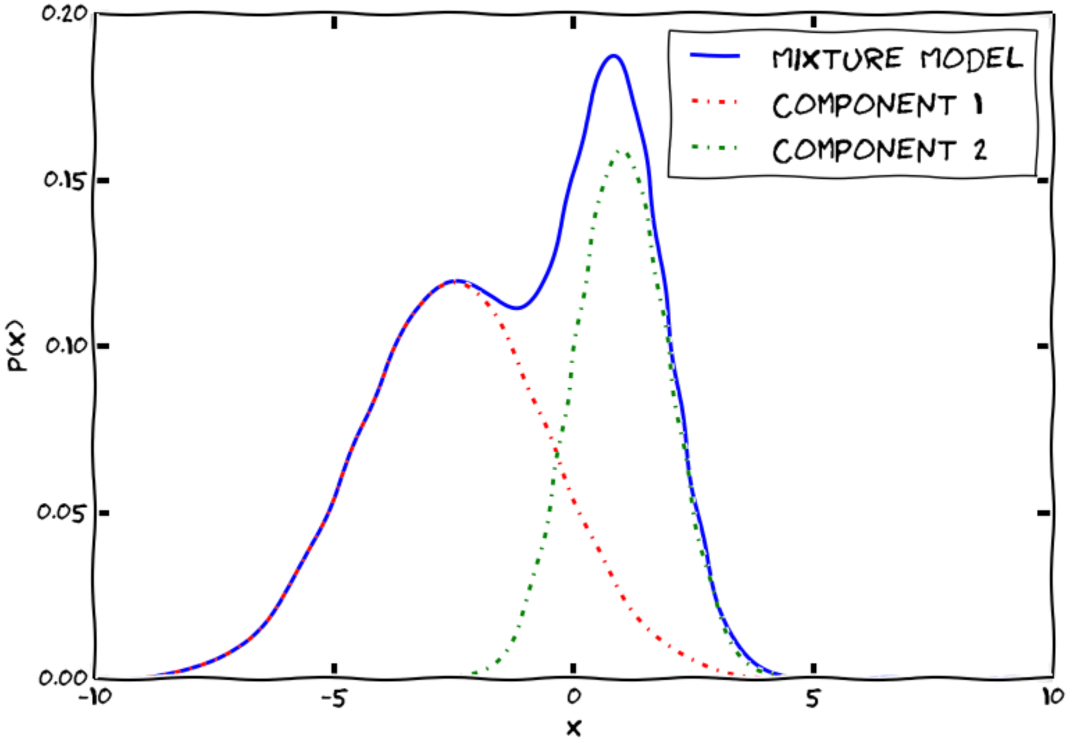
Faster and Sample Near-Optimal Algorithms for Proper Learning Mixtures of Gaussians

Constantinos Daskalakis, MIT

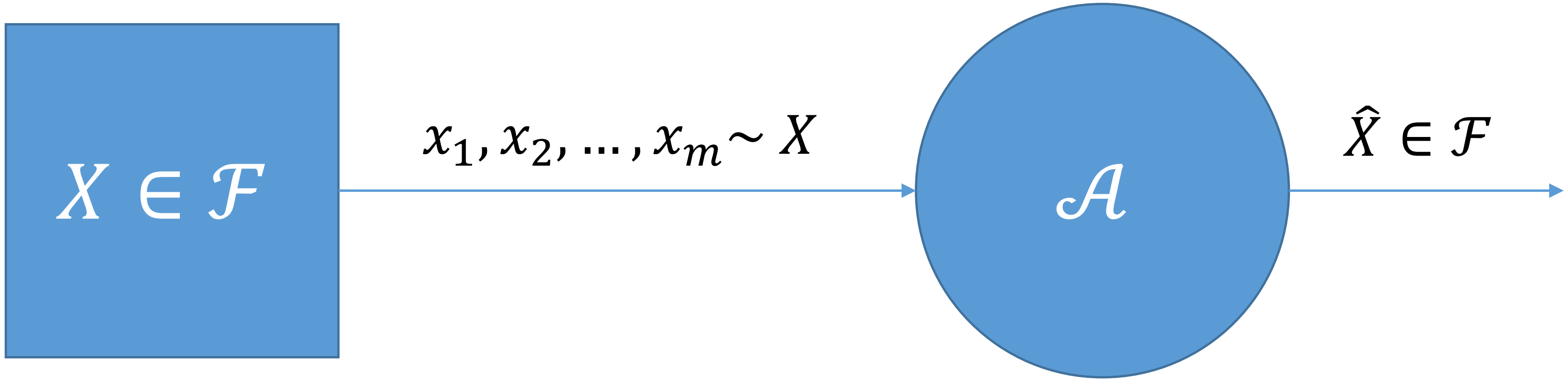
Gautam Kamath, MIT

What's a Gaussian Mixture Model (GMM)?

- Interpretation 1: PDF is a convex combination of Gaussian PDFs
  - $p(x) = \Sigma_i w_i \mathcal{N}\left(\mu_i, \sigma_i^2, x\right)$
- Interpretation 2: Several unlabeled Gaussian populations, mixed together
- Focus on mixtures of 2 Gaussians (2-GMM) in one dimension



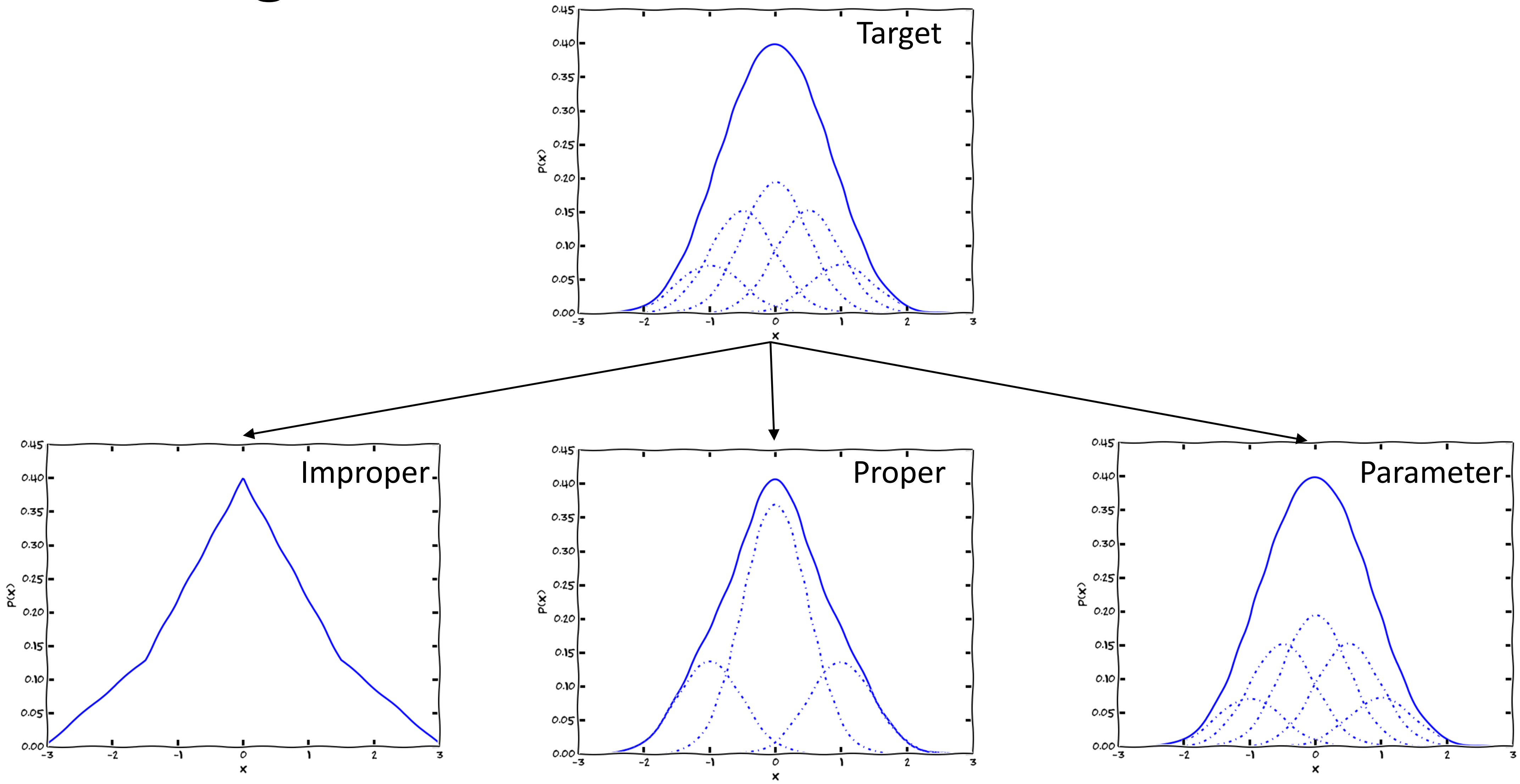
PAC (Proper) Learning Model



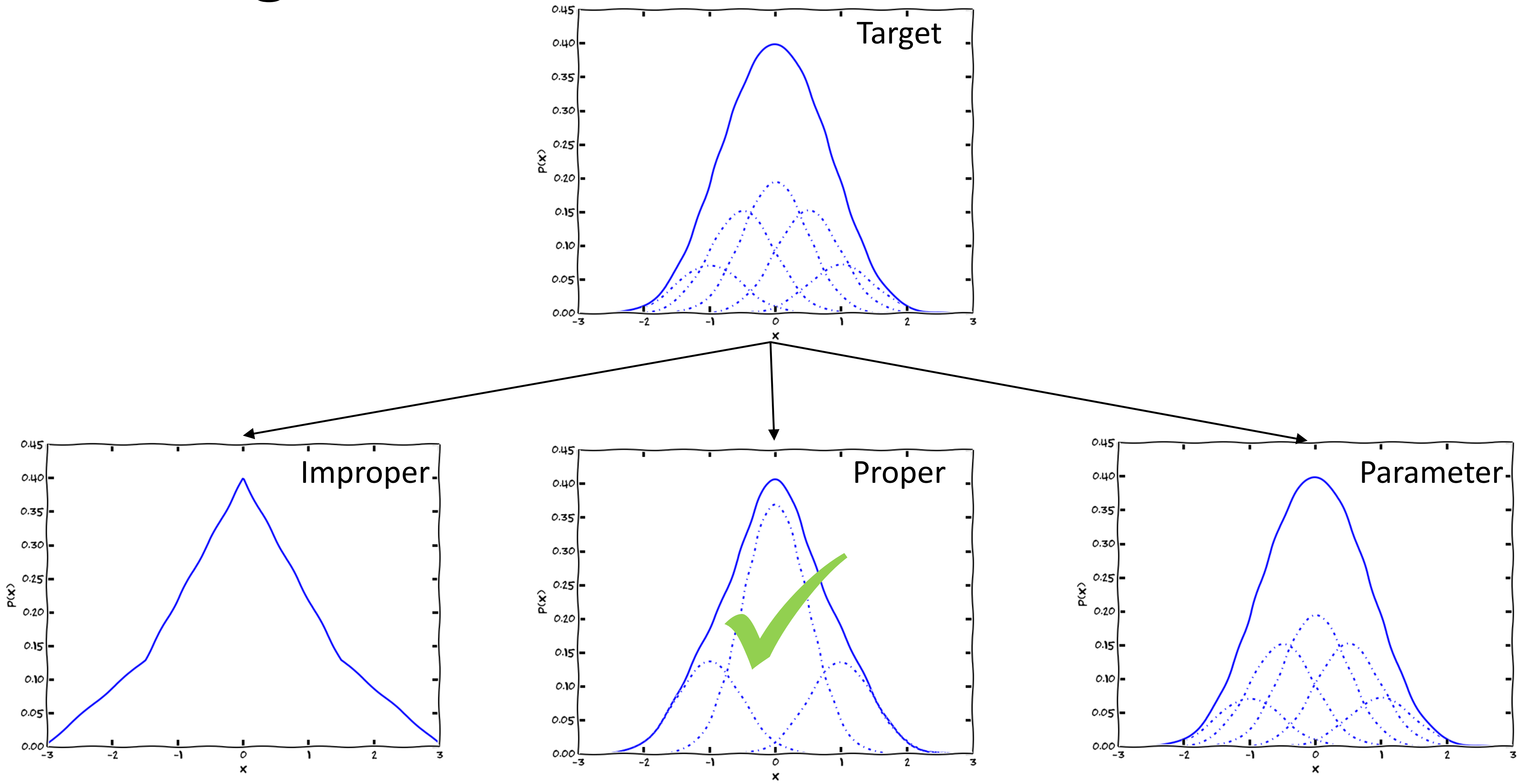
- Output (with high probability) a 2-GMM $\hat{X}$ which is close to X (in statistical distance*)
- Algorithm design goals:
  - Minimize sample size
  - Minimize time

*statistical distance = total variation distance = ½ × L^1 distance

Learning Goals



Learning Goals



Prior Work

Learning Model	Sample Complexity	Time Complexity
Improper Learning		
Proper Learning		
Parameter Estimation	$poly(1/\varepsilon)$	$poly(1/\varepsilon)$ [KMV10]

Prior Work

Learning Model	Sample Complexity	Time Complexity
Improper Learning	$\tilde{O}(1/\varepsilon^2)$	$poly(1/\varepsilon)$ [CDSS14]
Proper Learning		
Parameter Estimation	$poly(1/\varepsilon)$	$poly(1/\varepsilon)$ [KMV10]

Prior Work

Learning Model	Sample Complexity	Time Complexity
Improper Learning	$\tilde{O}(1/\varepsilon^2)$	$poly(1/\varepsilon)$ [CDSS14]
Proper Learning	$\tilde{O}(1/\varepsilon^2)$ $\boldsymbol{\tilde{O}(1/\varepsilon^2)}$	$\tilde{O}(1/\varepsilon^7)$ [AJOS14] $\boldsymbol{\tilde{O}(1/\varepsilon^5)}$ **[DK14]**
Parameter Estimation	$poly(1/\varepsilon)$	$poly(1/\varepsilon)$ [KMV10]

Prior Work

Learning Model	Sample Complexity	Time Complexity
Improper Learning	$\tilde{O}(1/\varepsilon^2)$	$poly(1/\varepsilon)$ [CDSS14]
Proper Learning	$\tilde{O}(1/\varepsilon^2)$ $\boldsymbol{\tilde{O}(1/\varepsilon^2)}$	$\tilde{O}(1/\varepsilon^7)$ [AJOS14] $\boldsymbol{\tilde{O}(1/\varepsilon^5)}$ **[DK14]**
Parameter Estimation	$poly(1/\varepsilon)$	$poly(1/\varepsilon)$ [KMV10]

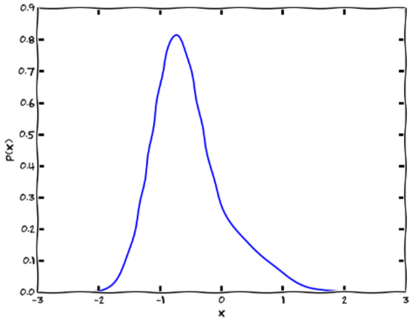
- Sample lower bounds:
  - Improper and Proper Learning: $\Omega(1/\varepsilon^2)$ [folklore]
  - Parameter Estimation: $\Omega(1/\varepsilon^6)$ [HP14]
    - Matching upper bound, but not immediately extendable to proper learning

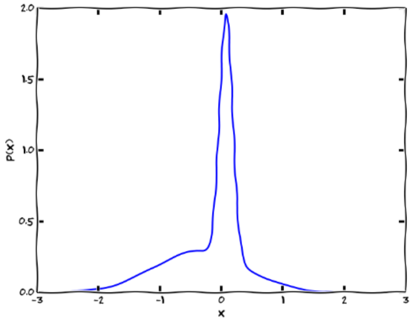
The Plan

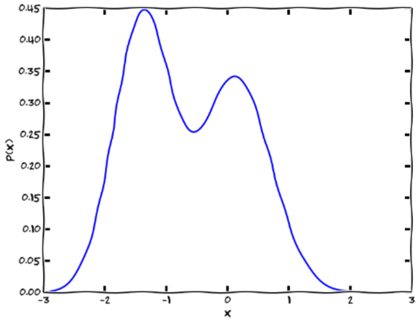
1. Generate a set of hypothesis GMMs
2. Pick a good candidate from the set

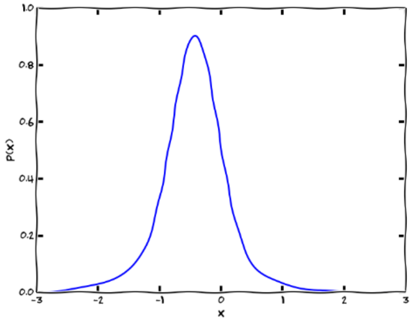
The Plan

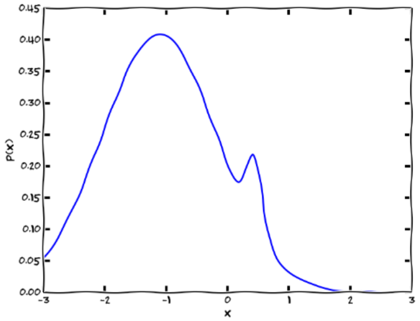
1. Generate a set of hypothesis GMMs
2. Pick a good candidate from the set





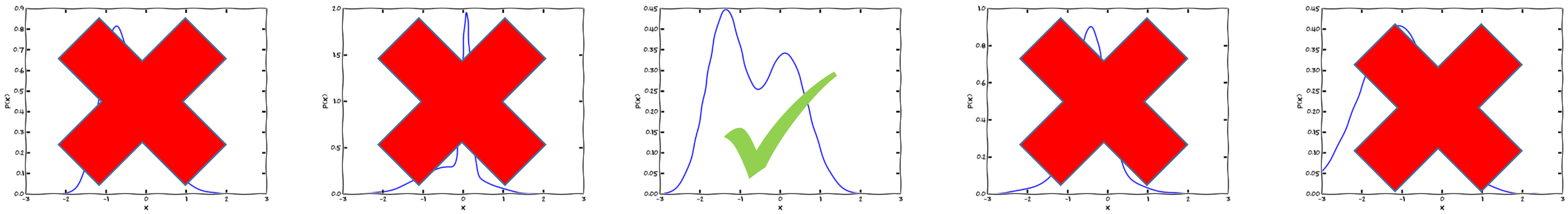






The Plan

1. Generate a set of hypothesis GMMs
2. Pick a good candidate from the set



Some Tools Along the Way

a) How to remove part of a distribution which we already know
b) How to robustly estimate parameters of a distribution
c) How to pick a good hypothesis from a pool of hypotheses

The Plan

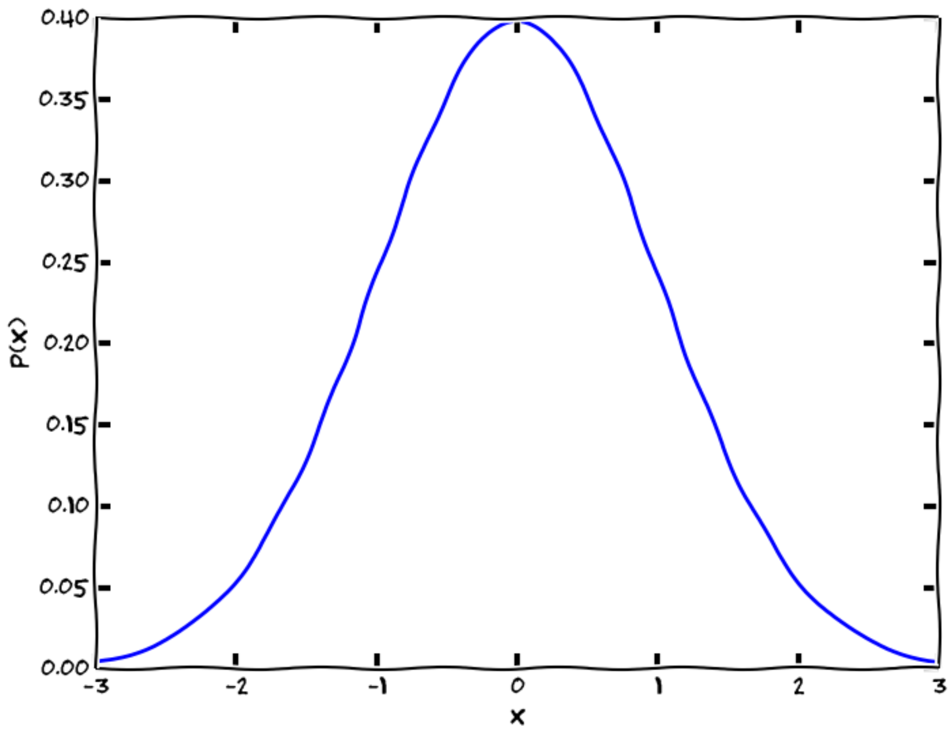
1. **Generate a set of hypothesis GMMs**
2. Pick a good candidate from the set

Who Do We Want In Our Pool?

- Hypothesis: $(\widehat{w}, \widehat{\mu_1}, \widehat{\sigma_1}, \widehat{\mu_2}, \widehat{\sigma_2})$
- Need at least one “good” hypothesis
- Parameters are close to true parameters
  - Implies desired statistical distance bound
- Want:
  - $|w \ - \widehat{w}| \leq \varepsilon$
  - $|\mu_i \ - \widehat{\mu_i}| \leq \varepsilon\sigma_i$
  - $|\sigma_i \ - \widehat{\sigma_i}| \leq \varepsilon\sigma_i$

Warm Up: Learning one Gaussian

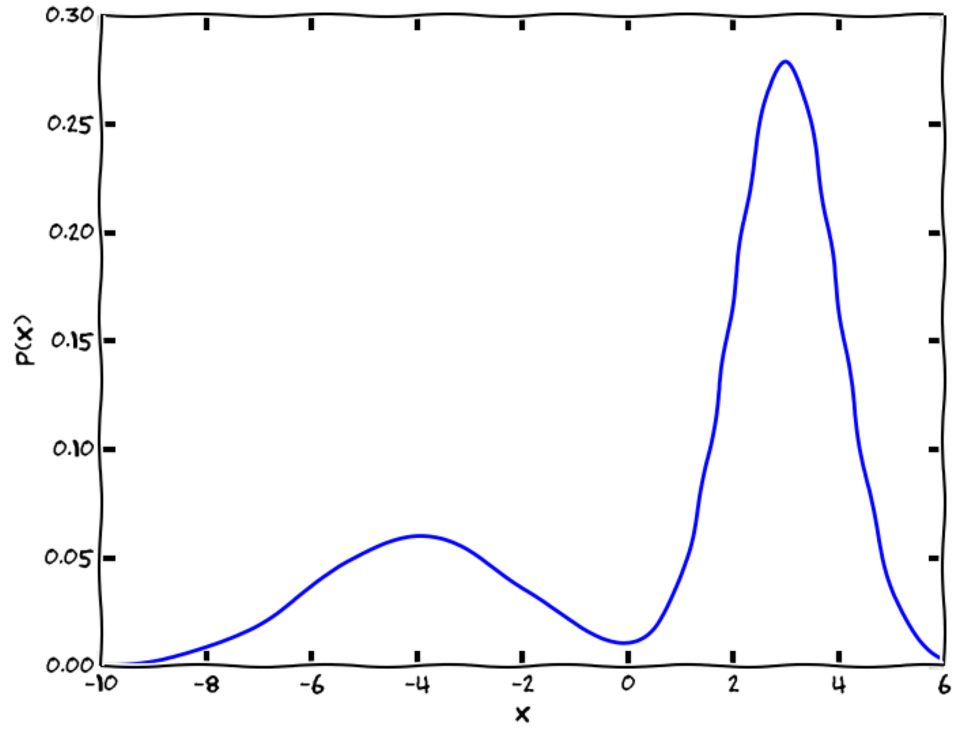
- Easy!
- $\hat{\mu}$ = sample mean
- $\widehat{\sigma^2}$ = sample variance



The Real Deal: Mixtures of Two Gaussians

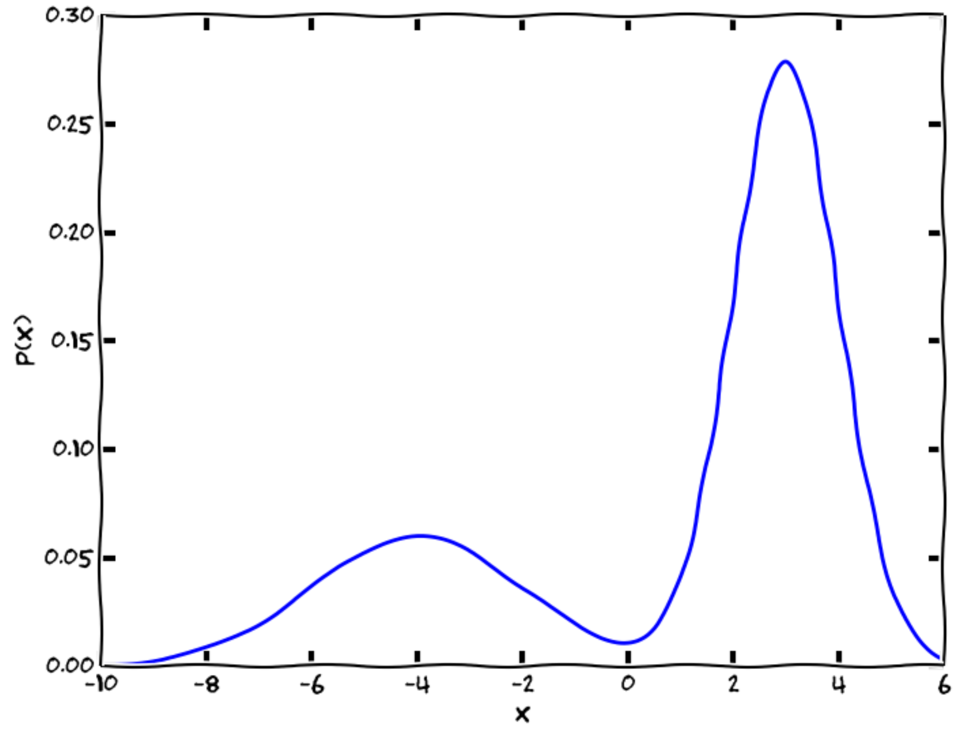
- Harder!
- Sample moments mix up samples from each component
- Plan:
  - Tall, skinny Gaussian* stands out – learn it first
  - Remove it from the mixture
  - Learn one Gaussian (easy?)

*Component with maximum $\frac{w_i}{\sigma_i}$



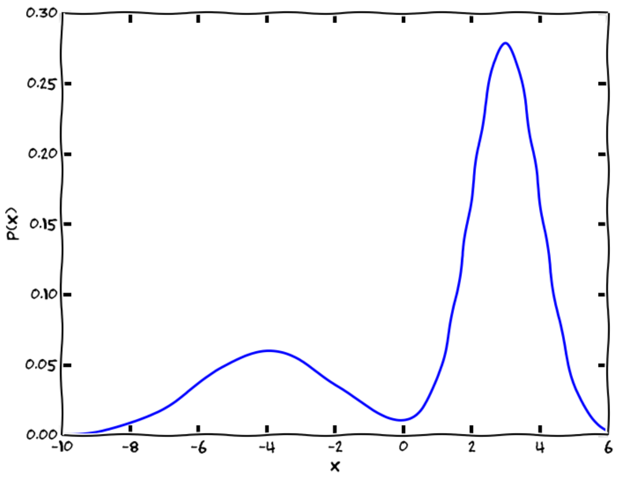
The First Component

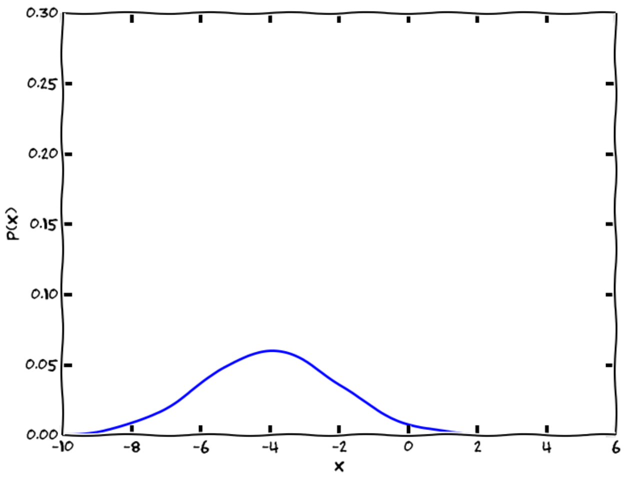
- Claim: Using $O\left(\frac{1}{\varepsilon^2}\right)$ sample size, can generate $\tilde{O}\left(\frac{1}{\varepsilon^3}\right)$ candidates $(\widehat{w}, \widehat{\mu_1}, \widehat{\sigma_1})$, at least one is close to the taller component
- If we knew which candidate was right, could we remove this component?



Some Tools Along the Way

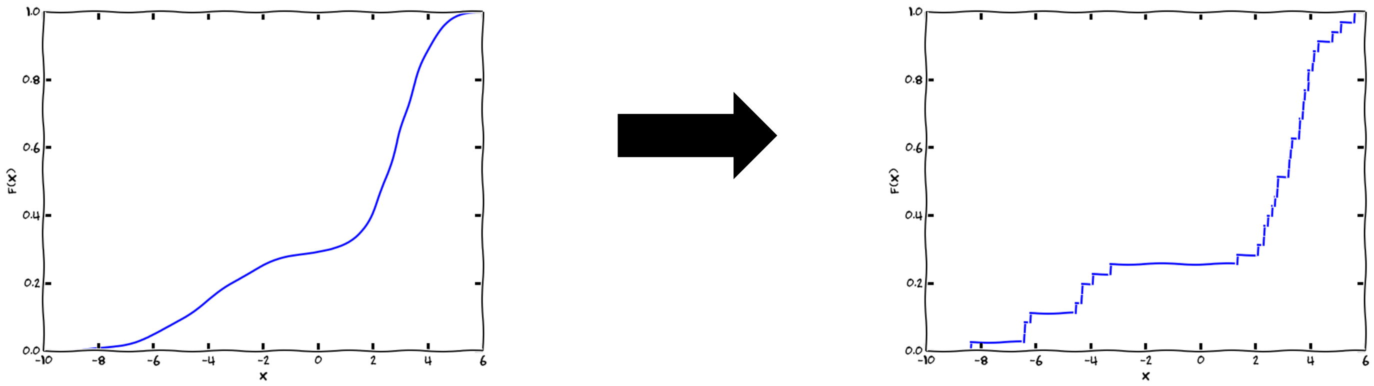
a) **How to remove part of a distribution which we already know**
b) How to robustly estimate parameters of a distribution
c) How to pick a good hypothesis from a pool of hypotheses



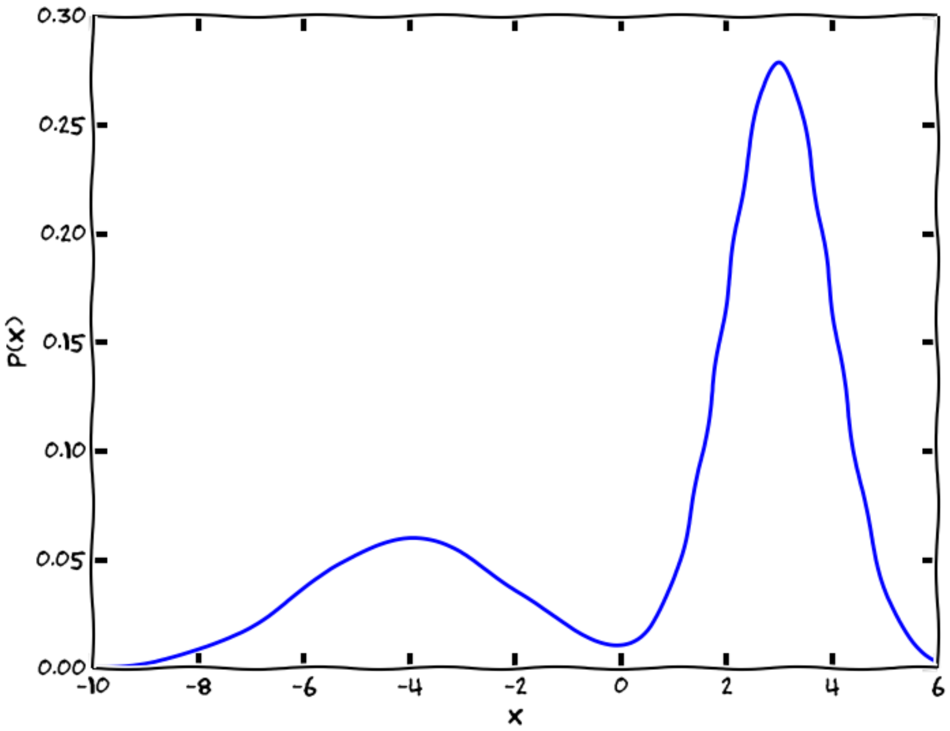


Dvoretzky–Kiefer–Wolfowitz (DKW) inequality

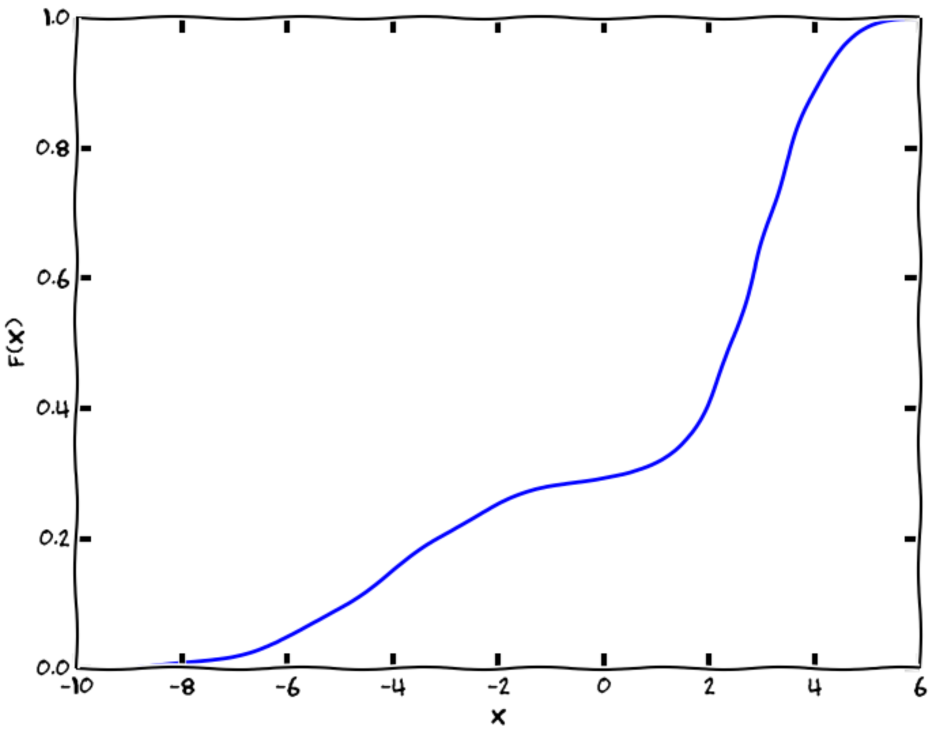
- Using sample of size $O\left(\frac{1}{\varepsilon^2}\right)$ from a distribution X, can output $\hat{X}$ such that $d_K\left(X, \hat{X}\right) \leq \varepsilon$
- Kolmogorov distance – CDFs of distributions are close in L^∞ distance
  - Weaker than statistical distance
- Works for *any* probability distribution!



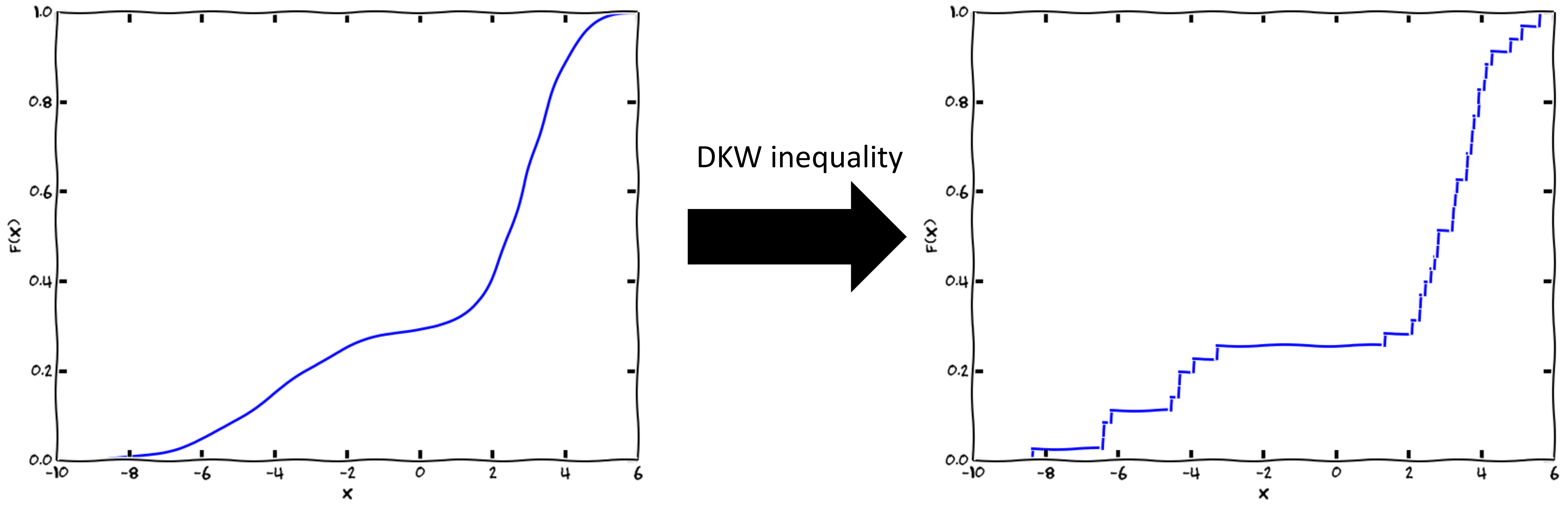
Subtracting out the known component



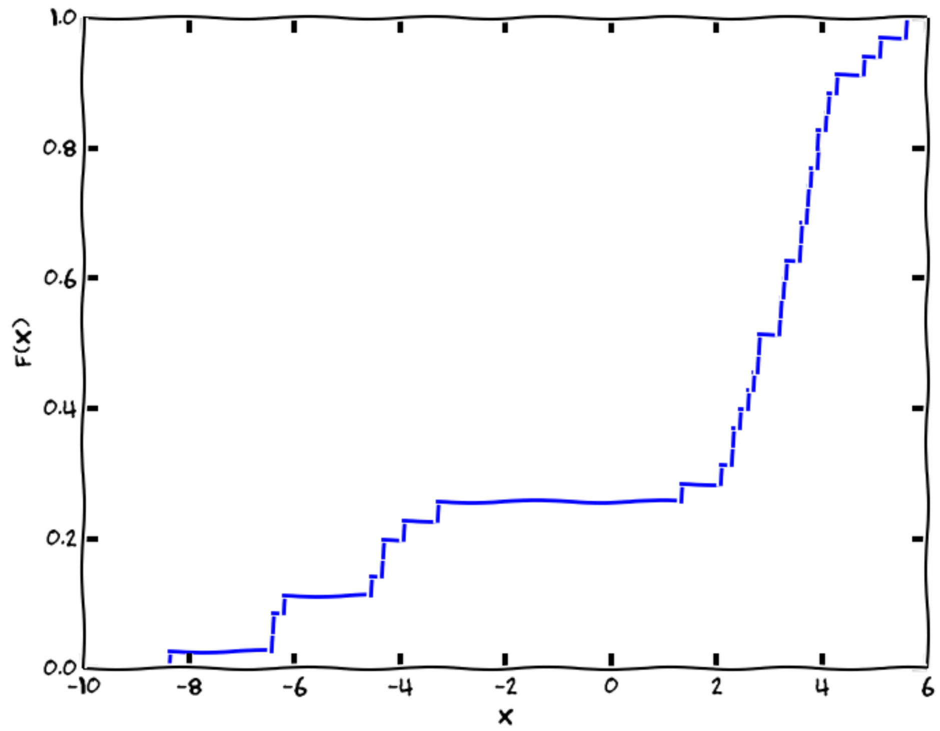
PDF to CDF



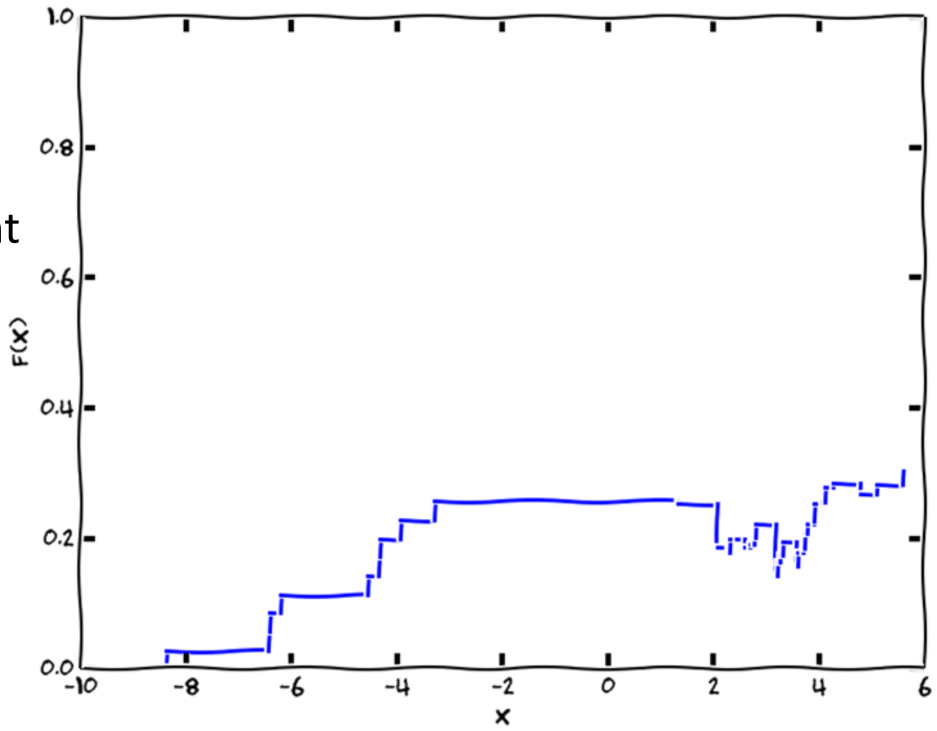
Subtracting out the known component



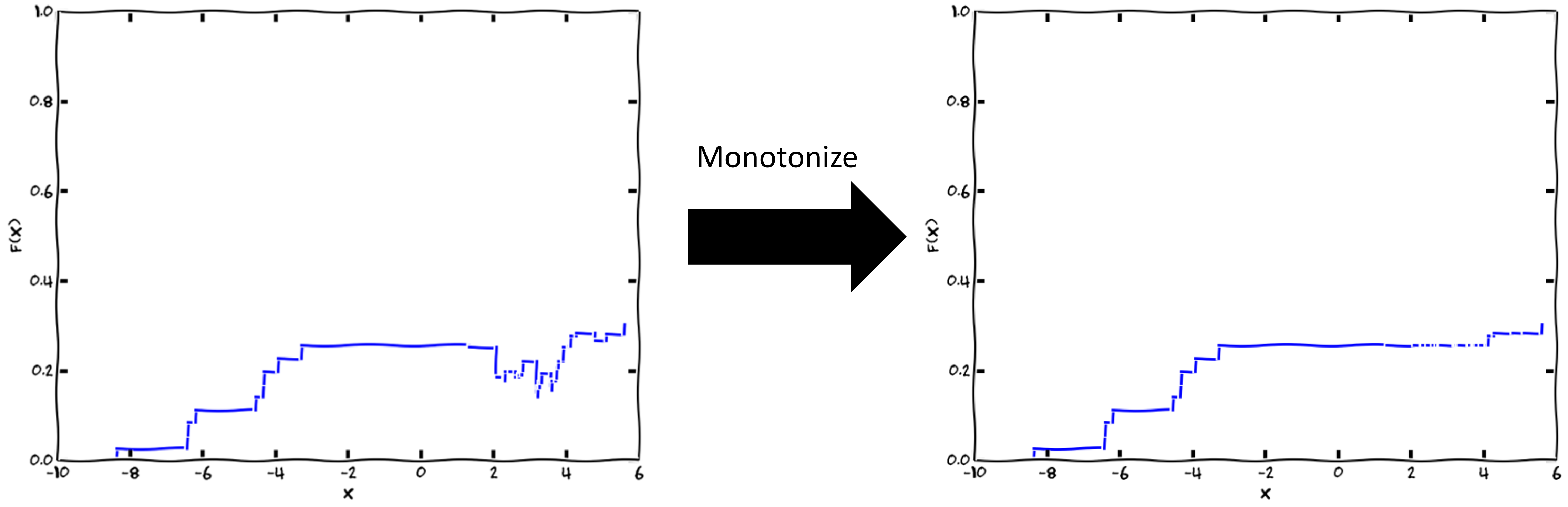
Subtracting out the known component



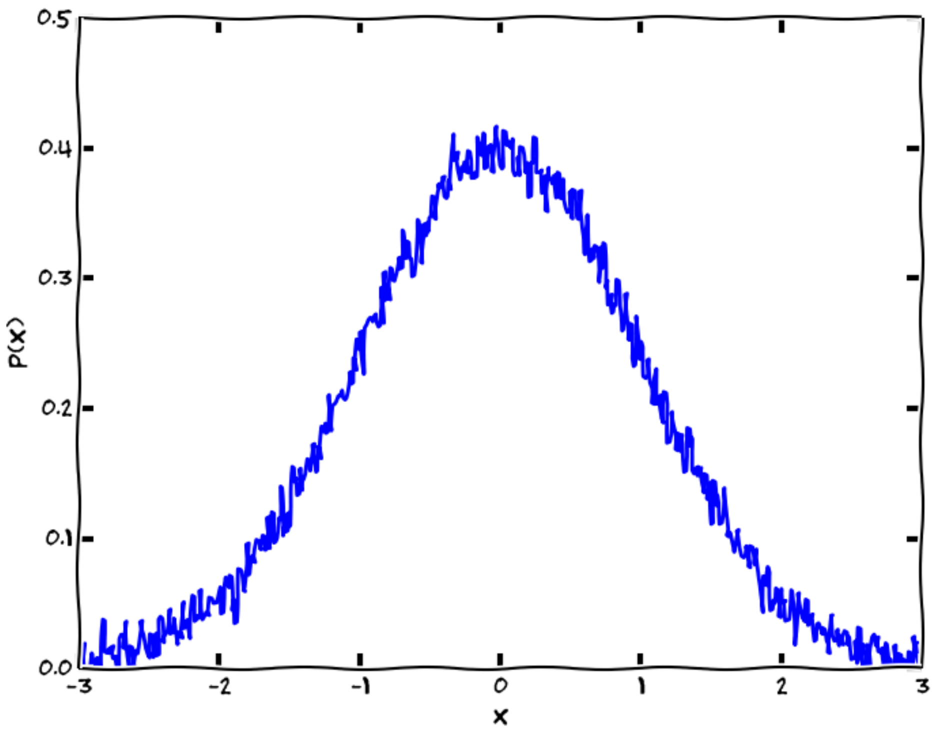
Subtract component



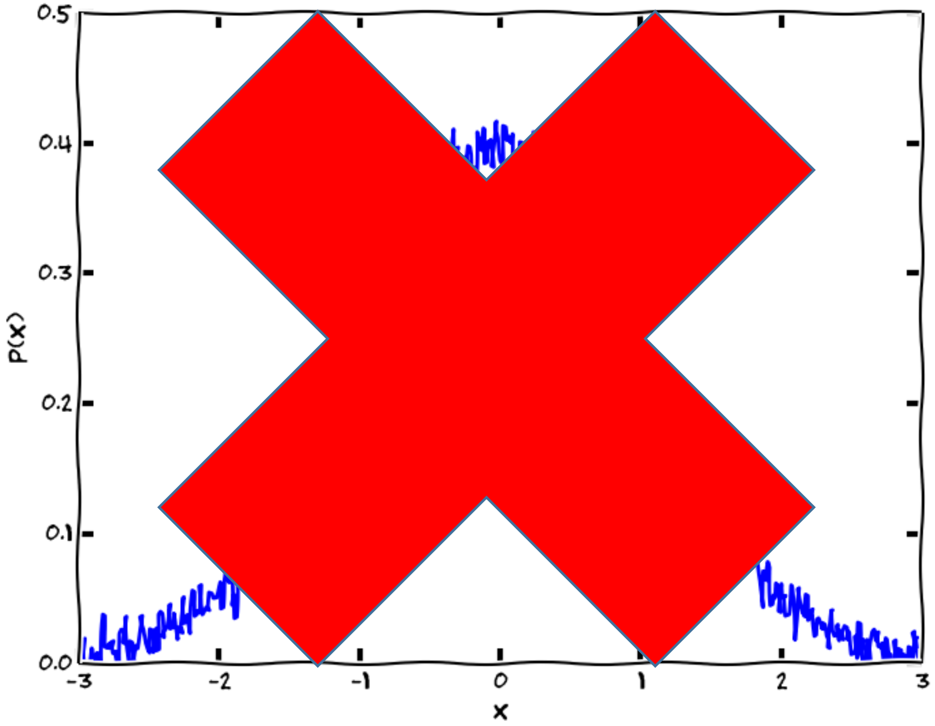
Subtracting out the known component



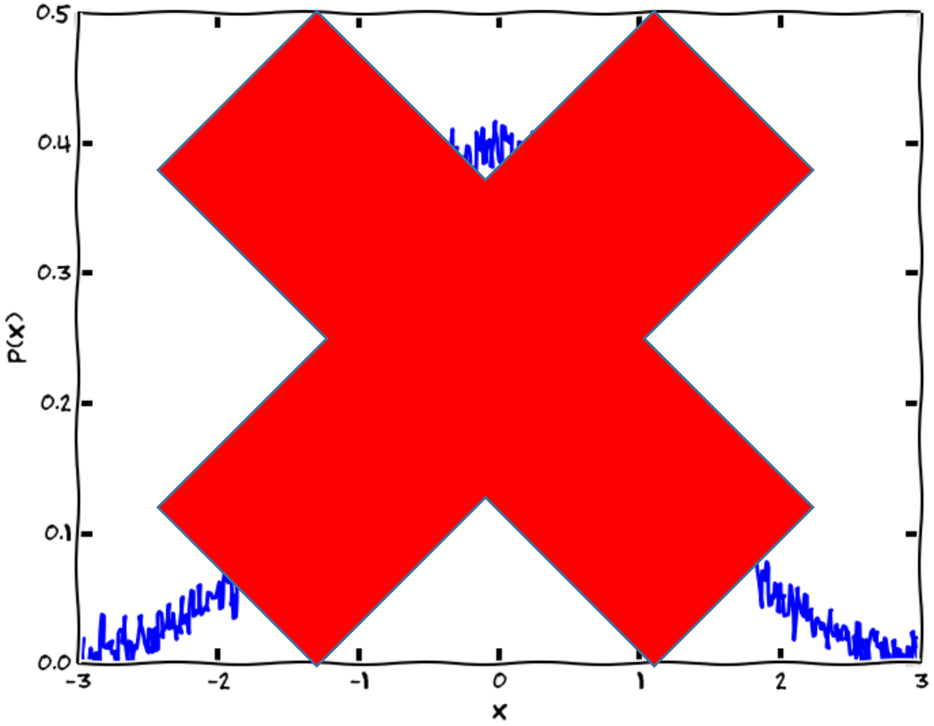
Warm Up (?): Learning one (almost) Gaussian

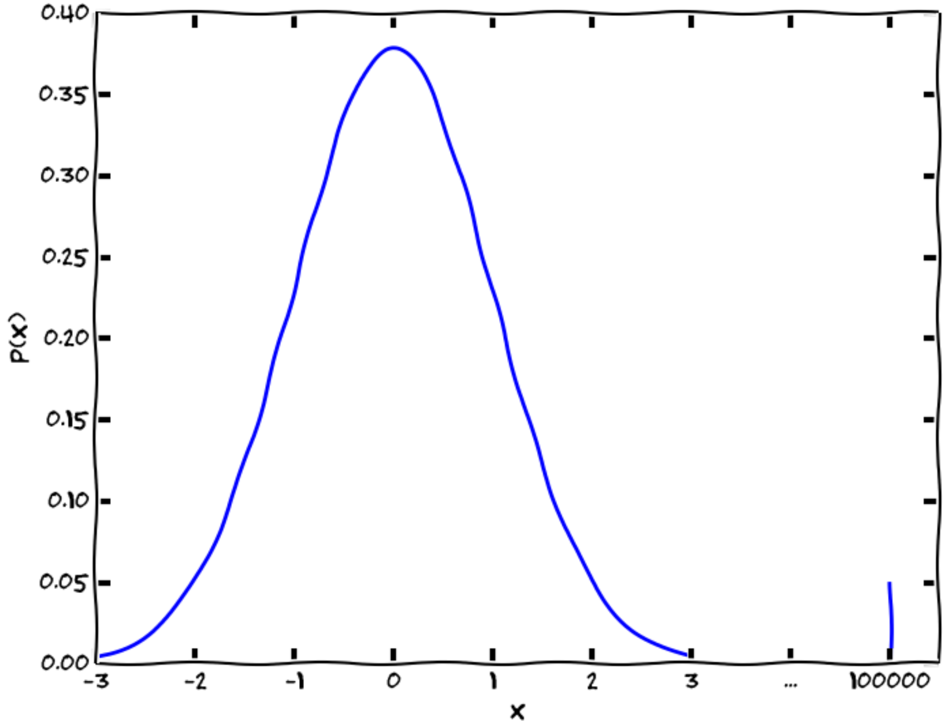


Warm Up (?): Learning one (almost) Gaussian



Warm Up (?): Learning one (almost) Gaussian



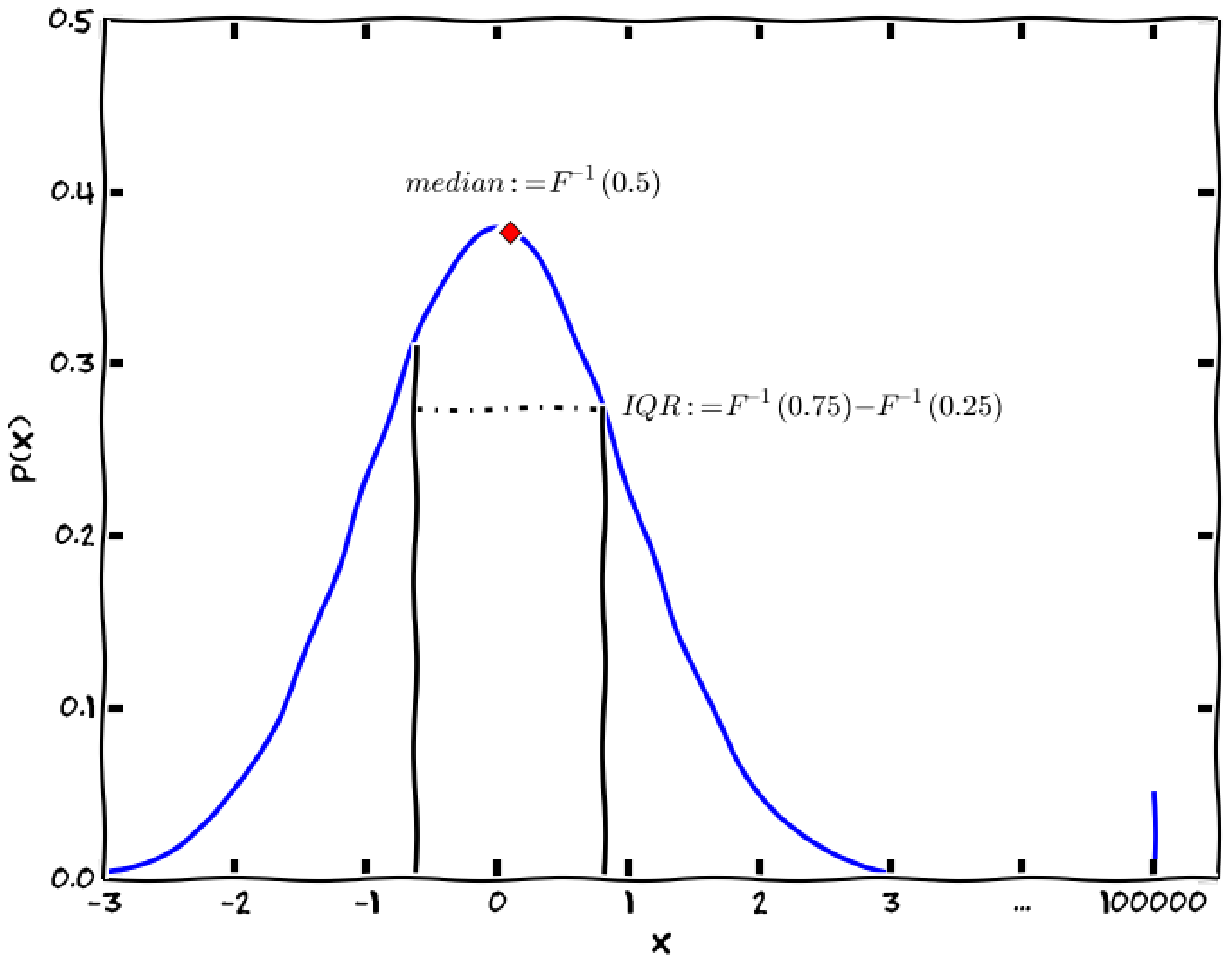


Some Tools Along the Way

a) How to remove part of a distribution which we already know
b) How to robustly estimate parameters of a distribution
c) How to pick a good hypothesis from a pool of hypotheses

Robust Statistics

- Broad field of study in statistics
- Median
- Interquartile range
- Recover original parameters (approximately), even for distributions at distance ε
- Entirely determined by the other component
  - Still $\tilde{O}\left(\frac{1}{\varepsilon^3}\right)$ candidates!



The Plan

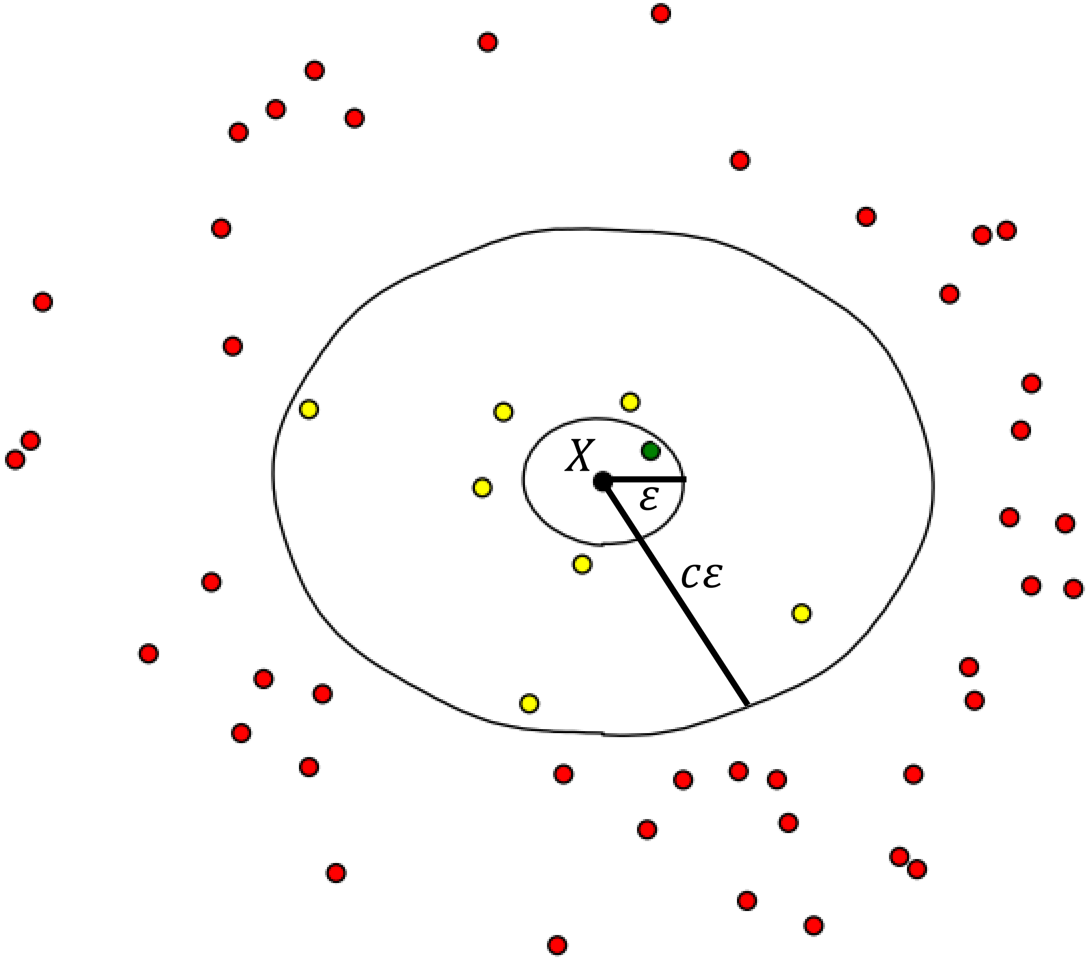
1. Generate a set of hypothesis GMMs
2. **Pick a good candidate from the set**

Some Tools Along the Way

a) How to remove part of a distribution which we already know
b) How to robustly estimate parameters of a distribution
c) How to pick a good hypothesis from a pool of hypotheses

Hypothesis Selection

- N candidate distributions
- At least one is ε-close to X (in statistical distance)
- Goal: Return candidate which is $O(\varepsilon)$-close to X



Hypothesis Selection

- Classical approaches [Yat85], [DL01]
  - Scheffé estimator, computation of the “Scheffé set” (potentially hard)
  - $O(N^2)$ time
- Acharya et al. [AJOS14]
  - Based on Scheffé estimator
  - $O(N \log N)$ time
- Our result [DK14]
  - General estimator, minimal access to hypotheses
  - $O(N \log N)$ time
  - Milder dependence on error probability

Hypothesis Selection

- Input:
  - Sample access to X and hypotheses $\mathcal{H} = \{H_1, \ldots, H_N\}$
  - PDF comparator for each pair H_i, H_j
  - Accuracy parameter ε , confidence parameter δ
- Output:
  - $H \in \mathcal{H}$
  - If there is a $H^* \in \mathcal{H}$ such that $d_{stat}(H^*, X) \leq \varepsilon$, then $d_{stat}(H, X) \leq O(\varepsilon)$ with probability $\geq 1 - \delta$
- Sample complexity: $O\left(\frac{\log 1/\delta}{\varepsilon^2} \log N\right)$
- Time complexity: $O\left(\frac{\log 1/\delta}{\varepsilon^2}\left(N \log N + \log^2 \frac{1}{\delta}\right)\right)$
- Expected time complexity: $O\left(\frac{N \log N/\delta}{\varepsilon^2}\right)$

Hypothesis Selection

- Naive: Tournament among candidate hypotheses; compare every pair; output hypothesis with most wins
- Us: Set up a single-elimination tournament
  - Issue: error doubles at every level of tree; $\log N$ levels $\rightarrow \Omega(2^{\log N}\,\varepsilon)$ error
  - Better analysis via double-window argument:
    - Great hypotheses: those within ε of target
    - Good hypotheses: those within 8ε of target
    - Bad hypotheses: the rest
  - Show: if density of good hypotheses small, error propagation won't happen
  - If density large, sub-sample $\sqrt{N}$ hypotheses; run naive tournament

Putting It All Together

- $N = \tilde{O}\left(\frac{1}{\varepsilon^3}\right)$ candidates
- Use hypothesis selection algorithm to pick one
- Sample complexity: $\tilde{O}(\log(1/\delta)\ /\varepsilon^2)$
- Time complexity: $\tilde{O}(\log^3(1/\delta)\ /\varepsilon^5)$

Open Questions

- Faster algorithms for 2-GMMs
- Time complexity of k-GMMs
- High dimensions

Bibliography

- [AJOS14] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theerta Suresh. Near-optimal-sample estimators for spherical Gaussian mixtures.
- [CDSS14] Siu On Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Efficient Density Estimation via Piecewise Polynomial Approximation.
- [DL01] Luc Devroye and Gabor Lugosi. Combinatorial Methods in Density Estimation.
- [HP14] Moritz Hardt and Eric Price. Sharp bounds for learning a mixture of two Gaussians.
- [KMV10] Adam Kalai, Ankur Moitra, Gregory Valiant. Efficiently Learning Mixtures of Two Gaussians.
- [Yat85] Yannis Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy.