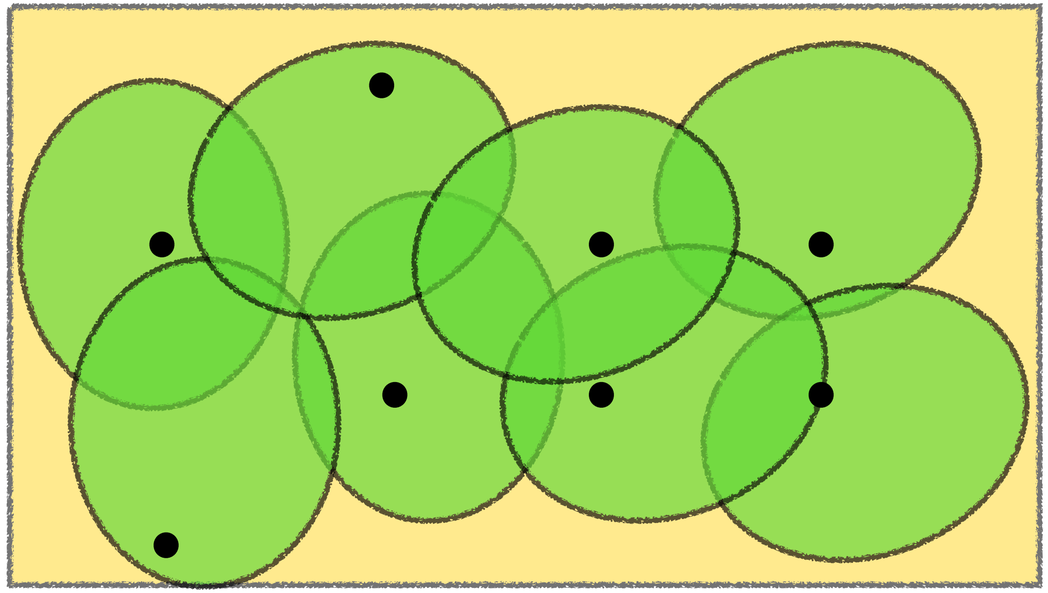


Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$$\mu = \mathrm{Unif}(\mathcal{X})$$



Smoothness Bounds Surprises

Lemma [BRS'24]: Let $p_1, \ldots, p_T$ be σ-smooth. Then for $\varepsilon > 0$ and $Z \in \mathcal{X}$,

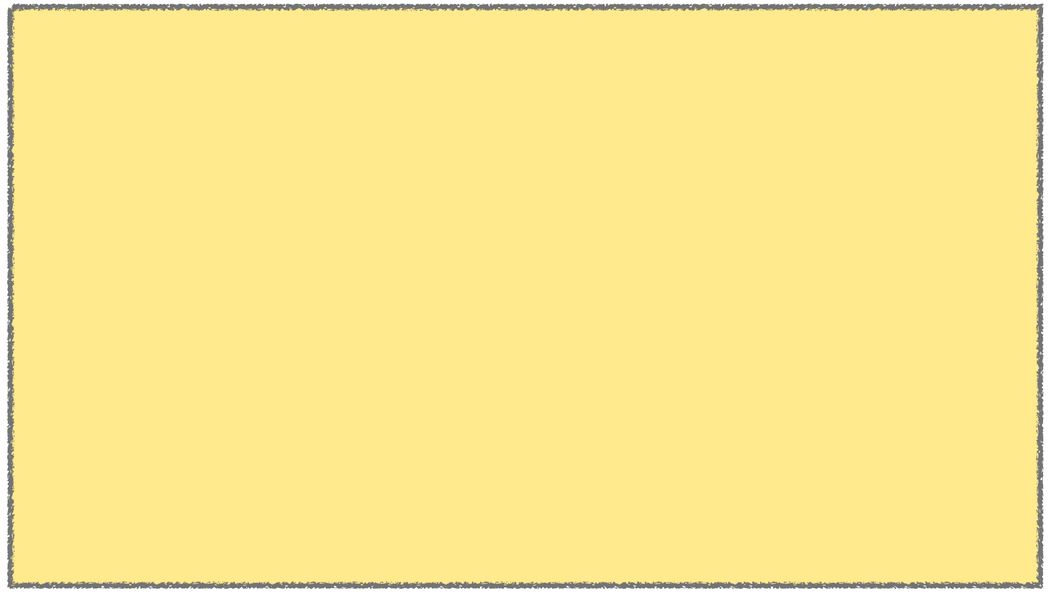
$$\left| \left\{ t \in [T] : p_t(Z) \geq \frac{2\log(T)}{\varepsilon} \cdot \frac{1}{t} \left(\frac{1}{\sigma} + \sum_{s=1}^{t-1} p_s(Z) \right) \right\} \right| \leq \varepsilon \cdot T.$$

Smoothness Bounds Surprises

Lemma [BRS'24]: Let $p_1, \ldots, p_T$ be σ-smooth. Then for $\varepsilon > 0$ and $Z \in \mathcal{X}$,

$$\left| \left\{ t \in [T] : p_t(Z) \geq \frac{2\log(T)}{\varepsilon} \cdot \frac{1}{t} \left(\frac{1}{\sigma} + \sum_{s=1}^{t-1} p_s(Z) \right) \right\} \right| \leq \varepsilon \cdot T.$$

$\mu = \mathrm{Unif}(\mathcal{X})$

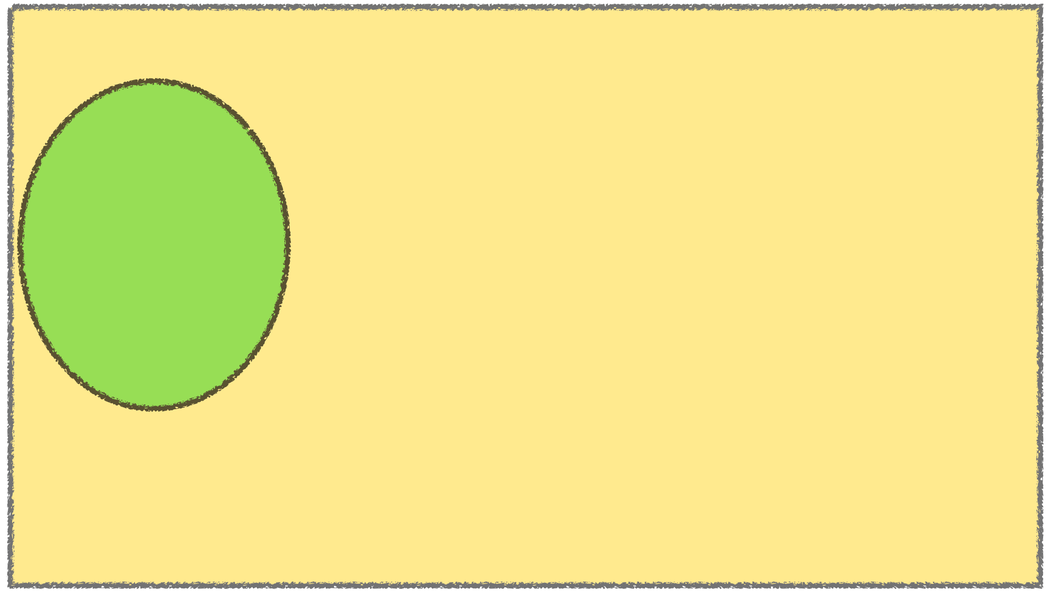


Smoothness Bounds Surprises

Lemma [BRS'24]: Let $p_1, \ldots, p_T$ be σ-smooth. Then for $\varepsilon > 0$ and $Z \in \mathcal{X}$,

$$\left| \left\{ t \in [T] : p_t(Z) \geq \frac{2\log(T)}{\varepsilon} \cdot \frac{1}{t} \left(\frac{1}{\sigma} + \sum_{s=1}^{t-1} p_s(Z) \right) \right\} \right| \leq \varepsilon \cdot T.$$

$\mu = \mathrm{Unif}(\mathcal{X})$

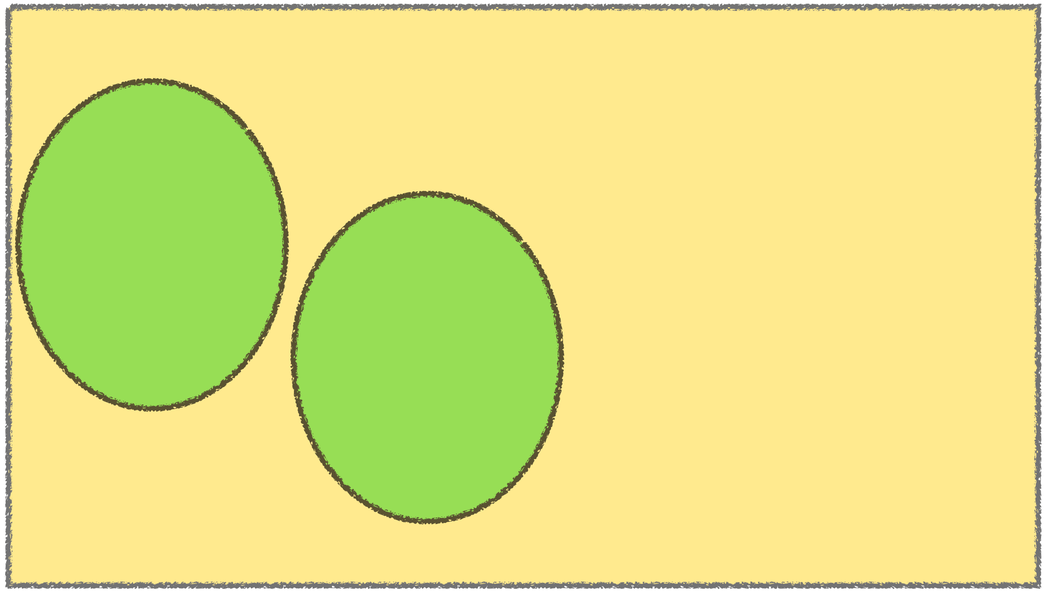


Smoothness Bounds Surprises

Lemma [BRS'24]: Let $p_1, \ldots, p_T$ be σ-smooth. Then for $\varepsilon > 0$ and $Z \in \mathcal{X}$,

$$\left| \left\{ t \in [T] : p_t(Z) \geq \frac{2\log(T)}{\varepsilon} \cdot \frac{1}{t} \left(\frac{1}{\sigma} + \sum_{s=1}^{t-1} p_s(Z) \right) \right\} \right| \leq \varepsilon \cdot T.$$

$$\mu = \mathrm{Unif}(\mathcal{X})$$

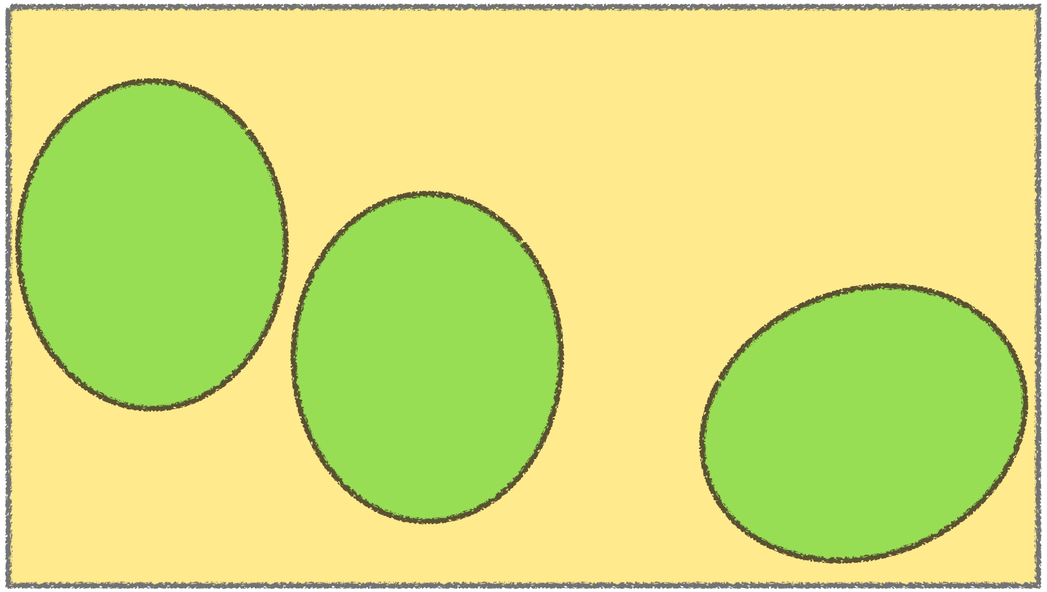


Smoothness Bounds Surprises

Lemma [BRS'24]: Let $p_1, \ldots, p_T$ be σ-smooth. Then for $\varepsilon > 0$ and $Z \in \mathcal{X}$,

$$\left| \left\{ t \in [T] : p_t(Z) \geq \frac{2\log(T)}{\varepsilon} \cdot \frac{1}{t} \left(\frac{1}{\sigma} + \sum_{s=1}^{t-1} p_s(Z) \right) \right\} \right| \leq \varepsilon \cdot T.$$

$$\mu = \mathrm{Unif}(\mathcal{X})$$

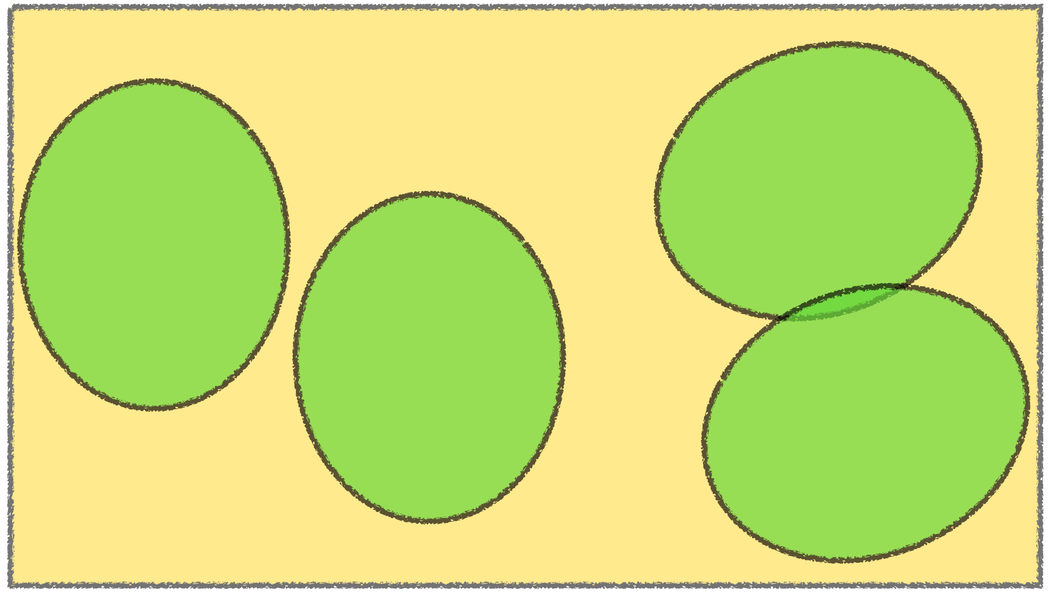


Smoothness Bounds Surprises

Lemma [BRS'24]: Let $p_1, \ldots, p_T$ be σ-smooth. Then for $\varepsilon > 0$ and $Z \in \mathcal{X}$,

$$\left| \left\{ t \in [T] : p_t(Z) \geq \frac{2\log(T)}{\varepsilon} \cdot \frac{1}{t} \left(\frac{1}{\sigma} + \sum_{s=1}^{t-1} p_s(Z) \right) \right\} \right| \leq \varepsilon \cdot T.$$

$$\mu = \mathrm{Unif}(\mathcal{X})$$

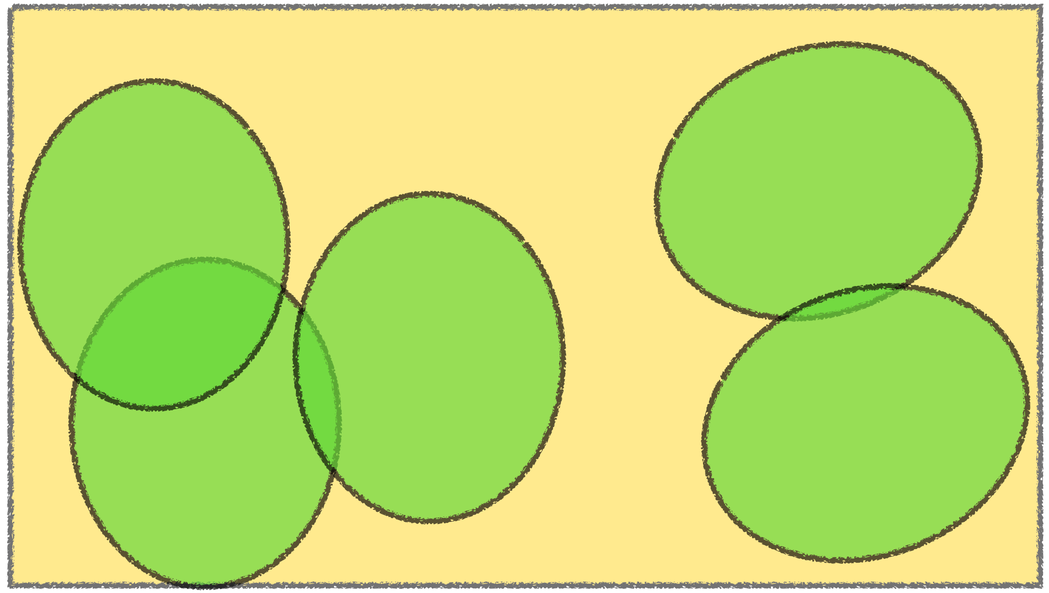


Smoothness Bounds Surprises

Lemma [BRS'24]: Let $p_1, \ldots, p_T$ be σ-smooth. Then for $\varepsilon > 0$ and $Z \in \mathcal{X}$,

$$\left| \left\{ t \in [T] : p_t(Z) \geq \frac{2\log(T)}{\varepsilon} \cdot \frac{1}{t} \left(\frac{1}{\sigma} + \sum_{s=1}^{t-1} p_s(Z) \right) \right\} \right| \leq \varepsilon \cdot T.$$

$$\mu = \mathrm{Unif}(\mathcal{X})$$

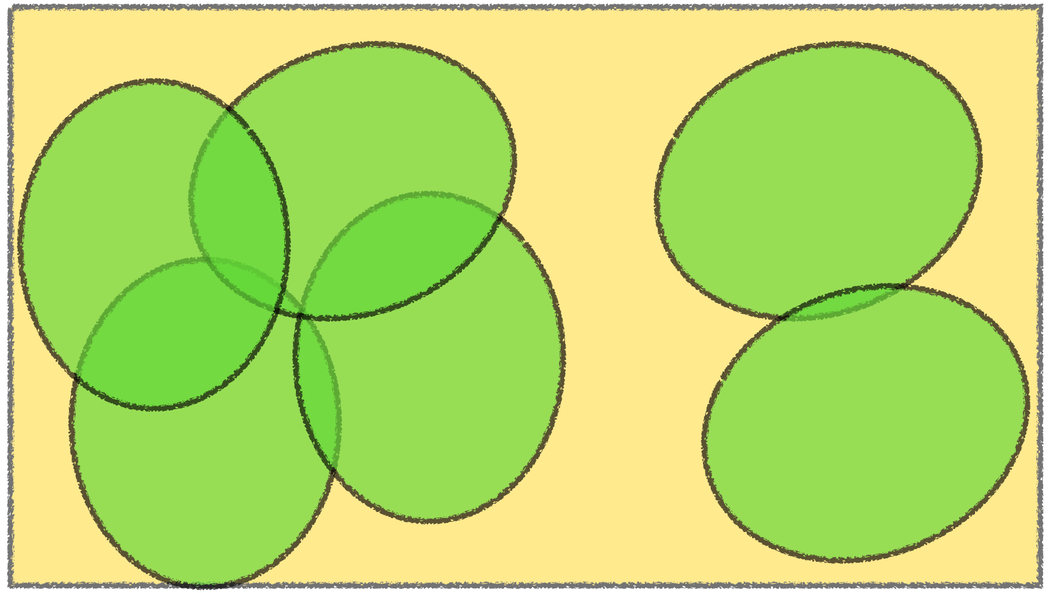


Smoothness Bounds Surprises

Lemma [BRS'24]: Let $p_1, \ldots, p_T$ be σ-smooth. Then for $\varepsilon > 0$ and $Z \in \mathcal{X}$,

$$\left| \left\{ t \in [T] : p_t(Z) \geq \frac{2\log(T)}{\varepsilon} \cdot \frac{1}{t} \left(\frac{1}{\sigma} + \sum_{s=1}^{t-1} p_s(Z) \right) \right\} \right| \leq \varepsilon \cdot T.$$

$$\mu = \mathrm{Unif}(\mathcal{X})$$

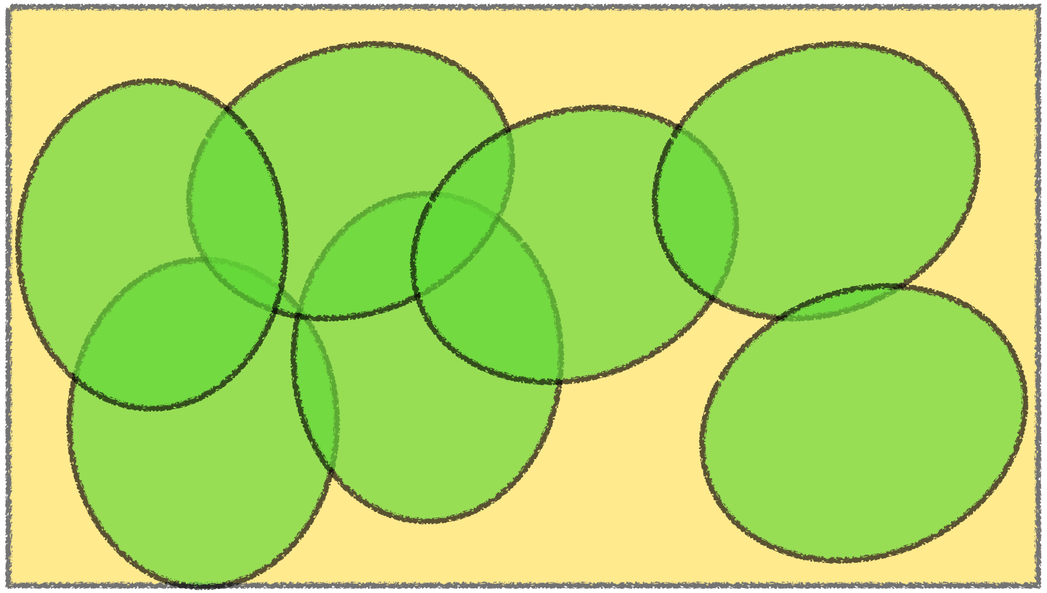


Smoothness Bounds Surprises

Lemma [BRS'24]: Let $p_1, \ldots, p_T$ be σ-smooth. Then for $\varepsilon > 0$ and $Z \in \mathcal{X}$,

$$\left| \left\{ t \in [T] : p_t(Z) \geq \frac{2\log(T)}{\varepsilon} \cdot \frac{1}{t} \left(\frac{1}{\sigma} + \sum_{s=1}^{t-1} p_s(Z) \right) \right\} \right| \leq \varepsilon \cdot T.$$

$$\mu = \mathrm{Unif}(\mathcal{X})$$

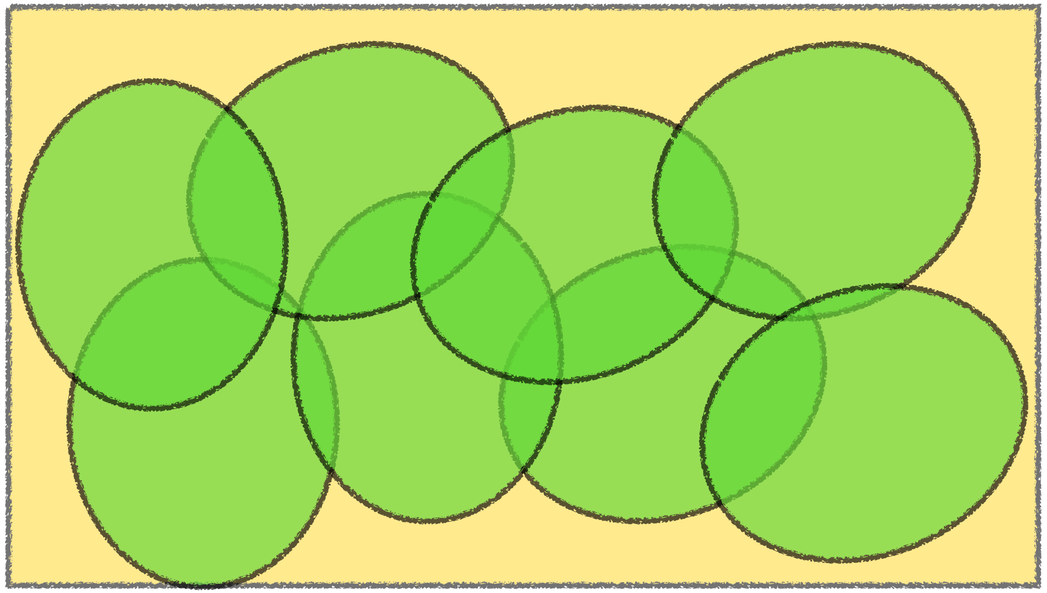


Smoothness Bounds Surprises

Lemma [BRS'24]: Let $p_1, \ldots, p_T$ be σ-smooth. Then for $\varepsilon > 0$ and $Z \in \mathcal{X}$,

$$\left| \left\{ t \in [T] : p_t(Z) \geq \frac{2\log(T)}{\varepsilon} \cdot \frac{1}{t} \left(\frac{1}{\sigma} + \sum_{s=1}^{t-1} p_s(Z) \right) \right\} \right| \leq \varepsilon \cdot T.$$

$$\mu = \mathrm{Unif}(\mathcal{X})$$



Smoothness Bounds Surprises

Lemma [BRS'24]: Let $p_1, \ldots, p_T$ be σ-smooth. Then for $\varepsilon > 0$ and $Z \in \mathcal{X}$,

$$\left| \left\{ t \in [T] : p_t(Z) \geq \frac{2\log(T)}{\varepsilon} \cdot \frac{1}{t} \left(\frac{1}{\sigma} + \sum_{s=1}^{t-1} p_s(Z) \right) \right\} \right| \leq \varepsilon \cdot T.$$

Let $\overline{p}_t = \frac{1}{t} \sum_{s=1}^{t} p_s$.

Smoothness Bounds Surprises

Lemma [BRS'24]: Let $p_1, \ldots, p_T$ be σ-smooth. Then for $\varepsilon > 0$ and $Z \in \mathcal{X}$,

$$\left| \left\{ t \in [T] : p_t(Z) \geq \frac{2\log(T)}{\varepsilon} \cdot \frac{1}{t} \left(\frac{1}{\sigma} + \sum_{s=1}^{t-1} p_s(Z) \right) \right\} \right| \leq \varepsilon \cdot T.$$

$$\text{Let } \overline{p}_t = \frac{1}{t} \sum_{s=1}^{t} p_s.$$

$$\text{Then, } p_t \lesssim \frac{\log(T)}{\sigma \cdot t} + \log(T) \cdot \overline{p}_{t-1} \text{ for most } t.$$

Smoothness Bounds Surprises

Lemma [BRS'24]: Let $p_1, \ldots, p_T$ be σ-smooth. Then for $\varepsilon > 0$ and $Z \in \mathcal{X}$,

$$\left| \left\{ t \in [T] : p_t(Z) \geq \frac{2\log(T)}{\varepsilon} \cdot \frac{1}{t} \left(\frac{1}{\sigma} + \sum_{s=1}^{t-1} p_s(Z) \right) \right\} \right| \leq \varepsilon \cdot T.$$

$$\text{Let } \overline{p}_t = \frac{1}{t} \sum_{s=1}^{t} p_s.$$

$$\text{Then, } p_t \lesssim \frac{\log(T)}{\sigma \cdot t} + \log(T) \cdot \overline{p}_{t-1} \text{ for most } t.$$

Smoothness Bounds Surprises

Lemma [BRS'24]: Let $p_1, \ldots, p_T$ be σ-smooth. Then for $\varepsilon > 0$ and $Z \in \mathcal{X}$,

$$\left| \left\{ t \in [T] : p_t(Z) \geq \frac{2\log(T)}{\varepsilon} \cdot \frac{1}{t} \left(\frac{1}{\sigma} + \sum_{s=1}^{t-1} p_s(Z) \right) \right\} \right| \leq \varepsilon \cdot T.$$

$$\mathbb{E}\left[(f_t(X_t) - f^\star(X_t))^2 \right] \lesssim \mathbb{E}\left[\frac{1}{t} \sum_{s=1}^{t-1} (f_t(X'_s) - f^\star(X'_s))^2 \right]$$

For most t!

Smoothness Bounds Surprises

Lemma [BRS'24]: Let $p_1, \ldots, p_T$ be σ-smooth. Then for $\varepsilon > 0$ and $Z \in \mathcal{X}$,

$$\left| \left\{ t \in [T] : p_t(Z) \geq \frac{2\log(T)}{\varepsilon} \cdot \frac{1}{t} \left(\frac{1}{\sigma} + \sum_{s=1}^{t-1} p_s(Z) \right) \right\} \right| \leq \varepsilon \cdot T.$$

Corollary [BRS'24]: Let $X_1, \ldots, X_T$ be σ-smooth and let f_t be **predictable**. Then,

$$\mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} (f_t(X_t) - f^\star(X_t))^2 \right] \lesssim \frac{\log(T)}{\sigma \cdot T} + \sqrt{\frac{1}{\sigma \cdot T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{t} \sum_{s=1}^{t-1} (f_t(X'_s) - f^\star(X'_s))^2 \right]}$$

Smoothness Bounds Surprises

$$\hat{f}_t \in \operatorname*{argmin}_{f \in \mathcal{F}} L_{t-1}(f) \qquad L_{t-1}(f) = \frac{1}{t-1} \sum_{t=1}^{t-1} \ell(f(X_s), f^\star(X_s))$$

$$\mathbb{E}\left[\mathrm{Err}_T\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} (f_t(X_t) - f^\star(X_t))^2\right]$$

Corollary [BRS'24]: Let $X_1, \ldots, X_T$ be σ-smooth and let f_t be predictable. Then,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} (f_t(X_t) - f^\star(X_t))^2\right] \lesssim \frac{\log(T)}{\sigma \cdot T} + \sqrt{\frac{1}{\sigma \cdot T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{t} \sum_{s=1}^{t-1} (f_t(X'_s) - f^\star(X'_s))^2\right]}$$

Smoothness Bounds Surprises

$$\hat{f}_t \in \operatorname*{argmin}_{f \in \mathcal{F}} L_{t-1}(f) \qquad L_{t-1}(f) = \frac{1}{t-1} \sum_{t=1}^{t-1} \ell(f(X_s), f^\star(X_s))$$

$$\mathbb{E}\left[\mathrm{Err}_T\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} (f_t(X_t) - f^\star(X_t))^2\right]$$

Corollary [BRS'24]: Let $X_1, \ldots, X_T$ be σ-smooth and let f_t be predictable. Then,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} (f_t(X_t) - f^\star(X_t))^2\right] \lesssim \frac{\log(T)}{\sigma \cdot T} + \sqrt{\frac{1}{\sigma \cdot T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{t} \sum_{s=1}^{t-1} (f_t(X'_s) - f^\star(X'_s))^2\right]}$$

Smoothness Bounds Surprises

$$\hat{f}_t \in \operatorname*{argmin}_{f \in \mathcal{F}} L_{t-1}(f) \qquad L_{t-1}(f) = \frac{1}{t-1} \sum_{t=1}^{t-1} \ell(f(X_s), f^\star(X_s))$$

$$\mathbb{E}\left[\mathrm{Err}_T\right] = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} (f_t(X_t) - f^\star(X_t))^2\right]$$

Corollary [BRS'24]: Let $X_1, \ldots, X_T$ be σ-smooth and let f_t be predictable. Then,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} (f_t(X_t) - f^\star(X_t))^2\right] \lesssim \frac{\log(T)}{\sigma \cdot T} + \sqrt{\frac{1}{\sigma \cdot T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \frac{1}{t} \sum_{s=1}^{t-1} (f_t(X'_s) - f^\star(X'_s))^2\right]}$$

Smoothness Bounds Surprises

Definition: For $X_1, \ldots, X_T$, a tangent sequence is some $X'_1, \ldots, X'_T$ such that for all $t \in [T]$,

$$X_t, X'_t \overset{iid}{\sim} p_t \,|\, \text{history}.$$

Corollary [BRS'24]: Let $X_1, \ldots, X_T$ be σ-smooth and let f_t be **predictable**. Then,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(f_t(X_t) - f^\star(X_t))^2\right] \lesssim \frac{\log(T)}{\sigma \cdot T} + \sqrt{\frac{1}{\sigma \cdot T} \cdot \mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{t}\sum_{s=1}^{t-1}(f_t(X'_s) - f^\star(X'_s))^2\right]}$$

Can we Extend to Smoothed Data?

Key facts used:

If data smooth, for fixed $f \in \mathscr{F}$, $\mathbb{E}\left[L_{t-1}(f)\right] \neq \mathbb{E}[\ell(f(X'_t), f^\star(X'_t))]$.

If data smooth, $\mathbb{E}\left[\sup_{g \in \mathscr{G}} \frac{1}{t} \sum_{s=1}^{t-1} g(X'_s)^2 - 2 \cdot g(X_s)^2\right] \lesssim \frac{\mathrm{comp}(\mathscr{G})}{t}$?

Can we Extend to Smoothed Data?

Key facts used:

If data smooth, for fixed $f \in \mathscr{F}$, $\mathbb{E}[\ell(f(X_t), f^\star(X_t))] \lesssim \mathbb{E}\left[L'_{t-1}(f)\right]$.

If data smooth, $\mathbb{E}\left[\sup_{g \in \mathscr{G}} \frac{1}{t} \sum_{s=1}^{t-1} g(X'_s)^2 - 2 \cdot g(X_s)^2\right] \lesssim \frac{\mathrm{comp}(\mathscr{G})}{t}$ **?**

Tutorial Outline

Part I

3. The Power of Empirical Risk Minimization

(a) Beyond Thresholds with the ERM

(b) Key Analysis Techniques

(i) Overall Framework.

(ii) Key Technique 1: Surprise Lemma

(iii) Key Technique 2: Coupling and Monotonicity

Smoothness Bounds Surprises

Definition: For $X_1, \ldots, X_T$, a tangent sequence is some $X'_1, \ldots, X'_T$ such that for all $t \in [T]$,

$$X_t, X'_t \overset{iid}{\sim} p_t \,|\, \text{history}.$$

Corollary [BRS'24]: Let $X_1, \ldots, X_T$ be σ-smooth and let f_t be **predictable**. Then,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(f_t(X_t) - f^\star(X_t))^2\right] \lesssim \frac{\log(T)}{\sigma \cdot T} + \sqrt{\frac{1}{\sigma \cdot T} \cdot \mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{t}\sum_{s=1}^{t-1}(f_t(X'_s) - f^\star(X'_s))^2\right]}$$

Can we Extend to Smoothed Data?

Key facts used:

If data smooth, for fixed $f \in \mathscr{F}$, $\mathbb{E}[\ell(f(X_t), f^\star(X_t))] \lesssim \mathbb{E}\left[L'_{t-1}(f)\right]$.

If data smooth, $\mathbb{E}\left[\sup_{g \in \mathscr{G}} \frac{1}{t} \sum_{s=1}^{t-1} g(X'_s)^2 - 2 \cdot g(X_s)^2\right] \lesssim \frac{\mathrm{comp}(\mathscr{G})}{t}$ **?**

If we Controlled Generalization Error

Corollary [BRS'24]: Let $X_1, \ldots, X_T$ be σ-smooth and let f_t be **predictable**. Then,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(f_t(X_t)-f^\star(X_t))^2\right] \lesssim \frac{\log(T)}{\sigma\cdot T}+\sqrt{\frac{1}{\sigma\cdot T}\cdot\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{t}\sum_{s=1}^{t-1}(f_t(X'_s)-f^\star(X'_s))^2\right]}$$

If $\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t-1}g(X'_s)^2-2\cdot g(X_s)^2\right] \lesssim \frac{\mathrm{comp}(\mathcal{G})}{t}$,

If we Controlled Generalization Error

Corollary [BRS'24]: Let $X_1, \ldots, X_T$ be σ-smooth and let f_t be **predictable**. Then,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(f_t(X_t)-f^\star(X_t))^2\right] \lesssim \frac{\log(T)}{\sigma\cdot T} + \sqrt{\frac{1}{\sigma\cdot T}\cdot \mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{t}\sum_{s=1}^{t-1}(f_t(X'_s)-f^\star(X'_s))^2\right]}$$

$$\mathbb{E}\left[\frac{1}{t}\sum_{s=1}^{t-1}(f_t(X'_s)-f^\star(X'_s))^2\right] \lesssim \mathbb{E}\left[\frac{1}{t}\sum_{s=1}^{t-1}(f_t(X_s)-f^\star(X_s))^2\right] + \frac{\mathrm{comp}(\mathscr{F})}{t}$$

If we Controlled Generalization Error

Corollary [BRS'24]: Let $X_1, \ldots, X_T$ be σ-smooth and let f_t be predictable. Then,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(f_t(X_t)-f^\star(X_t))^2\right] \lesssim \frac{\log(T)}{\sigma\cdot T}+\sqrt{\frac{1}{\sigma\cdot T}\cdot\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{t}\sum_{s=1}^{t-1}(f_t(X'_s)-f^\star(X'_s))^2\right]}$$

$$\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{t}\sum_{s=1}^{t-1}(f_t(X'_s)-f^\star(X'_s))^2\right] \lesssim \mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{t}\sum_{s=1}^{t-1}(f_t(X_s)-f^\star(X_s))^2\right] + \mathrm{comp}(\mathscr{F})\cdot\log(T)$$

If we Controlled Generalization Error

Corollary [BRS'24]: Let $X_1, \ldots, X_T$ be σ-smooth and let f_t be **predictable**. Then,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(f_t(X_t)-f^\star(X_t))^2\right] \lesssim \frac{\log(T)}{\sigma\cdot T}+\sqrt{\frac{1}{\sigma\cdot T}\cdot\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{t}\sum_{s=1}^{t-1}(f_t(X'_s)-f^\star(X'_s))^2\right]}$$

$$\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{t}\sum_{s=1}^{t-1}(f_t(X'_s)-f^\star(X'_s))^2\right] \lesssim \mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{t}\sum_{s=1}^{t-1}(f^\star(X_t)-f^\star(X_s))^2\right] + \mathrm{comp}(\mathscr{F})\cdot\log(T)$$

If we Controlled Generalization Error

Corollary [BRS'24]: Let $X_1, \ldots, X_T$ be σ-smooth and let f_t be predictable. Then,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(f_t(X_t)-f^\star(X_t))^2\right] \lesssim \frac{\log(T)}{\sigma\cdot T}+\sqrt{\frac{1}{\sigma\cdot T}\cdot\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{t}\sum_{s=1}^{t-1}(f_t(X'_s)-f^\star(X'_s))^2\right]}$$

$$\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{t}\sum_{s=1}^{t-1}(f_t(X'_s)-f^\star(X'_s))^2\right] \lesssim \mathbb{E}\left[\sum_{t=1}^{T}(f_t(X_t)-f^\star(X_t))^2\right]+\mathrm{comp}(\mathscr{F})\cdot\log(T)$$

If we Controlled Generalization Error

Corollary [BRS'24]: Let $X_1, \ldots, X_T$ be σ-smooth and let f_t be **predictable**. Then,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(f_t(X_t) - f^\star(X_t))^2\right] \lesssim \frac{\log(T)}{\sigma \cdot T} + \sqrt{\frac{1}{\sigma \cdot T} \cdot \mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{t}\sum_{s=1}^{t-1}(f_t(X'_s) - f^\star(X'_s))^2\right]}$$

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{t}\sum_{s=1}^{t-1}(f_t(X'_s) - f^\star(X'_s))^2\right] \lesssim \frac{\mathrm{comp}(\mathscr{F}) \cdot \log(T)}{T}$$

If we Controlled Generalization Error

Corollary [BRS'24]: Let $X_1, \ldots, X_T$ be σ-smooth and let f_t be **predictable**. Then,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(f_t(X_t)-f^\star(X_t))^2\right] \lesssim \frac{\log(T)}{\sigma\cdot T}+\sqrt{\frac{1}{\sigma\cdot T}\cdot\mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{t}\sum_{s=1}^{t-1}(f_t(X'_s)-f^\star(X'_s))^2\right]}$$

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(f_t(X_t)-f^\star(X_t))^2\right] \lesssim \frac{\log(T)}{\sigma\cdot T}+\sqrt{\frac{\mathrm{comp}(\mathscr{F})\cdot\log(T)}{\sigma\cdot T}}$$

Can we Extend to Smoothed Data?

Key facts used:

If data smooth, for fixed $f \in \mathscr{F}$, $\mathbb{E}\left[L_{t-1}(f)\right] \neq \mathbb{E}[\ell(f(X'_t), f^\star(X'_t))]$.

If data smooth, $\mathbb{E}\left[\sup_{g \in \mathscr{G}} \frac{1}{t} \sum_{s=1}^{t-1} g(X'_s)^2 - 2 \cdot g(X_s)^2\right] \lesssim \frac{\mathrm{comp}(\mathscr{G})}{t}$ **?**

Smoothness Allows Coupling

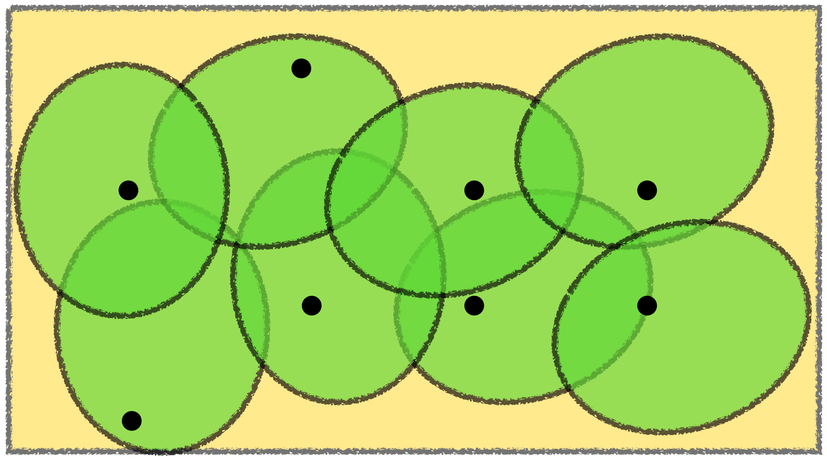
Lemma [HRS'21, BDGR'22]: For all t, there is a coupling between X_t and $Z_{t,1}, \ldots, Z_{t,k} \overset{\text{iid}}{\sim} \mu$ such that w.p. at least $1 - e^{-\sigma k}$, it holds that

$$X_t \in \{Z_{t,1}, \ldots, Z_{t,k}\}.$$

Smoothness Allows Coupling

Lemma [HRS'21, BDGR'22]: For all t, there is a coupling between X_t and $Z_{t,1}, \ldots, Z_{t,k} \overset{\text{iid}}{\sim} \mu$ such that w.p. at least $1 - e^{-\sigma k}$, it holds that

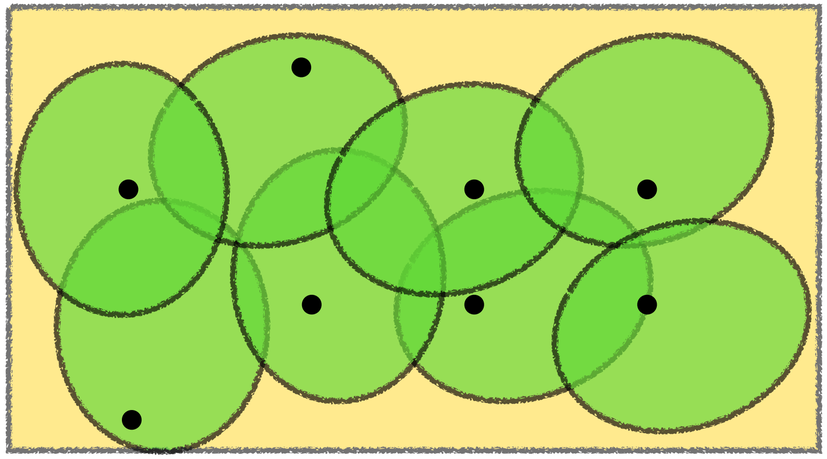
$$X_t \in \{Z_{t,1}, \ldots, Z_{t,k}\}.$$

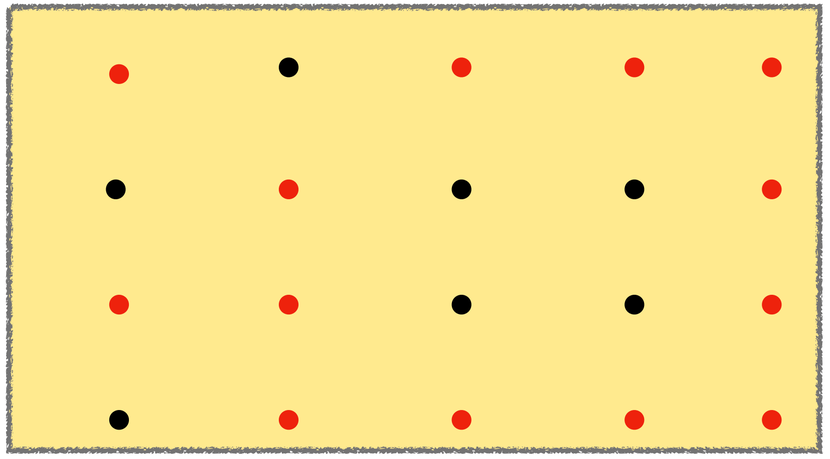


Smoothness Allows Coupling

Lemma [HRS'21, BDGR'22]: For all t, there is a coupling between X_t and $Z_{t,1}, \ldots, Z_{t,k} \overset{\text{iid}}{\sim} \mu$ such that w.p. at least $1 - e^{-\sigma k}$, it holds that

$$X_t \in \{Z_{t,1}, \ldots, Z_{t,k}\}.$$





Coupling Helps under Monotonicity

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t} g(X'_s)^2 - 2\cdot g(X_s)^2\right] \lesssim \mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t} \xi_s\cdot g(X_s) - \frac{g(X_s)^2}{2}\right]$$

Coupling Helps under Monotonicity

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t} g(X'_s)^2 - 2\cdot g(X_s)^2\right] \lesssim \mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\xi_s\cdot g(X_s) - \frac{g(X_s)^2}{2}\right]$$

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\xi_s\cdot g(X_s) - \frac{g(X_s)^2}{2}\right] \lesssim \mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\sum_{j=1}^{k}\xi_{s,j}\cdot g(Z_{s,j}) - \frac{g(Z_{s,j})^2}{2}\right]$$

Coupling Helps under Monotonicity

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t} g(X'_s)^2 - 2\cdot g(X_s)^2\right] \lesssim \mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\xi_s\cdot g(X_s) - \frac{g(X_s)^2}{2}\right]$$

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\xi_s\cdot g(X_s) - \frac{g(X_s)^2}{2}\right] \not\lesssim \mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\sum_{j=1}^{k}\xi_{s,j}\cdot g(Z_{s,j}) - \frac{g(Z_{s,j})^2}{2}\right]$$

Coupling Helps under Monotonicity

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t} g(X'_s)^2 - 2\cdot g(X_s)^2\right] \lesssim \mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\xi_s\cdot g(X_s) - \frac{g(X_s)^2}{2}\right]$$

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\xi_s\cdot g(X_s) - \frac{g(X_s)^2}{2}\right] \not\lesssim \mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\sum_{j=1}^{k}\xi_{s,j}\cdot g(Z_{s,j}) - \frac{g(Z_{s,j})^2}{2}\right]$$

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\xi_s\cdot g(X_s) - \frac{g(X_s)^2}{2}\right] \lesssim \log\mathbb{E}\left[\exp\left(\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\xi_s\cdot g(X_s) - \frac{g(X_s)^2}{2}\right)\right]$$

Wills Functional

Definition: $$\log W_T(\mathcal{F}) = \log \mathbb{E}\left[\exp\left(\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} \xi_t \cdot f(X_t) - \frac{f(X_t)^2}{2}\right)\right].$$

Wills Functional

Definition: $\log W_T(\mathscr{F}) = \log \mathbb{E}\left[\exp\left(\sup_{f\in\mathscr{F}} \sum_{t=1}^{T} \xi_t \cdot f(X_t) - \frac{f(X_t)^2}{2}\right)\right].$

Definition: Gaussian complexity $\mathscr{G}_T(\mathscr{F}) = \mathbb{E}\left[\sup_{f\in\mathscr{F}} \sum_{t=1}^{T} \xi_t \cdot f(X_t)\right].$

Wills Functional

Definition: $\log W_T(\mathcal{F}) = \log \mathbb{E}\left[\exp\left(\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} \xi_t \cdot f(X_t) - \frac{f(X_t)^2}{2}\right)\right].$

Definition: Gaussian complexity $\mathcal{G}_T(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{T} \xi_t \cdot f(X_t)\right].$

Wills Functional

Definition: $\log W_T(\mathscr{F}) = \log \mathbb{E}\left[\exp\left(\sup_{f\in\mathscr{F}} \sum_{t=1}^{T} \xi_t \cdot f(X_t) - \frac{f(X_t)^2}{2}\right)\right].$

Definition: Gaussian complexity $\mathscr{G}_T(\mathscr{F}) = \mathbb{E}\left[\sup_{f\in\mathscr{F}} \sum_{t=1}^{T} \xi_t \cdot f(X_t)\right].$

Theorem [M'23]: If ℓ is square loss and $\hat{f}$ is an ERM, then

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\log W_T(\mathscr{F})}{T} \lesssim \frac{\mathrm{vc}(\mathscr{F})}{T}.$$

Wills Functional

Definition: $$\log W_T(\mathscr{F}) = \log \mathbb{E}\left[\exp\left(\sup_{f\in\mathscr{F}} \sum_{t=1}^{T} \xi_t \cdot f(X_t) - \frac{f(X_t)^2}{2}\right)\right].$$

Theorem [M'23]: The Wills functional is monotone:

$$W_T(\mathscr{F}) \le W_{T+1}(\mathscr{F}).$$

Wills Functional

Definition: $\log W_T(\mathscr{F}) = \log \mathbb{E}\left[\exp\left(\sup_{f\in\mathscr{F}} \sum_{t=1}^{T} \xi_t \cdot f(X_t) - \frac{f(X_t)^2}{2}\right)\right].$

Theorem [M'23]: The Wills functional is monotone:

$$W_T(\mathscr{F}) \le W_{T+1}(\mathscr{F}).$$

$$\log \mathbb{E}\left[\exp\left(\sup_{g\in\mathscr{G}} \frac{1}{t}\sum_{s=1}^{t} \xi_s \cdot g(X_s) - \frac{g(X_s)^2}{2}\right)\right] \lesssim \log \mathbb{E}\left[\exp\left(\sup_{g\in\mathscr{G}} \frac{1}{t}\sum_{s=1}^{t}\sum_{j=1}^{k} \xi_{s,j} \cdot g(Z_{s,j}) - \frac{g(Z_{s,j})^2}{2}\right)\right]$$

Wills Functional

Definition: $\log W_T(\mathscr{F}) = \log \mathbb{E}\left[\exp\left(\sup_{f\in\mathscr{F}} \sum_{t=1}^{T} \xi_t \cdot f(X_t) - \frac{f(X_t)^2}{2}\right)\right].$

Theorem [M'23]: The Wills functional is monotone:

$$W_T(\mathscr{F}) \le W_{T+1}(\mathscr{F}).$$

Theorem [BRS'24]: If data smooth, then

$$\mathbb{E}\left[\sup_{g\in\mathscr{G}} \frac{1}{t}\sum_{s=1}^{t} g(X_s')^2 - 2\cdot g(X_s)^2\right] \lesssim \frac{\log W_{t/\sigma}(\mathscr{F})}{t}.$$

ERM Performance

Theorem [BRS'24]: If data are σ-smooth w.r.t. μ and f_t is ERM, then

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\log(T/\sigma)}{\sigma \cdot T} + \sqrt{\frac{1}{\sigma \cdot T} \cdot \log \mathbb{E}_\mu \left[W_{T \log(T)/\sigma}(\mathscr{F}) \right]}.$$

Theorem [BRS'24]: For all d there is $\mathscr{F}$ with $\mathrm{vc}(\mathscr{F}) \leq d$ and a realizable adversary such any algorithm (if μ is unknown) must pay

$$\mathbb{E}\left[\mathrm{Err}_T\right] \gtrsim \sqrt{\frac{d}{\sigma^{1/d} \cdot T}}.$$

ERM Performance

Theorem [BRS'24]: If data are σ-smooth w.r.t. μ and f_t is ERM, then

$$\mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \frac{\log(T/\sigma)}{\sigma \cdot T} + \sqrt{\frac{\mathrm{vc}(\mathscr{F}) \cdot \log(T/\sigma)}{\sigma \cdot T}}.$$

Theorem [BRS'24]: For all d there is $\mathscr{F}$ with $\mathrm{vc}(\mathscr{F}) \leq d$ and a realizable adversary such any algorithm (if μ is unknown) must pay

$$\mathbb{E}\left[\mathrm{Err}_T\right] \gtrsim \sqrt{\frac{d}{\sigma^{1/d} \cdot T}}.$$

Key Takeaways

Smoothed data bridges efficiency of statistical learning and robustness of online learning.

Technical tools:

(i) Surprise Lemma (compactness)

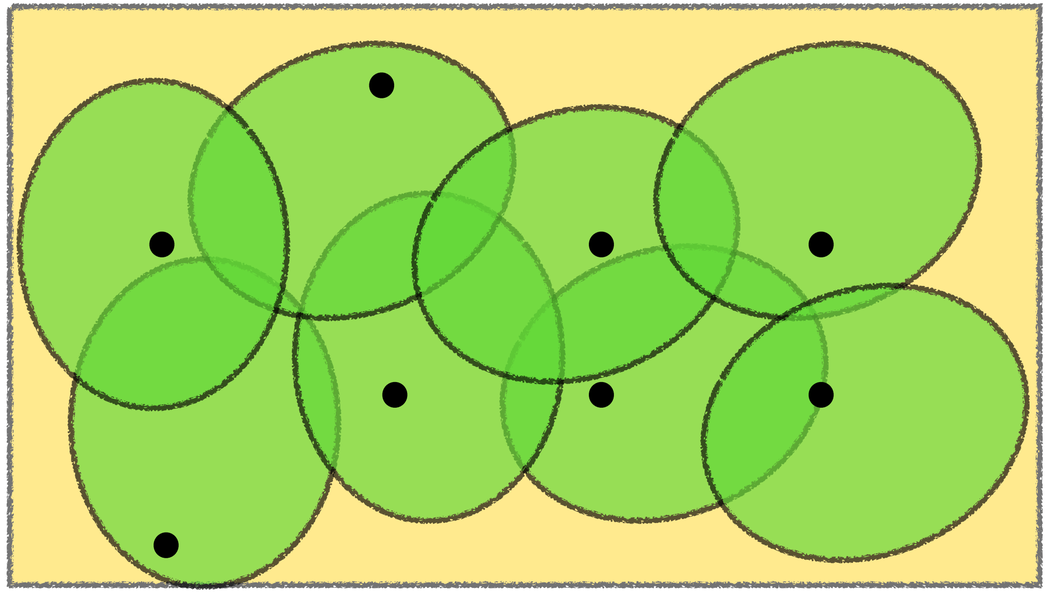
(ii) Coupling (rejection sampling)

Smoothed Data

Definition: For measures $p, \mu \in \Delta(\mathcal{X})$, p is σ-**smooth** with respect to μ if

$$\left\| \frac{dp}{d\mu} \right\|_\infty \leq \sigma^{-1}.$$

$$\mu = \mathrm{Unif}(\mathcal{X})$$



Key Takeaways

Smoothed data bridges efficiency of statistical learning and robustness of online learning.

Technical tools:

(i) Surprise Lemma (compactness)

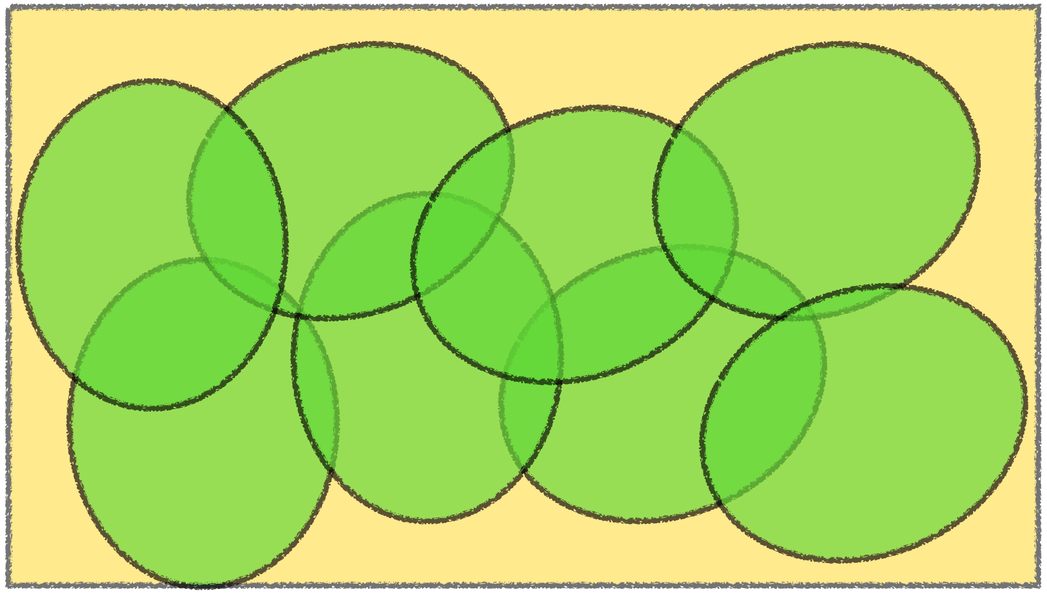
(ii) Coupling (rejection sampling)

Smoothness Bounds Surprises

Lemma [BRS'24]: Let $p_1, \ldots, p_T$ be σ-smooth. Then for $\varepsilon > 0$ and $Z \in \mathcal{X}$,

$$\left| \left\{ t \in [T] : p_t(Z) \geq \frac{2\log(T)}{\varepsilon} \cdot \frac{1}{t} \left(\frac{1}{\sigma} + \sum_{s=1}^{t-1} p_s(Z) \right) \right\} \right| \leq \varepsilon \cdot T.$$

$$\mu = \mathrm{Unif}(\mathcal{X})$$



Key Takeaways

Smoothed data bridges efficiency of statistical learning and robustness of online learning.

Technical tools:

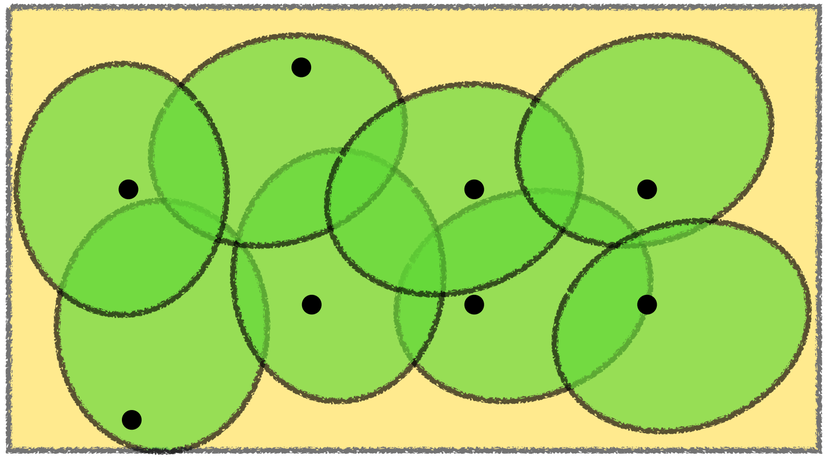
(i) Surprise Lemma (compactness)

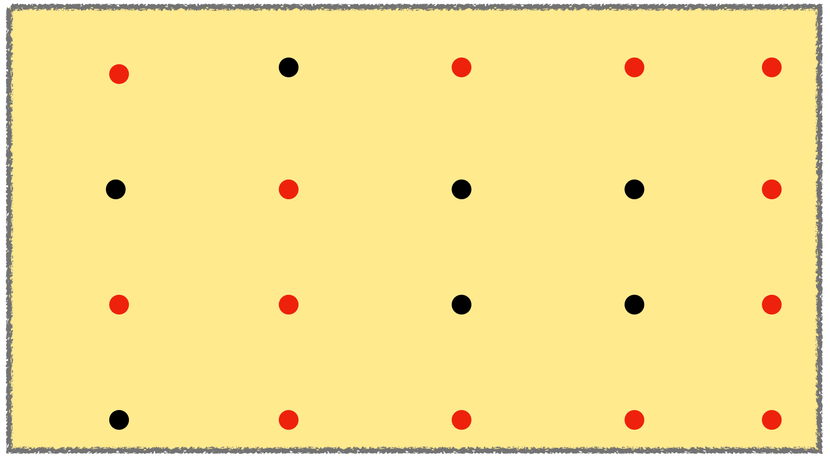
(ii) Coupling (rejection sampling)

Smoothness Allows Coupling

Lemma [HRS'21, BDGR'22]: For all t, there is a coupling between X_t and $Z_{t,1}, \ldots, Z_{t,k} \overset{\text{iid}}{\sim} \mu$ such that w.p. at least $1 - e^{-\sigma k}$, it holds that

$$X_t \in \{Z_{t,1}, \ldots, Z_{t,k}\}.$$





Key Takeaways

Smoothed data bridges efficiency of statistical learning and robustness of online learning.

Technical tools:

(i) Surprise Lemma (compactness)

(ii) Coupling (rejection sampling)

ERM Performance

Smoothed Online Learning

$$\sqrt{\frac{\mathsf{vc}(\mathscr{F})}{\sigma^{1/\mathsf{vc}(\mathscr{F})} \cdot T}} \lesssim \mathbb{E}\left[\mathrm{Err}_T\right] \lesssim \max\left(\sqrt{\frac{\mathsf{vc}(\mathscr{F}) \cdot \log(T/\sigma)}{\sigma \cdot T}}, \frac{\log(T/\sigma)}{\sigma \cdot T}\right).$$

Statistical Learning

$$\mathbb{E}\left[\mathrm{Err}_T\right] \asymp \frac{\mathsf{vc}(\mathscr{F}) \cdot \log(T)}{T}.$$

Key Takeaways

Smoothed data bridges efficiency of statistical learning and robustness of online learning.

Technical tools:

(i) Surprise Lemma (compactness)

(ii) Coupling (rejection sampling)

Tutorial Outline

Part I

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications

2. The Smoothed Model: Best of Both Worlds?

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications
2. The Smoothed Model: Best of Both Worlds?
3. **The Power of Empirical Risk Minimization**

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications
2. The Smoothed Model: Best of Both Worlds?
3. The Power of Empirical Risk Minimization

Part II

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications
2. The Smoothed Model: Best of Both Worlds?
3. The Power of Empirical Risk Minimization

Part II

1. Coupling Lemma

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications
2. The Smoothed Model: Best of Both Worlds?
3. The Power of Empirical Risk Minimization

Part II

1. Coupling Lemma
2. Handling Label Noise

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications
2. The Smoothed Model: Best of Both Worlds?
3. The Power of Empirical Risk Minimization

Part II

1. Coupling Lemma
2. Handling Label Noise
3. Oracle Efficiency: ERM Returns

Tutorial Outline

Part I

1. Statistical and Online Learning: Definitions and Applications
2. The Smoothed Model: Best of Both Worlds?
3. The Power of Empirical Risk Minimization

Part II

1. Coupling Lemma
2. Handling Label Noise
3. Oracle Efficiency: ERM Returns

Challenges of Adaptive Adversaries

Challenges of Adaptive Adversaries

- Main distinguishing factor from stochastic

Challenges of Adaptive Adversaries

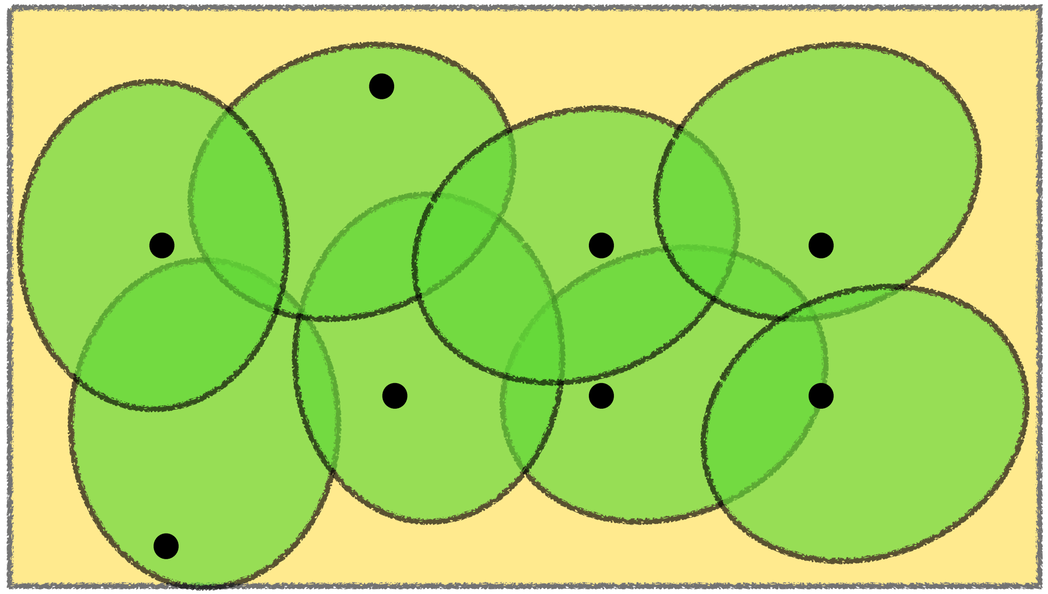
- Main distinguishing factor from stochastic
    - Distribution shifts

Challenges of Adaptive Adversaries

- Main distinguishing factor from stochastic
  - Distribution shifts
  - Adaptivity: $\mathscr{D}_t$ depends on $z_1, \ldots, z_{t-1}$

Challenges of Adaptive Adversaries

- Main distinguishing factor from stochastic
  - Distribution shifts
  - Adaptivity: $\mathscr{D}_t$ depends on $z_1, \ldots, z_{t-1}$



Extracting Stochasticity

Adaptive smooth sequences can be realized as subsequences of (slightly) longer IID sequences

Extracting Stochasticity via Coupling

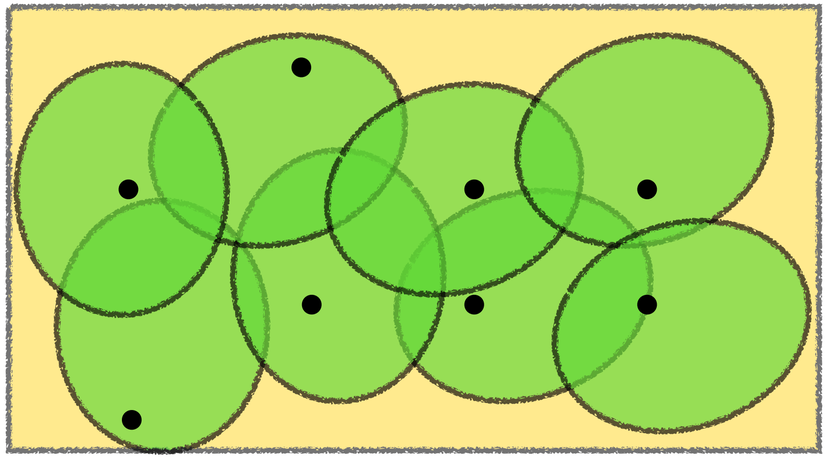
Lemma [HRS'21, BDGR'22]: For all t, there is a coupling between X_t and $Z_{t,1}, \ldots, Z_{t,k} \overset{\text{iid}}{\sim} \mu$ such that w.p. at least $1 - e^{-\sigma k}$, it holds that

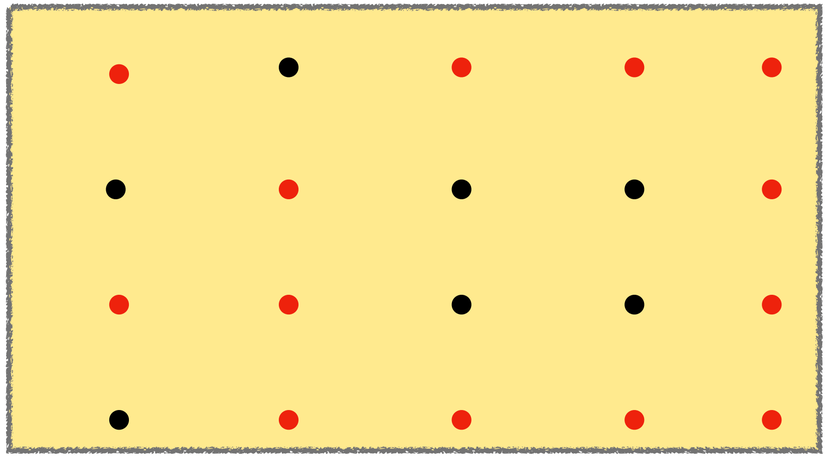
$$X_t \in \{Z_{t,1}, \ldots, Z_{t,k}\}.$$

Extracting Stochasticity via Coupling

Lemma [HRS'21, BDGR'22]: For all t, there is a coupling between X_t and $Z_{t,1}, \ldots, Z_{t,k} \overset{\text{iid}}{\sim} \mu$ such that w.p. at least $1 - e^{-\sigma k}$, it holds that

$$X_t \in \{Z_{t,1}, \ldots, Z_{t,k}\}.$$





Coupling Lemma and Rejection Sampling

Coupling Lemma and Rejection Sampling

Key idea: (Approximate) Rejection Sampling

Coupling Lemma and Rejection Sampling

Key idea: (Approximate) Rejection Sampling

Coupling Lemma “Algorithm”

Coupling Lemma and Rejection Sampling

Key idea: (Approximate) Rejection Sampling

Coupling Lemma "Algorithm"

Sample $\{Z_{i,k}\} \sim \mu$

Coupling Lemma and Rejection Sampling

Key idea: (Approximate) Rejection Sampling

Coupling Lemma "Algorithm"

Sample $\{Z_{i,k}\} \sim \mu$

$\mathscr{D}_i$ **be the** i**th distribution (depends on** $\{X_j\}_{j<i}$ **)**

Coupling Lemma and Rejection Sampling

Key idea: (Approximate) Rejection Sampling

Coupling Lemma "Algorithm"

Sample $\{Z_{i,k}\} \sim \mu$

$\mathscr{D}_i$ be the ith distribution (depends on $\{X_j\}_{j<i}$)

Scan through $\{Z_{i,k}\}_k$ and set $X_i = Z_{i,k}$ with prob $\sigma\mathscr{D}(Z_{i,k})/\mu(Z_{i,k})$

Coupling Lemma and Rejection Sampling

Key idea: (Approximate) Rejection Sampling

Coupling Lemma "Algorithm"

Sample $\{Z_{i,k}\} \sim \mu$

$\mathscr{D}_i$ be the ith distribution (depends on $\{X_j\}_{j<i}$)

Scan through $\{Z_{i,k}\}_k$ and set $X_i = Z_{i,k}$ with prob $\sigma\mathscr{D}(Z_{i,k})/\mu(Z_{i,k})$

If all the tosses fail sample from $\mathscr{D}_i$

Coupling Lemma and Rejection Sampling

Key idea: (Approximate) Rejection Sampling

Coupling Lemma "Algorithm"

Sample $\{Z_{i,k}\} \sim \mu$

$\mathscr{D}_i$ **be the** i**th distribution (depends on** $\{X_j\}_{j<i}$ **)**

Scan through $\{Z_{i,k}\}_k$ **and set** $X_i = Z_{i,k}$ **with prob** $\sigma\mathscr{D}(Z_{i,k})/\mu(Z_{i,k})$

≤ 1 by smoothness

If all the tosses fail sample from $\mathscr{D}_i$

Coupling Lemma and Rejection Sampling

Key idea: (Approximate) Rejection Sampling

Coupling Lemma "Algorithm"

Sample $\{Z_{i,k}\} \sim \mu$

$\mathscr{D}_i$ be the ith distribution (depends on $\{X_j\}_{j<i}$)

Scan through $\{Z_{i,k}\}_k$ and set $X_i = Z_{i,k}$ with prob $\sigma\mathscr{D}(Z_{i,k})/\mu(Z_{i,k})$

If all the tosses fail sample from $\mathscr{D}_i$

Coupling Lemma and Rejection Sampling

Key idea: (Approximate) Rejection Sampling

Coupling Lemma "Algorithm"

Sample $\{Z_{i,k}\} \sim \mu$

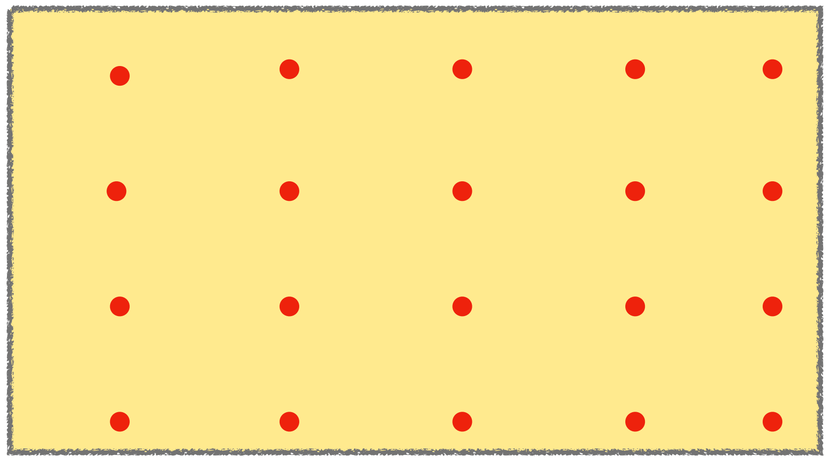
$\mathscr{D}_i$ **be the** i**th distribution (depends on** $\{X_j\}_{j<i}$ **)**

Scan through $\{Z_{i,k}\}_k$ **and set** $X_i = Z_{i,k}$ **with prob** $\sigma\mathscr{D}(Z_{i,k})/\mu(Z_{i,k})$

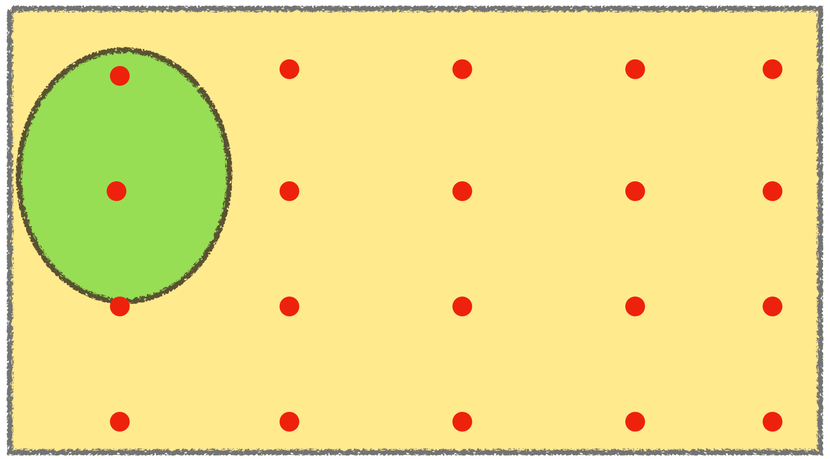
If all the tosses fail sample from $\mathscr{D}_i$

Warning: any learner cannot run this algorithm, since we don't know $\mathscr{D}_i$
We only have access to the "implicit" structure

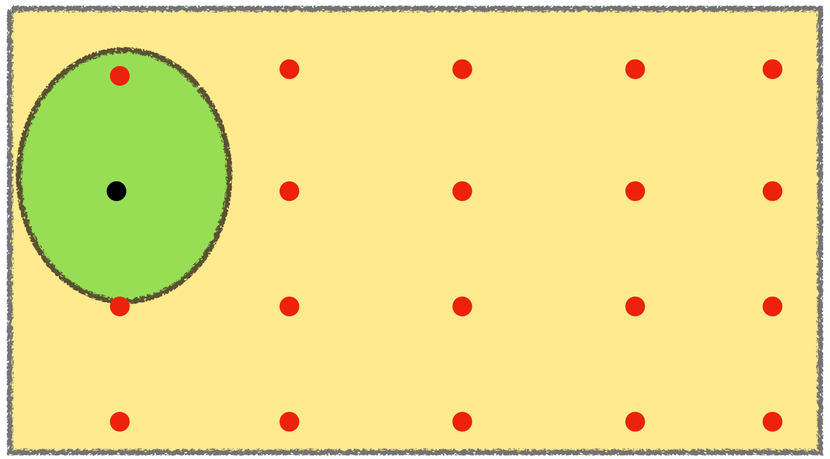
Coupling Lemma Visualized



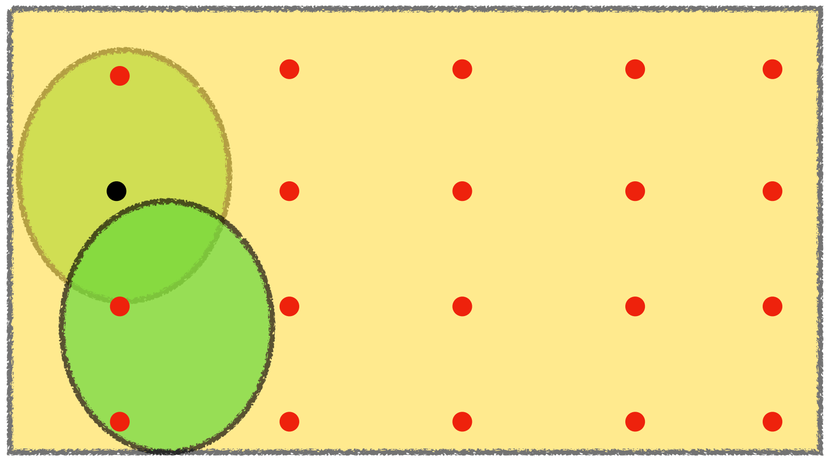
Coupling Lemma Visualized



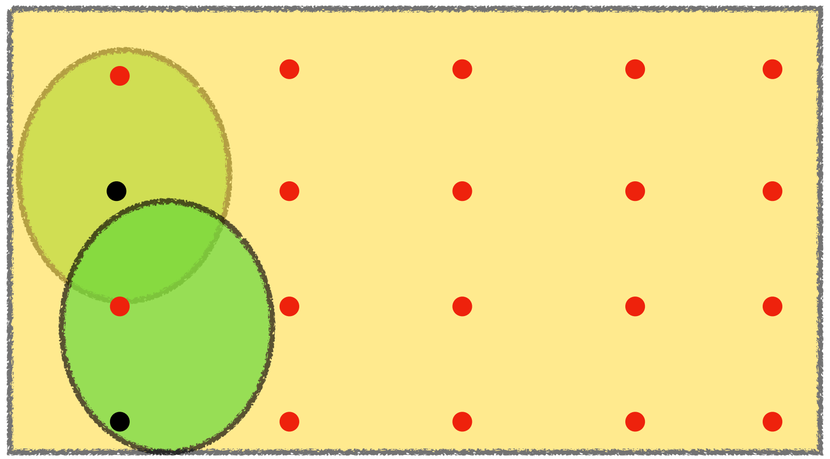
Coupling Lemma Visualized



Coupling Lemma Visualized

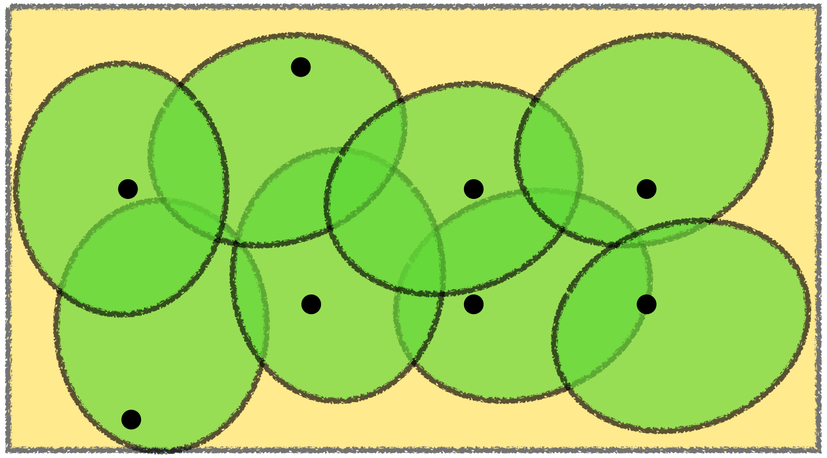


Coupling Lemma Visualized

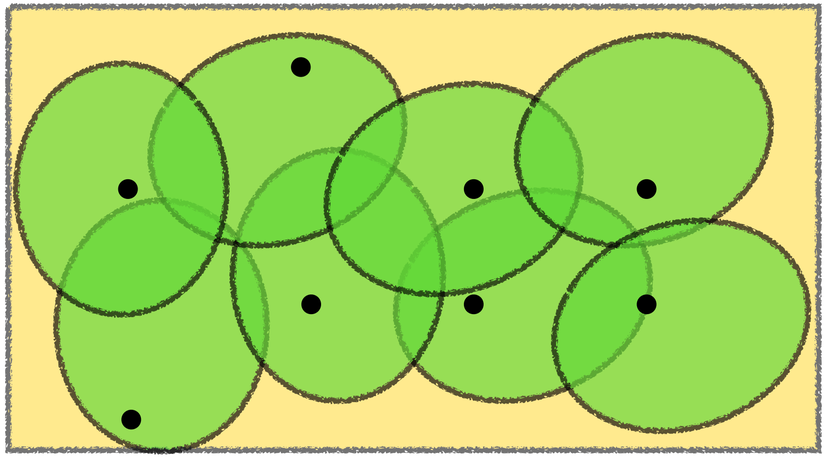


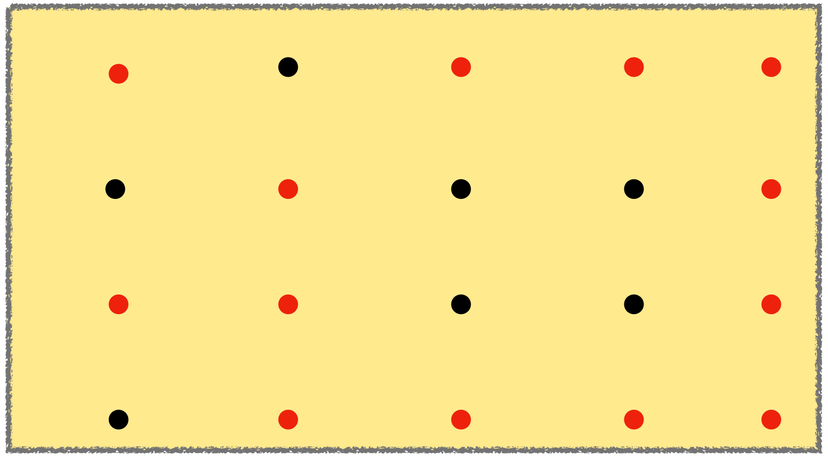
Coupling Lemma Visualized

Coupling Lemma Visualized



Coupling Lemma Visualized





Approximate Rejection Sampling Perspective

Approximate Rejection Sampling Perspective

Lemma [HRS'21, BDGR'22, BP'23]: For any pair of distributions μ_1, μ_2 as long as $n \geq F(\mu_1, \mu_2, \varepsilon)$, $\exists i^\star$ such that for $Z_1, \ldots, Z_n \sim \mu_2$, we have $d_{TV}(Z_{i^\star}, \mu_1) \leq \varepsilon$

Approximate Rejection Sampling Perspective

Lemma [HRS'21, BDGR'22, BP'23]: For any pair of distributions μ_1, μ_2 as long as $n \geq F(\mu_1, \mu_2, \varepsilon)$, $\exists i^\star$ such that for $Z_1, \ldots, Z_n \sim \mu_2$, we have $d_{TV}(Z_{i^\star}, \mu_1) \leq \varepsilon$

Depends on "distance" between μ_1 and μ_2

Approximate Rejection Sampling Perspective

Lemma [HRS'21, **B**DGR'22, **B**P'23]: For any pair of distributions μ_1, μ_2 as long as $n \geq F(\mu_1, \mu_2, \varepsilon)$, $\exists i^\star$ such that for $Z_1, \ldots, Z_n \sim \mu_2$, we have $d_{TV}(Z_{i^\star}, \mu_1) \leq \varepsilon$

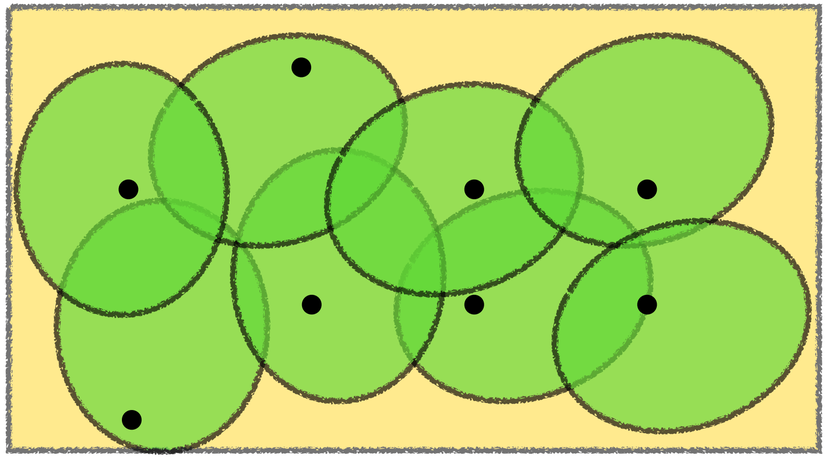
Depends on "distance" between μ_1 and μ_2

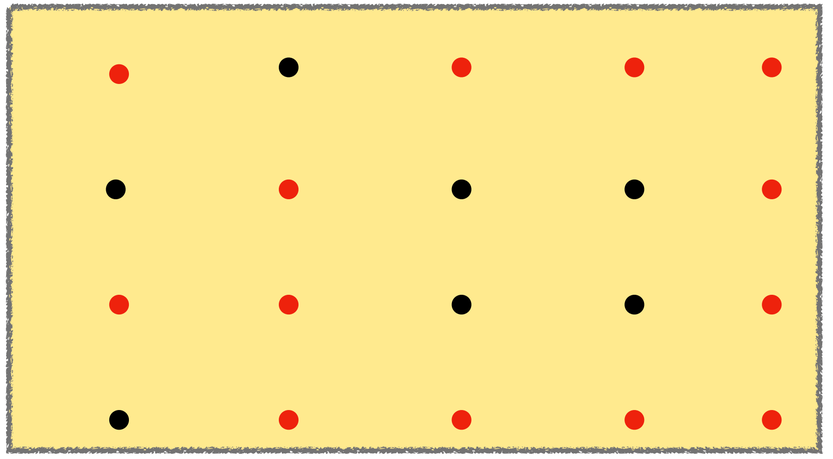
- Connections to channel coding [**B**P'23]
- Beyond uniformly bounded ratios: Extensions to generalizations of smoothness
- Applications: sampling from language models [H**B**F+'24, H**B**L+'25]

Extracting Stochasticity via Coupling

Lemma [HRS'21, BDGR'22]: For all t, there is a coupling between X_t and $Z_{t,1}, \ldots, Z_{t,k} \overset{\text{iid}}{\sim} \mu$ such that w.p. at least $1 - e^{-\sigma k}$, it holds that

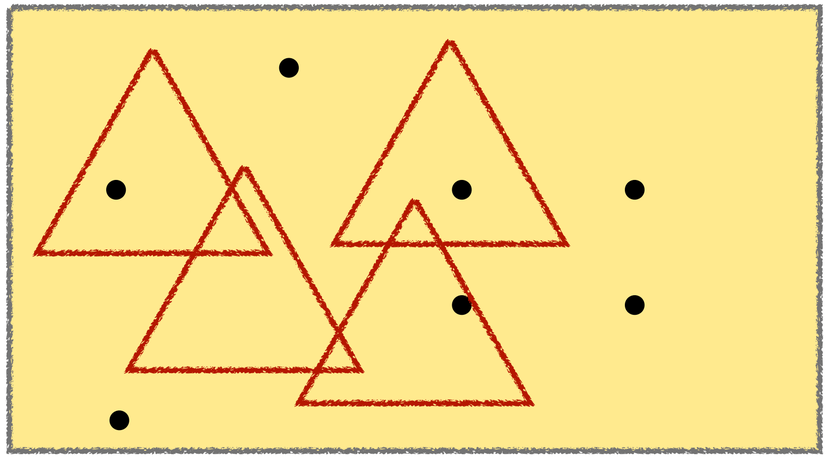
$$X_t \in \{Z_{t,1}, \ldots, Z_{t,k}\}.$$



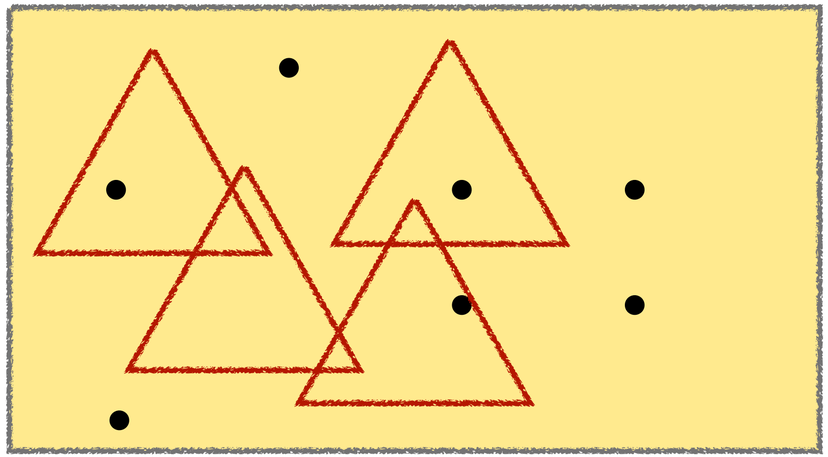


Bounding Bad Events by Coupling

Bounding Bad Events by Coupling

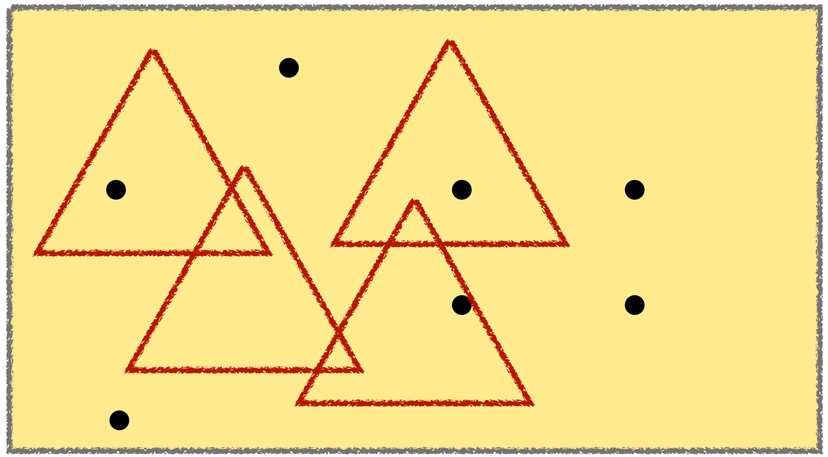


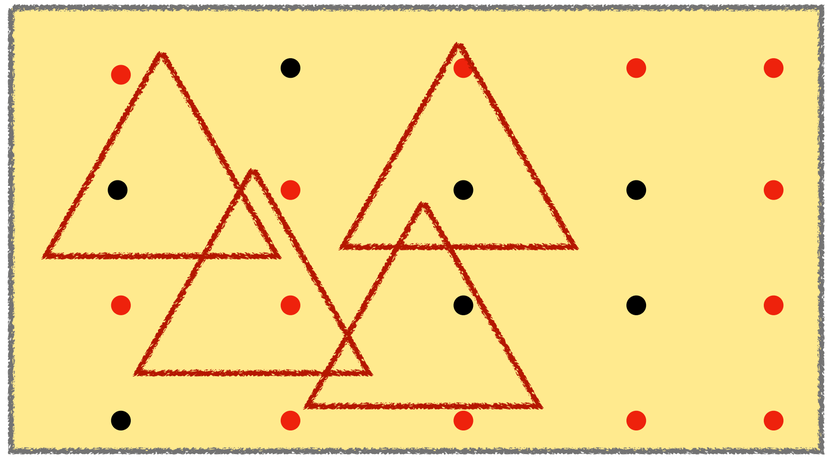
Bounding Bad Events by Coupling



Family of “bad” events $\mathscr{B}$ (corresponding to when algorithm or analysis fails)

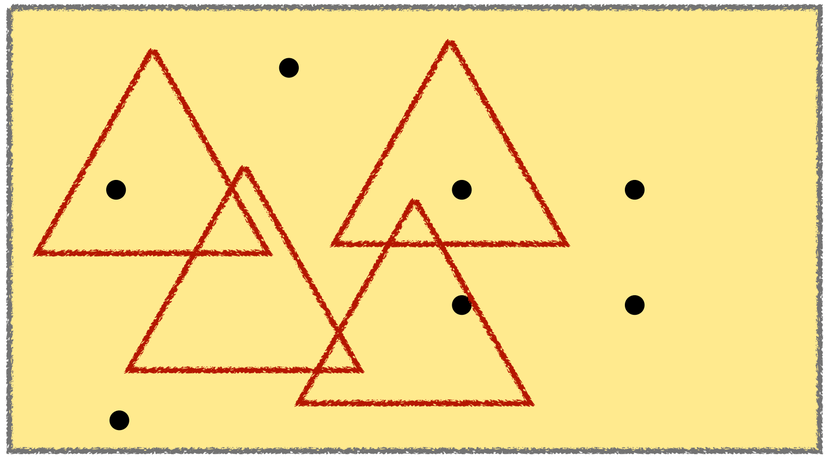
Bounding Bad Events by Coupling

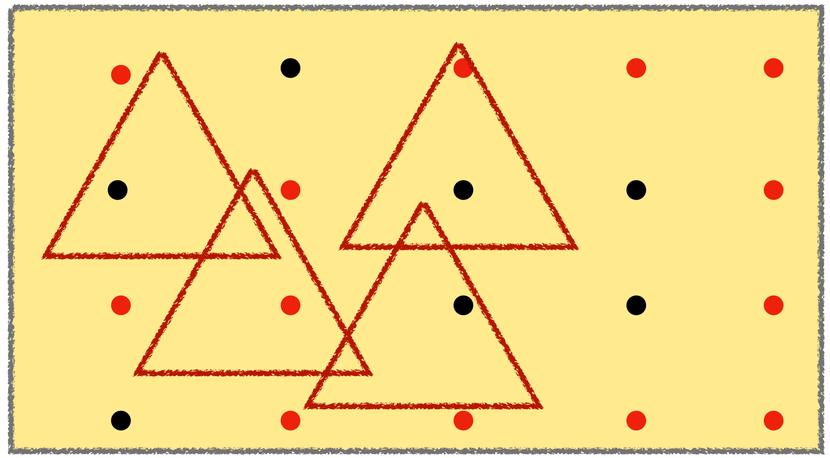




Family of “bad” events $\mathscr{B}$ (corresponding to when algorithm or analysis fails)

Bounding Bad Events by Coupling

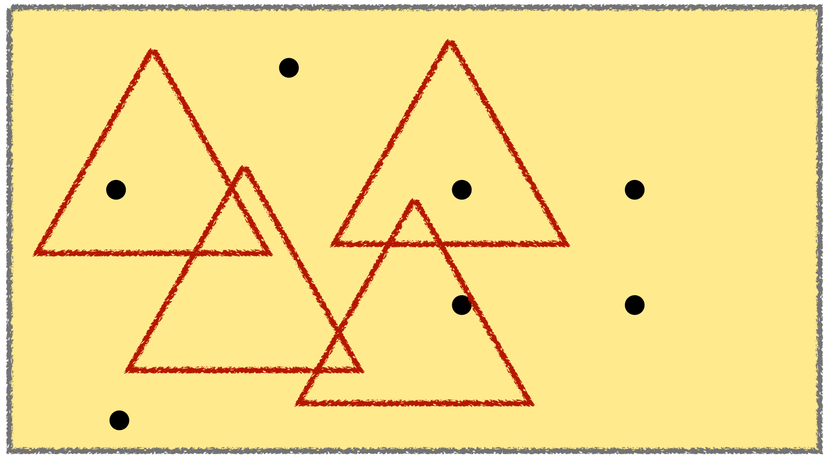


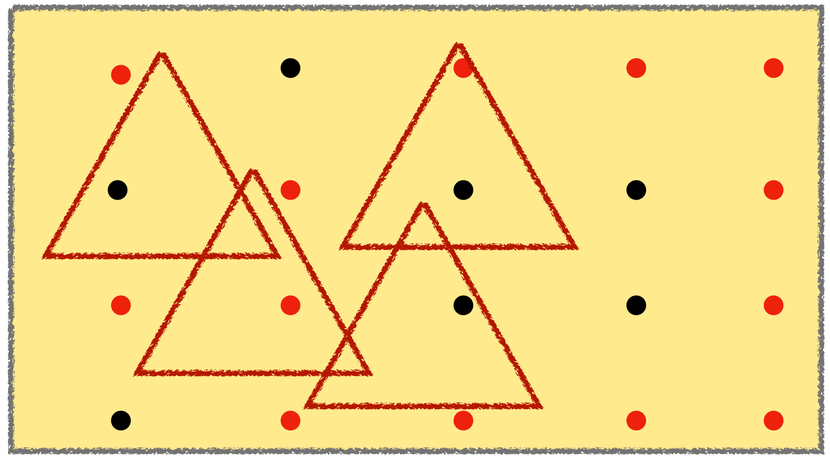


Family of “bad” events $\mathscr{B}$ (corresponding to when algorithm or analysis fails)

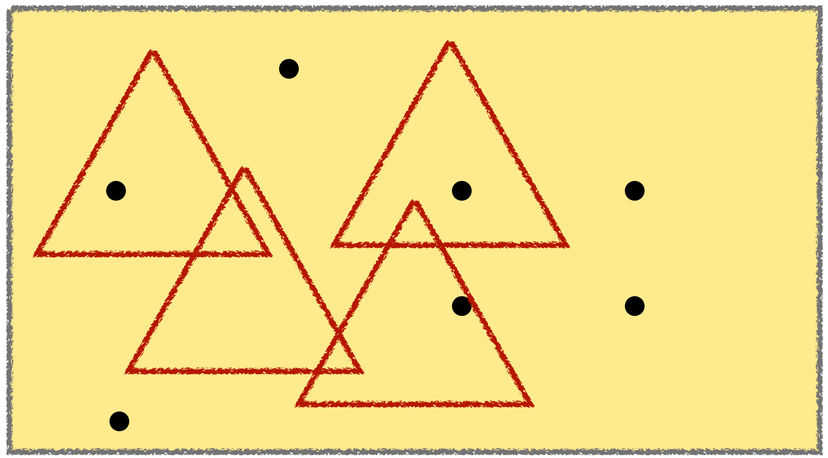
Coupling tells us that roughly $\Pr_{\text{smooth}}[\text{Bad}] \lesssim \Pr_{\text{IID}}[\text{Bad}]$

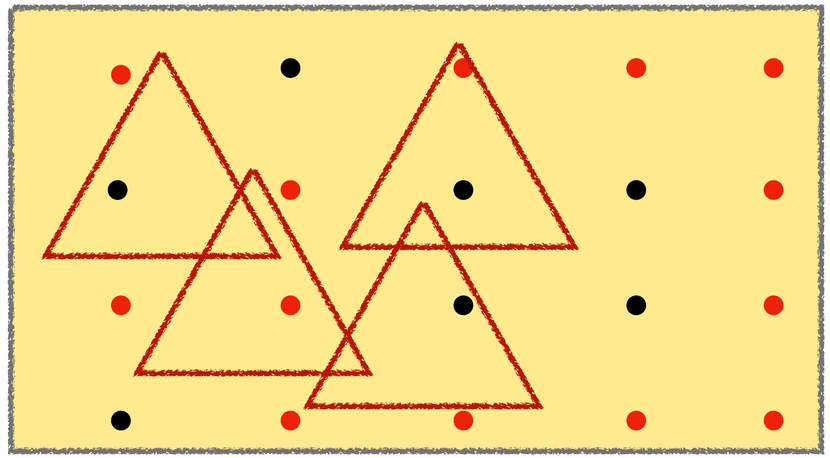
Coupling and Monotonicity





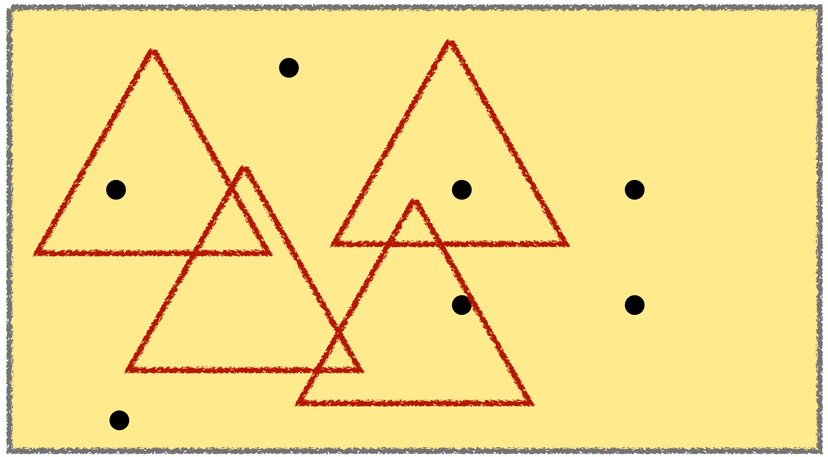
Coupling and Monotonicity

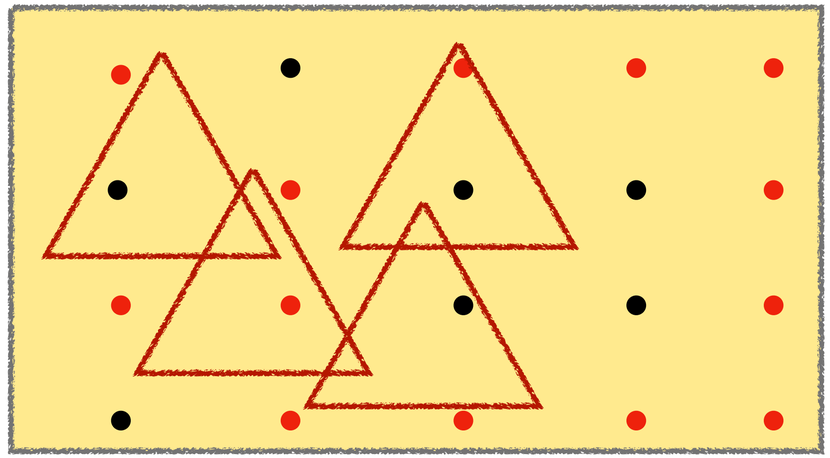




When is coupling useful?

Coupling and Monotonicity

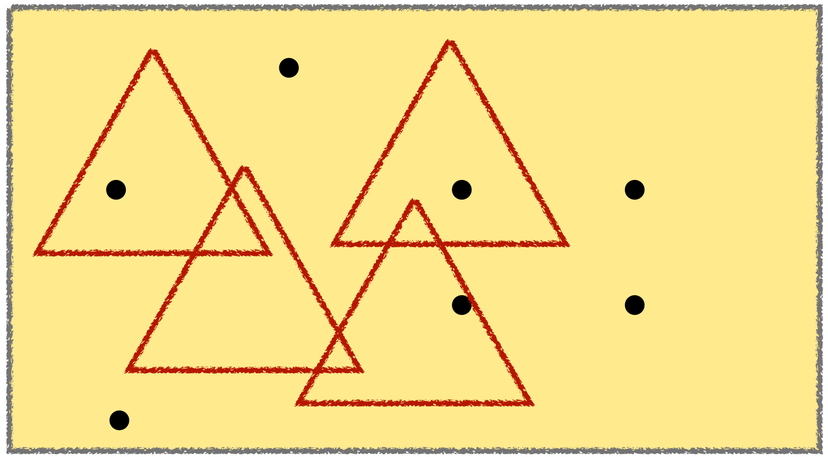


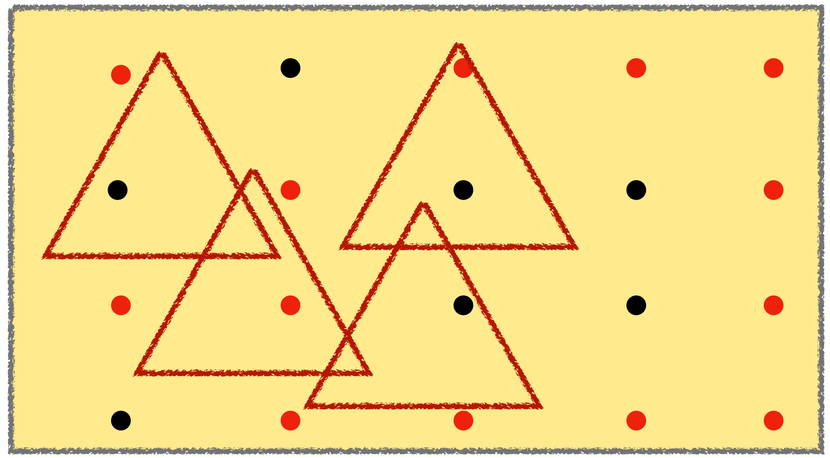


When is coupling useful?

Monotonicity (in terms of sample)

Coupling and Monotonicity



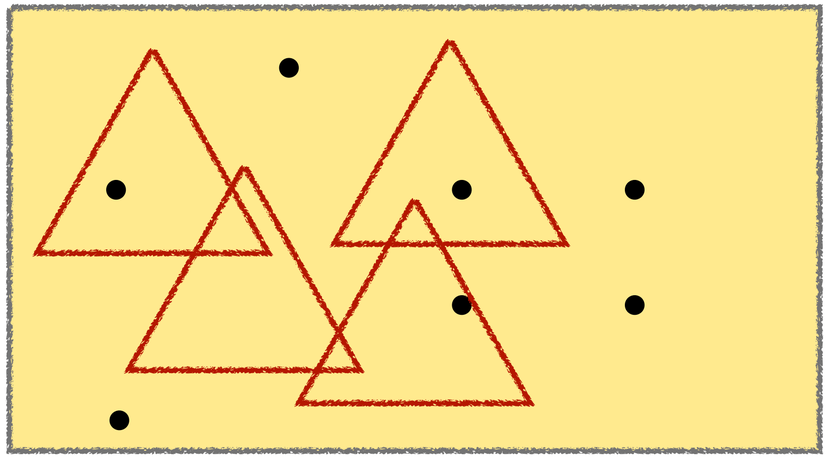


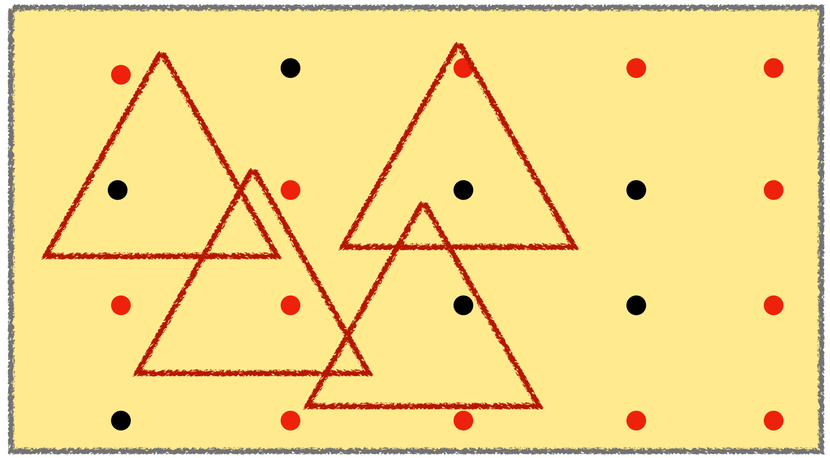
When is coupling useful?

Monotonicity (in terms of sample)

$$F : \text{Datasets} \to \mathbb{R}$$

Coupling and Monotonicity



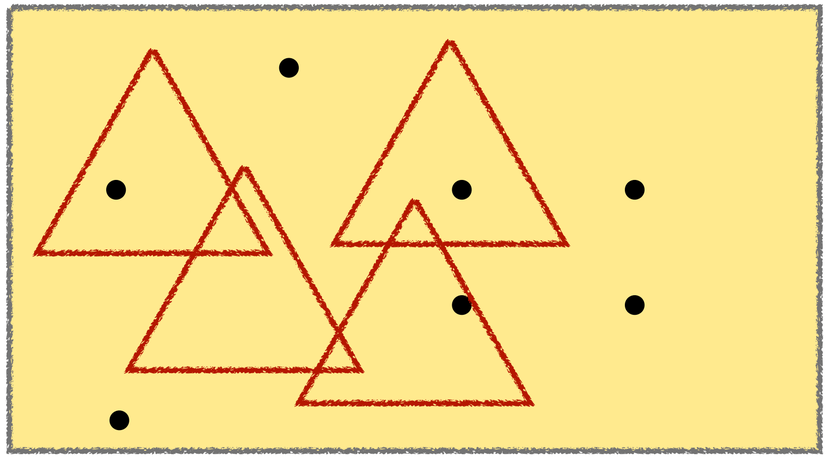


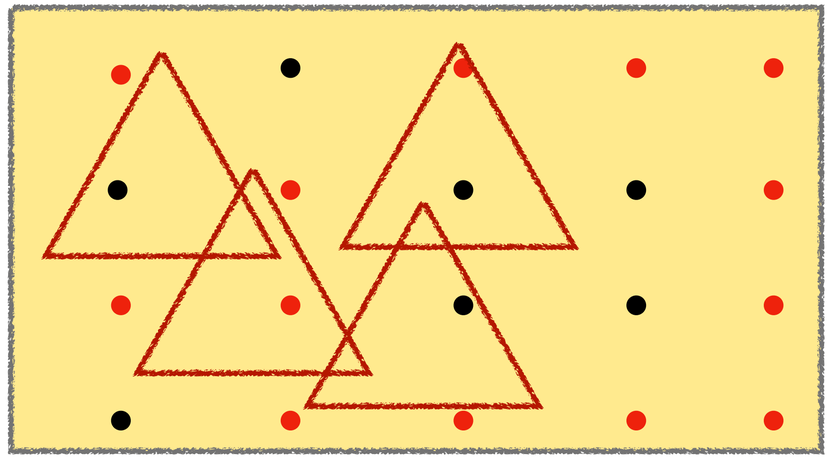
When is coupling useful?

Monotonicity (in terms of sample)

$$F : \text{Datasets} \to \mathbb{R} \qquad \{X_i\} \subset \{Z_i\} \implies \mathbb{E}F(\{X_i\}) \leq \mathbb{E}F(\{Z_i\})$$

Coupling and Monotonicity





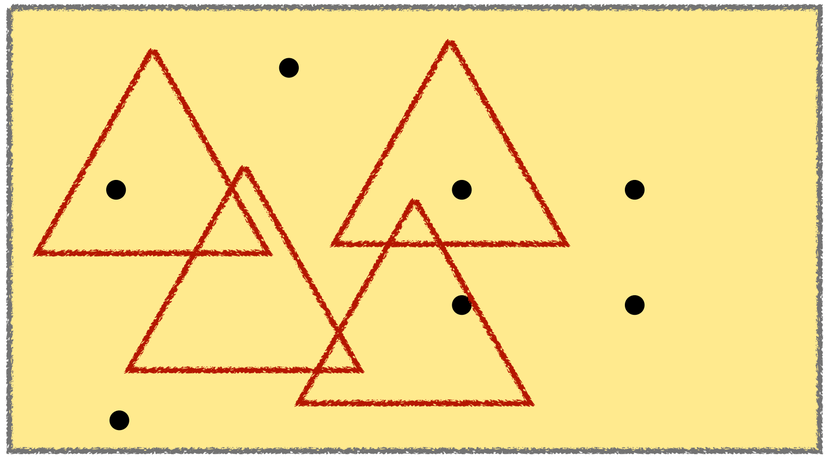
When is coupling useful?

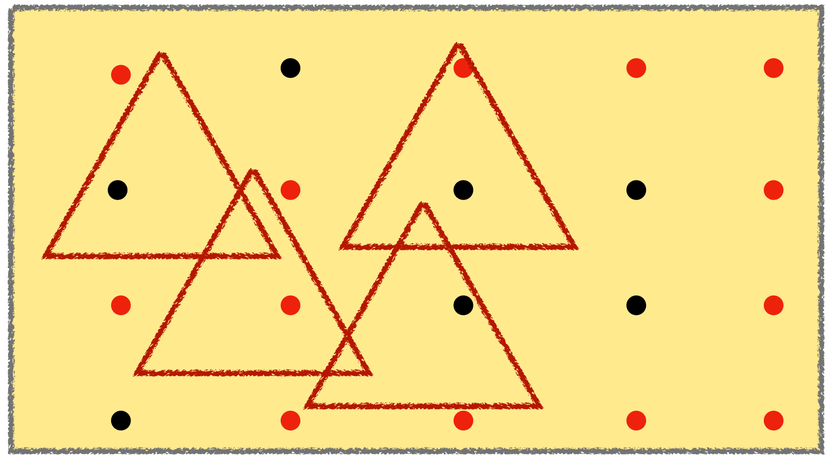
Monotonicity (in terms of sample)

$$F : \text{Datasets} \to \mathbb{R} \qquad \{X_i\} \subset \{Z_i\} \implies \mathbb{E}F(\{X_i\}) \leq \mathbb{E}F(\{Z_i\})$$

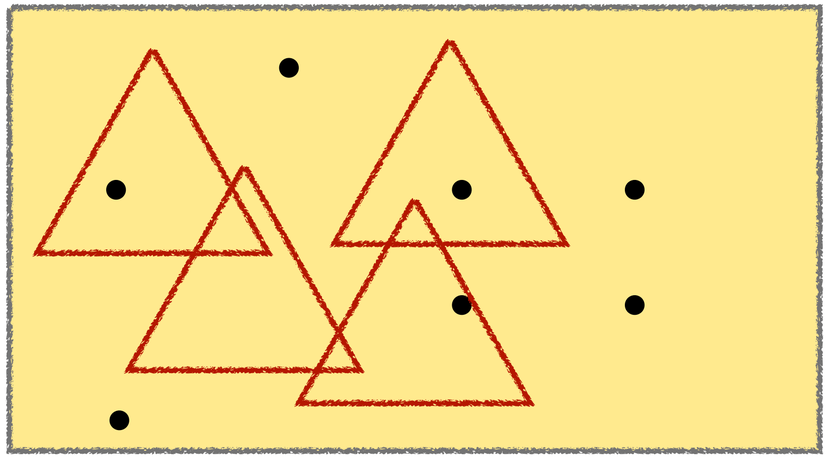
E.g. sums of positive functions, Rademacher complexity

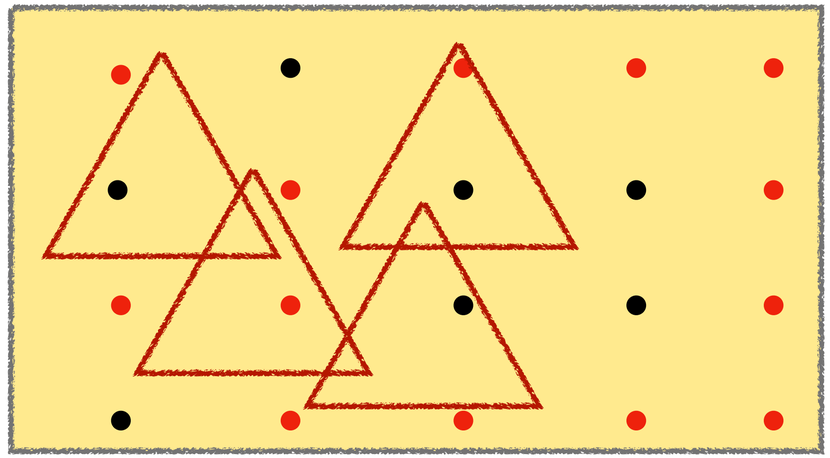
(De)coupling Inequality



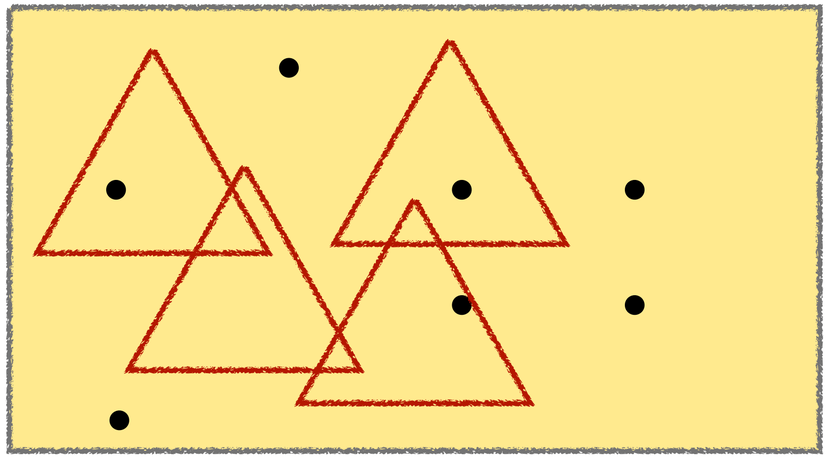


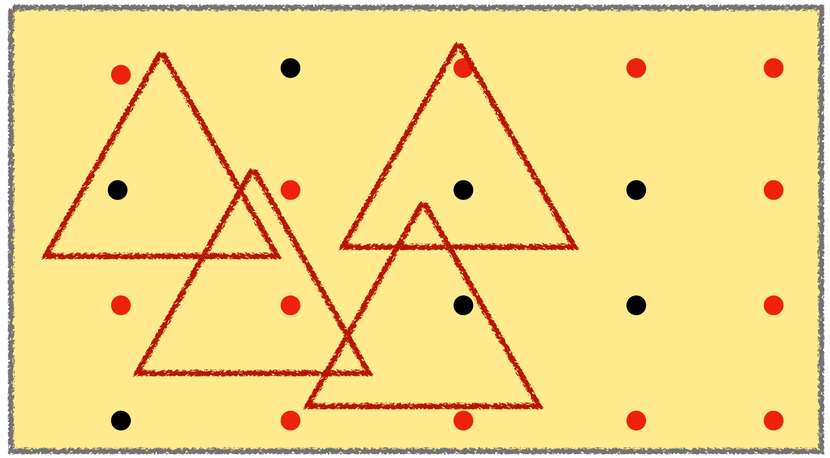
(De)coupling Inequality





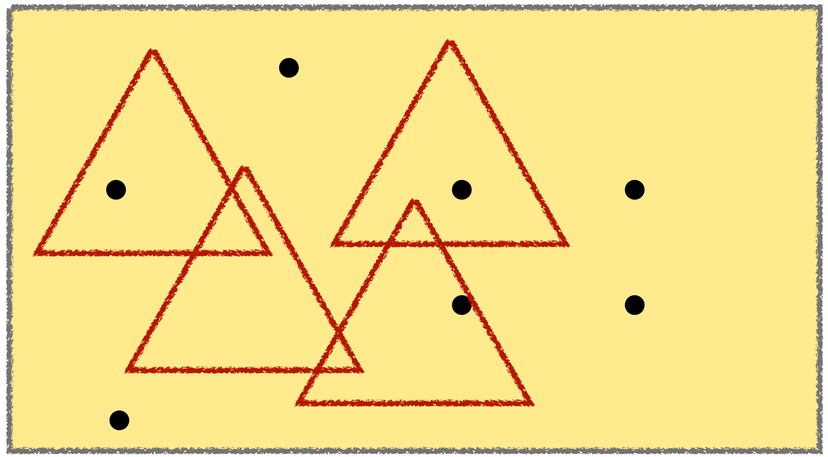
(De)coupling Inequality

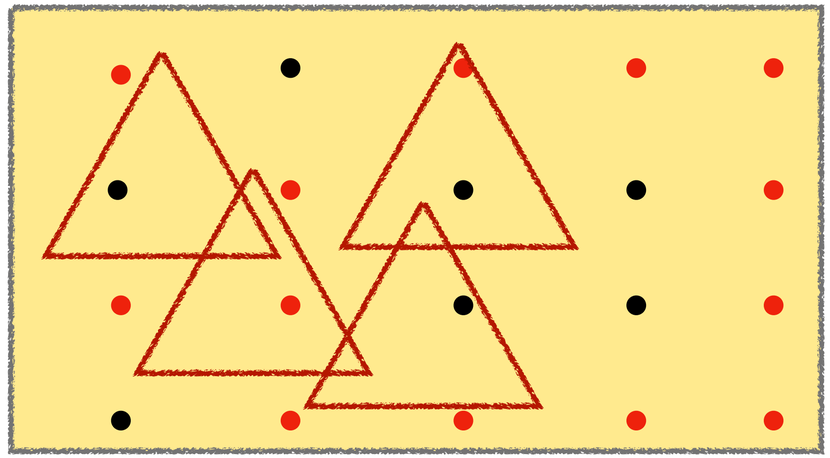




$X_1, \ldots, X_T$ adaptive smooth sequence and $\mathscr{B}$ be a family of positive functions

(De)coupling Inequality

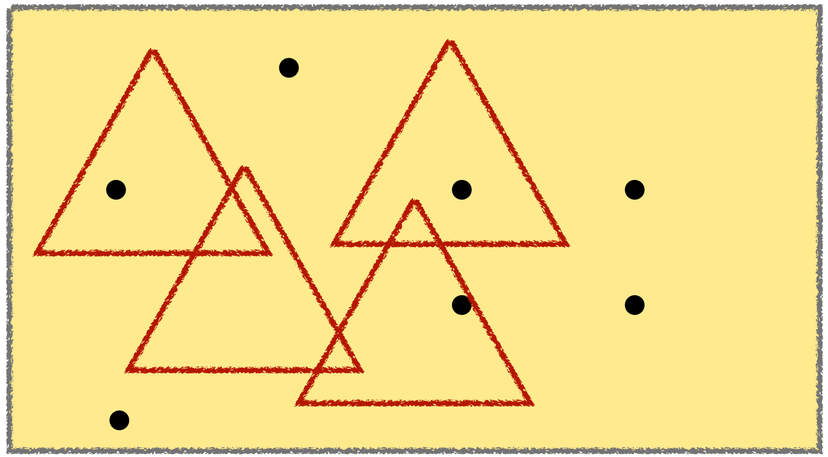


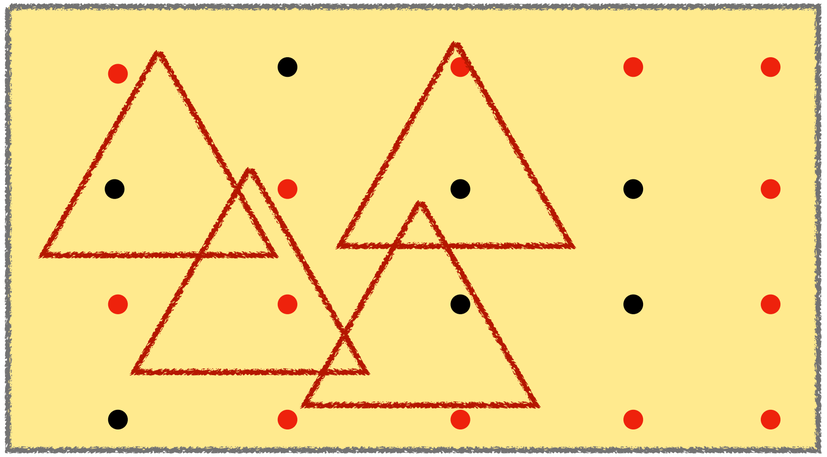


$X_1, \ldots, X_T$ adaptive smooth sequence and $\mathscr{B}$ be a family of positive functions

$$\mathbb{E}\left[\sup_{b\in\mathscr{B}} \sum_t b(X_t)\right] \leq \mathbb{E}_{Z_1,\ldots,Z_{Tk}\sim\mu}\left[\sup_{b\in\mathscr{B}} \sum_t b(Z_t)\right]$$

(De)coupling Inequality

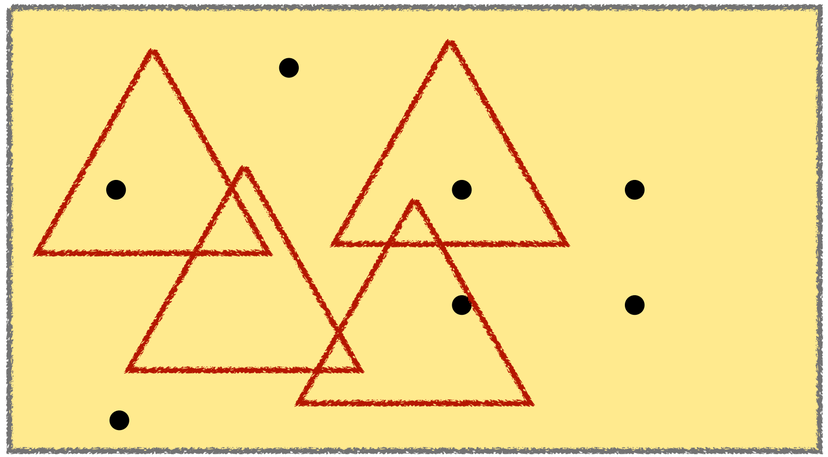


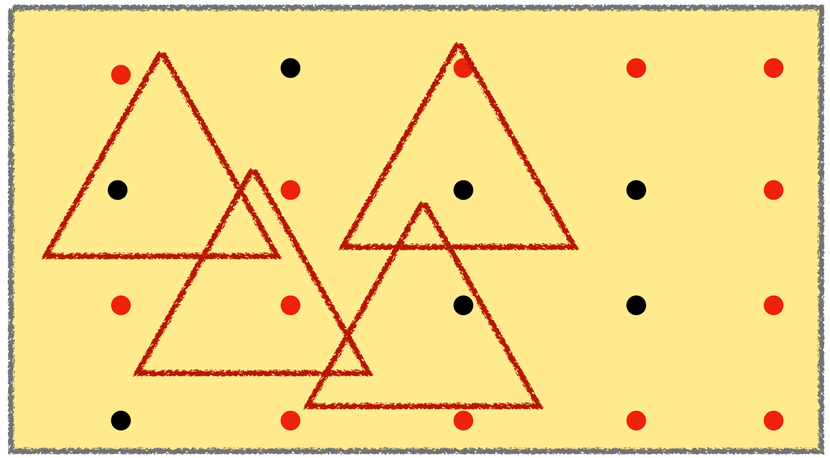


$X_1, \ldots, X_T$ adaptive smooth sequence and $\mathscr{B}$ be a family of positive functions

$$\mathbb{E}\left[\sup_{b\in\mathscr{B}} \sum_t b(X_t)\right] \leq \mathbb{E}_{Z_1,\ldots,Z_{Tk}\sim\mu}\left[\sup_{b\in\mathscr{B}} \sum_t b(Z_t)\right] + T^2 e^{-\sigma k}$$

(De)coupling Inequality



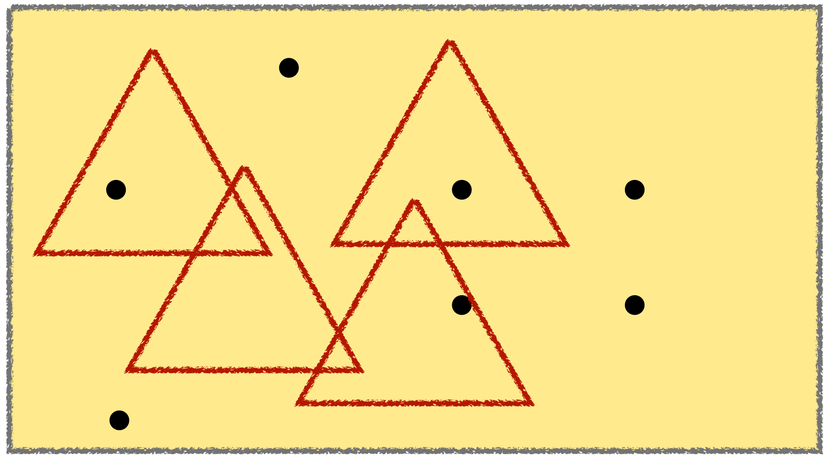


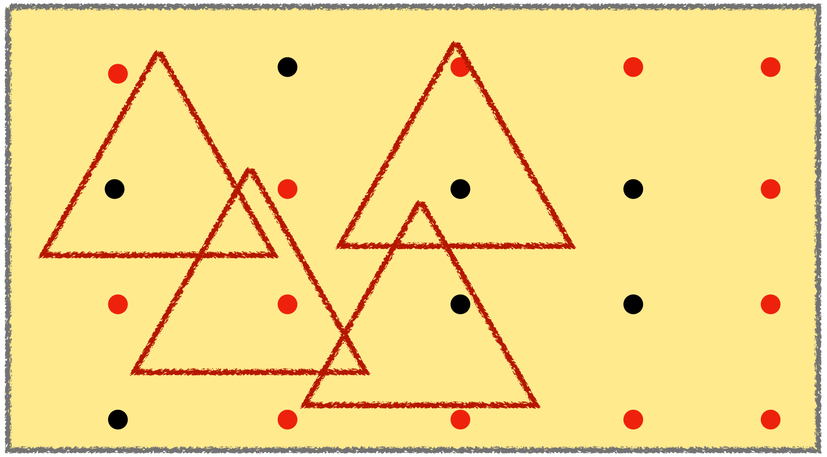
$X_1, \ldots, X_T$ adaptive smooth sequence and $\mathscr{B}$ be a family of positive functions

$$\mathbb{E}\left[\sup_{b\in\mathscr{B}} \sum_t b(X_t)\right] \leq \mathbb{E}_{Z_1,\ldots,Z_{Tk}\sim\mu}\left[\sup_{b\in\mathscr{B}} \sum_t b(Z_t)\right] + T^2 e^{-\sigma k}$$

(De)coupling inequality

(De)coupling Inequality





$X_1, \ldots, X_T$ adaptive smooth sequence and $\mathscr{B}$ be a family of positive functions

$$\mathbb{E}\left[\sup_{b\in\mathscr{B}} \sum_t b(X_t)\right] \leq \mathbb{E}_{Z_1,\ldots,Z_{Tk}\sim\mu}\left[\sup_{b\in\mathscr{B}} \sum_t b(Z_t)\right] + T^2 e^{-\sigma k}$$

Compare to naive change of measure: $\mathbb{E}\left[\sup_{b\in\mathscr{B}} \sum_t b(X_t)\right] \leq \sigma^{-T}\mathbb{E}_{Z_1,\ldots,Z_T\sim\mu}\left[\sup_{b\in\mathscr{B}} \sum_t b(Z_t)\right]$

What Objectives are Monotone?

What Objectives are Monotone?

Warning: Objectives that we care about are typically not directly monotone
E.g. Generalization, Regret, Discrepancy

What Objectives are Monotone?

Warning: Objectives that we care about are typically not directly monotone
E.g. Generalization, Regret, Discrepancy

Fortunately, typically, we reason about these objectives using monotone proxies
E.g. Rademacher complexity, Potential functions, Hereditary discrepancy

What Objectives are Monotone?

Warning: Objectives that we care about are typically not directly monotone
E.g. Generalization, Regret, Discrepancy

Fortunately, typically, we reason about these objectives using monotone proxies
E.g. Rademacher complexity, Potential functions, Hereditary discrepancy

We still need to be careful about what proxy we use!!

What Objectives are Monotone?

Warning: Objectives that we care about are typically not directly monotone
E.g. Generalization, Regret, Discrepancy

Fortunately, typically, we reason about these objectives using monotone proxies
E.g. Rademacher complexity, Potential functions, Hereditary discrepancy

We still need to be careful about what proxy we use!!

See for [HR**S**'21] a detailed discussion

Coupling Helps under Monotonicity

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t} g(X'_s)^2 - 2\cdot g(X_s)^2\right] \lesssim \mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t} \xi_s\cdot g(X_s) - \frac{g(X_s)^2}{2}\right]$$

Coupling Helps under Monotonicity

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t} g(X'_s)^2 - 2\cdot g(X_s)^2\right] \lesssim \mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\xi_s\cdot g(X_s) - \frac{g(X_s)^2}{2}\right]$$

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\xi_s\cdot g(X_s) - \frac{g(X_s)^2}{2}\right] \lesssim \mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\sum_{j=1}^{k}\xi_{s,j}\cdot g(Z_{s,j}) - \frac{g(Z_{s,j})^2}{2}\right]$$

Coupling Helps under Monotonicity

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t} g(X'_s)^2 - 2\cdot g(X_s)^2\right] \lesssim \mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\xi_s\cdot g(X_s) - \frac{g(X_s)^2}{2}\right]$$

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\xi_s\cdot g(X_s) - \frac{g(X_s)^2}{2}\right] \not\lesssim \mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\sum_{j=1}^{k}\xi_{s,j}\cdot g(Z_{s,j}) - \frac{g(Z_{s,j})^2}{2}\right]$$

Coupling Helps under Monotonicity

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t} g(X'_s)^2 - 2\cdot g(X_s)^2\right] \lesssim \mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\xi_s\cdot g(X_s) - \frac{g(X_s)^2}{2}\right]$$

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\xi_s\cdot g(X_s) - \frac{g(X_s)^2}{2}\right] \not\lesssim \mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\sum_{j=1}^{k}\xi_{s,j}\cdot g(Z_{s,j}) - \frac{g(Z_{s,j})^2}{2}\right]$$

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\xi_s\cdot g(X_s) - \frac{g(X_s)^2}{2}\right] \lesssim \log\mathbb{E}\left[\exp\left(\sup_{g\in\mathcal{G}}\frac{1}{t}\sum_{s=1}^{t}\xi_s\cdot g(X_s) - \frac{g(X_s)^2}{2}\right)\right]$$

Wills Functional is Monotone

Definition: $$\log W_T(\mathscr{F}) = \log \mathbb{E}\left[\exp\left(\sup_{f\in\mathscr{F}} \sum_{t=1}^{T} \xi_t \cdot f(X_t) - \frac{f(X_t)^2}{2}\right)\right].$$

Theorem [M'23]: The Wills functional is monotone:

$$W_T(\mathscr{F}) \le W_{T+1}(\mathscr{F}).$$

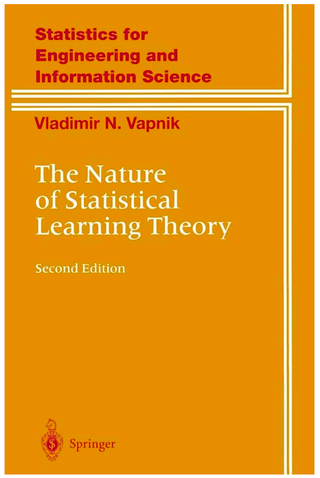
Wills Functional is Monotone

Definition: $\log W_T(\mathscr{F}) = \log \mathbb{E}\left[\exp\left(\sup_{f\in\mathscr{F}} \sum_{t=1}^{T} \xi_t \cdot f(X_t) - \frac{f(X_t)^2}{2}\right)\right].$

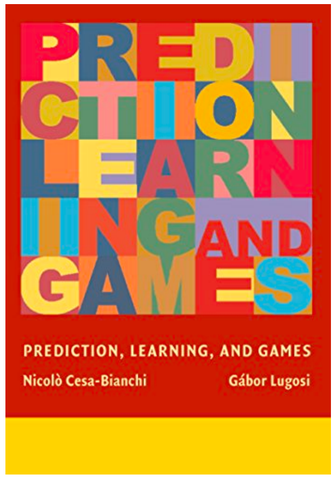
Theorem [M'23]: The Wills functional is monotone:

$$W_T(\mathscr{F}) \le W_{T+1}(\mathscr{F}).$$

$$\log \mathbb{E}\left[\exp\left(\sup_{g\in\mathscr{G}} \frac{1}{t}\sum_{s=1}^{t} \xi_s \cdot g(X_s) - \frac{g(X_s)^2}{2}\right)\right] \lesssim \log \mathbb{E}\left[\exp\left(\sup_{g\in\mathscr{G}} \frac{1}{t}\sum_{s=1}^{t}\sum_{j=1}^{k} \xi_{s,j} \cdot g(Z_{s,j}) - \frac{g(Z_{s,j})^2}{2}\right)\right]$$



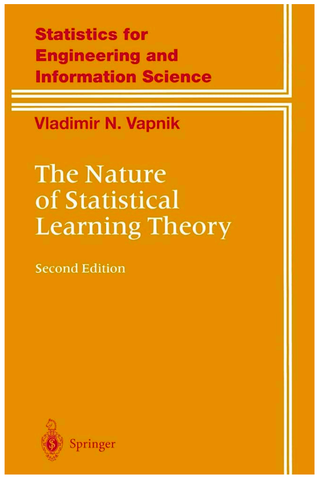
Difficulty of Learning



Smoothed data

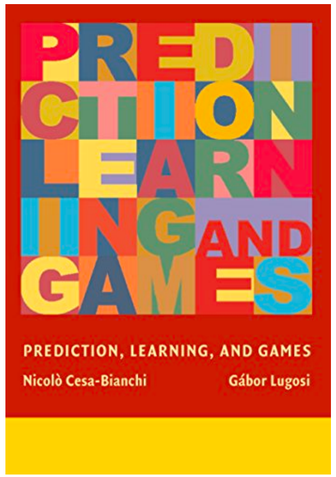
Statistical Learning

Online Learning



Coupling Lemma

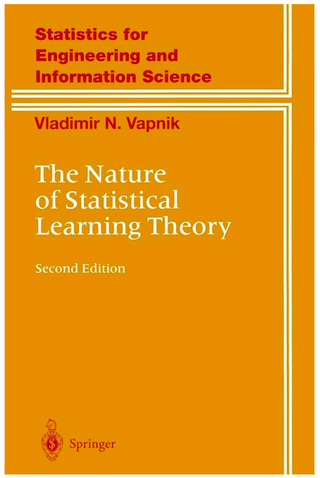
Difficulty of Learning



Smoothed data

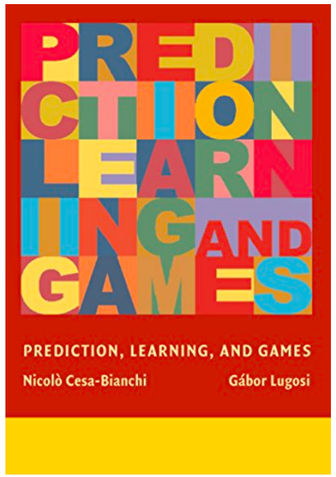
Statistical Learning

Online Learning



Coupling Lemma

Difficulty of Learning



Smoothed data

Statistical Learning

Online Learning

Story So Far

Story So Far

- Statistical Rates for ERM for well-specified

Story So Far

- Statistical Rates for ERM for well-specified

Theorem [BRS'24]: If data are σ-smooth w.r.t. μ and f_t is ERM, then

$$\mathbb{E}[\mathrm{Err}_T] \approx \sqrt{\frac{\mathrm{vc}(\mathscr{F})}{\sigma \cdot T}}.$$

Story So Far

- Statistical Rates for ERM for well-specified

Theorem [BRS'24]: If data are σ-smooth w.r.t. μ and f_t is ERM, then

$$\mathbb{E}[\mathrm{Err}_T] \approx \sqrt{\frac{\mathrm{vc}(\mathscr{F})}{\sigma \cdot T}}.$$

- Surprise Lemma

Story So Far

- Statistical Rates for ERM for well-specified

Theorem [BRS'24]: If data are σ-smooth w.r.t. μ and f_t is ERM, then

$$\mathbb{E}[\mathrm{Err}_T] \approx \sqrt{\frac{\mathrm{vc}(\mathscr{F})}{\sigma \cdot T}}.$$

- Surprise Lemma
- Coupling Lemma

Remainder of Talk

Remainder of Talk

- Does knowledge of μ help?

Remainder of Talk

- Does knowledge of μ help?
- What if there were label noise?

Remainder of Talk

- Does knowledge of μ help?
- What if there were label noise?
- Computational efficiency?

Smoothed Online Learning

1. We get T data points X_t smooth w.r.t μ and Y_t generated arbitrarily.

2. We have access to a model class $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: For each t, return $f_t \in \mathscr{F}$ with small regret

$$\mathbb{E}\left[\mathrm{Reg}_T\right] = \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \ell(f_t(X_t), Y_t) - \inf_{f \in \mathscr{F}} \sum_{t=1}^{T} \ell(f(X_t), Y_t)\right].$$

Smoothed Online Learning

1. We get T data points X_t smooth w.r.t μ and Y_t generated arbitrarily.

2. We have access to a model class $\mathscr{F} : \mathscr{X} \to \mathscr{Y}$.

Goal: For each t, return $f_t \in \mathscr{F}$ with small regret

$$\mathbb{E}\left[\mathrm{Reg}_T\right] = \frac{1}{T} \cdot \mathbb{E}\left[\sum_{t=1}^{T} \ell(f_t(X_t), Y_t) - \inf_{f \in \mathscr{F}} \sum_{t=1}^{T} \ell(f(X_t), Y_t)\right].$$

IID $X's$ are still easy to learn with arbitrary labels
Bounded in terms of the VC dimension

Categories of Smoothed Online Learning

Categories of Smoothed Online Learning

Label Noise

No noise $\Longrightarrow$ Realizable

Arbitrary noise $\Longrightarrow$ Agnostic

Intermediate models: Well-specified, RCN, Massart, …

Categories of Smoothed Online Learning

Label Noise

No noise $\Longrightarrow$ Realizable

Arbitrary noise $\Longrightarrow$ Agnostic

Intermediate models: Well-specified, RCN, Massart, …

Knowledge of μ

"Known" typically means sample access

Categories of Smoothed Online Learning

Label Noise

No noise $\Longrightarrow$ Realizable

Arbitrary noise $\Longrightarrow$ Agnostic

Intermediate models: Well-specified, RCN, Massart, …

	Known	**Unknown**
Realizable		
Agnostic		

Knowledge of μ

"Known" typically means sample access

Bounds for Smoothed Online Learning

	Known	Unknown
Realizable	$T^{-1}d\log(T/\sigma)$	$\sqrt{T^{-1}d\sigma^{-1}}$
Agnostic	$\sqrt{T^{-1}d\log(T/\sigma)}$	$\sqrt{T^{-1}d\sigma^{-1}}$

Bounds for Smoothed Online Learning

	Known	Unknown
Realizable	$T^{-1}d\log(T/\sigma)$	$\sqrt{T^{-1}d\sigma^{-1}}$
Agnostic	$\sqrt{T^{-1}d\log(T/\sigma)}$	$\sqrt{T^{-1}d\sigma^{-1}}$

Statistical Bound with Known Base Measure

Statistical Bound with Known Base Measure

Theorem [HRS'21]: For known base measure smoothed online learning we have

$$\textbf{Agnostic: } \mathbb{E}[\mathrm{Reg}_T] \approx \sqrt{\frac{\mathrm{vc}(\mathscr{F}) \cdot \log(T/\sigma)}{T}}$$

$$\textbf{Realizable: } \mathbb{E}[\mathrm{Reg}_T] \approx \frac{\mathrm{vc}(\mathscr{F}) \cdot \log(T/\sigma)}{T}$$

Achieving the Statistical Bound: Agnostic Case

Achieving the Statistical Bound: Agnostic Case

Focus on known μ in the binary classification setting

Achieving the Statistical Bound: Agnostic Case

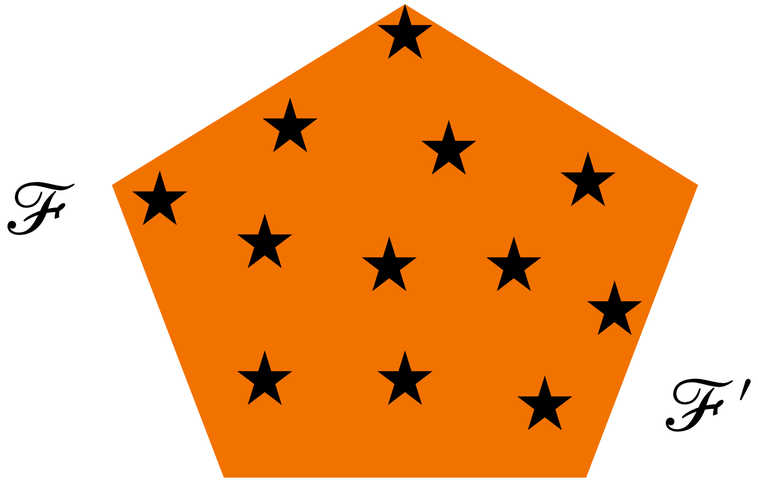
Focus on known μ in the binary classification setting

Definition: $\mathscr{F}' \subset \mathscr{F}$ **is an** ϵ**-net under** μ **if for all** $f \in \mathscr{F}$**, there exists** $f' \in \mathscr{F}'$ **such that** $\Pr_{x \sim \mu}\left[f(x) \neq f'(x)\right] \leq \epsilon$.

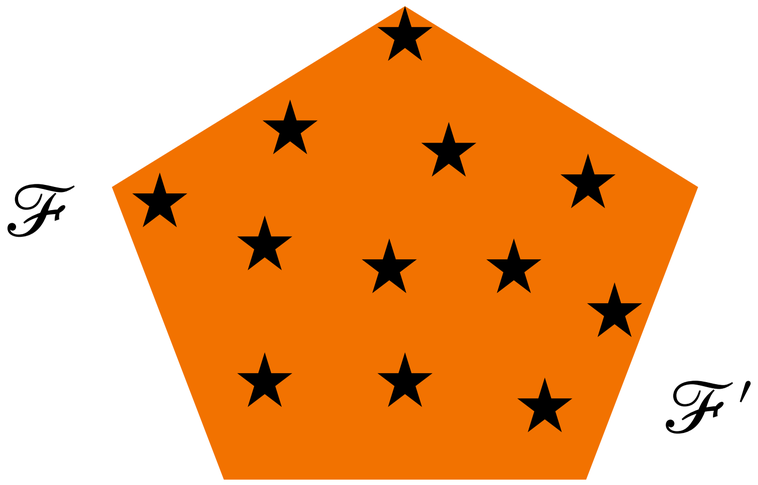
Achieving the Statistical Bound: Agnostic Case

Focus on known μ in the binary classification setting

Definition: $\mathscr{F}' \subset \mathscr{F}$ **is an** ϵ**-net under** μ **if for all** $f \in \mathscr{F}$**, there exists** $f' \in \mathscr{F}'$ **such that** $\Pr_{x \sim \mu}\left[f(x) \neq f'(x)\right] \leq \epsilon$.

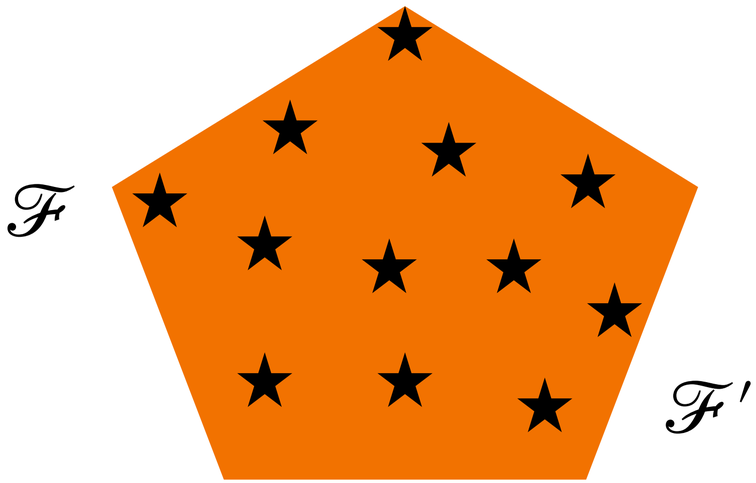


Achieving the Statistical Bound: Agnostic Case



Achieving the Statistical Bound: Agnostic Case

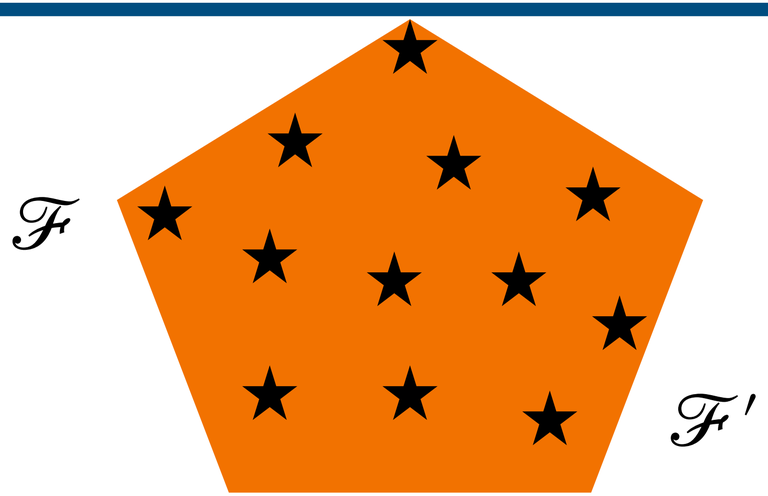
Theorem [D'66,H'87]**: VC class** d **implies** $\log|\mathscr{F}'| \lesssim d\log(1/\epsilon)$



Achieving the Statistical Bound: Agnostic Case

Theorem [D'66,H'87]**: VC class** d **implies** $\log|\mathscr{F}'| \lesssim d\log(1/\epsilon)$

Algorithm: Construct $\mathscr{F}'$ (can be done using samples from μ).
Play experts on $\mathscr{F}'$



Achieving the Statistical Bound: Agnostic Case

Achieving the Statistical Bound: Agnostic Case

Theorem [V'87,HLW'87]**: Regret with respect to best expert in** $\mathscr{F}'$

$$\sqrt{\frac{\log|\mathscr{F}'|}{T}} \lesssim \sqrt{\frac{d\log(1/\epsilon)}{T}}\,.$$

Achieving the Statistical Bound: Agnostic Case

Theorem [V'87,HLW'87]**: Regret with respect to best expert in** $\mathscr{F}'$

$$\sqrt{\frac{\log|\mathscr{F}'|}{T}} \lesssim \sqrt{\frac{d\log(1/\epsilon)}{T}}\,.$$

- How does this relate to regret with respect to $\mathscr{F}$?

Achieving the Statistical Bound: Agnostic Case

Theorem [V'87,HLW'87]**: Regret with respect to best expert in** $\mathscr{F}'$

$$\sqrt{\frac{\log|\mathscr{F}'|}{T}} \lesssim \sqrt{\frac{d\log(1/\epsilon)}{T}}\,.$$

- How does this relate to regret with respect to $\mathscr{F}$?
- We need to bound

$$\mathbb{E}_{X_i\sim\mathscr{D}_i}\left[\sup_{f\in\mathscr{F}}\inf_{f'\in\mathscr{F}'}\sum_{i=1}^{T}\mathbb{I}\left[f(X_i)\neq f'(X_i)\right]\right]$$

Smoothness $\Longrightarrow$ any fixed f, $\mathbb{E}\left[\mathbb{I}\left[f(X_i)\neq f'(X_i)\right]\right] \leq \epsilon\sigma^{-1}$

Achieving the Statistical Bound: Agnostic Case

Achieving the Statistical Bound: Agnostic Case

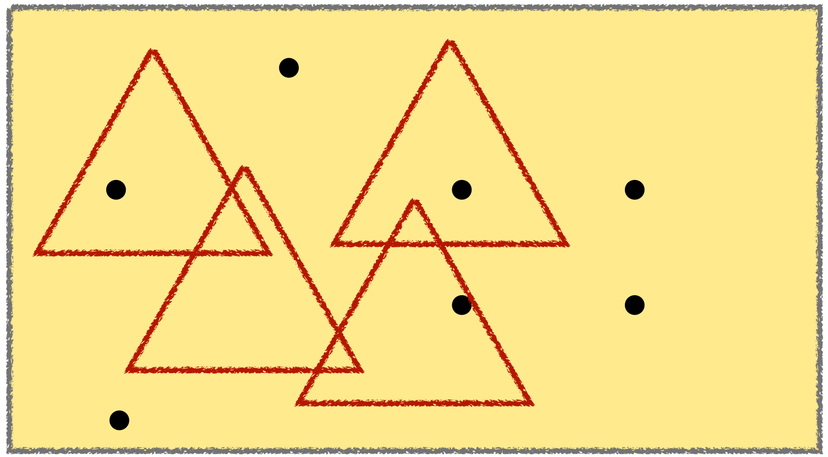
- **Main challenge:** For adaptive $\mathscr{D}_i$ depends on the previous draws, X_j are not independent $\Longrightarrow$ can't apply VC theorem/symmetrization

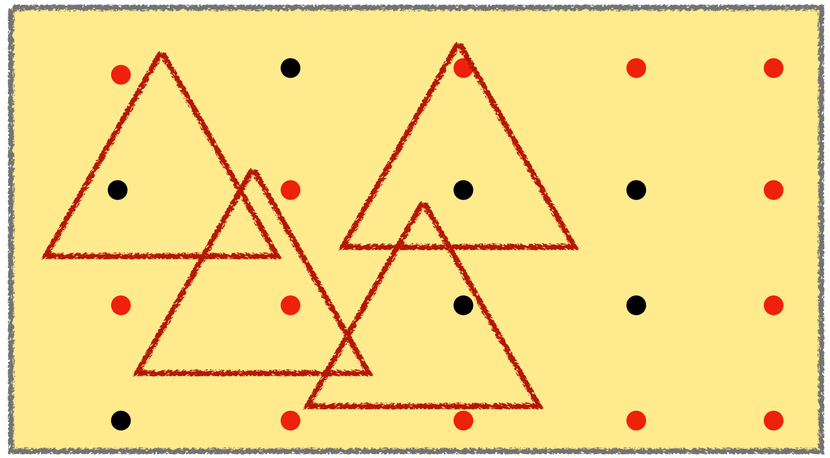
Achieving the Statistical Bound: Agnostic Case

- **Main challenge:** For adaptive $\mathscr{D}_i$ depends on the previous draws, X_j are not independent $\Longrightarrow$ can't apply VC theorem/symmetrization
- In particular, need to bound the following empirical process
  Let $\mathscr{B}$ be a VC class (of positive functions) such that $b \in \mathscr{B}$ has $\mathbb{E}_\mu b \leq \epsilon$ and let $X_1, \ldots, X_T$ be generated from an adaptive sequence of smooth distributions

$$\mathbb{E}\left[\sup_{b\in\mathscr{B}} \sum_i b(X_i)\right]$$

(De)coupling Inequality





$X_1, \ldots, X_T$ adaptive smooth sequence and $\mathscr{B}$ be a family of positive functions

$$\mathbb{E}\left[\sup_{b\in\mathscr{B}} \sum_t b(X_t)\right] \leq \mathbb{E}_{Z_1,\ldots,Z_{Tk}\sim\mu}\left[\sup_{b\in\mathscr{B}} \sum_t b(Z_t)\right] + T^2 e^{-\sigma k}$$

(De)coupling inequality

Completing the Proof

Completing the Proof

- Let $\mathscr{B}$ be a VC class (of positive functions) such that $b \in \mathscr{B}$ has $\mathbb{E}_\mu b \leq \epsilon$ and let $X_1, \ldots, X_T$ be generated from an adaptive sequence of smooth distributions

$$\mathbb{E}\left[\sup_{b \in \mathscr{B}} \sum_i b(X_i)\right]$$

Completing the Proof

- Let $\mathscr{B}$ be a VC class (of positive functions) such that $b \in \mathscr{B}$ has $\mathbb{E}_\mu b \leq \epsilon$ and let $X_1, \ldots, X_T$ be generated from an adaptive sequence of smooth distributions

$$\mathbb{E}\left[\sup_{b\in\mathscr{B}} \sum_i b(X_i)\right]$$

- Apply coupling lemma,

$$\mathbb{E}\sup_{b\in\mathscr{B}} \sum_i b(X_i) \lesssim \mathbb{E}\sup_{b\in\mathscr{B}} \sum_{i,j} b(Z_{i,j})$$

Completing the Proof

- Let $\mathscr{B}$ be a VC class (of positive functions) such that $b \in \mathscr{B}$ has $\mathbb{E}_\mu b \leq \epsilon$ and let $X_1, \ldots, X_T$ be generated from an adaptive sequence of smooth distributions

$$\mathbb{E}\left[\sup_{b\in\mathscr{B}} \sum_i b(X_i)\right]$$

- Apply coupling lemma,

$$\mathbb{E}\sup_{b\in\mathscr{B}} \sum_i b(X_i) \lesssim \mathbb{E}\sup_{b\in\mathscr{B}} \sum_{i,j} b(Z_{i,j})$$

- $Z_{i,j}$ independent $\Longrightarrow$ apply VC theorem/symmetrization

Completing the Proof

Completing the Proof

- VC theorem implies

$$\mathbb{E}_{x_i \sim \mathscr{D}_i}\left[\sup_{f \in \mathscr{F}} \inf_{f' \in \mathscr{F}'} \sum_{i=1}^{T} \mathbb{I}\left[f(x_t) \neq f'(x_t)\right]\right] \leq \epsilon\sigma^{-1}T + \sqrt{T\epsilon\sigma^{-1} \cdot \mathsf{vc}(\mathscr{F})}$$

Completing the Proof

- VC theorem implies

$$\mathbb{E}_{x_i \sim \mathscr{D}_i}\left[\sup_{f \in \mathscr{F}} \inf_{f' \in \mathscr{F}'} \sum_{i=1}^{T} \mathbb{I}\left[f(x_t) \neq f'(x_t)\right]\right] \leq \epsilon\sigma^{-1}T + \sqrt{T\epsilon\sigma^{-1} \cdot \mathsf{vc}(\mathscr{F})}$$

- Recall: Regret with respect to best expert in $\mathscr{F}'$: $\sqrt{d\log(1/\epsilon)/T}$

Completing the Proof

- VC theorem implies

$$\mathbb{E}_{x_i \sim \mathscr{D}_i}\left[\sup_{f \in \mathscr{F}} \inf_{f' \in \mathscr{F}'} \sum_{i=1}^{T} \mathbb{I}\left[f(x_t) \neq f'(x_t)\right]\right] \leq \epsilon \sigma^{-1} T + \sqrt{T \epsilon \sigma^{-1} \cdot \mathsf{vc}(\mathscr{F})}$$

- Recall: Regret with respect to best expert in $\mathscr{F}'$: $\sqrt{d \log(1/\epsilon)/T}$
- Setting $\epsilon = \sigma T^{-1}$ gives regret bound

Completing the Proof

- VC theorem implies

$$\mathbb{E}_{x_i \sim \mathscr{D}_i} \left[\sup_{f \in \mathscr{F}} \inf_{f' \in \mathscr{F}'} \sum_{i=1}^{T} \mathbb{I}\left[f(x_t) \neq f'(x_t) \right] \right] \leq \epsilon \sigma^{-1} T + \sqrt{T \epsilon \sigma^{-1} \cdot \mathrm{vc}(\mathscr{F})}$$

- Recall: Regret with respect to best expert in $\mathscr{F}'$: $\sqrt{d \log(1/\epsilon)/T}$
- Setting $\epsilon = \sigma T^{-1}$ gives regret bound

Naive change of measure on the sequence would have paid σ^{-T}

Completing the Proof

- VC theorem implies

$$\mathbb{E}_{x_i \sim \mathscr{D}_i}\left[\sup_{f\in\mathscr{F}}\inf_{f'\in\mathscr{F}'}\sum_{i=1}^{T}\mathbb{I}\left[f(x_t)\neq f'(x_t)\right]\right] \leq \epsilon\sigma^{-1}T + \sqrt{T\epsilon\sigma^{-1}\cdot \mathrm{vc}(\mathscr{F})}$$

Bernstein

Important to get log dependence

- Recall: Regret with respect to best expert in $\mathscr{F}'$: $\sqrt{d\log(1/\epsilon)/T}$
- Setting $\epsilon = \sigma T^{-1}$ gives regret bound

Naive change of measure on the sequence would have paid σ^{-T}

Main Theorem

Main Theorem

Theorem [HRS'21]: Known base measure smoothed online learning we have

$$\mathbb{E}[\mathrm{Reg}_T] \approx \sqrt{\frac{\mathrm{vc}(\mathscr{F}) \cdot \log(T/\sigma)}{T}}.$$

Main Theorem

Theorem [HRS'21]: Known base measure smoothed online learning we have

$$\mathbb{E}[\mathrm{Reg}_T] \approx \sqrt{\frac{\mathrm{vc}(\mathscr{F}) \cdot \log(T/\sigma)}{T}}.$$

- This can be extended to non-parametric classes (essentially whenever covering numbers are bounded) [BDGR'22, HHSY'22]

Main Theorem

Theorem [HRS'21]: Known base measure smoothed online learning we have

$$\mathbb{E}[\mathrm{Reg}_T] \approx \sqrt{\frac{\mathrm{vc}(\mathscr{F}) \cdot \log(T/\sigma)}{T}}.$$

- This can be extended to non-parametric classes (essentially whenever covering numbers are bounded) [BDGR'22, HHSY'22]
- Handling the nonparametric case needs different ideas (Distributional Sequential Rademacher complexity)

Main Theorem

Theorem [HRS'21]: Known base measure smoothed online learning we have

$$\mathbb{E}[\mathrm{Reg}_T] \approx \sqrt{\frac{\mathrm{vc}(\mathscr{F}) \cdot \log(T/\sigma)}{T}}.$$

- This can be extended to non-parametric classes (essentially whenever covering numbers are bounded) [BDGR'22, HHSY'22]
- Handling the nonparametric case needs different ideas (Distributional Sequential Rademacher complexity)
- Whether a "natural" covering-based "algorithm" exists is an interesting open question

Bounds for Smoothed Online Learning

Same idea but with a first order algorithm

	Known	Unknown
Realizable	$T^{-1}d\log(T/\sigma)$	$\sqrt{T^{-1}d\sigma^{-1}}$
Agnostic	$\sqrt{T^{-1}d\log(T/\sigma)}$	$\sqrt{T^{-1}d\sigma^{-1}}$

Bounds for Smoothed Online Learning

	Known	Unknown
Realizable	$T^{-1}d\log(T/\sigma)$	$\sqrt{T^{-1}d\sigma^{-1}}$
Agnostic	$\sqrt{T^{-1}d\log(T/\sigma)}$	$\sqrt{T^{-1}d\sigma^{-1}}$

Part 1

Algorithm for Unknown Base Measure

Algorithm for Unknown Base Measure

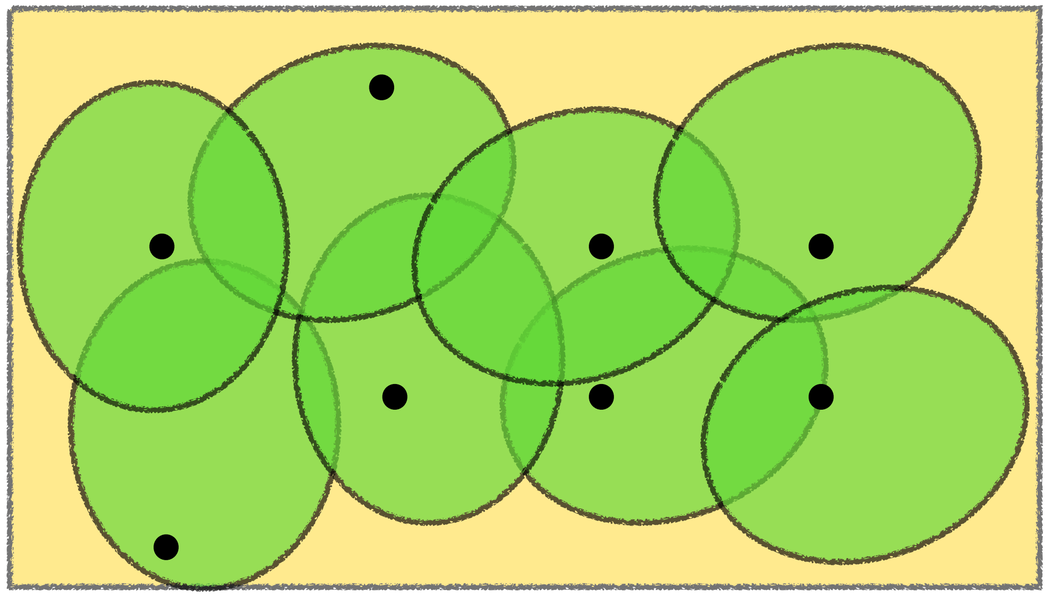
- When μ is not known, we can't construct a net for $\mathcal{F}$!

Algorithm for Unknown Base Measure

- When μ is not known, we can't construct a net for $\mathcal{F}$!
- The only "clue" we have about μ is from the realized samples.
  But not enough samples to "learn" μ. In fact, not necessarily identifiable

Algorithm for Unknown Base Measure

- When μ is not known, we can't construct a net for $\mathcal{F}$!
- The only "clue" we have about μ is from the realized samples.
  But not enough samples to "learn" μ. In fact, not necessarily identifiable



Algorithm for Unknown Base Measure

- When μ is not known, we can't construct a net for $\mathcal{F}$!
- The only "clue" we have about μ is from the realized samples.
  But not enough samples to "learn" μ. In fact, not necessarily identifiable
- Surprise Lemma to the rescue

Algorithm for Unknown Base Measure

- When μ is not known, we can't construct a net for $\mathscr{F}$!
- The only "clue" we have about μ is from the realized samples.
  But not enough samples to "learn" μ. In fact, not necessarily identifiable
- Surprise Lemma to the rescue

$$\textbf{Let } \overline{p}_t = \frac{1}{t}\sum_{s=1}^{t} p_s.$$

$$\textbf{Then, } p_t \lesssim \frac{\log(T)}{\sigma \cdot t} + \log(T) \cdot \overline{p}_{t-1} \textbf{ for most } t.$$

Algorithm for Unknown Base Measure

- When μ is not known, we can't construct a net for $\mathscr{F}$!
- The only "clue" we have about μ is from the realized samples.
  But not enough samples to "learn" μ. In fact, not necessarily identifiable
- Surprise Lemma to the rescue
- Instead of likelihood ratio, keep track of number of times a net on the historical data is not a good representation of $\mathscr{F}$ for future data
- With a clever epoching idea [B'25] gets $\sqrt{d\sigma^{-1}T}$ rate.

Bounds for Smoothed Online Learning

	Known	Unknown
Realizable	$T^{-1}d\log(T/\sigma)$	$\sqrt{T^{-1}d\sigma^{-1}}$
Agnostic	$\sqrt{T^{-1}d\log(T/\sigma)}$	$\sqrt{T^{-1}d\sigma^{-1}}$

Efficiency

Efficiency

- So far, we have not talked too much about efficiency

Efficiency

- So far, we have not talked too much about efficiency
- What is the right notion of efficiency here?

Efficiency

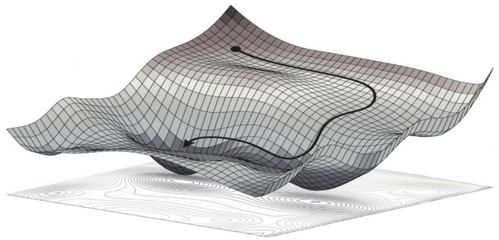
- So far, we have not talked too much about efficiency
- What is the right notion of efficiency here?
    - We want to reason about **arbitrary** concept classes

Efficiency

- So far, we have not talked too much about efficiency
- What is the right notion of efficiency here?
    - We want to reason about **arbitrary** concept classes
- **Oracle Efficiency**: Assume access to optimization “oracle” for class

Efficiency

- So far, we have not talked too much about efficiency
- What is the right notion of efficiency here?
    - We want to reason about **arbitrary** concept classes
- **Oracle Efficiency**: Assume access to optimization “oracle” for class



Deep learning

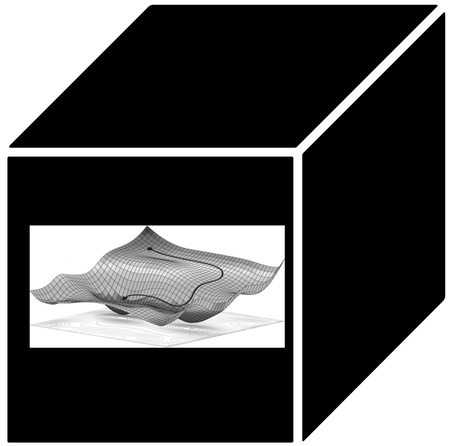
SAT Solvers

Integer Programming

Oracle Efficiency

Empirical Risk Minimization

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} L_T(f) \qquad L_T(f) = \frac{1}{T} \sum_{t=1}^{T} \ell(f(X_t), Y_t)$$



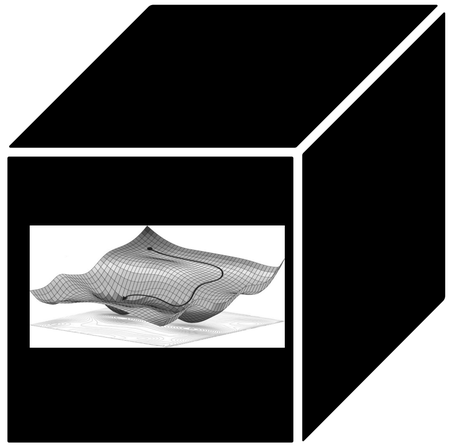
ERM is sufficient* for statistical learning

Can we efficiently reduce online learning to statistical learning?

Oracle Efficiency

Empirical Risk Minimization

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} L_T(f) \qquad L_T(f) = \frac{1}{T} \sum_{t=1}^{T} \ell(f(X_t), Y_t)$$



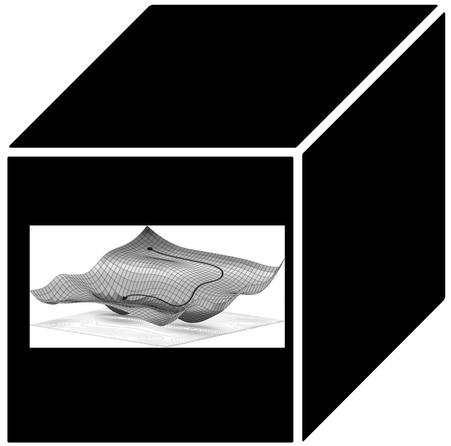
Can we efficiently reduce online learning to statistical learning?

Without smoothness, Oracle efficiency not achievable [HK'16]

Oracle Efficiency

Empirical Risk Minimization

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} L_T(f) \qquad L_T(f) = \frac{1}{T} \sum_{t=1}^{T} \ell(f(X_t), Y_t)$$

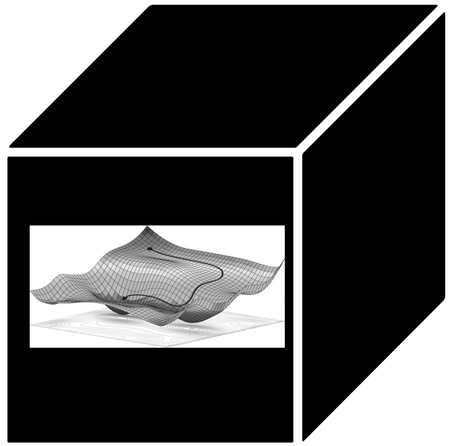


Can we efficiently reduce online learning to statistical learning?

Oracle Efficiency

Empirical Risk Minimization

$$\hat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} L_T(f) \qquad L_T(f) = \frac{1}{T} \sum_{t=1}^{T} \ell(f(X_t), Y_t)$$



Can we efficiently reduce online learning to statistical learning?

With smoothness, Oracle efficiency is achievable

Bounds for Efficient Smoothed Online Learning

	Known	Unknown
Realizable (Efficiency)	$T^{-1}d\log(T/\sigma)$	$\sqrt{T^{-1}d\sigma^{-1}}$
	$\sqrt{T^{-1}d\sigma^{-1}}$	$\sqrt{T^{-1}d\sigma^{-1}}$
Agnostic (Efficient)	$\sqrt{T^{-1}d\log(T/\sigma)}$	$\sqrt{T^{-1}d\sigma^{-1}}$
	$\sqrt{T^{-1}d\sigma^{-1}}$	???

Bounds for Efficient Smoothed Online Learning

	Known	Unknown
Realizable (Efficiency)	$T^{-1}d\log(T/\sigma)$	$\sqrt{T^{-1}d\sigma^{-1}}$
	$\sqrt{T^{-1}d\sigma^{-1}}$	$\sqrt{T^{-1}d\sigma^{-1}}$
Agnostic (Efficient)	$\sqrt{T^{-1}d\log(T/\sigma)}$	$\sqrt{T^{-1}d\sigma^{-1}}$
	$\sqrt{T^{-1}d\sigma^{-1}}$	???

Bounds for Efficient Smoothed Online Learning

	Known	Unknown
Realizable (Efficiency)	$T^{-1}d\log(T/\sigma)$	$\sqrt{T^{-1}d\sigma^{-1}}$
	$\sqrt{T^{-1}d\sigma^{-1}}$	$\sqrt{T^{-1}d\sigma^{-1}}$
Agnostic (Efficient)	$\sqrt{T^{-1}d\log(T/\sigma)}$	$\sqrt{T^{-1}d\sigma^{-1}}$
	$\sqrt{T^{-1}d\sigma^{-1}}$	???

Oracle Efficiency with Known Measure

Oracle Efficiency with Known Measure

- μ is known and access to ERM oracle

Oracle Efficiency with Known Measure

- μ is known and access to ERM oracle
- Algorithmic framework: Follow-the-perturbed leader [KV'05]

Oracle Efficiency with Known Measure

- μ is known and access to ERM oracle
- Algorithmic framework: Follow-the-perturbed leader [KV'05]
- Historical data: $S_t = \{(X_i, Y_i)\}_{i \le t}$

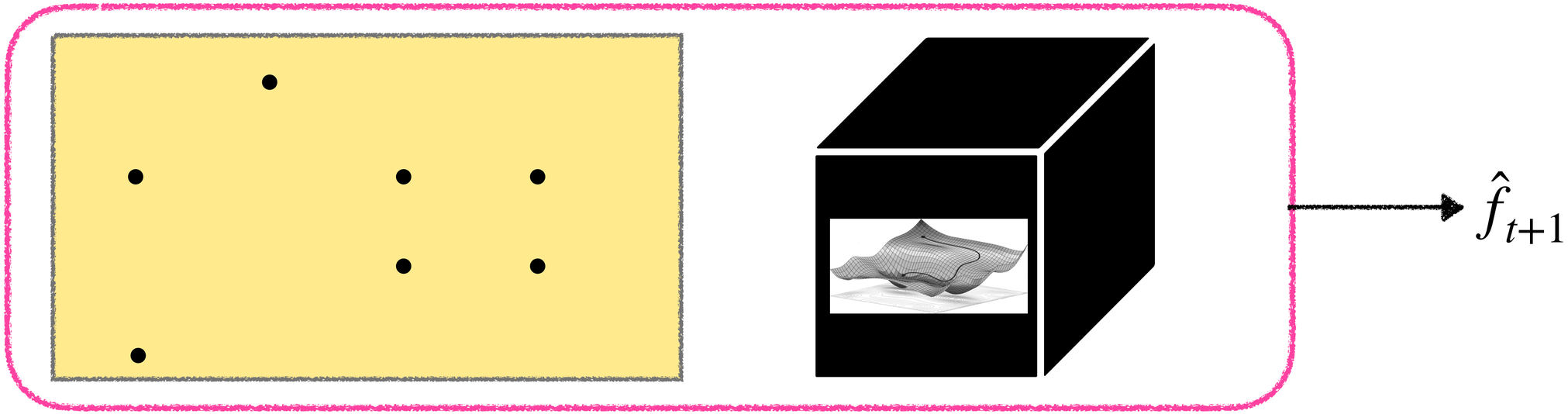
Oracle Efficiency with Known Measure

- μ is known and access to ERM oracle
- Algorithmic framework: Follow-the-perturbed leader [KV'05]
- Historical data: $S_t = \{(X_i, Y_i)\}_{i \le t}$
- Given X_{t+1} make a prediction for Y_{t+1}

Oracle Efficiency with Known Measure

- μ is known and access to ERM oracle
- Algorithmic framework: Follow-the-perturbed leader [KV'05]
- Historical data: $S_t = \{(X_i, Y_i)\}_{i \le t}$
- Given X_{t+1} make a prediction for Y_{t+1}

Historical Data S_{t-1}



$\hat{f}_{t+1}$

Oracle Efficiency with Known Measure

Algorithm: Sample $\{Z_i\} \sim \mu$. Label at random

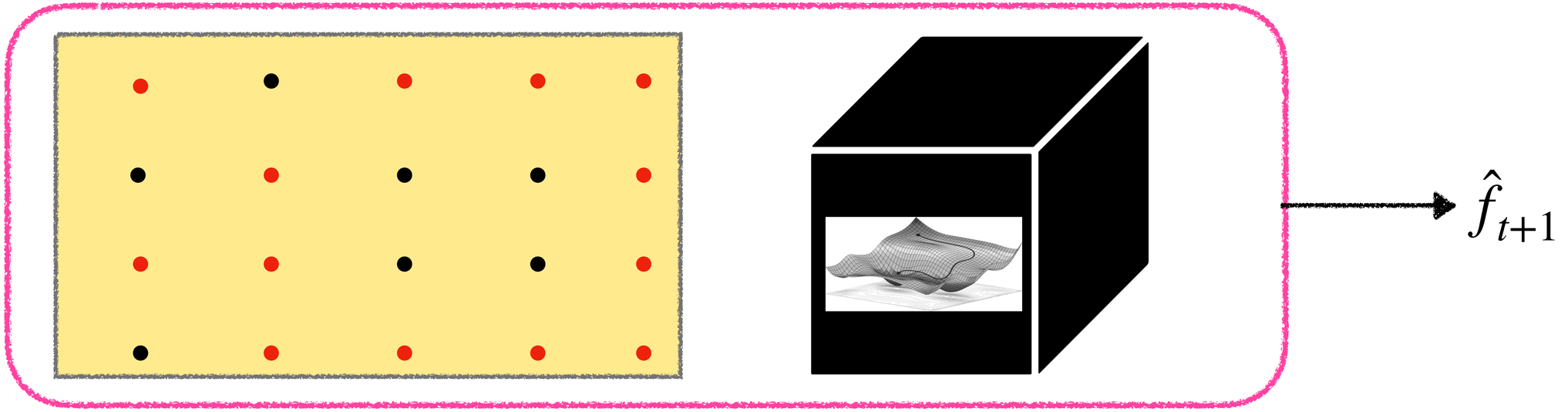
Run ERM on Historical data: $S_t = \{(X_i, Y_i)\}_{i \leq t} \cup$ Hallucinated data $\{Z_i, \tilde{Y}_i\}$

Oracle Efficiency with Known Measure

Algorithm: Sample $\{Z_i\} \sim \mu$. Label at random

Run ERM on Historical data: $S_t = \{(X_i, Y_i)\}_{i \le t} \cup$ Hallucinated data $\{Z_i, \tilde{Y}_i\}$

Historical Data $S_{t-1} \cup$ Hallucinated Data



$\hat{f}_{t+1}$

Analysis: Stability

Algorithm: Sample $\{Z_i\} \sim \mu$. Label at random

Run ERM on Historical data: $S_t = \{(X_i, Y_i)\}_{i \leq t} \cup$ Hallucinated data $\{Z_i, \tilde{Y}_i\}$

Analysis: Stability

Algorithm: Sample $\{Z_i\} \sim \mu$. Label at random

Run ERM on Historical data: $S_t = \{(X_i, Y_i)\}_{i \le t} \cup$ Hallucinated data $\{Z_i, \tilde{Y}_i\}$

In typical analysis of FTPL-type algorithms, we look at stability

Analysis: Stability

Algorithm: Sample $\{Z_i\} \sim \mu$. Label at random

Run ERM on Historical data: $S_t = \{(X_i, Y_i)\}_{i \le t} \cup$ Hallucinated data $\{Z_i, \tilde{Y}_i\}$

In typical analysis of FTPL-type algorithms, we look at stability

$$\mathbb{E}_{X_t \sim \mathscr{D}_t} \left[\mathbb{E}_{\hat{f}_t \sim \mathrm{ALG}_t} \ell(\hat{f}_t(X_t), Y_t) - \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1}} \ell(\hat{f}_{t+1}(X_t), Y_t) \right]$$

Analysis: Stability

Algorithm: Sample $\{Z_i\} \sim \mu$. Label at random

Run ERM on Historical data: $S_t = \{(X_i, Y_i)\}_{i \le t} \cup$ Hallucinated data $\{Z_i, \tilde{Y}_i\}$

In typical analysis of FTPL-type algorithms, we look at stability

$$\mathbb{E}_{X_t \sim \mathscr{D}_t}\left[\mathbb{E}_{\hat{f}_t \sim \mathrm{ALG}_t} \ell(\hat{f}_t(X_t), Y_t) - \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1}} \ell(\hat{f}_{t+1}(X_t), Y_t)\right]$$

(X_t, Y_t) in "training data"

Analysis: Stability

Algorithm: Sample $\{Z_i\} \sim \mu$. Label at random

Run ERM on Historical data: $S_t = \{(X_i, Y_i)\}_{i \le t} \cup$ Hallucinated data $\{Z_i, \tilde{Y}_i\}$

In typical analysis of FTPL-type algorithms, we look at stability

$$\mathbb{E}_{X_t \sim \mathscr{D}_t} \left[\mathbb{E}_{\hat{f}_t \sim \mathrm{ALG}_t} \ell(\hat{f}_t(X_t), Y_t) - \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1}} \ell(\hat{f}_{t+1}(X_t), Y_t) \right]$$

(X_t, Y_t) in "training data"

Observation: connection to Rademacher/Gaussian processes is due to the Hallucinated data having random signs

Analysis: Stability Decomposition

Analysis: Stability Decomposition

Theorem

Analysis: Stability Decomposition

Theorem

$$\mathbb{E}_{X_t \sim \mathscr{D}_t}\left[\mathbb{E}_{\hat{f}_t \sim \mathrm{ALG}_t} \ell(\hat{f}_t(X_t), Y_t) - \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1}} \ell(\hat{f}_{t+1}(X_t), Y_t)\right] \leq$$

Analysis: Stability Decomposition

Theorem

$$\mathbb{E}_{X_t \sim \mathscr{D}_t}\left[\mathbb{E}_{\hat{f}_t \sim \mathrm{ALG}_t} \ell(\hat{f}_t(X_t), Y_t) - \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1}} \ell(\hat{f}_{t+1}(X_t), Y_t)\right] \leq$$

$$\mathrm{TV}(\mathrm{ALG}_t, \mathbb{E}_{z_t \sim \mathscr{D}_t}[\mathrm{ALG}_{t+1}])$$

Analysis: Stability Decomposition

Theorem

$$\mathbb{E}_{X_t \sim \mathscr{D}_t}\left[\mathbb{E}_{\hat{f}_t \sim \mathrm{ALG}_t} \ell(\hat{f}_t(X_t), Y_t) - \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1}} \ell(\hat{f}_{t+1}(X_t), Y_t)\right] \leq$$

$$\mathrm{TV}(\mathrm{ALG}_t, \mathbb{E}_{z_t \sim \mathscr{D}_t}[\mathrm{ALG}_{t+1}])$$

$$+ \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1} X_t, X_t' \sim \mathscr{D}_t}\left[\ell(\hat{f}_{t+1}(X_t), Y_t) - \ell(\hat{f}_{t+1}(X_t'), Y_t')\right]$$

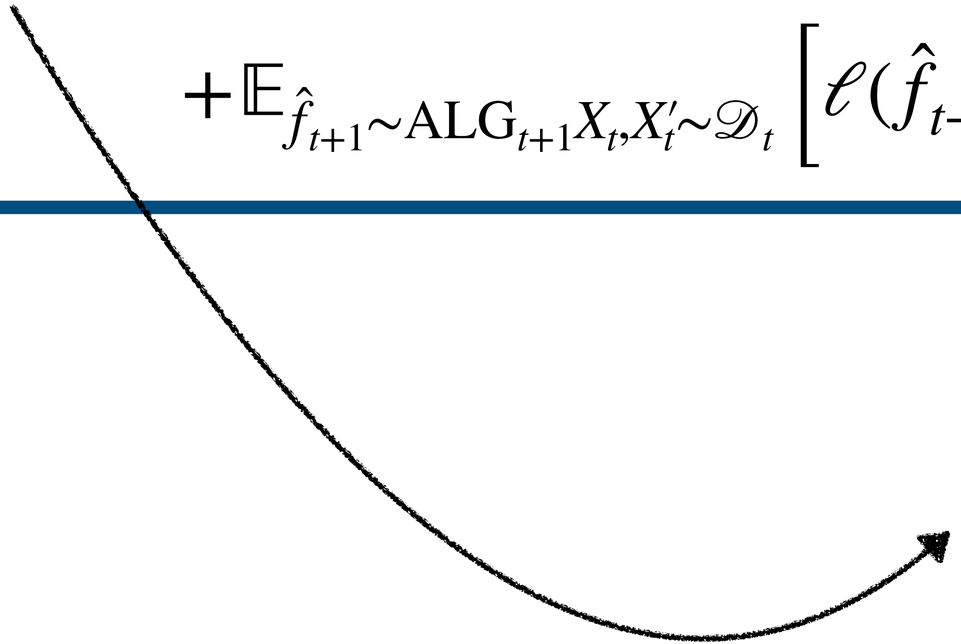
Analysis: Stability Decomposition

Theorem

$$\mathbb{E}_{X_t \sim \mathscr{D}_t}\left[\mathbb{E}_{\hat{f}_t \sim \mathrm{ALG}_t} \ell(\hat{f}_t(X_t), Y_t) - \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1}} \ell(\hat{f}_{t+1}(X_t), Y_t)\right] \leq$$

$$\mathrm{TV}(\mathrm{ALG}_t, \mathbb{E}_{z_t \sim \mathscr{D}_t}[\mathrm{ALG}_{t+1}])$$

$$+ \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1} X_t, X_t' \sim \mathscr{D}_t}\left[\ell(\hat{f}_{t+1}(X_t), Y_t) - \ell(\hat{f}_{t+1}(X_t'), Y_t')\right]$$



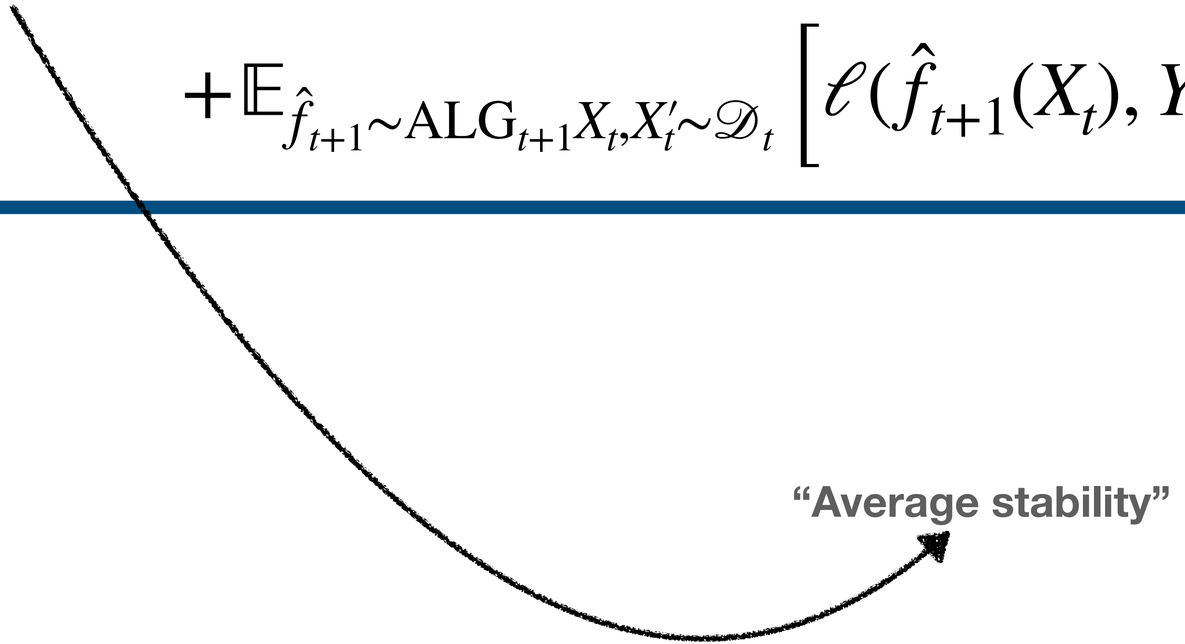
Analysis: Stability Decomposition

Theorem

$$\mathbb{E}_{X_t \sim \mathscr{D}_t}\left[\mathbb{E}_{\hat{f}_t \sim \mathrm{ALG}_t} \ell(\hat{f}_t(X_t), Y_t) - \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1}} \ell(\hat{f}_{t+1}(X_t), Y_t)\right] \leq$$

$$\mathrm{TV}(\mathrm{ALG}_t, \mathbb{E}_{z_t \sim \mathscr{D}_t}[\mathrm{ALG}_{t+1}])$$

$$+ \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1} X_t, X_t' \sim \mathscr{D}_t}\left[\ell(\hat{f}_{t+1}(X_t), Y_t) - \ell(\hat{f}_{t+1}(X_t'), Y_t')\right]$$



"Average stability"

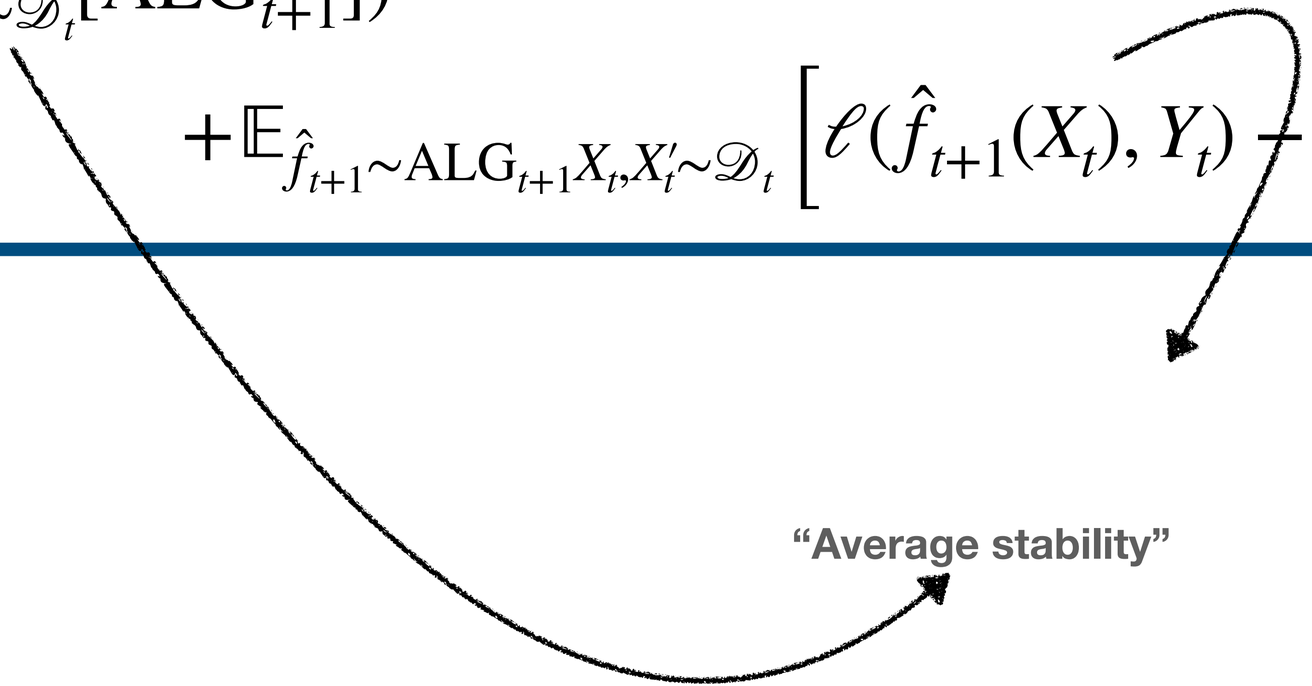
Analysis: Stability Decomposition

Theorem

$$\mathbb{E}_{X_t \sim \mathscr{D}_t}\left[\mathbb{E}_{\hat{f}_t \sim \mathrm{ALG}_t} \ell(\hat{f}_t(X_t), Y_t) - \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1}} \ell(\hat{f}_{t+1}(X_t), Y_t)\right] \leq$$

$$\mathrm{TV}(\mathrm{ALG}_t, \mathbb{E}_{z_t \sim \mathscr{D}_t}[\mathrm{ALG}_{t+1}])$$

$$+ \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1} X_t, X_t' \sim \mathscr{D}_t}\left[\ell(\hat{f}_{t+1}(X_t), Y_t) + \ell(\hat{f}_{t+1}(X_t'), Y_t')\right]$$



Analysis: Stability Decomposition

Theorem

$$\mathbb{E}_{X_t \sim \mathscr{D}_t}\left[\mathbb{E}_{\hat{f}_t \sim \mathrm{ALG}_t} \ell(\hat{f}_t(X_t), Y_t) - \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1}} \ell(\hat{f}_{t+1}(X_t), Y_t)\right] \leq$$

$$\mathrm{TV}(\mathrm{ALG}_t, \mathbb{E}_{z_t \sim \mathscr{D}_t}[\mathrm{ALG}_{t+1}])$$

$$+ \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1} X_t, X_t' \sim \mathscr{D}_t}\left[\ell(\hat{f}_{t+1}(X_t), Y_t) + \ell(\hat{f}_{t+1}(X_t'), Y_t')\right]$$

"Generalization error" under resampling from smooth distribution

"Average stability"

Analysis: Stability Decomposition

Theorem

$$\mathbb{E}_{X_t \sim \mathscr{D}_t}\left[\mathbb{E}_{\hat{f}_t \sim \mathrm{ALG}_t} \ell(\hat{f}_t(X_t), Y_t) - \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1}} \ell(\hat{f}_{t+1}(X_t), Y_t)\right] \leq$$

$$\mathrm{TV}(\mathrm{ALG}_t, \mathbb{E}_{z_t \sim \mathscr{D}_t}[\mathrm{ALG}_{t+1}])$$

$$+ \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1} X_t, X_t' \sim \mathscr{D}_t}\left[\ell(\hat{f}_{t+1}(X_t), Y_t) + \ell(\hat{f}_{t+1}(X_t'), Y_t')\right]$$

"Generalization error" under resampling from smooth distribution

"Average stability"

Analysis using stability of emp processes

Faster rate under binary using Ingster-Suslina+poisson

Analysis: Stability Decomposition

Theorem

$$\mathbb{E}_{X_t \sim \mathscr{D}_t}\left[\mathbb{E}_{\hat{f}_t \sim \mathrm{ALG}_t} \ell(\hat{f}_t(X_t), Y_t) - \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1}} \ell(\hat{f}_{t+1}(X_t), Y_t)\right] \leq$$

$$\mathrm{TV}(\mathrm{ALG}_t, \mathbb{E}_{z_t \sim \mathscr{D}_t}[\mathrm{ALG}_{t+1}])$$

$$+ \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1} X_t, X_t' \sim \mathscr{D}_t}\left[\ell(\hat{f}_{t+1}(X_t), Y_t) + \ell(\hat{f}_{t+1}(X_t'), Y_t')\right]$$

"Generalization error" under resampling from smooth distribution

"Average stability"

Analysis using stability of emp processes

Analysis: Stability Decomposition

Theorem

$$\mathbb{E}_{X_t \sim \mathscr{D}_t}\left[\mathbb{E}_{\hat{f}_t \sim \mathrm{ALG}_t} \ell(\hat{f}_t(X_t), Y_t) - \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1}} \ell(\hat{f}_{t+1}(X_t), Y_t)\right] \leq$$

$$\mathrm{TV}(\mathrm{ALG}_t, \mathbb{E}_{z_t \sim \mathscr{D}_t}[\mathrm{ALG}_{t+1}])$$

$$+ \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1} X_t, X_t' \sim \mathscr{D}_t}\left[\ell(\hat{f}_{t+1}(X_t), Y_t) + \ell(\hat{f}_{t+1}(X_t'), Y_t')\right]$$

"Generalization error" under resampling from smooth distribution using coupling to extract a subsequence from the hallucinated examples that "looks" like $\mathscr{D}_{t+1}$

"Average stability"

Analysis using stability of emp processes

Analysis: Stability Decomposition

Theorem

$$\mathbb{E}_{X_t \sim \mathscr{D}_t}\left[\mathbb{E}_{\hat{f}_t \sim \mathrm{ALG}_t} \ell(\hat{f}_t(X_t), Y_t) - \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1}} \ell(\hat{f}_{t+1}(X_t), Y_t)\right] \leq$$

$$\mathrm{TV}(\mathrm{ALG}_t, \mathbb{E}_{z_t \sim \mathscr{D}_t}[\mathrm{ALG}_{t+1}])$$

$$+ \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1} X_t, X_t' \sim \mathscr{D}_t}\left[\ell(\hat{f}_{t+1}(X_t), Y_t) - \ell(\hat{f}_{t+1}(X_t'), Y_t')\right]$$

Analysis: Stability Decomposition

Theorem

$$\mathbb{E}_{X_t \sim \mathscr{D}_t}\left[\mathbb{E}_{\hat{f}_t \sim \mathrm{ALG}_t} \ell(\hat{f}_t(X_t), Y_t) - \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1}} \ell(\hat{f}_{t+1}(X_t), Y_t)\right] \leq$$

$$\mathrm{TV}(\mathrm{ALG}_t, \mathbb{E}_{z_t \sim \mathscr{D}_t}[\mathrm{ALG}_{t+1}])$$

$$+ \mathbb{E}_{\hat{f}_{t+1} \sim \mathrm{ALG}_{t+1} X_t, X_t' \sim \mathscr{D}_t}\left[\ell(\hat{f}_{t+1}(X_t), Y_t) - \ell(\hat{f}_{t+1}(X_t'), Y_t')\right]$$

Both steps crucially use smoothness and coupling arguments

Oracle Efficiency with Known Measure

Algorithm: Sample $\{Z_i\} \sim \mu$. Label at random

Run ERM on Historical data: $S_t = \{(X_i, Y_i)\}_{i \leq t} \cup$ Hallucinated data $\{Z_i, \tilde{Y}_i\}$

Oracle Efficiency with Known Measure

Algorithm: Sample $\{Z_i\} \sim \mu$. Label at random

Run ERM on Historical data: $S_t = \{(X_i, Y_i)\}_{i \leq t} \cup$ Hallucinated data $\{Z_i, \tilde{Y}_i\}$

Coupling lemma as an algorithmic method to generate synthetic data: Accounts for uncertainty “worry” about bad events under IID

Oracle Efficiency with Known Measure

Algorithm: Sample $\{Z_i\} \sim \mu$. Label at random

Run ERM on Historical data: $S_t = \{(X_i, Y_i)\}_{i \leq t} \cup$ Hallucinated data $\{Z_i, \tilde{Y}_i\}$

Coupling lemma as an algorithmic method to generate synthetic data: Accounts for uncertainty "worry" about bad events under IID

Key technical contribution: Technique for algorithmic generalization for data from "unseen" distributions

Oracle Efficiency with Known Measure

Algorithm: Sample $\{Z_i\} \sim \mu$. Label at random

Run ERM on Historical data: $S_t = \{(X_i, Y_i)\}_{i \leq t} \cup$ Hallucinated data $\{Z_i, \tilde{Y}_i\}$

Coupling lemma as an algorithmic method to generate synthetic data: Accounts for uncertainty "worry" about bad events under IID

Key technical contribution: Technique for algorithmic generalization for data from "unseen" distributions

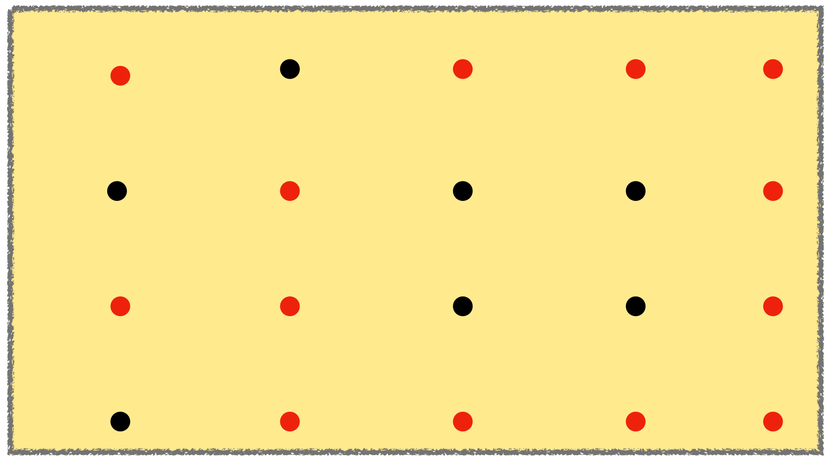
Coupling relates stability of the algorithm to that of Rademacher/Gaussian processes

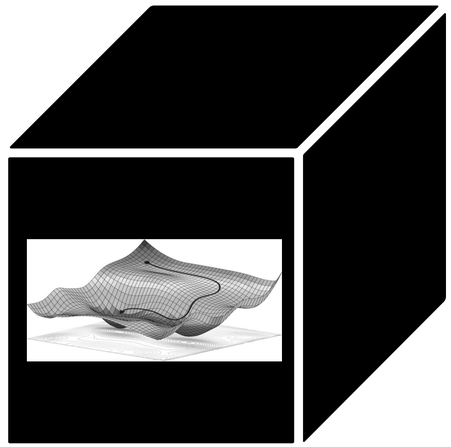
Oracle Efficiency with Known Measure

Theorem [HHSY'21, BDGR'21]: Known base measure oracle efficient smoothed online learning we have

$$\mathbb{E}[\mathrm{Reg}_T] \lesssim \sqrt{\frac{\mathrm{vc}(\mathscr{F})}{\sigma T}}$$

Historical Data S_{t-1} ∪ Hallucinated Data





$\hat{f}_{t+1}$

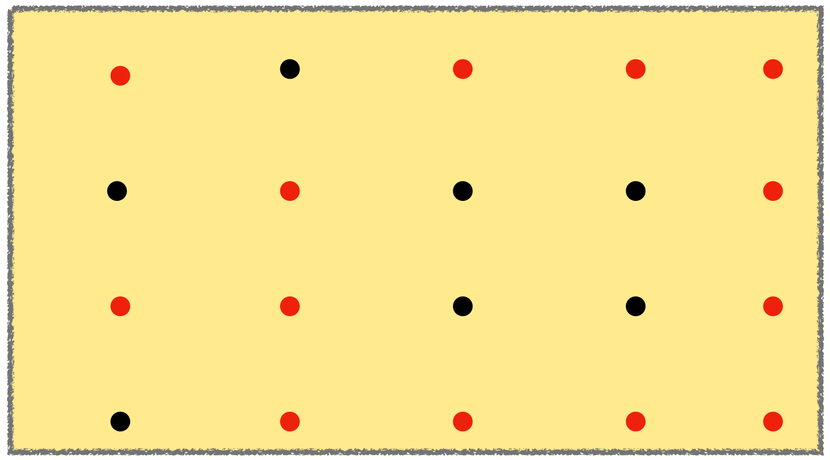
Oracle Efficiency with Known Measure

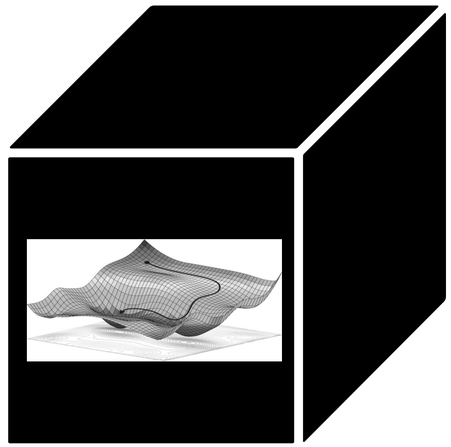
Theorem [HHSY'21,BDGR'21]: Known base measure oracle efficient smoothed online learning we have

$$\mathbb{E}[\mathrm{Reg}_T] \lesssim \sqrt{\frac{\mathrm{vc}(\mathscr{F})}{\sigma T}}$$

Rate can be improved to $\sigma^{-1/4}$ for binary classification using "Poissonization" [HHSY'21]

Historical Data S_{t-1} ∪ Hallucinated Data





$\hat{f}_{t+1}$

Computational Lower Bounds

Computational Lower Bounds

Theorem [HHSY'21,BDGR'21]: Any proper algorithm that has $o\left(\sqrt{Td\sqrt{\sigma}}\right)$ regret in smoothed online learning needs $\sqrt{d/\sigma}$ oracle calls.

Computational Lower Bounds

Theorem [HHSY'21, BDGR'21]: Any proper algorithm that has $o\left(\sqrt{Td\sqrt{\sigma}}\right)$ regret in smoothed online learning needs $\sqrt{d/\sigma}$ oracle calls.

Note that the statistical algorithm requires exponential in d time

Computational Lower Bounds

Theorem [HHSY'21,BDGR'21]: Any proper algorithm that has $o\left(\sqrt{Td\sqrt{\sigma}}\right)$ **regret in smoothed online learning needs** $\sqrt{d/\sigma}$ **oracle calls.**

Note that the statistical algorithm requires exponential in d time

Can we do better than running experts on a (large) net? Or are there matching lower bounds

Bounds for Efficient Smoothed Online Learning

	Known	Unknown
Realizable (Efficiency)	$T^{-1}d\log(T/\sigma)$	$\sqrt{T^{-1}d\sigma^{-1}}$
	$\sqrt{T^{-1}d\sigma^{-1}}$	$\sqrt{T^{-1}d\sigma^{-1}}$
Agnostic (Efficient)	$\sqrt{T^{-1}d\log(T/\sigma)}$	$\sqrt{T^{-1}d\sigma^{-1}}$
	$\sqrt{T^{-1}d\sigma^{-1}}$	???

Other Applications

- Statistical and Computational Equivalence between Statistical Learning and Smoothed Online Learning [HR**S**'21, HH**S**Y'22, **B**DGR'22, **B**R**S**'24, **B**P'23]
- Private Learning with public data [HR**S**'20, **B**BD**S**W'24, **BS**'25]
- Online Discrepancy minimization [HR**S**'21]
- Data-driven Algorithm design [HR**S**'21]
- Bandits, RL, Robotics [**B**ST'22, **B**S'22, **B**SR'24, **B**DGR'22]
- Equilibria Computation in General Games [DGH**S**'23]

Key Takeaways

Smoothed data bridges efficiency of statistical learning and robustness of online learning.

Key Takeaways

Smoothed data bridges efficiency of statistical learning and robustness of online learning.

Technical tools:

Key Takeaways

Smoothed data bridges efficiency of statistical learning and robustness of online learning.

Technical tools:

(i) Surprise Lemma (compactness)

Key Takeaways

Smoothed data bridges efficiency of statistical learning and robustness of online learning.

Technical tools:

(i) Surprise Lemma (compactness)

(ii) Coupling (rejection sampling)

Open Problems

When can we get the fast rate?

	Known	Unknown
Realizable	$T^{-1}d\log(T/\sigma)$	$\sqrt{T^{-1}d\sigma^{-1}}$
(Efficiency)	$\sqrt{T^{-1}d\sigma^{-1}}$	$\sqrt{T^{-1}d\sigma^{-1}}$
Agnostic	$\sqrt{T^{-1}d\log(T/\sigma)}$	$\sqrt{T^{-1}d\sigma^{-1}}$
(Efficient)	$\sqrt{T^{-1}d\sigma^{-1}}$	???

Can we get oracle efficiency at all?

Open Problems

How fundamental is the worse dependence on σ?

	Known	Unknown
Realizable	$T^{-1}d\log(T/\sigma)$	$\sqrt{T^{-1}d\sigma^{-1}}$
(Efficiency)	$\sqrt{T^{-1}d\sigma^{-1}}$	$\sqrt{T^{-1}d\sigma^{-1}}$
Agnostic	$\sqrt{T^{-1}d\log(T/\sigma)}$	$\sqrt{T^{-1}d\sigma^{-1}}$
(Efficient)	$\sqrt{T^{-1}d\sigma^{-1}}$	???

Broader Open Problems

Broader Open Problems

Algorithmic

Broader Open Problems

Algorithmic

What is a good oracle model
for modern ML?

Broader Open Problems

Algorithmic

What is a good oracle model
for modern ML?

E.g. Oracles for sampling,
LLMs

Broader Open Problems

Algorithmic

What is a good oracle model for modern ML?

E.g. Oracles for sampling, LLMs

Epistemic

Broader Open Problems

Algorithmic

What is a good oracle model for modern ML?

E.g. Oracles for sampling, LLMs

Epistemic

What is the best way to capture the relation of the past and the future?

Broader Open Problems

Algorithmic

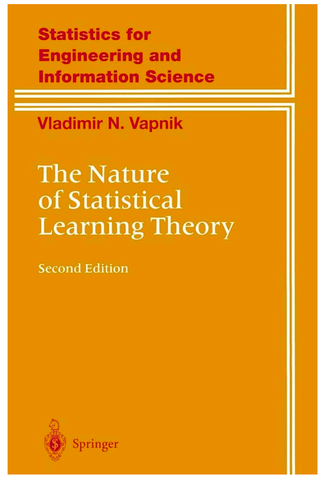
What is a good oracle model for modern ML?

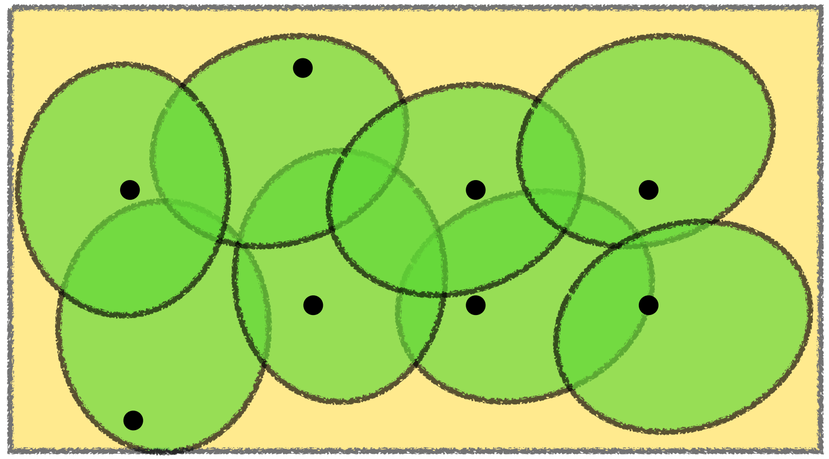
E.g. Oracles for sampling, LLMs

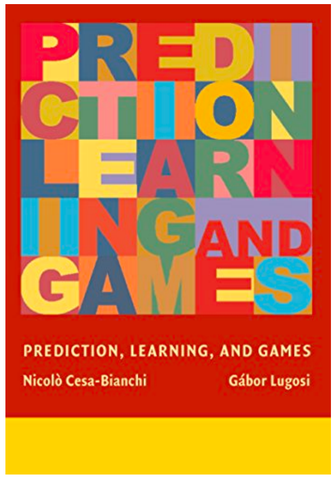
Epistemic

What is the best way to capture the relation of the past and the future?

E.g. Abstention, relaxed benchmarks







Smoothed data

Statistical Learning

Online Learning

Thank you