Local Self-Tuning in Nonparametric Regression

Samory Kpotufe
ORFE, Princeton University
Local Regression

Data: \( \{(X_i, Y_i)\}_{i=1}^n \), \( Y = f(X) + \text{noise} \) 
\( f \in \text{nonparametric } \mathcal{F} \), i.e. \( \dim(\mathcal{F}) = \infty \).

Learn: 
\( f_n(x) = \text{avg } (Y_i) \) of Neighbors(\( x \)).
(e.g. \( k \)-NN, kernel, or tree-based reg.)

Quite basic \( \implies \) common in modern applications.

Sensitive to choice of Neighbors(\( x \)): \( k \), band. \( h \), tree cell size.
Goal: choose Neighbors(\( x \)) optimally!
Local Regression

Data: \( \{(X_i, Y_i)\}_{i=1}^{n} \), \( Y = f(X) + \text{noise} \)
\( f \in \text{nonparametric} \ F, \text{i.e. } \dim(F) = \infty \).

Learn:
\( f_n(x) = \text{avg} \ (Y_i) \text{ of Neighbors}(x) \).
(e.g. \(k\)-NN, kernel, or tree-based reg.)

Quite basic \(\implies\) common in modern applications.

Sensitive to choice of Neighbors\( (x) \): \( k \), band. \( h \), tree cell size.
Goal: choose Neighbors\( (x) \) optimally!
Local Regression

Data: \( \{(X_i, Y_i)\}_{i=1}^{n} \), \( Y = f(X) + \text{noise} \)
\( f \in \text{nonparametric } \mathcal{F} \), i.e. \( \text{dim}(\mathcal{F}) = \infty \).

Learn:
\( f_n(x) = \text{avg } (Y_i) \) of Neighbors(\( x \)).
(e.g. \( k \)-NN, kernel, or tree-based reg.)

Quite basic \( \Rightarrow \) common in modern applications.

Sensitive to choice of Neighbors(\( x \)): \( k \), band. \( h \), tree cell size.
Goal: choose Neighbors(\( x \)) optimally!
Local Regression

Data: \( \{(X_i, Y_i)\}_{i=1}^{n} \), \( Y = f(X) + \text{noise} \)
\( f \in \text{nonparametric } \mathcal{F} \), i.e. \( \dim(\mathcal{F}) = \infty \).

Learn:
\( f_n(x) = \text{avg} (Y_i) \) of Neighbors\((x)\).
(e.g. \( k\)-NN, kernel, or tree-based reg.)

Quite basic \( \implies \) common in modern applications.

Sensitive to choice of Neighbors\((x)\): \( k \), band. \( h \), tree cell size.
Goal: choose Neighbors\((x)\) optimally!
Performance depends on problem parameters

Performance would depend on \( \dim(X) \) and how fast \( f \) varies ... Suppose \( X \in \mathbb{R}^D \), and \( \forall x, x', \quad |f(x) - f(x')| \leq \lambda \|x - x'\|^\alpha \).

Performance measure: \( \|f_n - f\|_{2,P_X}^2 = \mathbb{E}_X |f_n(X) - f(X)|^2 \).

Minimax global performance \( \text{(Stone 80-82)} \)

\[
\|f_n - f\|_{2,P_X}^2 \propto \lambda^{2D/(2\alpha+D)} \cdot n^{-2\alpha/(2\alpha+D)}.
\]
Performance depends on problem parameters

Performance would depend on $\text{dim}(X)$ and how fast $f$ varies ... 
Suppose $X \in \mathbb{R}^D$, and $\forall x, x', \quad |f(x) - f(x')| \leq \lambda \|x - x'\|^{\alpha}$.

Performance measure: $\|f_n - f\|_{2,P_X}^2 = \mathbb{E}_X |f_n(X) - f(X)|^2$.

Minimax global performance \textit{(Stone 80-82)}

$$\|f_n - f\|_{2,P_X}^2 \propto \lambda^{2D/(2\alpha+D)} \cdot n^{-2\alpha/(2\alpha+D)}.$$
Performance depends on problem parameters

Performance would depend on $\dim(X)$ and how fast $f$ varies ...

Suppose $X \in \mathbb{R}^D$, and $\forall x, x', \quad |f(x) - f(x')| \leq \lambda \|x - x'\|^\alpha$.

Performance measure: $\|f_n - f\|_{2,P_X}^2 \doteq \mathbb{E}_X |f_n(X) - f(X)|^2$.

Minimax global performance (Stone 80-82)

$\|f_n - f\|_{2,P_X}^2 \propto \lambda^{2D/(2\alpha+D)} \cdot n^{-2\alpha/(2\alpha+D)}$. 
Performance depends on problem parameters

Performance would depend on $\text{dim}(X)$ and how fast $f$ varies ...
Suppose $X \in \mathbb{R}^D$, and $\forall x, x'$, $|f(x) - f(x')| \leq \lambda \|x - x'\|^\alpha$.

Performance measure: $\|f_n - f\|_{2,P_X}^2 = \mathbb{E}_X |f_n(X) - f(X)|^2$.

Minimax global performance (Stone 80-82)

$\|f_n - f\|_{2,P_X}^2 \propto \lambda^{2D/(2\alpha+D)} \cdot n^{-2\alpha/(2\alpha+D)}$. 

Some milder situations for $X \in \mathbb{R}^D$

$f$ is quite smooth, $f$ is sparse, $f$ is additive, ... 

Of interest here: $X$ has low intrinsic dimension $d \ll D$. 
Some milder situations for $X \in \mathbb{R}^D$

$f$ is quite smooth, $f$ is sparse, $f$ is additive, ...

Of interest here: $X$ has low intrinsic dimension $d \ll D$. 
\( \mathcal{X} \subset \mathbb{R}^D \) but has low intrinsic dimension \( d \ll D \)

Linear data
$\mathcal{X} \subset \mathbb{R}^D$ but has low intrinsic dimension $d \ll D$
\( X \subset \mathbb{R}^D \) but has low intrinsic dimension \( d \ll D \)
\( X \subset \mathbb{R}^D \) but has low intrinsic dimension \( d \ll D \)

Common approach: **Dimension reduction or estimation.**
Dimension reduction/estimation increases tuning!

Recent Alternative:

$f_n$ operates in $\mathbb{R}^D$ but adapts to the unknown $d$ of $\mathcal{X}$.

We want:

$$\|f_n - f\|_{2,P_X}^2 \lesssim n^{-1/Cd} \ll n^{-1/CD}$$
Dimension reduction/estimation increases tuning!

**Recent Alternative:**

$f_n$ operates in $\mathbb{R}^D$ but adapts to the unknown $d$ of $X$.

We want: $\|f_n - f\|_{2,P_X}^2 \lesssim n^{-1/Cd} \ll n^{-1/CD}$
Dimension reduction/estimation increases tuning!

Recent Alternative:
\( f_n \) operates in \( \mathbb{R}^D \) but adapts to the unknown \( d \) of \( \mathcal{X} \).

We want: \( \| f_n - f \|_{2, P_X}^2 \lesssim n^{-1/Cd} \ll n^{-1/CD} \)
Some work on adaptivity to intrinsic dimension:

Adaptivity to intrinsic $d$

**Main insight:** Key algorithmic quantities depend on $d$, not on $D$.

$$\text{For Lipschitz } f, \quad \|f_{n,\epsilon} - f\|_{2,P_X}^2 \approx \frac{\epsilon^{-d}}{n} + \epsilon^2.$$  

Cross-validate over $\epsilon$ for a good rate in terms of $d$. 
Adaptivity to intrinsic $d$

Main insight: Key algorithmic quantities depend on $d$, not on $D$.

Kernel reg.: Avg. **mass of a ball** of radius $\epsilon$ is approx. $\epsilon^d$

For Lipschitz $f$, $\|f_{n,\epsilon} - f\|_{2,P_X}^2 \approx \frac{\epsilon^{-d}}{n} + \epsilon^2$.

Cross-validate over $\epsilon$ for a good rate in terms of $d$. 
Adaptivity to intrinsic $d$

Main insight: Key algorithmic quantities depend on $d$, not on $D$.

RPtree: **Number of cells** of diameter $\epsilon$ is approx. $\epsilon^{-d}$.

For Lipschitz $f$, $\|f_n,\epsilon - f\|_{2,P_X}^2 \approx \frac{\epsilon^{-d}}{n} + \epsilon^2$.

Cross-validate over $\epsilon$ for a good rate in terms of $d$. 
Adaptivity to intrinsic $d$

Main insight: Key algorithmic quantities depend on $d$, not on $D$.

RPtree: **Number of cells** of diameter $\epsilon$ is approx. $\epsilon^{-d}$.

For Lipschitz $f$, $\| f_{n,\epsilon} - f \|_{2, P_X}^2 \approx \frac{\epsilon^{-d}}{n} + \epsilon^2$.

Cross-validate over $\epsilon$ for a good rate in terms of $d$. 
**Adaptivity to intrinsic** $d$

**Main insight:** Key algorithmic quantities depend on $d$, not on $D$.

**RPtree:** **Number of cells** of diameter $\epsilon$ is approx. $\epsilon^{-d}$.

For Lipschitz $f$, $\|f_{n,\epsilon} - f\|_{2,P_X}^2 \approx \frac{\epsilon^{-d}}{n} + \epsilon^2$.

**Cross-validate over $\epsilon$ for a good rate in terms of $d$.**
Insights help with tuning under time-constraints.

**Main Idea:** compress data in a way that respects structure of $\mathcal{X}$.

- Faster Kernel regression. [Kpo. 2009]
- Fast online tree-regression. [Kpo. and Orabona 2013]

$O(\log n)$ time vs. $O(n)$, + regression rates remain optimal!
Insights help with tuning under time-constraints.

**Main Idea:** compress data in a way that respects structure of $\mathcal{X}$.

- Faster Kernel regression. [Kpo. 2009]
- Fast online tree-regression. [Kpo. and Orabona 2013]

$O(\log n)$ time vs. $O(n)$, + regression rates remain optimal!
Insights help with tuning under time-constraints.

**Main Idea:** compress data in a way that respects structure of $\mathcal{X}$.

- Faster Kernel regression. [Kpo. 2009]
- Fast online tree-regression. [Kpo. and Orabona 2013]

$O(\log n)$ time vs. $O(n)$, + regression rates remain optimal!
Insights help with tuning under time-constraints.

**Main Idea:** compress data in a way that respects structure of \( \mathcal{X} \).

- Faster Kernel regression. [Kpo. 2009]
- Fast online tree-regression. [Kpo. and Orabona 2013]

\[ O(\log n) \text{ time vs. } O(n) , \text{ + regression rates remain optimal!} \]
So far, we have viewed $d$ as a global characteristic of $\mathcal{X}$ ...
Problem complexity is likely to depend on location!

Choose Neighbors$(x)$ adaptively so that:

$$|f_n(x) - f(x)|^2 \propto \lambda_x^{2d_x/(2\alpha_x+d_x)} \cdot n^{-2\alpha_x/(2\alpha_x+d_x)}.$$
Problem complexity is likely to depend on location!

Space $\mathcal{X}$, $d = d(x)$

Function $f$, $\alpha = \alpha(x)$

Choose Neighbors($x$) adaptively so that:

$$|f_n(x) - f(x)|^2 \propto \lambda_x^{2d_x/(2\alpha_x + d_x)} \cdot n^{-2\alpha_x/(2\alpha_x + d_x)}.$$
Problem complexity is likely to depend on location!

Choose Neighbors($x$) adaptively so that:

$$|f_n(x) - f(x)|^2 \propto \lambda_x^{2d_x/(2\alpha_x+d_x)} \cdot n^{-2\alpha_x/(2\alpha_x+d_x)}.$$

Choose Neighbors($x$): Cannot cross-validate locally at $x$!
I. Local notions of smoothness and dimension.

II. Local adaptivity to dimension: $k$-NN example.

III. Full local adaptivity: kernel example.
Local smoothness

Use local Hölder parameters $\lambda = \lambda(x), \alpha = \alpha(x)$ on $B(x, r)$:
For all $x' \in B(x, r)$, $|f(x) - f(x')| \leq \lambda \rho(x, x')^\alpha$.

$f(x) = x^\alpha$ is flatter at $x = 0$ as $\alpha$ is increased.
Local dimension

*Figure*: $d$-dimensional balls centered at $x$.

**Volume growth**: $\text{vol}(B(x, r)) = C \cdot r^d = \epsilon^{-d} \cdot \text{vol}(B(x, \epsilon r))$.

If $P_X$ is $\mathcal{U}(B(x, r))$, then $P_X(B(x, r)) \leq \epsilon^{-d} \cdot P_X(B(x, \epsilon r))$.

**Def.**: $P_X$ is $(C, d)$-homogeneous on $B(x, r)$ if $\forall r' \leq r, \epsilon > 0$, $P_X(B(x, r')) \leq C\epsilon^{-d} \cdot P_X(B(x, \epsilon r'))$. 
**Local dimension**

*Figure*: $d$-dimensional balls centered at $x$.

**Volume growth**: \( \text{vol}(B(x, r)) = C \cdot r^d = \epsilon^{-d} \cdot \text{vol}(B(x, \epsilon r)) \).

If \( P_X \) is \( U(B(x, r)) \), then \( P_X(B(x, r)) \lesssim \epsilon^{-d} \cdot P_X(B(x, \epsilon r)) \).

**Def.**: \( P_X \) is \((C, d)\)-homogeneous on \( B(x, r) \) if \( \forall r' \leq r, \epsilon > 0 \),

\[
P_X(B(x, r')) \leq C \epsilon^{-d} \cdot P_X(B(x, \epsilon r')).
\]
Local dimension

**Figure**: $d$-dimensional balls centered at $x$.

**Volume growth**: $\text{vol}(B(x, r)) = C \cdot r^d = \epsilon^{-d} \cdot \text{vol}(B(x, \epsilon r))$.

If $P_X$ is $\mathcal{U}(B(x, r))$, then $P_X(B(x, r)) \lesssim \epsilon^{-d} \cdot P_X(B(x, \epsilon r))$.

**Def.**: $P_X$ is $(C, d)$-homogeneous on $B(x, r)$ if $\forall r' \leq r, \epsilon > 0$, $P_X(B(x, r')) \leq C \epsilon^{-d} \cdot P_X(B(x, \epsilon r'))$. 
Local dimension

Figure: $d$-dimensional balls centered at $x$.

Volume growth: $\text{vol}(B(x,r)) = C \cdot r^d = \epsilon^{-d} \cdot \text{vol}(B(x,\epsilon r))$.

If $P_X$ is $\mathcal{U}(B(x,r))$, then $P_X(B(x,r)) \lesssim \epsilon^{-d} \cdot P_X(B(x,\epsilon r))$.

Def.: $P_X$ is $(C,d)$-homogeneous on $B(x,r)$ if $\forall r' \leq r, \epsilon > 0$, $P_X(B(x,r')) \leq C\epsilon^{-d} \cdot P_X(B(x,\epsilon r'))$. 
The growth of $P_X$ can capture the intrinsic dimension in $B(x)$.

Location of query $x$ matters!

Size of neighborhood $B$ matters!
The growth of $P_X$ can capture the intrinsic dimension in $B(x)$.

Location of query $x$ matters!

Size of neighborhood $B$ matters!
The growth of $P_X$ can capture the intrinsic dimension in $B(x)$.

Location of query $x$ matters!

Size of neighborhood $B$ matters!
Size of neighborhood $B$ matters!

For $k$-NN, or kernel reg, size of $B$ depends on $n$ and ($k$ or $h$).
Size of neighborhood $B$ matters!

For $k$-NN, or kernel reg, size of $B$ depends on $n$ and ($k$ or $h$).
The growth of $P_X(B)$ can capture the intrinsic dimension locally.
The growth of $P_X(B)$ can capture the intrinsic dimension locally.

$\mathcal{X}$ can be a collection of subspaces of various dimensions.
Intrinsic $d$ tightly captures the minimax rate:

**Theorem:** Consider a metric measure space $(\mathcal{X}, \rho, \mu)$, such that for all $x \in \mathcal{X}, r > 0, \epsilon > 0$, we have $\mu(B(x, r)) \approx \epsilon^{-d} \mu(B(x, \epsilon r))$. Then, for any regressor $f_n$, there exists $P_{X,Y}$, where $P_X = \mu$ and $f(x) = \mathbb{E} Y|x$ is $\lambda$-Lipschitz, such that

$$\mathbb{E}_{P^n_{X,Y}} \|f_n - f\|_{2,\mu}^2 \gtrsim \lambda^{2d/(2+d)} \cdot n^{-2/(2+d)}.$$
Intrinsic \( d \) tightly captures the minimax rate:

**Theorem:** Consider a metric measure space \((\mathcal{X}, \rho, \mu)\), such that for all \( x \in \mathcal{X}, r > 0, \epsilon > 0 \), we have \( \mu(B(x, r)) \approx \epsilon^{-d} \mu(B(x, \epsilon r)) \). Then, for any regressor \( f_n \), there exists \( P_{X,Y} \), where \( P_X = \mu \) and \( f(x) = \mathbb{E} Y | x \) is \( \lambda \)-Lipschitz, such that

\[
\mathbb{E}_{P_X^n, Y} \| f_n - f \|_{2, \mu}^2 \gtrsim \lambda^{2d/(2+d)} \cdot n^{-2/(2+d)}.
\]
I. Local notions of smoothness and dimension.

II. Local adaptivity to dimension: $k$-NN example.

III. Full local adaptivity: kernel example.
Main Assumptions:

- \( X \in \text{metric space } (\mathcal{X}, \rho) \).
- \( P_X \) is locally homogeneous with unknown \( d(x) \).
- \( f \) is \( \lambda \)-Lipschitz on \( \mathcal{X} \), i.e. \( \alpha = 1 \).

**k-NN regression:** \( f_n(x) = \text{weighted avg } (Y_i) \text{ of } k-\text{NN}(x) \).

Suppose \( \mathcal{X} \subset \mathbb{R}^D \), the learner operates in \( \mathbb{R}^D \)!
No dimensionality reduction, no dimension estimation!
Main Assumptions:

- $X \in$ metric space $(\mathcal{X}, \rho)$.
- $P_X$ is locally homogeneous with unknown $d(x)$.
- $f$ is $\lambda$-Lipschitz on $\mathcal{X}$, i.e. $\alpha = 1$.

$k$-NN regression: $f_n(x) =$ weighted avg $(Y_i)$ of $k$-NN$(x)$.

Suppose $\mathcal{X} \subset \mathbb{R}^D$, the learner operates in $\mathbb{R}^D$!
**No dimensionality reduction, no dimension estimation!**
**Bias-Variance tradeoff**

\[
\mathbb{E}_{(X_i,Y_i) \sim \mathcal{D}^n} |f_n(x) - f(x)|^2 = \mathbb{E} |f_n(x) - \mathbb{E} f_n(x)|^2 + |\mathbb{E} f_n(x) - f(x)|^2.
\]
General intuition:

Fix, \( n \gtrsim k \gtrsim \log n \), and let \( x \in \text{region } B \) of dimension \( d \).

Rate of convergence of \( f_n(x) \) depends on:

- (Variance of \( f_n(x) \)) \( \approx 1/k \).
- (Bias of \( f_n(x) \)) \( \approx r_k(x) \).

We have:

\[
\| f_n(x) - f(x) \|_2 \lesssim \frac{1}{k} + r_k(x)
\]

It turns out: \( r_k(x) \approx (k/n)^{1/d} \), where \( d = d(B) \).
General intuition:

Fix, \( n \gtrsim k \gtrsim \log n \), and let \( x \in \text{region } B \) of dimension \( d \).

Rate of convergence of \( f_n(x) \) depends on:

- (Variance of \( f_n(x) \)) \( \approx 1/k \).
- (Bias of \( f_n(x) \)) \( \approx r_k(x) \).

We have: \( |f_n(x) - f(x)|^2 \lesssim \frac{1}{k} + r_k(x)^2 \).

It turns out: \( r_k(x) \approx (k/n)^{1/d} \), where \( d = d(B) \).
General intuition:

Fix, $n \gtrsim k \gtrsim \log n$, and let $x \in$ region $B$ of dimension $d$.

Rate of convergence of $f_n(x)$ depends on:

- (Variance of $f_n(x)$) $\approx 1/k$.
- (Bias of $f_n(x)$) $\approx r_k(x)$.

We have: $|f_n(x) - f(x)|^2 \lesssim \frac{1}{k} + r_k(x)^2$.

It turns out: $r_k(x) \approx (k/n)^{1/d}$, where $d = d(B)$.
General intuition:

Fix, $n \gtrsim k \gtrsim \log n$, and let $x \in \text{region } B$ of dimension $d$.

Rate of convergence of $f_n(x)$ depends on:

- (Variance of $f_n(x)$) $\approx 1/k$.
- (Bias of $f_n(x)$) $\approx r_k(x)$.

We have: $|f_n(x) - f(x)|^2 \lesssim \frac{1}{k} + r_k(x)^2$.

It turns out: $r_k(x) \approx (k/n)^{1/d}$, where $d = d(B)$. 
General intuition:

Fix, $n \gtrsim k \gtrsim \log n$, and let $x \in \text{region } B$ of dimension $d$.

Rate of convergence of $f_n(x)$ depends on:

- (Variance of $f_n(x)$) $\approx 1/k$.
- (Bias of $f_n(x)$) $\approx r_k(x)$.

We have: $|f_n(x) - f(x)|^2 \lesssim \frac{1}{k} + r_k(x)^2$.

It turns out: $r_k(x) \approx (k/n)^{1/d}$, where $d = d(B)$.
Choosing $k$ locally at $x$ - Intuition

**Remember:** Cross-valid. or dim. estimation at $x$ are impractical.

Instead:

Main technical hurdle: intrinsic dimension might vary with $k$. 

![Graph showing the relationship between $k$, $k(x)$, and $r_k^2(x)$](image-url)
Choosing $k$ locally at $x$: Intuition

**Remember:** Cross-valid. or dim. estimation at $x$ are impractical.

Instead:

Main technical hurdle: intrinsic dimension might vary with $k$. 
Choosing \( k(x) \) - Result

**Theorem:** Suppose \( k(x) \) is chosen as above. The following holds \textit{w.h.p. simultaneously} for all \( x \).

Consider any \( B \) centered at \( x \), s.t. \( P_X(B) \gtrsim n^{-1/3} \). Suppose \( P_X \) is \((C,d)\)-homogeneous on \( B \). We have

\[
|f_n(x) - f(x)|^2 \lesssim \lambda^2 \left( \frac{C \ln n}{nP_X(B)} \right)^{2/(2+d)}.
\]

As \( n \to \infty \) the claim applies to any \( B \) centered at \( x \), \( P_X(B) \neq 0 \).
**Theorem:** Suppose $k(x)$ is chosen as above. The following holds \textit{w.h.p. simultaneously} for all $x$.

Consider any $B$ centered at $x$, s.t. $P_X(B) \gtrsim n^{-1/3}$. Suppose $P_X$ is $(C,d)$-homogeneous on $B$. We have

$$|f_n(x) - f(x)|^2 \lesssim \lambda^2 \left( \frac{C \ln n}{nP_X(B)} \right)^{2/(2+d)}.$$

As $n \to \infty$ the claim applies to any $B$ centered at $x$, $P_X(B) \neq 0$. 

---

*Choosing $k(x)$ - Result*
I. Local notions of smoothness and dimension.

II. Local adaptivity to dimension: \( k \)-NN example.

III. Full local adaptivity: kernel example.
(Recent work with Vikas Garg)
Main Assumptions:

- $X \in$ metric space $(\mathcal{X}, \rho)$ of diameter 1.
- $P_X$ is locally homogeneous with unknown $d(x)$.
- $f$ is locally Hölder with unknown $\lambda(x), \alpha(x)$.

Kernel regression: $f_n(x) = \text{weighted avg} \ (Y_i) \ \text{for} \ X_i \ \text{in} \ B_\rho(x, h)$. 
Main Assumptions:

- \( X \in \text{metric space} (\mathcal{X}, \rho) \) of diameter 1.
- \( P_X \) is locally homogeneous with unknown \( d(x) \).
- \( f \) is locally Hölder with unknown \( \lambda(x), \alpha(x) \).

Kernel regression: \( f_n(x) = \text{weighted avg} \ (Y_i) \) for \( X_i \) in \( B_\rho(x, h) \).
General intuition:

Fix, $0 < h < 1$, and let $x \in \text{region } B$ of dimension $d$.

Rate of convergence of $f_n(x)$ depends on:

- **(Variance of $f_n(x)$)** $\approx \frac{1}{n h(x)}$.
- **(Bias of $f_n(x)$)** $\approx h^{2\alpha}$.

We have:

$$|f_n(x) - f(x)|^2 \gtrapprox \frac{1}{n h(x)} h^{2\alpha}.$$

It turns out: $n h(x) \approx n h^d$, where $d = d(B)$. 

General intuition:

Fix, $0 < h < 1$, and let $x \in$ region $B$ of dimension $d$.

Rate of convergence of $f_n(x)$ depends on:

- (Variance of $f_n(x)$) $\approx 1/n_h(x)$.
- (Bias of $f_n(x)$) $\approx h^{2\alpha}$.

We have: $|f_n(x) - f(x)|^2 \lesssim \frac{1}{n_h(x)} + h^{2\alpha}$.

It turns out: $n_h(x) \approx nh^d$, where $d = d(B)$. 
General intuition:

Fix, $0 < h < 1$, and let $x \in \text{region } B$ of dimension $d$.

Rate of convergence of $f_n(x)$ depends on:

- \textbf{(Variance of } f_n(x)\textbf{)} $\approx \frac{1}{n h(x)}$.
- \textbf{(Bias of } f_n(x)\textbf{)} $\approx h^{2\alpha}$.

We have: $|f_n(x) - f(x)|^2 \lesssim \frac{1}{n h(x)} + h^{2\alpha}$.

It turns out: $n_h(x) \approx n h^d$, where $d = d(B)$.
General intuition:

Fix, $0 < h < 1$, and let $x \in$ region $B$ of dimension $d$.

Rate of convergence of $f_n(x)$ depends on:

- (Variance of $f_n(x)$) $\approx 1/n_h(x)$.
- (Bias of $f_n(x)$) $\approx h^{2\alpha}$.

We have: $|f_n(x) - f(x)|^2 \lesssim \frac{1}{n_h(x)} + h^{2\alpha}$.

It turns out: $n_h(x) \approx nh^d$, where $d = d(B)$. 
**General intuition:**

Fix, $0 < h < 1$, and let $x \in \text{region } B$ of dimension $d$.

Rate of convergence of $f_n(x)$ depends on:

- *(Variance of $f_n(x)$)* $\approx 1/n_h(x)$.
- *(Bias of $f_n(x)$)* $\approx h^{2\alpha}$.

We have: $|f_n(x) - f(x)|^2 \lesssim \frac{1}{n_h(x)} + h^{2\alpha}$.

It turns out: $n_h(x) \approx nh^d$, where $d = d(B)$. 
From the previous intuition

Suppose we know \( \alpha(x) \) but not \( d(x) \).

Monitor \( \frac{1}{n_h(x)} \) and \( h^{2\alpha} \).

Picking \( h_d(x) : |f_n(x) - f(x)|^2 \approx \text{err}(h^*) \lesssim n^{-2\alpha/(2\alpha+d)} \).
From the previous intuition

Suppose we know $\alpha(x)$ but not $d(x)$.

Monitor $\frac{1}{n_h(x)}$ and $h^{2\alpha}$.

Picking $h_d(x)$: $|f_n(x) - f(x)|^2 \approx \text{err}(h^*) \lesssim n^{-2\alpha/(2\alpha + d)}$. 
From Lepski

Suppose we know \( d(x) \) but not \( \alpha(x) \).

Intuition:

For every \( h < h^* \), \( \frac{1}{nh^d} > h^{2\alpha} \) therefore for such \( h \)

\[
|f_n(h; x) - f(x)|^2 \lesssim \frac{1}{nh^d} + h^{2\alpha} \leq 2 \frac{1}{nh^d}.
\]
From Lepski

Suppose we know $d(x)$ but not $\alpha(x)$.

Intuition:

For every $h < h^*$, $\frac{1}{nh^d} > h^{2\alpha}$ therefore for such $h$

$$|f_n(h; x) - f(x)|^2 \lesssim \frac{1}{nh^d} + h^{2\alpha} \leq 2 \frac{1}{nh^d}.$$. 
From Lepski

Suppose we know $d(x)$ but not $\alpha(x)$.

All intervals $\left[ f_n(h; x) \pm \sqrt{2 \frac{1}{nh^d}} \right]$, $h < h^*$ must intersect!

Picking $h_\alpha(x)$: $|f_n(x) - f(x)|^2 \approx \text{err}(h^*) \lesssim n^{-\frac{2\alpha}{2\alpha+d}}$. 
From Lepski

Suppose we know $d(x)$ but not $\alpha(x)$.

All intervals $\left[ f_n(h; x) \pm \sqrt{\frac{2}{nh^d}} \right], h < h^*$ must intersect!

Picking $h_\alpha(x)$: $|f_n(x) - f(x)|^2 \approx \text{err}(h^*) \lesssim n^{-2\alpha/(2\alpha + d)}$. 
Combine Lepski with previous intuition

We know neither $d$ nor $\alpha$.

All intervals $\left[f_n(h; x) \pm \sqrt{2\frac{1}{n_h(x)}}\right], h < h_d$ must intersect!

Picking $h_{\alpha,d}(x): |f_n(x) - f(x)|^2 \approx \text{err}(h_d) \approx \text{err}(h^*) \lesssim n^{-2\alpha/(2\alpha+d)}$. 
Combine Lepski with previous intuition

We know neither $d$ nor $\alpha$.

All intervals $\left[ f_n(h; x) \pm \sqrt{\frac{1}{2n h(x)}} \right]$, $h < h_d$ must intersect!

Picking $h_{\alpha,d}(x)$: $\left| f_n(x) - f(x) \right|^2 \approx \text{err}(h_d) \approx \text{err}(h^*) \lesssim n^{-2\alpha/(2\alpha+d)}$. 
Choosing $h_{\alpha,d}(x)$ - Result

**Tightness assumption on $d(x)$:** $\exists r_0, \forall x \in \mathcal{X}, \exists C, C', d$ such that $\forall r \leq r_0$, $C r^d \leq P_X(B(x,r)) \leq C' r^d$.

**Theorem:** Suppose $h_{\alpha,d}(x)$ is chosen as described. Let $n \geq N(r_0)$. The following holds w.h.p. simultaneously for all $x$. Let $d, \alpha, \lambda$ be the local problem parameters on $B(x,r_0)$. We have

$$|f_n(x) - f(x)|^2 \leq \lambda^{2d/(2\alpha+d)} \left( \frac{\ln n}{n} \right)^{2\alpha/(2\alpha+d)}.$$

The rate is optimal.
Choosing $h_{\alpha,d}(x)$ - Result

**Tightness assumption on $d(x)$**: \( \exists r_0, \forall x \in \mathcal{X}, \exists C, C', d \text{ such that } \forall r \leq r_0, \quad C r^d \leq P_X(B(x,r)) \leq C' r^d. \)

**Theorem**: Suppose $h_{\alpha,d}(x)$ is chosen as described. Let $n \geq N(r_0)$. The following holds w.h.p. simultaneously for all $x$. Let $d, \alpha, \lambda$ be the local problem parameters on $B(x,r_0)$. We have

\[
|f_n(x) - f(x)|^2 \lesssim \lambda^{2d/(2\alpha+d)} \left( \frac{\ln n}{n} \right)^{2\alpha/(2\alpha+d)}.
\]

The rate is optimal.
Simulation on data with mixed spatial complexity

... the approach works, but should be made more efficient!
Simulation on data with mixed spatial complexity

... the approach works, but should be made more efficient!
Future direction:
Extend self-tuning to tree-based kernel implementations.
Initial experiments with tree-based kernel:

Without CValidation: automatically detect interval containing $h^*$. 
**Future direction:**
Extend self-tuning to data streaming setting.
Initial streaming experiments:

Robustness to increasing intrinsic dimension.
Future direction: *Adaptive error bands.*

Would lead to more local sampling strategies.
TAKE HOME MESSAGE:

- We can adapt to intrinsic $d(X')$ without preprocessing.
- Local-learners can self-tune optimally to local $d(x)$ and $\alpha(x)$.

Results extend to plug-in classification!

Many potential future directions!
TAKE HOME MESSAGE:

- We can adapt to intrinsic $d(X)$ without preprocessing.
- Local-learners can self-tune optimally to local $d(x)$ and $\alpha(x)$.

Results extend to plug-in classification!

Many potential future directions!
TAKE HOME MESSAGE:

- We can adapt to intrinsic \( d(\mathcal{X}) \) without preprocessing.
- Local-learners can self-tune optimally to local \( d(x) \) and \( \alpha(x) \).

Results extend to plug-in classification!

Many potential future directions!
Thank you!