# Scalable Non-Parametric Statistical Estimation

Aymeric DIEULEVEUT

ENS Paris, INRIA

February 6, 2017

**Statistics**

Statistical model

Performance measure

Estimator

Convergence: $F(\#obs)$

**Statistics**

Statistical model

Performance measure

Estimator

Convergence: $F(\#obs)$

**Optimization**

Minimize a given function

Algorithm focused

Scales with dimension and observations

Convergence: $F(\#\text{iter})$

**Statistics**

Statistical model

Performance measure

Estimator

Convergence: $F(\#obs)$

**Accurate & Efficient**

Scalable estimators with optimal statistical properties

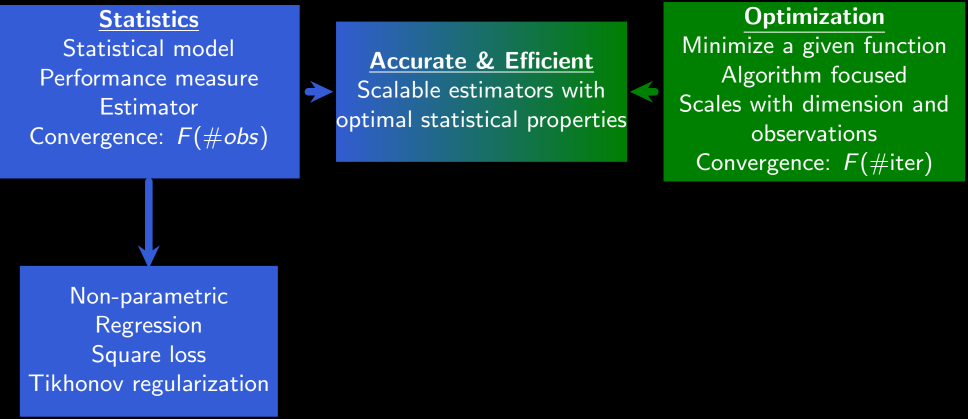
**Optimization**

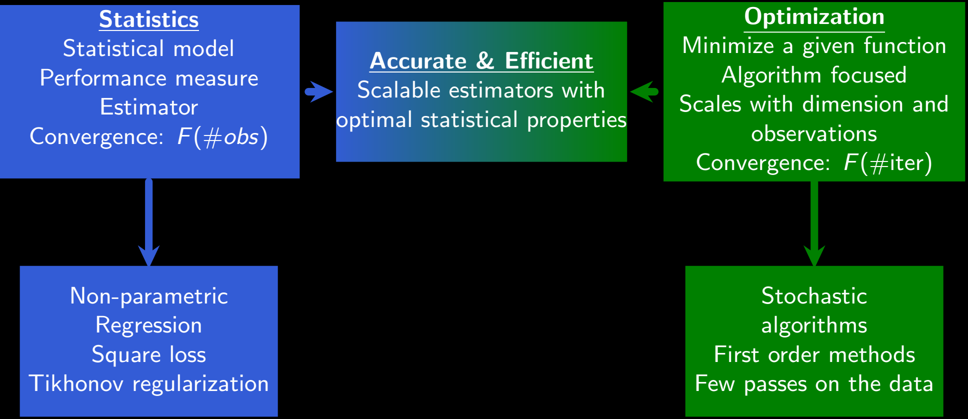
Minimize a given function

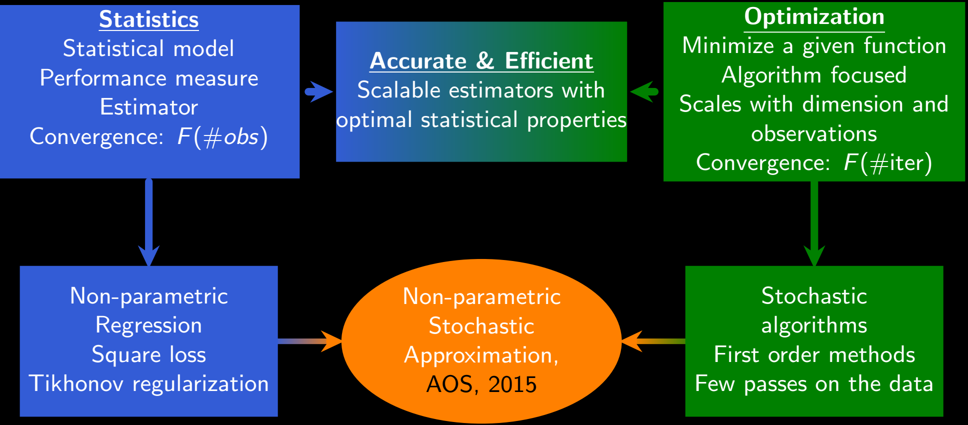
Algorithm focused

Scales with dimension and observations

Convergence: $F(\#\text{iter})$


Statistics
Statistical model
Performance measure
Estimator
Convergence: $F(\#obs)$
Accurate & Efficient
Scalable estimators with
optimal statistical properties
Optimization
Minimize a given function
Algorithm focused
Scales with dimension and
observations
Convergence: $F(\#\text{iter})$
Non-parametric
Regression
Square loss
Tikhonov regularization


Statistics
Statistical model
Performance measure
Estimator
Convergence: $F(\#obs)$
Accurate & Efficient
Scalable estimators with optimal statistical properties
Optimization
Minimize a given function
Algorithm focused
Scales with dimension and observations
Convergence: $F(\#\text{iter})$
Non-parametric
Regression
Square loss
Tikhonov regularization
Stochastic algorithms
First order methods
Few passes on the data


Statistics
Statistical model
Performance measure
Estimator
Convergence: $F(\#obs)$
Accurate & Efficient
Scalable estimators with
optimal statistical properties
Optimization
Minimize a given function
Algorithm focused
Scales with dimension and
observations
Convergence: $F(\#\text{iter})$
Non-parametric
Regression
Square loss
Tikhonov regularization
Non-parametric
Stochastic
Approximation,
AOS, 2015
Stochastic
algorithms
First order methods
Few passes on the data

## Non-parametric Stochastic Approximation with large step sizes 1/2.

Aymeric Dieuleveut & Francis Bach, *in the Annals of Statistics*, 2015.

Random design least-squares regression.

## Non-parametric Stochastic Approximation with large step sizes 1/2.

Aymeric Dieuleveut & Francis Bach, *in the Annals of Statistics*, 2015.

Random design least-squares regression.

$$\varepsilon(f) := \mathbb{E}_{(X,Y)}\left[(f(X)-Y)^2\right].$$

## Non-parametric Stochastic Approximation with large step sizes 1/2.

Aymeric Dieuleveut & Francis Bach, *in the Annals of Statistics*, 2015.

Random design least-squares regression.

$$\varepsilon(f) := \mathbb{E}_{(X,Y)}\left[(f(X)-Y)^2\right].$$

Within a reproducing kernel Hilbert space $\mathcal{H}$:

$$\min_{f\in\mathcal{H}} \varepsilon(f).$$

$(x_i, y_i)$ i.i.d. observations.

## Non-parametric Stochastic Approximation with large step sizes 1/2.

Aymeric Dieuleveut & Francis Bach, *in the Annals of Statistics*, 2015.

Random design least-squares regression.

$$\varepsilon(f) := \mathbb{E}_{(X,Y)}\left[(f(X) - Y)^2\right].$$

Within a reproducing kernel Hilbert space $\mathcal{H}$:

$$\min_{f\in\mathcal{H}} \varepsilon(f).$$

$(x_i, y_i)$ i.i.d. observations.

Sequence of estimators $f_t \in \mathcal{H}$.

## Non-parametric Stochastic Approximation with large step sizes 1/2.

Aymeric Dieuleveut & Francis Bach, *in the Annals of Statistics*, 2015.

Random design least-squares regression.

$$\varepsilon(f) := \mathbb{E}_{(X,Y)}\left[(f(X) - Y)^2\right].$$

Within a reproducing kernel Hilbert space $\mathcal{H}$:

$$\min_{f \in \mathcal{H}} \varepsilon(f).$$

$(x_i, y_i)$ i.i.d. observations.

Sequence of estimators $f_t \in \mathcal{H}$.
Update after each observation.

## Non-parametric Stochastic Approximation with large step sizes 1/2.

Aymeric Dieuleveut & Francis Bach, *in the Annals of Statistics*, 2015.

Random design least-squares regression.

$$\varepsilon(f) := \mathbb{E}_{(X,Y)}\left[(f(X)-Y)^2\right].$$

Within a reproducing kernel Hilbert space $\mathcal{H}$:

$$\min_{f\in\mathcal{H}} \varepsilon(f).$$

$(x_i, y_i)$ i.i.d. observations.

Sequence of estimators $f_t \in \mathcal{H}$.
Update after each observation.
Using unbiased gradients of the loss function:

## Non-parametric Stochastic Approximation with large step sizes 1/2.

Aymeric Dieuleveut & Francis Bach, *in the Annals of Statistics*, 2015.

Random design least-squares regression.

$$\varepsilon(f) := \mathbb{E}_{(X,Y)}\left[(f(X) - Y)^2\right].$$

Within a reproducing kernel Hilbert space $\mathcal{H}$:

$$\min_{f \in \mathcal{H}} \varepsilon(f).$$

$(x_i, y_i)$ i.i.d. observations.

Sequence of estimators $f_t \in \mathcal{H}$.
Update after each observation.
Using unbiased gradients of the loss function:

$$f_{t+1} = f_t - \gamma_t(f_t(x_t) - y_t)K_{x_t},$$

where: $K$ is the kernel,
$K_x = K(x, \cdot)$.

# Non-parametric Stochastic Approximation with large step sizes 1/2.

Aymeric Dieuleveut & Francis Bach, *in the Annals of Statistics*, 2015.

Random design least-squares regression.

$$\varepsilon(f) := \mathbb{E}_{(X,Y)}\left[(f(X)-Y)^2\right].$$

Within a reproducing kernel Hilbert space $\mathcal{H}$:

$$\min_{f\in\mathcal{H}} \varepsilon(f).$$

$(x_i, y_i)$ i.i.d. observations.

Sequence of estimators $f_t \in \mathcal{H}$.
Update after each observation.
Using unbiased gradients of the loss function:

$$f_{t+1} = f_t - \gamma_t (f_t(x_t) - y_t) K_{x_t},$$

where: $K$ is the kernel,

$$K_x = K(x,\cdot).$$

$\looparrowright$ Stochastic Approximation.

## Non-parametric Stochastic Approximation with large step sizes 1/2.

Aymeric Dieuleveut & Francis Bach, *in the Annals of Statistics*, 2015.

Random design least-squares regression.

$$\varepsilon(f) := \mathbb{E}_{(X,Y)}\left[(f(X) - Y)^2\right].$$

Within a reproducing kernel Hilbert space $\mathcal{H}$:

$$\min_{f \in \mathcal{H}} \varepsilon(f).$$

$(x_i, y_i)$ i.i.d. observations.

Sequence of estimators $f_t \in \mathcal{H}$.
Update after each observation.
Using unbiased gradients of the loss function:

$$f_{t+1} = f_t - \gamma_t (f_t(x_t) - y_t) K_{x_t},$$

where: $K$ is the kernel,
$K_x = K(x, \cdot)$.

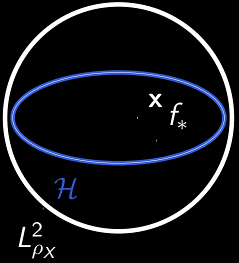
$\looparrowright$ Stochastic Approximation.

Depending on assumptions on:

- the Gaussian complexity of the unit ball of the kernel space,
- the smoothness in $\mathcal{H}$ of the optimal predictor $f_*(X) = \mathbb{E}[Y|X]$.

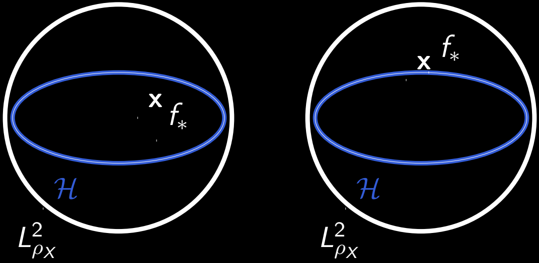
# Non-parametric Stochastic Approximation with large step sizes 2/2.

Aymeric Dieuleveut & Francis Bach, *in the Annals of Statistics*, 2015.



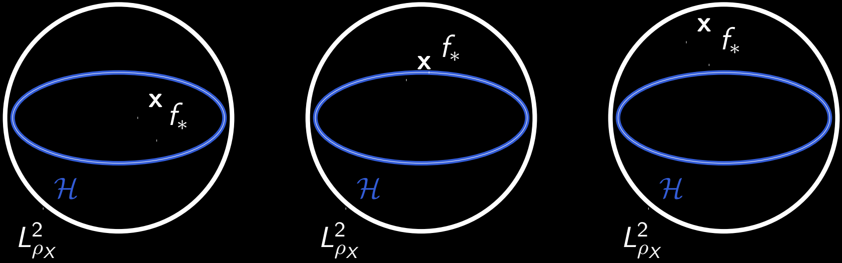
## Non-parametric Stochastic Approximation with large step sizes 2/2.

Aymeric Dieuleveut & Francis Bach, *in the Annals of Statistics*, 2015.



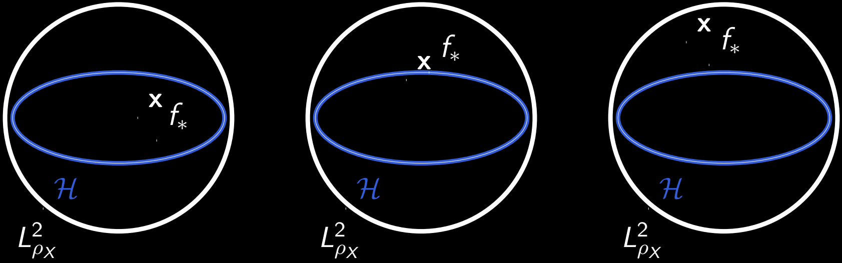
# Non-parametric Stochastic Approximation with large step sizes 2/2.

Aymeric Dieuleveut & Francis Bach, *in the Annals of Statistics*, 2015.



# Non-parametric Stochastic Approximation with large step sizes 2/2.

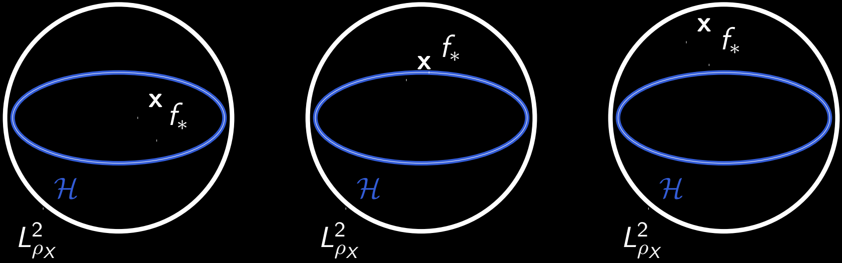
Aymeric Dieuleveut & Francis Bach, *in the Annals of Statistics*, 2015.




**Theorem: Averaged, unregularized, least mean squares algorithm, with large step sizes,** gets **Statistical optimal rate of convergence.**

## Non-parametric Stochastic Approximation with large step sizes 2/2.

Aymeric Dieuleveut & Francis Bach, *in the Annals of Statistics*, 2015.

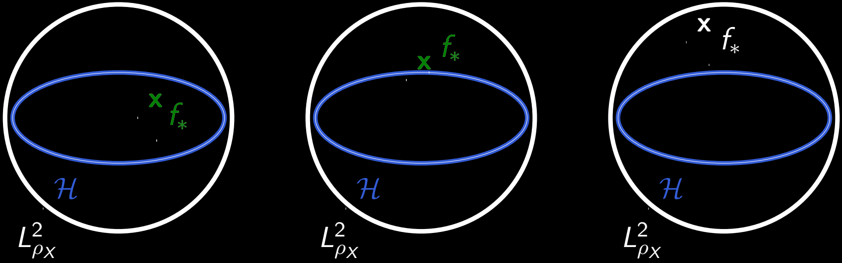



**Theorem: Averaged, unregularized, least mean squares algorithm, with large step sizes,** gets **Statistical optimal rate of convergence.**

- Recovers the finite dimension situation with rate $O\left(\frac{\sigma^2 d}{n}\right)$.
- Optimal rates in both the well-specified regime and some situations of the mis-specified.

## Non-parametric Stochastic Approximation with large step sizes 2/2.

Aymeric Dieuleveut & Francis Bach, *in the Annals of Statistics*, 2015.

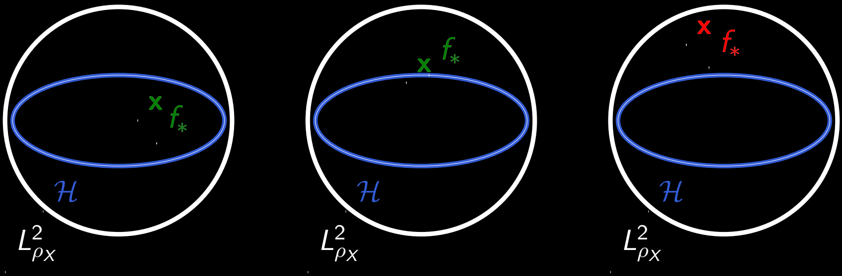



**Theorem: Averaged, unregularized, least mean squares algorithm, with large step sizes,** gets **Statistical optimal rate of convergence.**

- Recovers the finite dimension situation with rate $O\left(\frac{\sigma^2 d}{n}\right)$.
- Optimal rates in both the well-specified regime and some situations of the mis-specified.

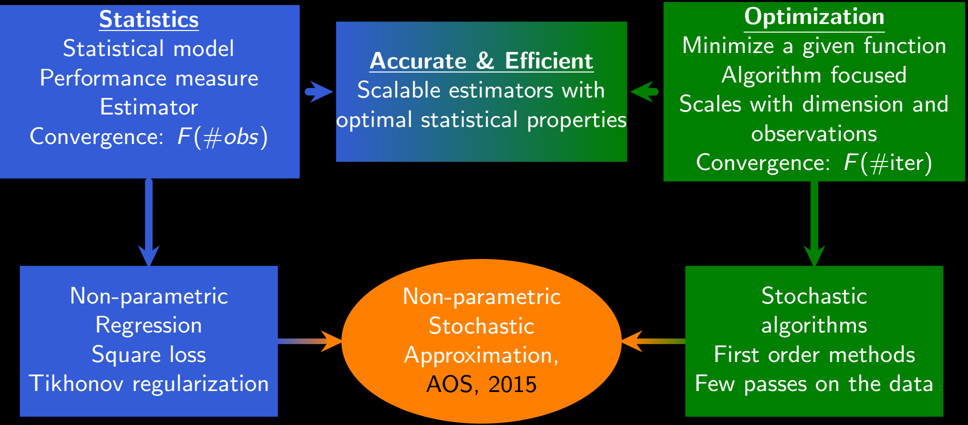
# Non-parametric Stochastic Approximation with large step sizes 2/2.

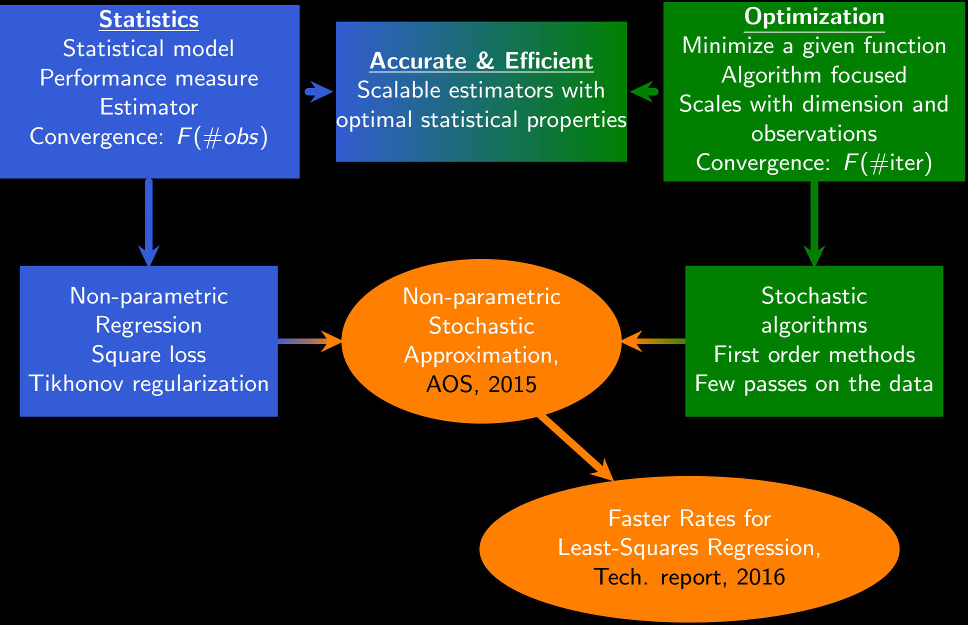
Aymeric Dieuleveut & Francis Bach, *in the Annals of Statistics*, 2015.




**Theorem: Averaged, unregularized, least mean squares algorithm, with large step sizes,** gets **Statistical optimal rate of convergence.**

- Recovers the finite dimension situation with rate $O\left(\frac{\sigma^2 d}{n}\right)$.
- Optimal rates in both the well-specified regime and some situations of the mis-specified.


Statistics
Statistical model
Performance measure
Estimator
Convergence: $F(\#obs)$
Accurate & Efficient
Scalable estimators with
optimal statistical properties
Optimization
Minimize a given function
Algorithm focused
Scales with dimension and
observations
Convergence: $F(\#\text{iter})$
Non-parametric
Regression
Square loss
Tikhonov regularization
Non-parametric
Stochastic
Approximation,
AOS, 2015
Stochastic
algorithms
First order methods
Few passes on the data


Statistics
Statistical model
Performance measure
Estimator
Convergence: $F(\#obs)$
Accurate & Efficient
Scalable estimators with
optimal statistical properties
Optimization
Minimize a given function
Algorithm focused
Scales with dimension and
observations
Convergence: $F(\#\mathrm{iter})$
Non-parametric
Regression
Square loss
Tikhonov regularization
Non-parametric
Stochastic
Approximation,
AOS, 2015
Stochastic
algorithms
First order methods
Few passes on the data
Faster Rates for
Least-Squares Regression,
Tech. report, 2016

## Harder, Better, Faster, Stronger Convergence Rates for Least Squares Regression

Aymeric Dieuleveut, Nicolas Flammarion & Francis Bach, Technical report, 2016.

Classical tradeoff: a Bias term and a Variance term appear.

- The bias is the hardness of forgetting the initial condition.
- The variance is linked with the statistical hardness of the problem.

## Harder, Better, Faster, Stronger Convergence Rates for Least Squares Regression

Aymeric Dieuleveut, Nicolas Flammarion & Francis Bach, Technical report, 2016.

Classical tradeoff: a Bias term and a Variance term appear.

- The bias is the hardness of forgetting the initial condition.
- The variance is linked with the statistical hardness of the problem.

Lower bounds:

- Optimal first order algorithm forgets initial conditions as $\Omega\left(\frac{\|\theta_0-\theta_*\|^2}{t^2}\right)$
- Optimal statistical estimation is $\Omega\left(\frac{\sigma^2 d}{n}\right)$,

## Harder, Better, Faster, Stronger Convergence Rates for Least Squares Regression

Aymeric Dieuleveut, Nicolas Flammarion & Francis Bach, Technical report, 2016.

Classical tradeoff: a Bias term and a Variance term appear.

- The bias is the hardness of forgetting the initial condition.
- The variance is linked with the statistical hardness of the problem.

Lower bounds:

- Optimal first order algorithm forgets initial conditions as $\Omega\left(\frac{\|\theta_0-\theta_*\|^2}{t^2}\right)$
- Optimal statistical estimation is $\Omega\left(\frac{\sigma^2 d}{n}\right)$,
- Single pass over the data: $t = n$.

## Harder, Better, Faster, Stronger Convergence Rates for Least Squares Regression

Aymeric Dieuleveut, Nicolas Flammarion & Francis Bach, Technical report, 2016.

Classical tradeoff: a Bias term and a Variance term appear.

- The bias is the hardness of forgetting the initial condition.
- The variance is linked with the statistical hardness of the problem.

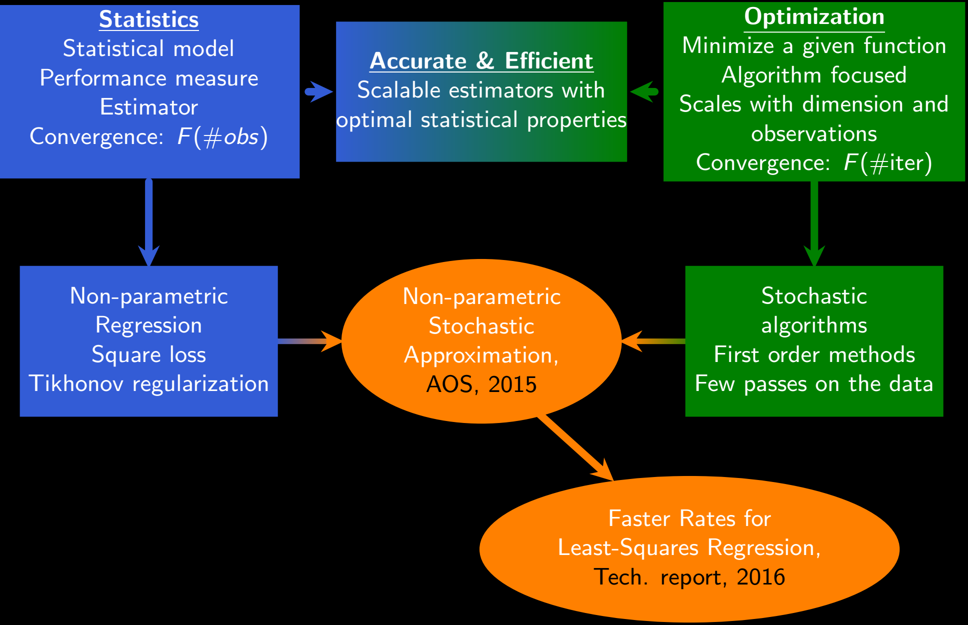
Lower bounds:

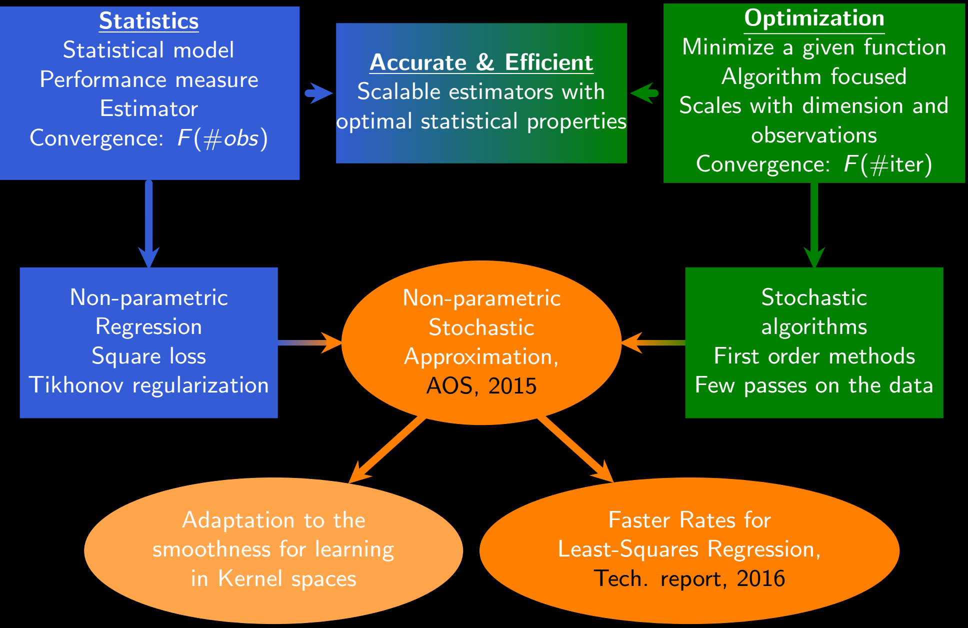
- Optimal first order algorithm forgets initial conditions as $\Omega\left(\frac{\|\theta_0-\theta_*\|^2}{t^2}\right)$
- Optimal statistical estimation is $\Omega\left(\frac{\sigma^2 d}{n}\right)$,
- Single pass over the data: $t = n$.

New algorithm, based on Nesterov acceleration:

↬ Both optimal terms: $\mathbb{E}\left[\varepsilon(\bar{\theta}_n) - \varepsilon(\bar{\theta}_*)\right] \leq \frac{L\|\theta_0-\theta_*\|^2}{n^2} + \frac{\sigma^2 d}{n}$.

↬ Improves convergence rate for mis-specified non-parametric regression.


Statistics
Statistical model
Performance measure
Estimator
Convergence: $F(\#obs)$
Accurate & Efficient
Scalable estimators with
optimal statistical properties
Optimization
Minimize a given function
Algorithm focused
Scales with dimension and
observations
Convergence: $F(\#\text{iter})$
Non-parametric
Regression
Square loss
Tikhonov regularization
Non-parametric
Stochastic
Approximation,
AOS, 2015
Stochastic
algorithms
First order methods
Few passes on the data
Faster Rates for
Least-Squares Regression,
Tech. report, 2016


Statistics
Statistical model
Performance measure
Estimator
Convergence: $F(\#obs)$
Accurate & Efficient
Scalable estimators with
optimal statistical properties
Optimization
Minimize a given function
Algorithm focused
Scales with dimension and
observations
Convergence: $F(\#\text{iter})$
Non-parametric
Regression
Square loss
Tikhonov regularization
Non-parametric
Stochastic
Approximation,
AOS, 2015
Stochastic
algorithms
First order methods
Few passes on the data
Adaptation to the
smoothness for learning
in Kernel spaces
Faster Rates for
Least-Squares Regression,
Tech. report, 2016

**Statistics**
Statistical model
Performance measure
Estimator
Convergence: $F(\#obs)$

**Accurate & Efficient**
Scalable estimators with
optimal statistical properties

**Optimization**
Minimize a given function
Algorithm focused
Scales with dimension and
observations
Convergence: $F(\#\text{iter})$

**Statistics**

Statistical model
Performance measure
Estimator
Convergence: $F(\#obs)$

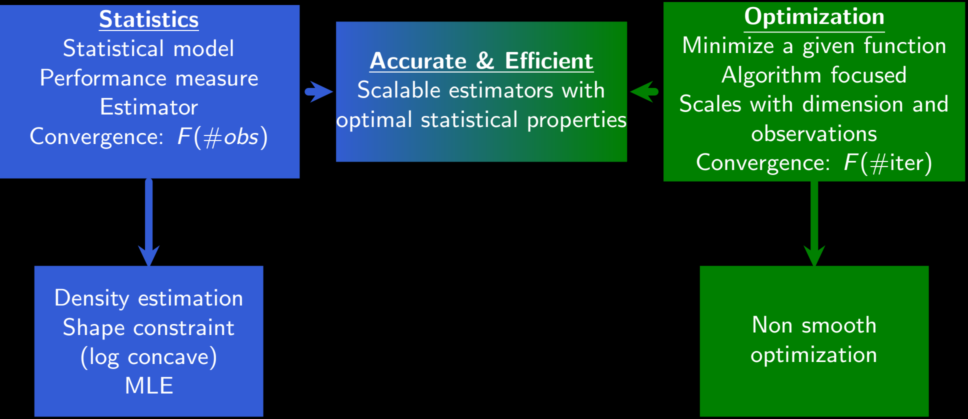
**Accurate & Efficient**

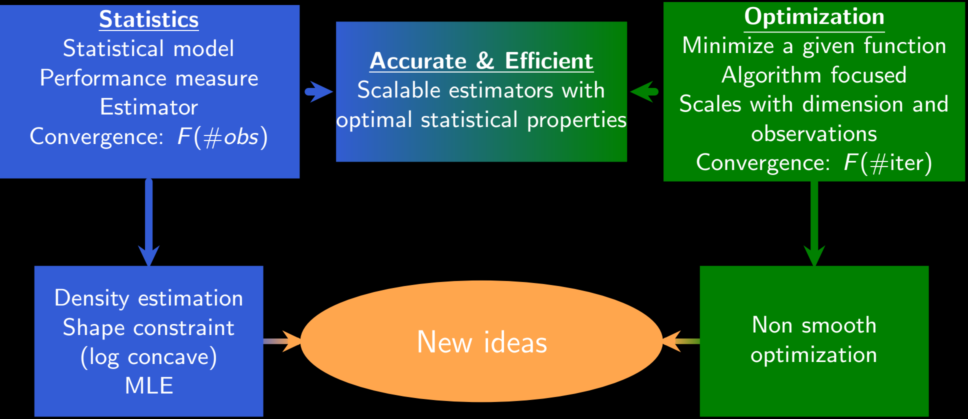
Scalable estimators with
optimal statistical properties

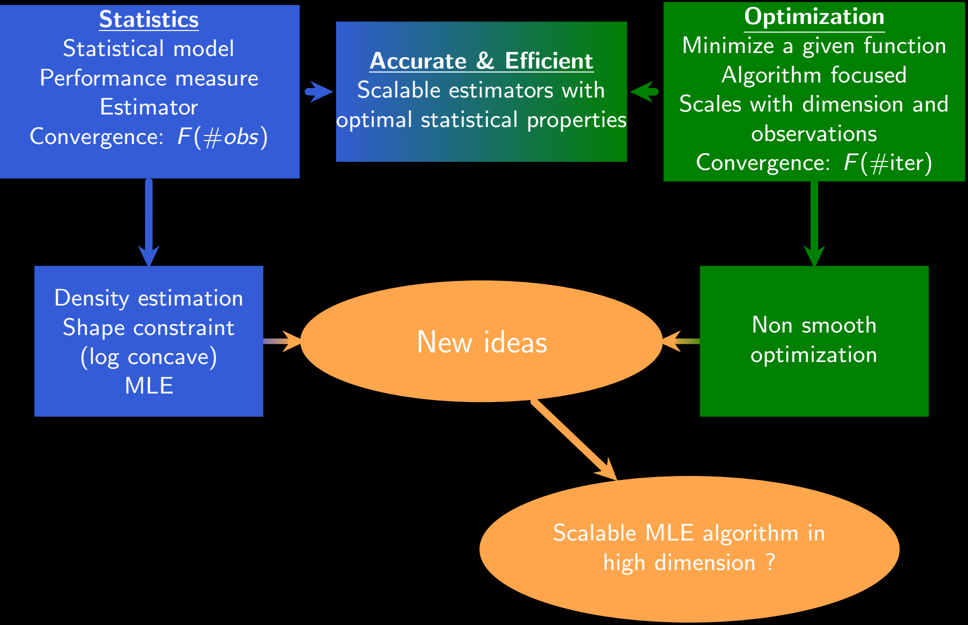
**Optimization**

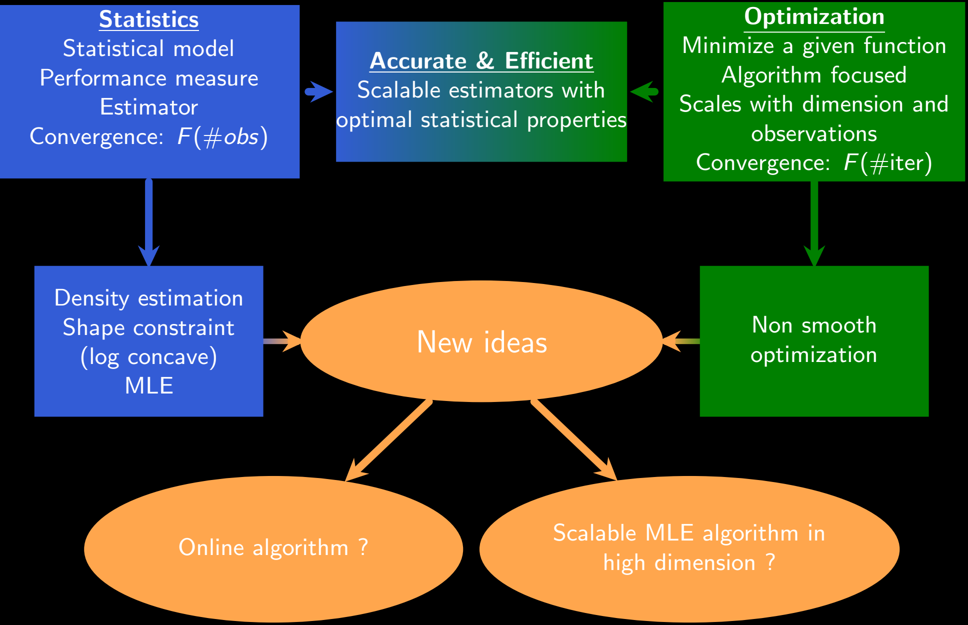
Minimize a given function
Algorithm focused
Scales with dimension and
observations
Convergence: $F(\#\text{iter})$

Density estimation
Shape constraint
(log concave)
MLE


Statistics
Statistical model
Performance measure
Estimator
Convergence: $F(\#obs)$
Accurate & Efficient
Scalable estimators with
optimal statistical properties
Optimization
Minimize a given function
Algorithm focused
Scales with dimension and
observations
Convergence: $F(\#\text{iter})$
Density estimation
Shape constraint
(log concave)
MLE
Non smooth
optimization


Statistics
Statistical model
Performance measure
Estimator
Convergence: $F(\#obs)$
Accurate & Efficient
Scalable estimators with
optimal statistical properties
Optimization
Minimize a given function
Algorithm focused
Scales with dimension and
observations
Convergence: $F(\#\text{iter})$
Density estimation
Shape constraint
(log concave)
MLE
New ideas
Non smooth
optimization


Statistics
Statistical model
Performance measure
Estimator
Convergence: $F(\#obs)$
Accurate & Efficient
Scalable estimators with
optimal statistical properties
Optimization
Minimize a given function
Algorithm focused
Scales with dimension and
observations
Convergence: $F(\#\text{iter})$
Density estimation
Shape constraint
(log concave)
MLE
New ideas
Non smooth
optimization
Scalable MLE algorithm in
high dimension ?


Statistics
Statistical model
Performance measure
Estimator
Convergence: $F(\#obs)$
Accurate & Efficient
Scalable estimators with
optimal statistical properties
Optimization
Minimize a given function
Algorithm focused
Scales with dimension and
observations
Convergence: $F(\#\text{iter})$
Density estimation
Shape constraint
(log concave)
MLE
New ideas
Non smooth
optimization
Online algorithm ?
Scalable MLE algorithm in
high dimension ?