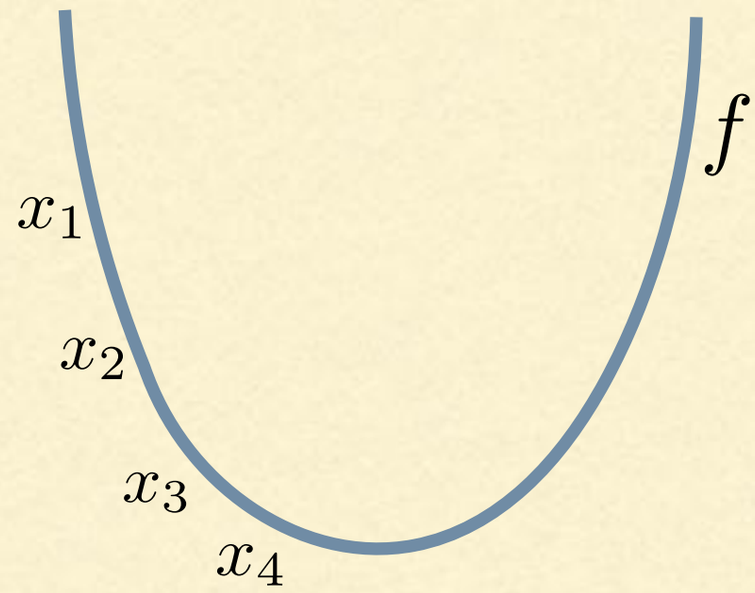
NON-STATIONARY STOCHASTIC OPTIMIZATION UNDER LOCAL TEMPORAL AND SPATIAL CHANGES

Xi Chen, **Yining Wang** and Yu-Xiang Wang
Carnegie Mellon University and New York University

STATIONARY OPTIMIZATION

- The classical setting of optimizing a convex function
  - Objective: find the minimum of convex $f : \mathcal{X} \to \mathbb{R}^d$
  - Query x_t , receive feedback $f(x_t)$ or $\nabla f(x_t)$ with stochastic noise ε_t
- Example: stochastic gradient descent

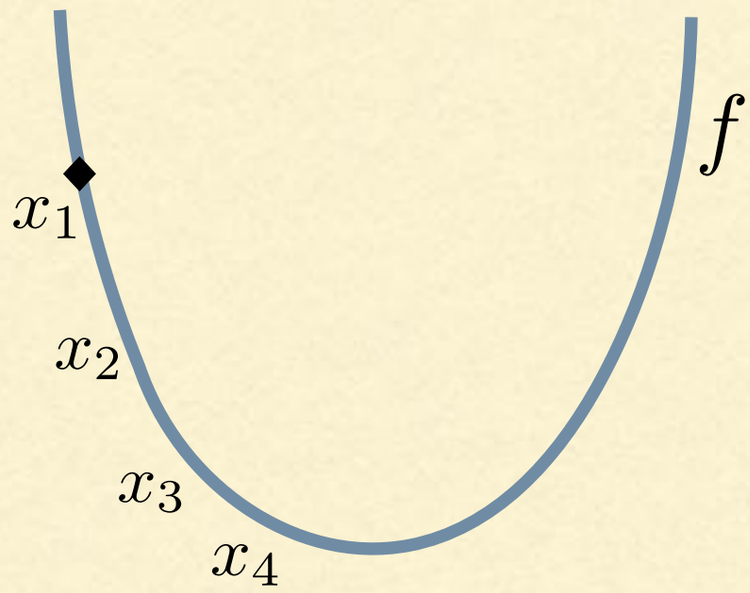
$$x_{t+1} = P_{\mathcal{X}}\left[x_t - \eta_t(\nabla f(x_t) + \varepsilon_t)\right]$$



STATIONARY OPTIMIZATION

- The classical setting of optimizing a convex function
  - Objective: find the minimum of convex $f : \mathcal{X} \to \mathbb{R}^d$
  - Query x_t , receive feedback $f(x_t)$ or $\nabla f(x_t)$ with stochastic noise ε_t
- Example: stochastic gradient descent

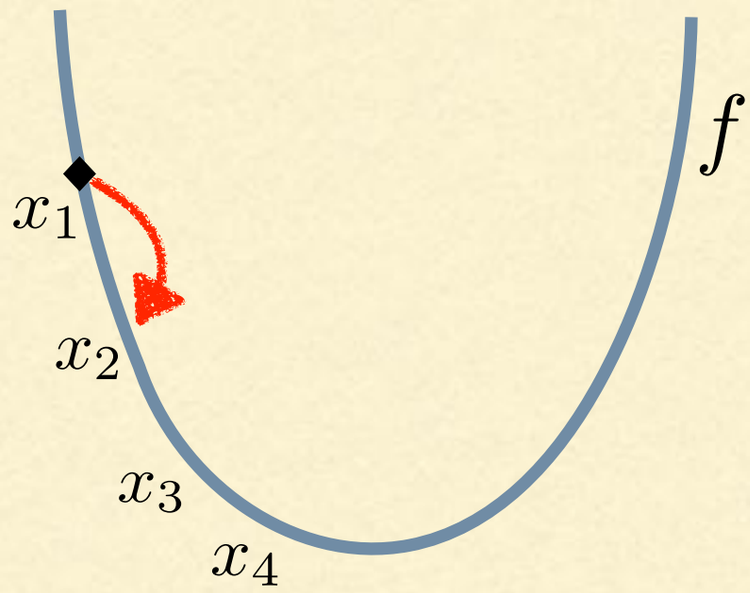
$$x_{t+1} = P_{\mathcal{X}}\left[x_t - \eta_t(\nabla f(x_t) + \varepsilon_t)\right]$$



STATIONARY OPTIMIZATION

- The classical setting of optimizing a convex function
  - Objective: find the minimum of convex $f : \mathcal{X} \to \mathbb{R}^d$
  - Query x_t , receive feedback $f(x_t)$ or $\nabla f(x_t)$ with stochastic noise ε_t
- Example: stochastic gradient descent

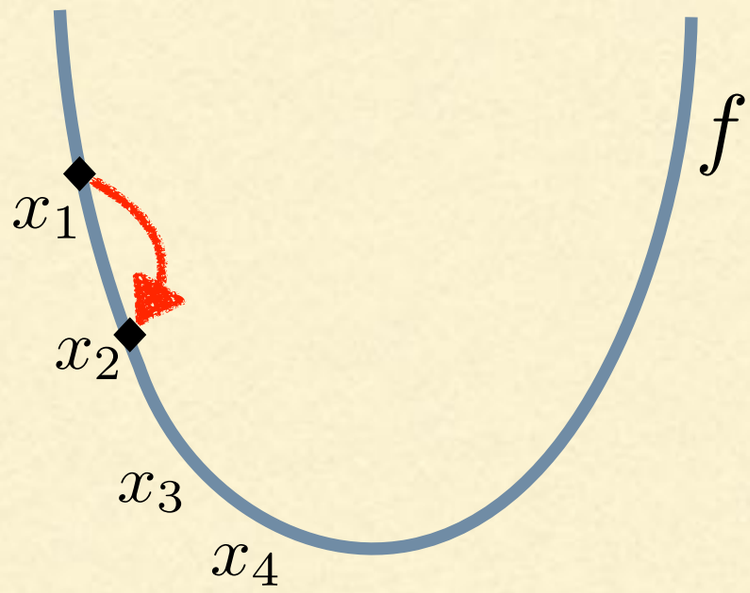
$$x_{t+1} = P_{\mathcal{X}}\left[x_t - \eta_t(\nabla f(x_t) + \varepsilon_t)\right]$$



STATIONARY OPTIMIZATION

- The classical setting of optimizing a convex function
  - Objective: find the minimum of convex $f : \mathcal{X} \to \mathbb{R}^d$
  - Query x_t , receive feedback $f(x_t)$ or $\nabla f(x_t)$ with stochastic noise ε_t
- Example: stochastic gradient descent

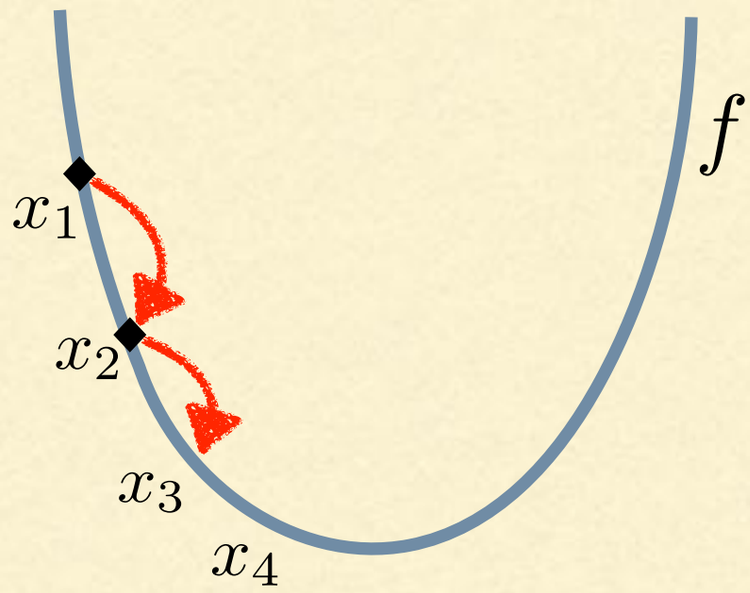
$$x_{t+1} = P_{\mathcal{X}}\left[x_t - \eta_t(\nabla f(x_t) + \varepsilon_t)\right]$$



STATIONARY OPTIMIZATION

- The classical setting of optimizing a convex function
  - Objective: find the minimum of convex $f : \mathcal{X} \to \mathbb{R}^d$
  - Query x_t , receive feedback $f(x_t)$ or $\nabla f(x_t)$ with stochastic noise ε_t
- Example: stochastic gradient descent

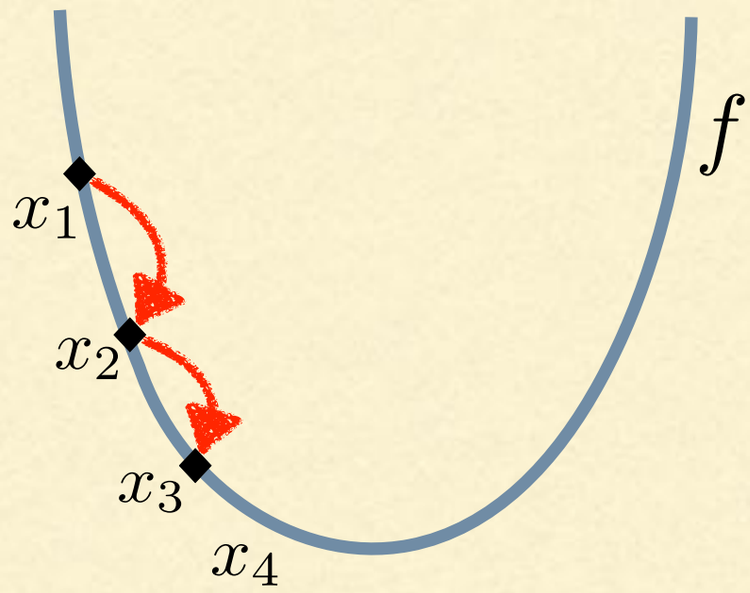
$$x_{t+1} = P_{\mathcal{X}}\left[x_t - \eta_t(\nabla f(x_t) + \varepsilon_t)\right]$$



STATIONARY OPTIMIZATION

- The classical setting of optimizing a convex function
  - Objective: find the minimum of convex $f : \mathcal{X} \to \mathbb{R}^d$
  - Query x_t , receive feedback $f(x_t)$ or $\nabla f(x_t)$ with stochastic noise ε_t
- Example: stochastic gradient descent

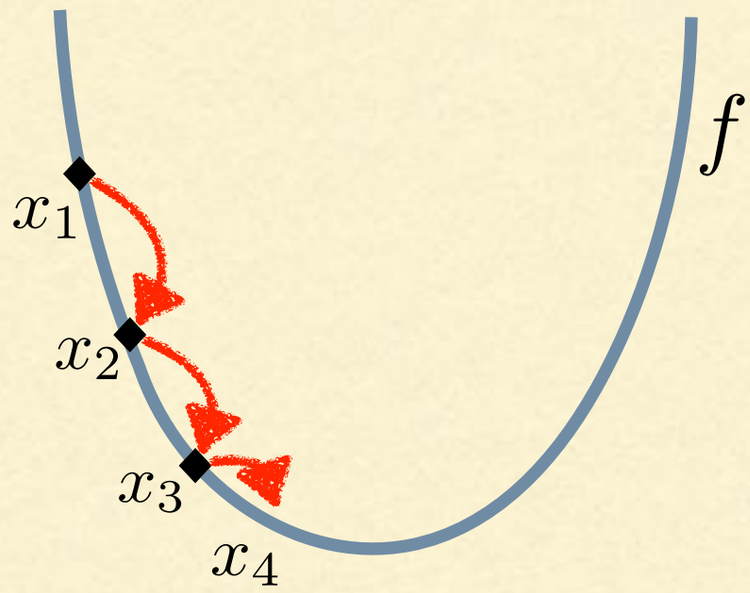
$$x_{t+1} = P_{\mathcal{X}}\left[x_t - \eta_t(\nabla f(x_t) + \varepsilon_t)\right]$$



STATIONARY OPTIMIZATION

- The classical setting of optimizing a convex function
  - Objective: find the minimum of convex $f : \mathcal{X} \to \mathbb{R}^d$
  - Query x_t , receive feedback $f(x_t)$ or $\nabla f(x_t)$ with stochastic noise ε_t
- Example: stochastic gradient descent

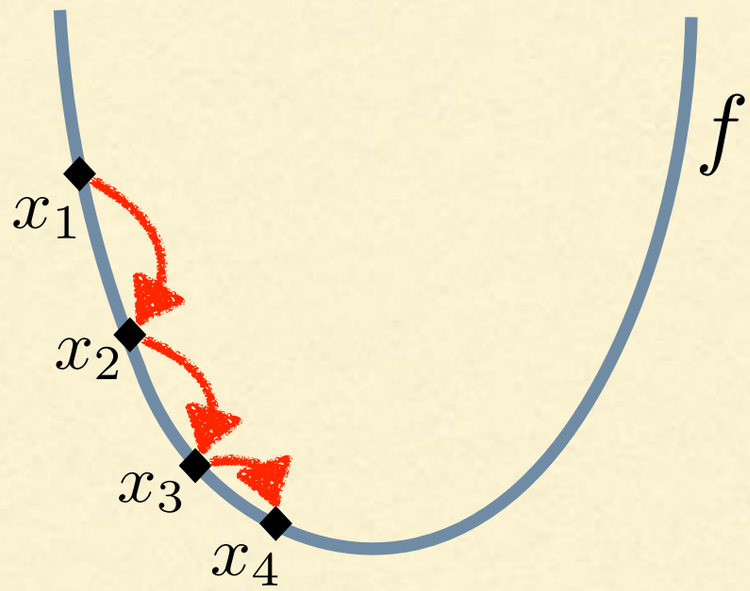
$$x_{t+1} = P_{\mathcal{X}}\left[x_t - \eta_t(\nabla f(x_t) + \varepsilon_t)\right]$$



STATIONARY OPTIMIZATION

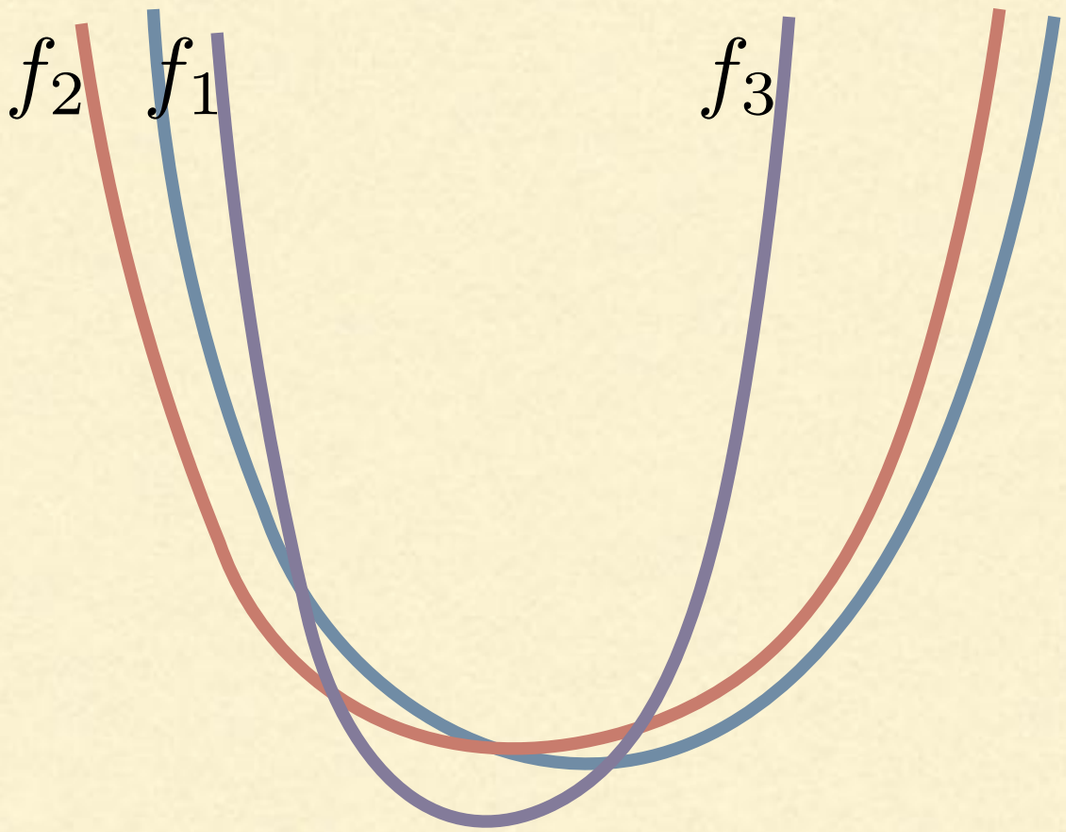
- The classical setting of optimizing a convex function
  - Objective: find the minimum of convex $f : \mathcal{X} \to \mathbb{R}^d$
  - Query x_t , receive feedback $f(x_t)$ or $\nabla f(x_t)$ with stochastic noise ε_t
- Example: stochastic gradient descent

$$x_{t+1} = P_{\mathcal{X}}\left[x_t - \eta_t(\nabla f(x_t) + \varepsilon_t)\right]$$



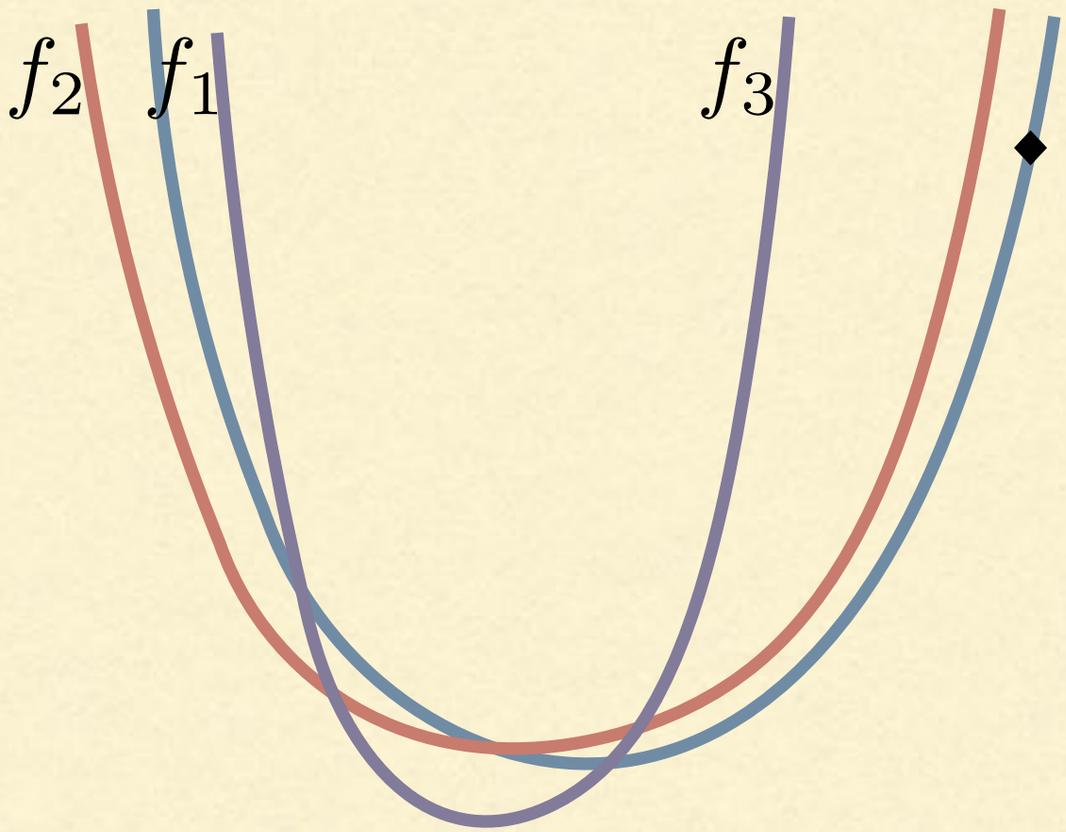
DYNAMIC OPTIMIZATION

- In many applications, f_t changes with time t
  - *Examples:* dynamic pricing, online recommendation systems, portfolio selection, simulation optimization
- Objective: minimize **regret:**
  - Difference between
  - *What the algorithm achieves*, and
  - *What an "oracle" solution can achieve*



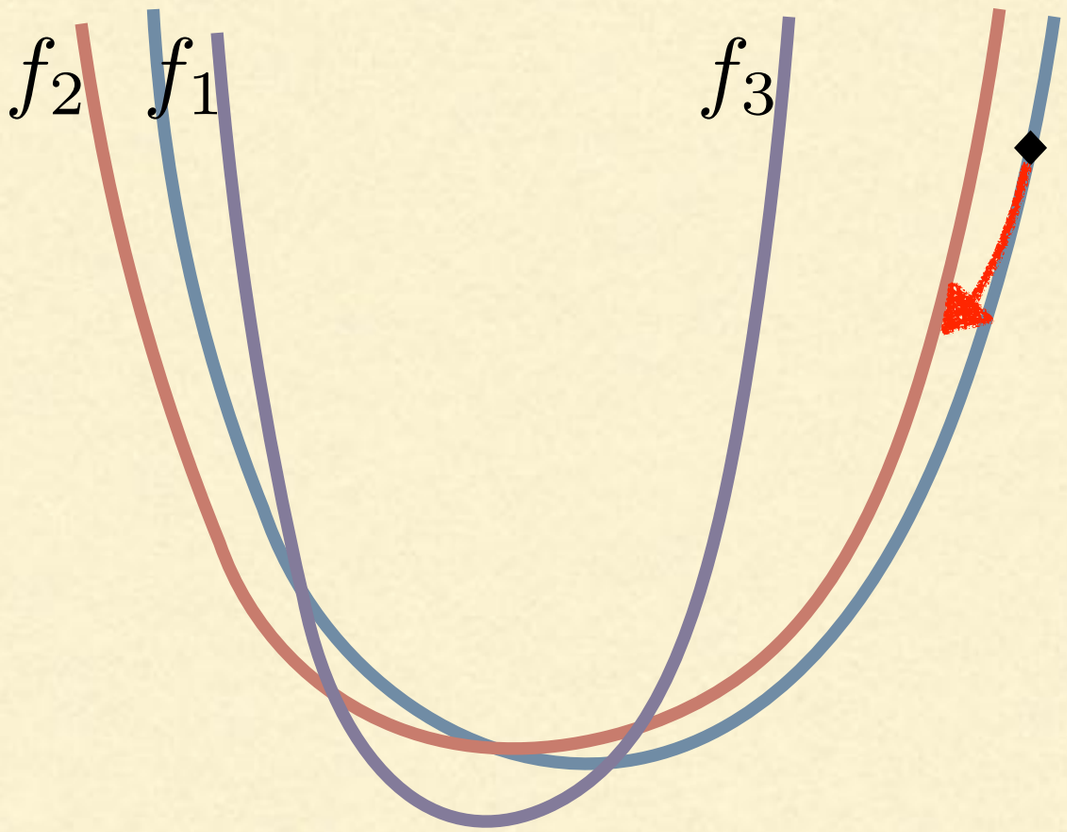
DYNAMIC OPTIMIZATION

- In many applications, f_t changes with time t
  - *Examples:* dynamic pricing, online recommendation systems, portfolio selection, simulation optimization
- Objective: minimize **regret:**
  - Difference between
  - *What the algorithm achieves*, and
  - *What an "oracle" solution can achieve*



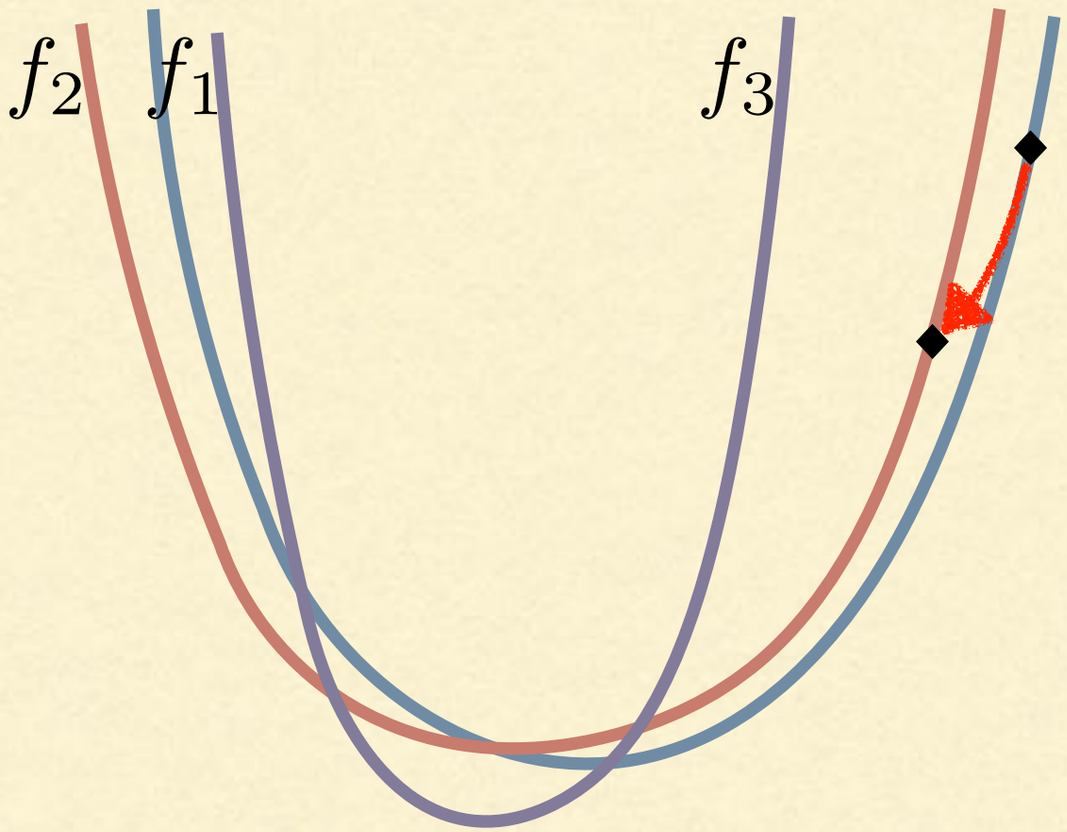
DYNAMIC OPTIMIZATION

- In many applications, f_t changes with time t
  - *Examples:* dynamic pricing, online recommendation systems, portfolio selection, simulation optimization
- Objective: minimize **regret:**
  - Difference between
  - *What the algorithm achieves*, and
  - *What an "oracle" solution can achieve*



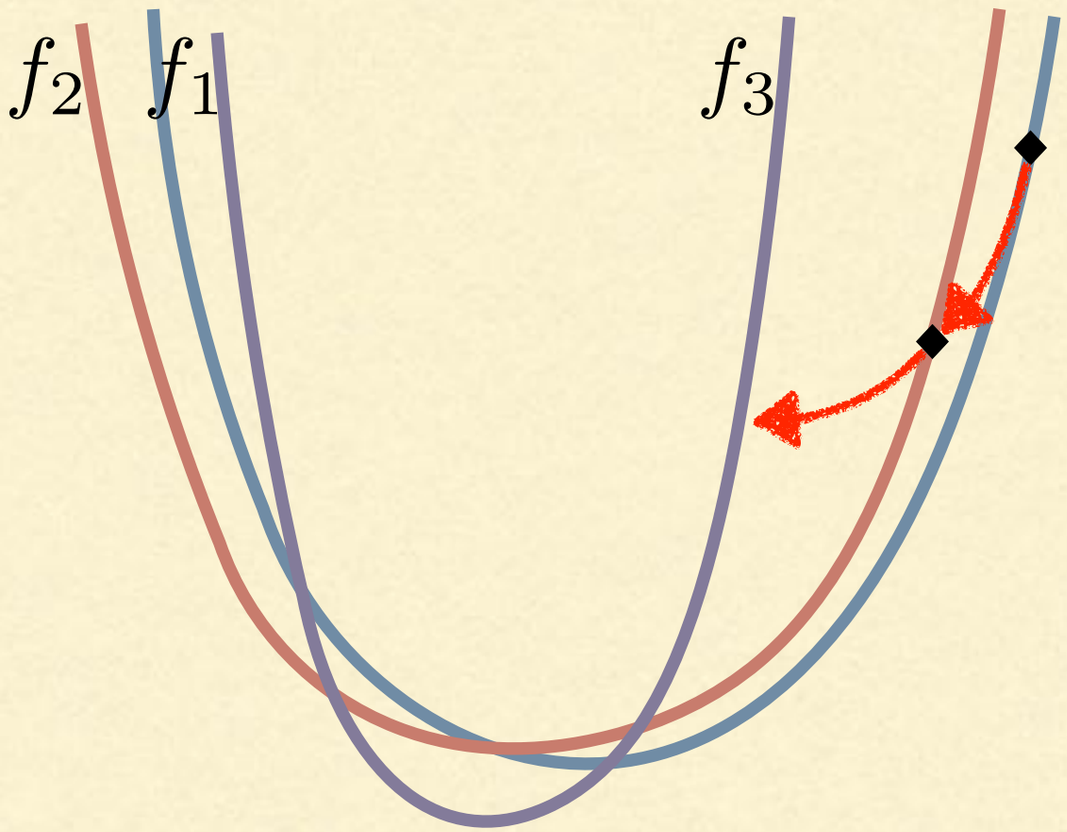
DYNAMIC OPTIMIZATION

- In many applications, f_t changes with time t
  - *Examples:* dynamic pricing, online recommendation systems, portfolio selection, simulation optimization
- Objective: minimize **regret:**
  - Difference between
  - *What the algorithm achieves*, and
  - *What an "oracle" solution can achieve*



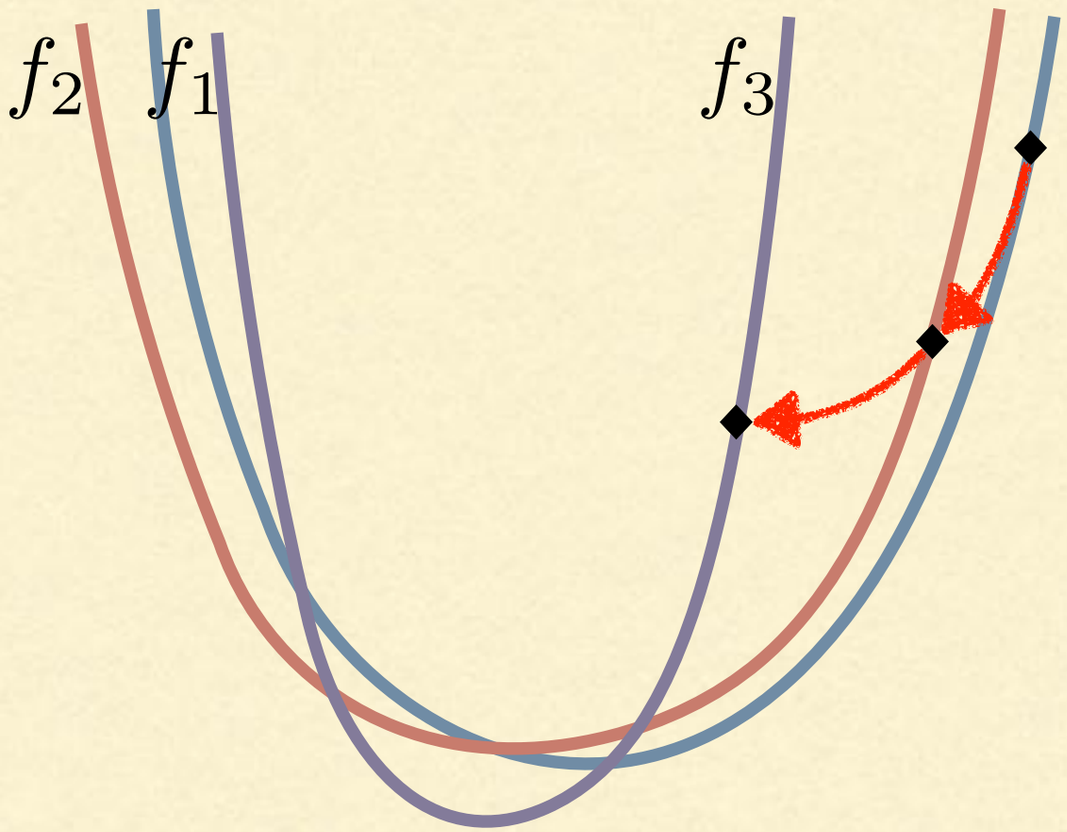
DYNAMIC OPTIMIZATION

- In many applications, f_t changes with time t
  - *Examples:* dynamic pricing, online recommendation systems, portfolio selection, simulation optimization
- Objective: minimize **regret:**
  - Difference between
  - *What the algorithm achieves*, and
  - *What an "oracle" solution can achieve*



DYNAMIC OPTIMIZATION

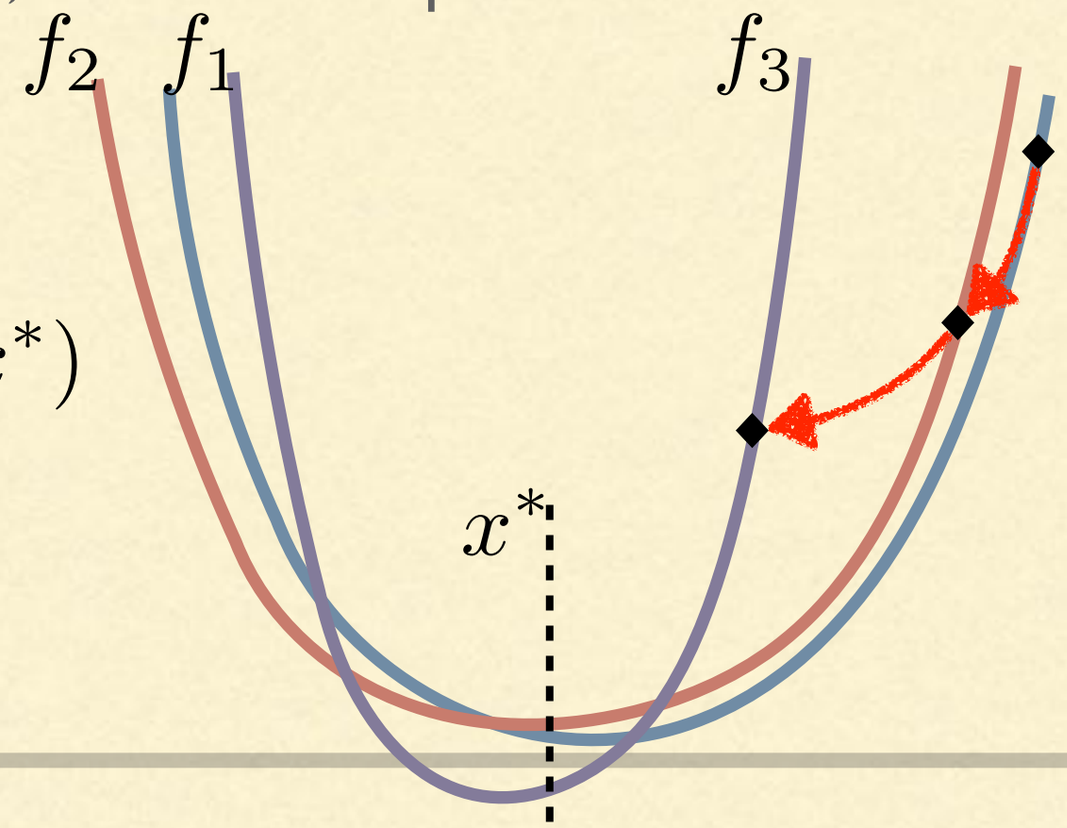
- In many applications, f_t changes with time t
  - *Examples:* dynamic pricing, online recommendation systems, portfolio selection, simulation optimization
- Objective: minimize **regret:**
  - Difference between
  - *What the algorithm achieves*, and
  - *What an "oracle" solution can achieve*



STATIONARY VS DYNAMIC REGRETS

- Stationary regret: competing against a stationary oracle x^*
  - Advantages: no assumptions on function sequence
  - Disadvantages: weak notion of "oracle", unnatural for practical use

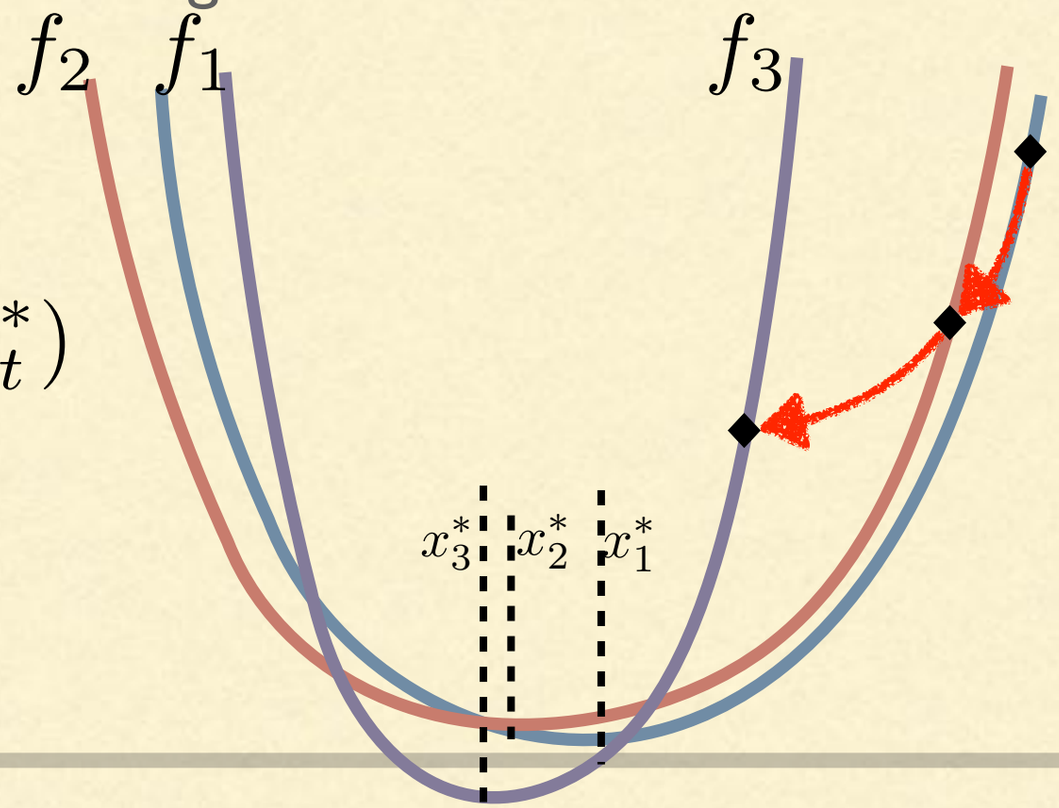
$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(x_t)\right] - \inf_{x^* \in \mathcal{X}} \sum_{t=1}^{T} f_t(x^*)$$



STATIONARY VS DYNAMIC REGRETS

- Dynamic regret: competing against a dynamic oracle x_t^*
  - Advantages: strong notion of "oracle", intuitive concept.
  - Disadvantages: requires assumptions on "changes" of functions

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(x_t)\right] - \sum_{t=1}^{T} \inf_{x_t^* \in \mathcal{X}} f_t(x_t^*)$$



KNOWN RESULTS

$$M \cdot \mathbf{I}_d \preceq \nabla^2 f_t \preceq L \cdot \mathbf{I}_d$$

- Suppose the functions are strongly convex and smooth
- Stationary regret results
  - $\nabla f_t(\cdot)$ oracle: SGD achieves $\mathcal{O}(\log T)$ *well-known; e.g., Hazan' 16*
  - $f_t(\cdot)$ oracle: EGS achieves $\mathcal{O}(\sqrt{T})$ *Nemirovski & Yudin'83, Flaxman et al.' 04, Agarwal et al.' 10*

KNOWN RESULTS

$$M \cdot \mathbf{I}_d \preceq \nabla^2 f_t \preceq L \cdot \mathbf{I}_d$$

- Suppose the functions are strongly convex and smooth
- Dynamic regret results Besbes et al.' 15
  - Assumptions: $\frac{1}{T}\sum_{t=1}^{T-1} \|f_{t+1} - f_t\|_\infty \leq V_T$
  - $\nabla f_t(\cdot)$ oracle: $\mathcal{O}(V_T^{1/2} \cdot T)$
  - $f_t(\cdot)$ oracle: $\mathcal{O}(V_T^{1/3} \cdot T)$

KNOWN RESULTS

$$M \cdot \mathbf{I}_d \preceq \nabla^2 f_t \preceq L \cdot \mathbf{I}_d$$

- Suppose the functions are strongly convex and smooth
- Dynamic regret results Besbes et al.' 15
  - Assumptions: $\frac{1}{T}\sum_{t=1}^{T-1} \|f_{t+1} - f_t\|_\infty \leq V_T$
  - $\nabla f_t(\cdot)$ oracle: $\mathcal{O}(V_T^{1/2} \cdot T)$
  - $f_t(\cdot)$ oracle: $\mathcal{O}(V_T^{1/3} \cdot T)$

KNOWN RESULTS

$$M \cdot \mathbf{I}_d \preceq \nabla^2 f_t \preceq L \cdot \mathbf{I}_d$$

- Suppose the functions are strongly convex and smooth
- Dynamic regret results Besbes et al.' 15
- Assumptions: $\frac{1}{T}\sum_{t=1}^{T-1} \|f_{t+1} - f_t\|_\infty \leq V_T$
- $\nabla f_t(\cdot)$ oracle: $\mathcal{O}(V_T^{1/2} \cdot T)$
- $f_t(\cdot)$ oracle: $\mathcal{O}(V_T^{1/3} \cdot T)$

KNOWN RESULTS

$$M \cdot \mathbf{I}_d \preceq \nabla^2 f_t \preceq L \cdot \mathbf{I}_d$$

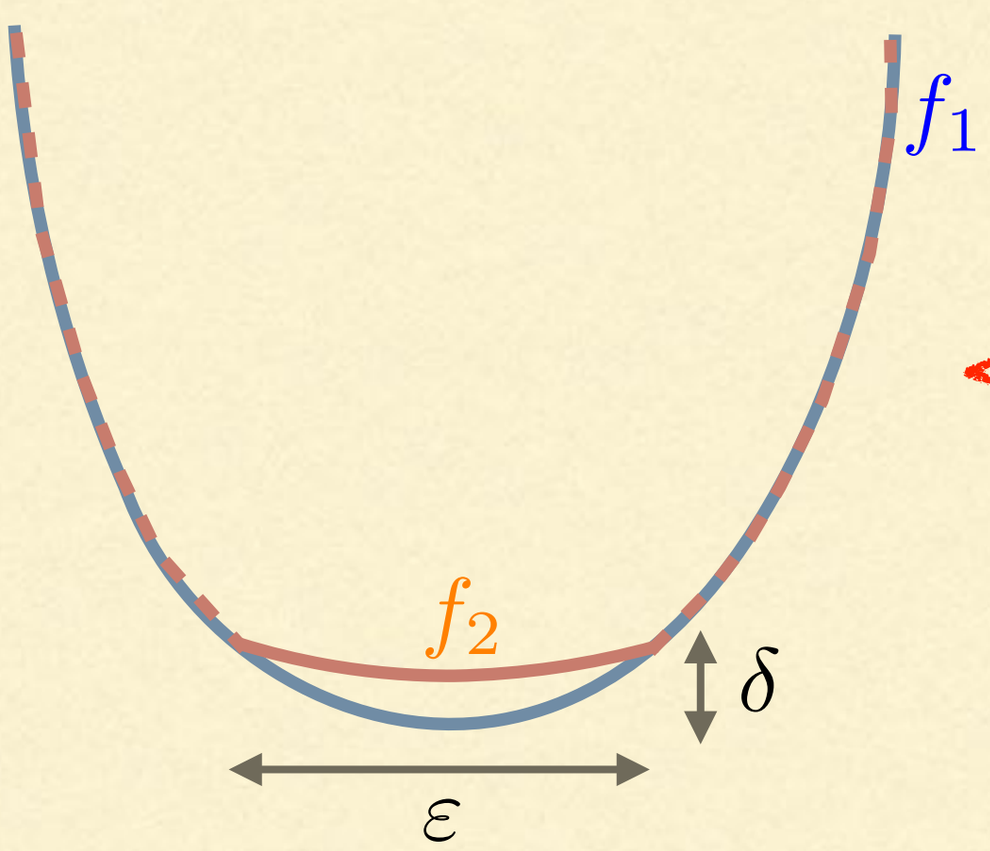
- Suppose the functions are strongly convex and smooth
- Dynamic regret results Besbes et al.' 15
- Assumptions: $\frac{1}{T}\sum_{t=1}^{T-1} \|f_{t+1} - f_t\|_\infty \leq V_T$
- $\nabla f_t(\cdot)$ oracle: $\mathcal{O}(V_T^{1/2} \cdot T)$

A little bit restrictive ...

- $f_t(\cdot)$ oracle: $\mathcal{O}(V_T^{1/3} \cdot T)$

LOCAL CHANGES

- Local spatial changes

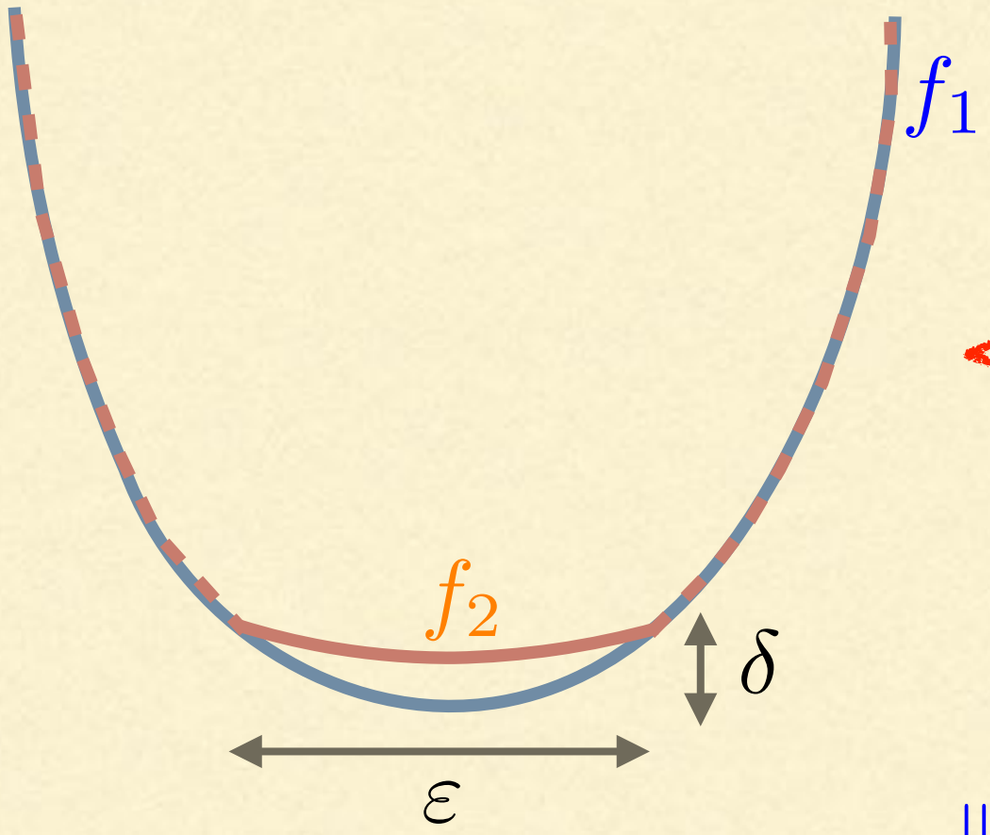


$$\|f_1 - f_2\|_\infty \approx \delta$$

$$\|f_1 - f_2\|_p \approx \varepsilon^{1/p} \cdot \delta$$

LOCAL CHANGES

- Local spatial changes



$$\|f_1 - f_2\|_\infty \approx \delta$$

$$\|f_1 - f_2\|_p \approx \varepsilon^{1/p} \cdot \delta$$

$$\|f_1 - f_2\|_p \ll \|f_1 - f_2\|_\infty$$

LOCAL CHANGES

- Local temporal changes

$$\max_{1 \le t \le T-1} \|f_{t+1} - f_t\| \approx \|f - g\|$$

f, f, f, g, f, f, f

t = 1 —> T

$$\left[\frac{1}{T} \sum_{t=1}^{T-1} \|f_{t+1} - f_t\|^q \right]^{1/q} \approx T^{-1/q} \cdot \|f - g\|$$

LOCAL CHANGES

- Local temporal changes

$$\max_{1 \le t \le T-1} \|f_{t+1} - f_t\| \approx \|f - g\|$$

f, f, f, g, f, f, f

t = 1 —> T

$$\left[\frac{1}{T}\sum_{t=1}^{T-1} \|f_{t+1} - f_t\|^q\right]^{1/q} \approx T^{-1/q} \cdot \|f - g\|$$

$$\left[\frac{1}{T}\sum_{t=1}^{T-1} \|f_{t+1} - f_t\|^q\right]^{1/q} \ll \max_{1 \le t \le T-1} \|f_{t+1} - f_t\|$$

LOCAL CHANGES

$$\mathrm{Var}_{p,q}(\boldsymbol{f}) := \left[\frac{1}{T}\sum_{t=1}^{T-1} \|f_{t+1} - f_t\|_p^q\right]^{1/q} \leq V_T$$

- $p \in [1, \infty]$ measures the *local spatial changes* of **f**
- $q \in [1, \infty]$ measures the *local temporal changes* of **f**

Question: what is the optimal regret in V_T?

MAIN RESULTS

- Our main results and comparison with Besbes et al.' 15

	OUR PAPER $p, q \in [1, \infty]$	BESBES ET AL.'15 $p = \infty, q = 1$
$\nabla f_t(\cdot)$ ORACLE	$\mathcal{O}(V_T^{2p/(4p+d)} \cdot T)$	$\mathcal{O}(V_T^{1/2} \cdot T)$
$f_t(\cdot)$ ORACLE	$\mathcal{O}(V_T^{2p/(6p+d)} \cdot T)$	$\mathcal{O}(V_T^{1/3} \cdot T)$

NOTE d is domain dimension: $\mathcal{X} \subseteq \mathbb{R}^d$

REMARKS

	OUR PAPER $p, q \in [1, \infty]$	BESBES ET AL.'15 $p = \infty, q = 1$
$\nabla f_t(\cdot)$ ORACLE	$\mathcal{O}(V_T^{2p/(4p+d)} \cdot T)$	$\mathcal{O}(V_T^{1/2} \cdot T)$
$f_t(\cdot)$ ORACLE	$\mathcal{O}(V_T^{2p/(6p+d)} \cdot T)$	$\mathcal{O}(V_T^{1/3} \cdot T)$

Curse of Dimensionality

1. Regret scales *exponentially* with domain dimension *d*
2. Such dependency becomes milder as $p \to \infty$ and disappears for $p = \infty$

REMARKS

	OUR PAPER $p, q \in [1, \infty]$	BESBES ET AL.'15 $p = \infty, q = 1$
$\nabla f_t(\cdot)$ ORACLE	$\mathcal{O}(V_T^{2p/(4p+d)} \cdot T)$	$\mathcal{O}(V_T^{1/2} \cdot T)$
$f_t(\cdot)$ ORACLE	$\mathcal{O}(V_T^{2p/(6p+d)} \cdot T)$	$\mathcal{O}(V_T^{1/3} \cdot T)$

Automatic Temporal Adaptivity

1. The optimal regret does *not* depend on parameter q
2. This may not be true for the region $q < 1$

PROOF OF UPPER BOUND

- The *re-starting* procedure: (Besbes et al.'15)
  - Let *A* be either SGD or EGS algorithm

$$\underbrace{f_1, \cdots, f_{\overline{b}_1}}_{B_1}, \underbrace{f_{\underline{b}_2}, \cdots, f_{\overline{b}_2}}_{B_2}, \cdots, \underbrace{f_{\underline{b}_J}, \cdots, f_T}_{B_J}$$

  - Run *A* on $B_1, B_2, \ldots, B_J$ independently.
  - Idea: combine stationary regret and perturbation constraint to upper bound dynamic regret

PROOF OF UPPER BOUND

$$\underbrace{f_1, \cdots, f_{\bar{b}_1}}_{B_1}, \underbrace{f_{\underline{b}_2}, \cdots, f_{\bar{b}_2}}_{B_2}, \cdots, \underbrace{f_{\underline{b}_J}, \cdots, f_T}_{B_J}$$

- Fix B_l. Consider the following decomposition of dynamic regret:

$$\underline{f_t(x_t) - f_t(x_t^*)} = \underline{f_t(x_t) - f_t(x_1^*)} + \underline{f_t(x_1^*) - f_t(x_t^*)}$$

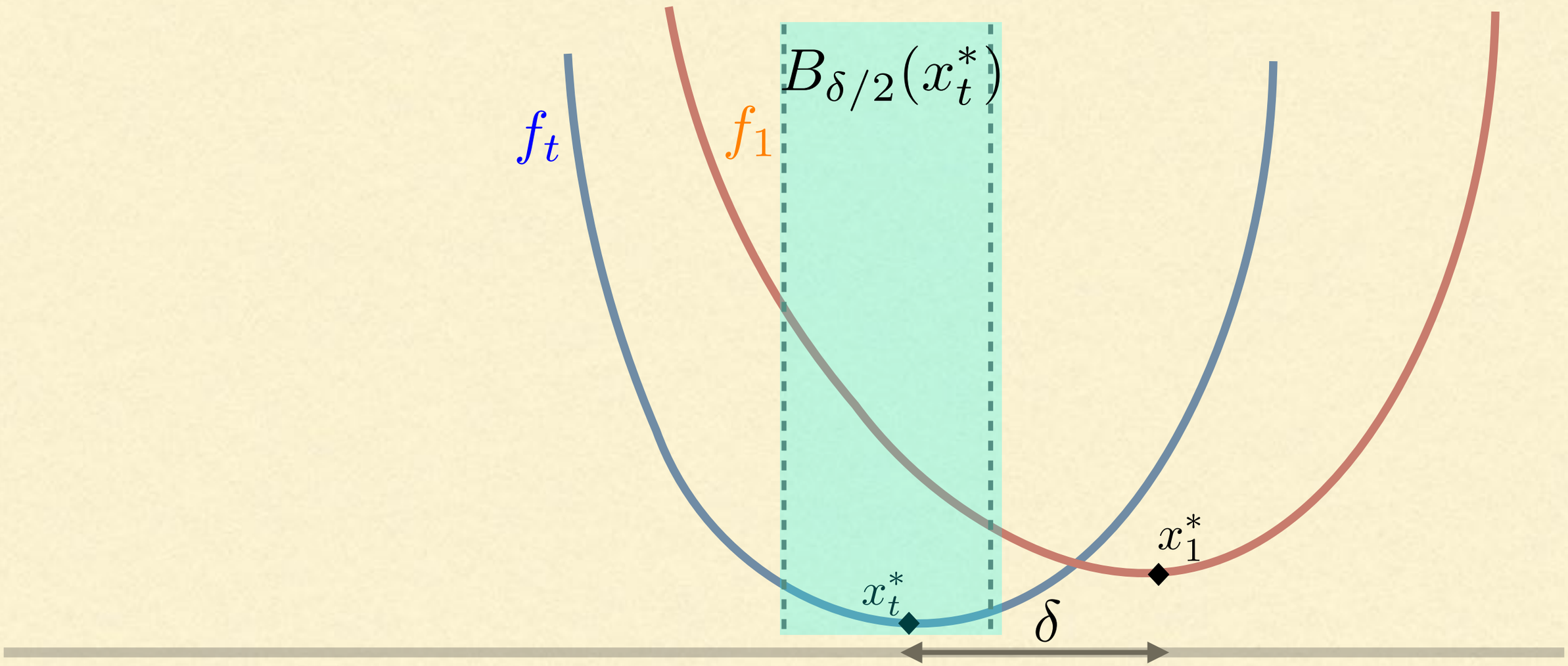
dynamic regret — stationary regret — perturbation

- Stationary regret: $\mathcal{O}(J \cdot \sqrt{T/J})$
- Perturbation: easy if $p = \infty$: $f_t(x_1^*) - f_t(x_t^*) \leq 2\|f_t - f_1\|_\infty$

Besbes et. al.'s approach

PROOF OF UPPER BOUND

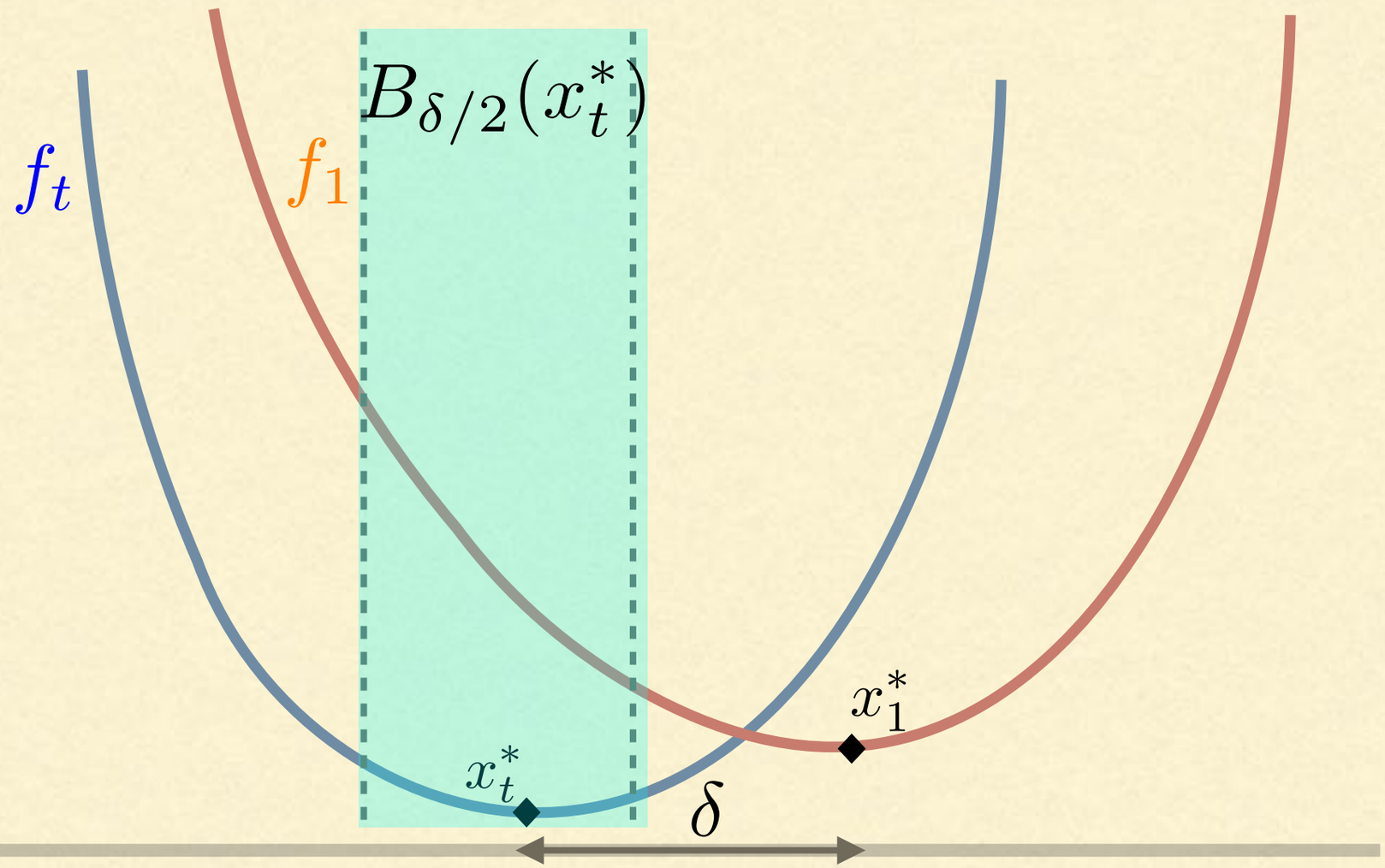
- Key perturbation lemma for $p \in [1, \infty)$



PROOF OF UPPER BOUND

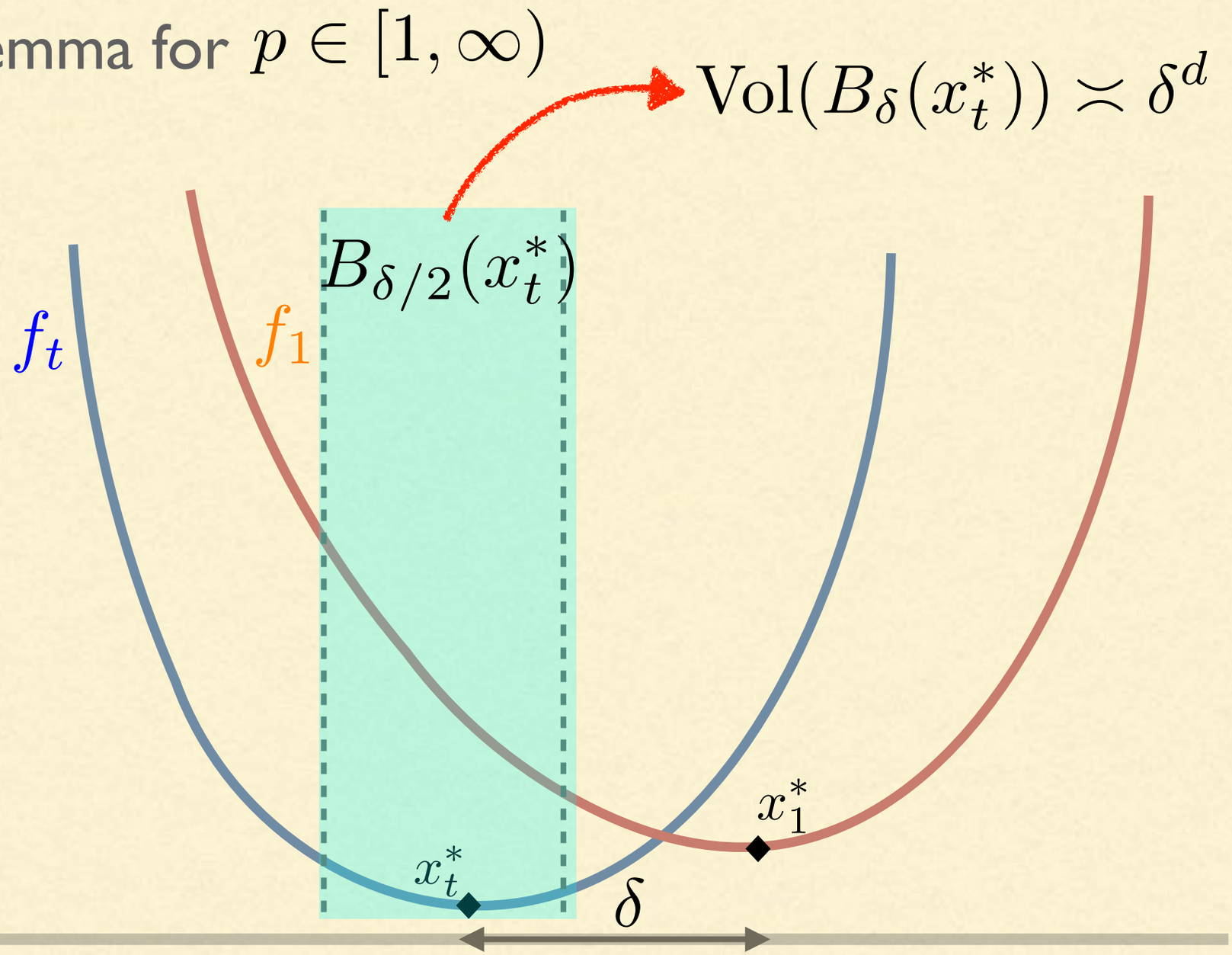
- Key perturbation lemma for $p \in [1, \infty)$

$$\mathrm{Vol}(B_\delta(x_t^*)) \asymp \delta^d$$



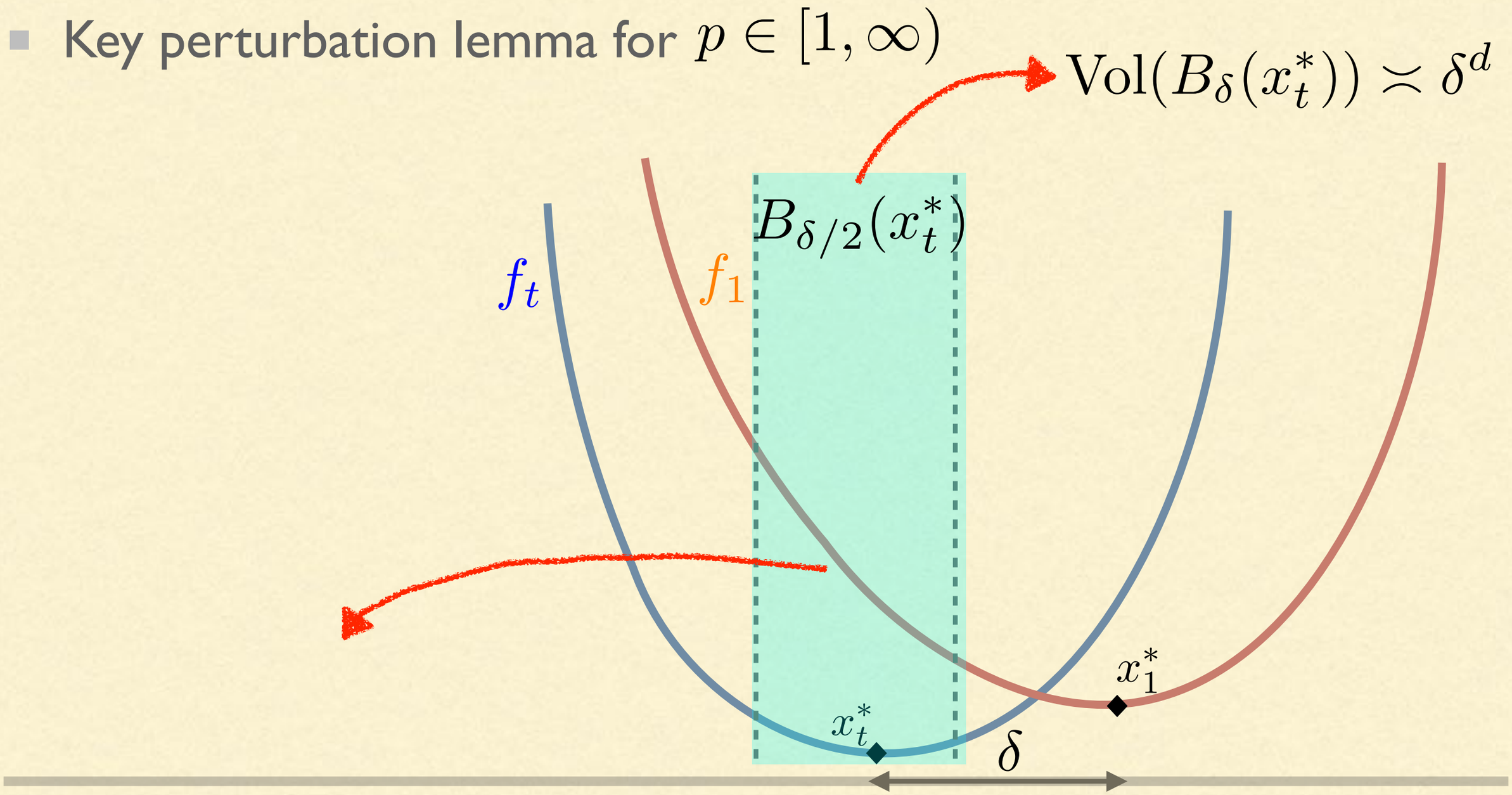
PROOF OF UPPER BOUND

- Key perturbation lemma for $p \in [1, \infty)$



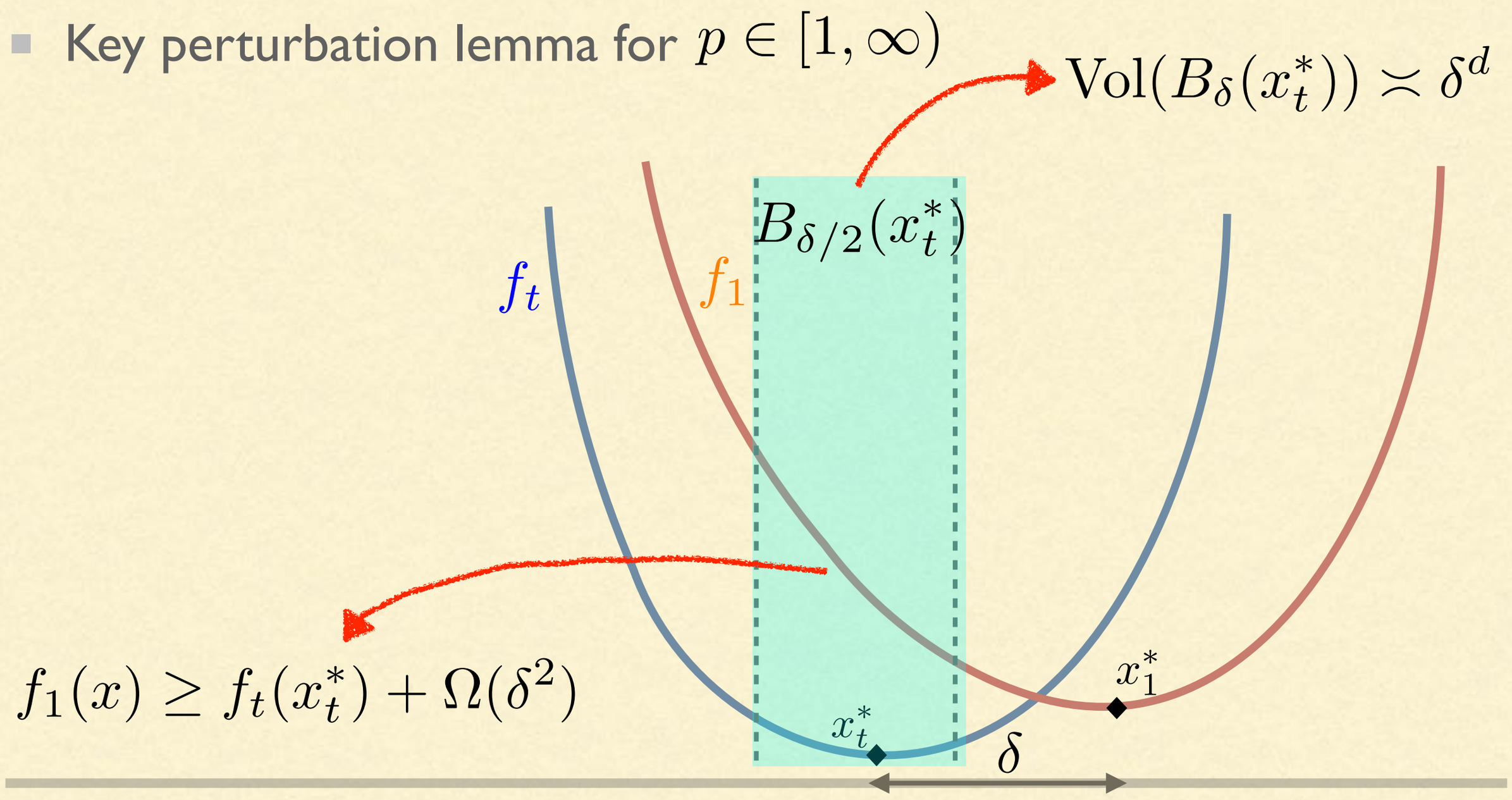
PROOF OF UPPER BOUND

- Key perturbation lemma for $p \in [1, \infty)$



PROOF OF UPPER BOUND

- Key perturbation lemma for $p \in [1, \infty)$



PROOF OF UPPER BOUND

- Key perturbation lemma for $p \in [1, \infty)$

$$\mathrm{Vol}(B_\delta(x_t^*)) \asymp \delta^d \qquad f_1(x) \geq f_t(x_t^*) + \Omega(\delta^2)$$

- Conclusion: $\|x_t - x_1\|_2 = \mathcal{O}(\|f_t - f_1\|_p^{p/(2p+d)})$

$$\left|f_t(x_t^*) - f_t(x_1^*)\right| = \mathcal{O}(\|f_t - f_1\|_p^{2p/(2p+d)})$$

PROOF OF UPPER BOUND

- Bounding the perturbation terms:

$r = 2p/(2p+d)$

$$\text{perturbation of } B_1 \;\leq \sum_{t=1}^{b} \left| f_t(x_t^*) - f_t(x_1^*) \right| \;\lesssim \sum_{t=1}^{b} \|f_t - f_1\|_p^r$$

$$\leq b \cdot \left[\sum_{t=1}^{b-1} \|f_{t+1} - f_t\|_p \right]^r \leq b^{1+r-r/q} \left[\sum_{t=1}^{b-1} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

$$\text{total perturbation} \leq b^{1+r-r/q} \sum_{\ell=1}^{J} \left[\sum_{t=\underline{b}_\ell}^{\bar{b}_\ell} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

$$\leq J^{1-r/q} b^{1+r-r/q} \left[\sum_{t=1}^{T-1} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

PROOF OF UPPER BOUND

- Bounding the perturbation terms:

$r = 2p/(2p+d)$

$$\text{perturbation of } B_1 \;\le \sum_{t=1}^{b} \left| f_t(x_t^*) - f_t(x_1^*) \right| \;\lesssim \sum_{t=1}^{b} \|f_t - f_1\|_p^r$$

$$\le b \cdot \left[\sum_{t=1}^{b-1} \|f_{t+1} - f_t\|_p \right]^r \le b^{1+r-r/q} \left[\sum_{t=1}^{b-1} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

$$\text{total perturbation} \le b^{1+r-r/q} \sum_{\ell=1}^{J} \left[\sum_{t=\underline{b}_\ell}^{\bar{b}_\ell} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

$$\le J^{1-r/q} b^{1+r-r/q} \left[\sum_{t=1}^{T-1} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

PROOF OF UPPER BOUND

- Bounding the perturbation terms:

$r = 2p/(2p+d)$

$$\text{perturbation of } B_1 \;\leq \sum_{t=1}^{b} \left| f_t(x_t^*) - f_t(x_1^*) \right| \;\lesssim \sum_{t=1}^{b} \|f_t - f_1\|_p^r$$

$$\leq b \cdot \left[\sum_{t=1}^{b-1} \|f_{t+1} - f_t\|_p \right]^r \leq b^{1+r-r/q} \left[\sum_{t=1}^{b-1} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

$$\text{total perturbation} \leq b^{1+r-r/q} \sum_{\ell=1}^{J} \left[\sum_{t=\underline{b}_\ell}^{\bar{b}_\ell} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

$$\leq J^{1-r/q} b^{1+r-r/q} \left[\sum_{t=1}^{T-1} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

PROOF OF UPPER BOUND

- Bounding the perturbation terms:

$r = 2p/(2p+d)$

$$\text{perturbation of } B_1 \;\le \sum_{t=1}^{b} \left| f_t(x_t^*) - f_t(x_1^*) \right| \;\lesssim \sum_{t=1}^{b} \|f_t - f_1\|_p^r$$

$$\le b \cdot \left[\sum_{t=1}^{b-1} \|f_{t+1} - f_t\|_p \right]^r \le b^{1+r-r/q} \left[\sum_{t=1}^{b-1} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

total perturbation

$$\le b^{1+r-r/q} \sum_{\ell=1}^{J} \left[\sum_{t=\underline{b}_\ell}^{\bar{b}_\ell} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

Holder's ineq.

$$\le J^{1-r/q} b^{1+r-r/q} \left[\sum_{t=1}^{T-1} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

PROOF OF UPPER BOUND

- Bounding the perturbation terms:

$r = 2p/(2p+d)$

$$\text{perturbation of } B_1 \;\le \sum_{t=1}^{b} \left| f_t(x_t^*) - f_t(x_1^*) \right| \;\lesssim \sum_{t=1}^{b} \|f_t - f_1\|_p^r$$

$$\le b \cdot \left[\sum_{t=1}^{b-1} \|f_{t+1} - f_t\|_p \right]^r \le b^{1+r-r/q} \left[\sum_{t=1}^{b-1} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

$$\text{total perturbation} \le b^{1+r-r/q} \sum_{\ell=1}^{J} \left[\sum_{t=\underline{b}_\ell}^{\bar{b}_\ell} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

Holder's ineq.

$$\le J^{1-r/q} b^{1+r-r/q} \left[\sum_{t=1}^{T-1} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

PROOF OF UPPER BOUND

- Bounding the perturbation terms:

$r = 2p/(2p+d)$

$$\text{perturbation of } B_1 \;\le \sum_{t=1}^{b} \left| f_t(x_t^*) - f_t(x_1^*) \right| \;\lesssim \sum_{t=1}^{b} \|f_t - f_1\|_p^r$$

$$\le b \cdot \left[\sum_{t=1}^{b-1} \|f_{t+1} - f_t\|_p \right]^r \le b^{1+r-r/q} \left[\sum_{t=1}^{b-1} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

$$\text{total perturbation} \le b^{1+r-r/q} \sum_{\ell=1}^{J} \left[\sum_{t=\underline{b}_\ell}^{\bar{b}_\ell} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

Holder's ineq.

$$\le J^{1-r/q} b^{1+r-r/q} \left[\sum_{t=1}^{T-1} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

PROOF OF UPPER BOUND

- Bounding the perturbation terms:

$r = 2p/(2p+d)$

$$\text{perturbation of } B_1 \;\le \sum_{t=1}^{b} \left| f_t(x_t^*) - f_t(x_1^*) \right| \;\lesssim \sum_{t=1}^{b} \|f_t - f_1\|_p^r$$

$$\le b \cdot \left[\sum_{t=1}^{b-1} \|f_{t+1} - f_t\|_p \right]^r \le b^{1+r-r/q} \left[\sum_{t=1}^{b-1} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

$$\text{total perturbation} \le b^{1+r-r/q} \sum_{\ell=1}^{J} \left[\sum_{t=\underline{b}_\ell}^{\bar{b}_\ell} \|f_{t+1} - f_t\|_p^q \right]^{r/q} \qquad [\mathrm{Var}_{p,q}(\mathbf{f})]^r = V_T^r$$

Holder's ineq.

$$\le J^{1-r/q} b^{1+r-r/q} \left[\sum_{t=1}^{T-1} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

PROOF OF UPPER BOUND

- Bounding the perturbation terms:

$r = 2p/(2p+d)$

$$\text{perturbation of } B_1 \;\le \sum_{t=1}^{b} \left| f_t(x_t^*) - f_t(x_1^*) \right| \;\lesssim \sum_{t=1}^{b} \|f_t - f_1\|_p^r$$

$$\le b \cdot \left[\sum_{t=1}^{b-1} \|f_{t+1} - f_t\|_p \right]^r \le b^{1+r-r/q} \left[\sum_{t=1}^{b-1} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

$$\text{total perturbation} \le b^{1+r-r/q} \sum_{\ell=1}^{J} \left[\sum_{t=\underline{b}_\ell}^{\bar{b}_\ell} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

Holder's ineq.

$$\le J^{1-r/q} b^{1+r-r/q} \left[\sum_{t=1}^{T-1} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

$[\mathrm{var}_{p,q}(\mathbf{f})]^r = V_T^r$

PROOF OF UPPER BOUND

- Bounding the perturbation terms:

$r = 2p/(2p+d)$

$$\text{perturbation of } B_1 \;\le \sum_{t=1}^{b} \left| f_t(x_t^*) - f_t(x_1^*) \right| \;\lesssim \sum_{t=1}^{b} \|f_t - f_1\|_p^r$$

$$\le b \cdot \left[\sum_{t=1}^{b-1} \|f_{t+1} - f_t\|_p \right]^r \le b^{1+r-r/q} \left[\sum_{t=1}^{b-1} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

$$\text{total perturbation} \le b^{1+r-r/q} \sum_{\ell=1}^{J} \left[\sum_{t=\underline{b}_\ell}^{\bar{b}_\ell} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

Holder's ineq.

$$\le J^{1-r/q} b^{1+r-r/q} \left[\sum_{t=1}^{T-1} \|f_{t+1} - f_t\|_p^q \right]^{r/q}$$

$[Var_{p,q}(\mathbf{f})]^r = V_T^r$

PROOF OF UPPER BOUND

$$\underbrace{f_1, \cdots, f_{\overline{b}_1}}_{B_1}, \underbrace{f_{\underline{b}_2}, \cdots, f_{\overline{b}_2}}_{B_2}, \cdots, \underbrace{f_{\underline{b}_J}, \cdots, f_T}_{B_J}$$

Δ_T

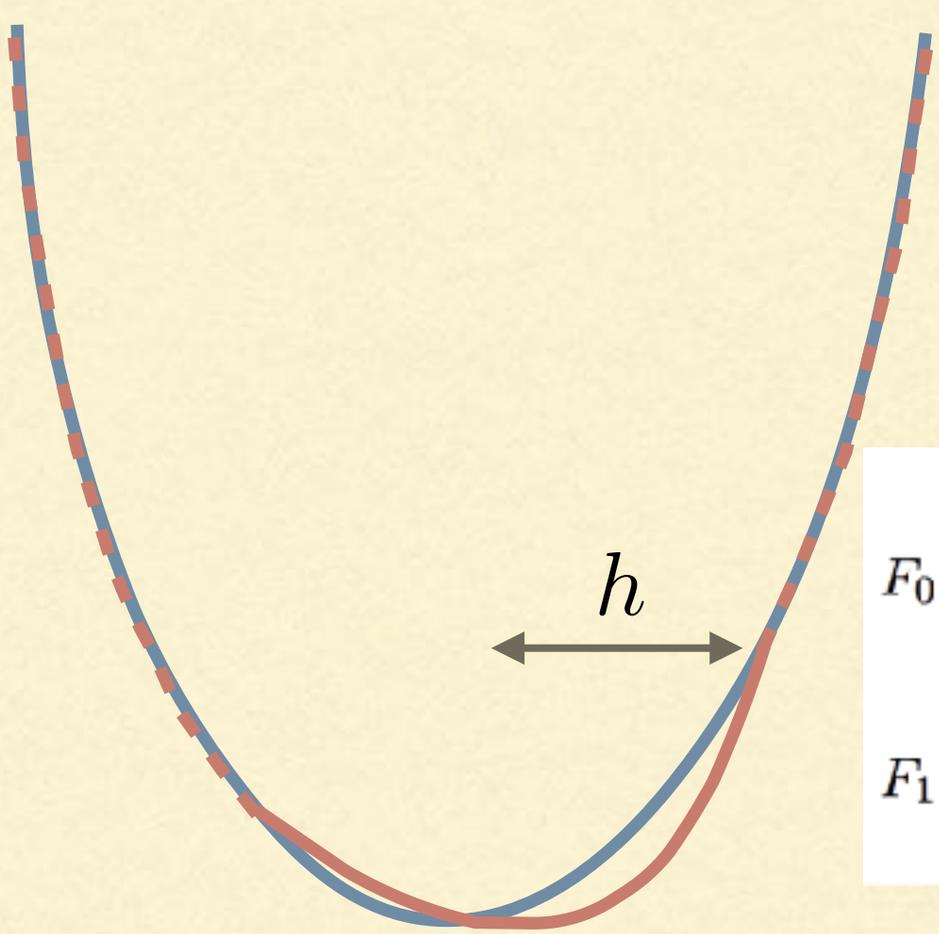
$$r = 2p/(2p+d)$$

- Stationary regret: $\mathcal{O}(J \cdot \sqrt{\Delta_T})$
- Total perturbation: $\mathcal{O}(J^{1-r/q} \Delta_T^{1+r-r/q} V_T^r)$

$$\Delta_T \asymp V_T^{-2r/(2r+1)} \longrightarrow \text{dynamic regret: } \mathcal{O}(V_T^{2p/(6p+d)} \cdot T)$$

PROOF IDEA OF LOWER BOUND

- Proof strategy: find "adversary examples" that are hard to distinguish from samples, but have different minimizers.



$$F_0(x) := \begin{cases} x^2, & 0 \le x < \sqrt{h}; \\ \frac{4}{\sqrt{h}}x^3 - 11x^2 + 12\sqrt{h}x - 4h, & \sqrt{h} \le x < 2\sqrt{h}; \\ 8(x - \sqrt{h})^2, & 2\sqrt{h} \le x \le 1. \end{cases}$$

$$F_1(x) := \begin{cases} (x - \sqrt{h})^2, & 0 \le x < \sqrt{h}; \\ 8(x - \sqrt{h})^2, & \sqrt{h} \le x \le 1. \end{cases}$$

PROOF IDEA OF LOWER BOUND

- Formally, find two function sequences **f** and **g** such that:
  - $\mathrm{KL}(\boldsymbol{f}\|\boldsymbol{g})$ is small $\mathcal{O}(h^2 T)$
  - $\mathrm{Var}_{p,q}(\boldsymbol{f},\boldsymbol{g})$ is small $\mathcal{O}(h^{2p/(2p+d)}/\Delta_T)$
  - $\chi(\boldsymbol{f},\boldsymbol{g}) = \sum_{t=1}^{T} \inf_{x\in\mathcal{X}} \{f_t(x) - f_t^*, g_t(x) - g_t^*\}$ is large $\Omega(hT)$

$$h \asymp V_T^{2p/(6p+d)}$$

$$\Delta_T \asymp T \cdot V_T^{-4p/(6p+d)}$$

→ lower bound: $\Omega(V_T^{2p/(6p+d)} \cdot T)$

OPEN QUESTIONS

- Adaptivity:
  - Our choice of J (#. of intervals) requires knowledge of p, q and V_T.

    Is it possible to design adaptive interval length rules?

- General convexity
  - Our analysis requires the function to be strongly convex and smooth.

    Is it possible to remove both assumptions?