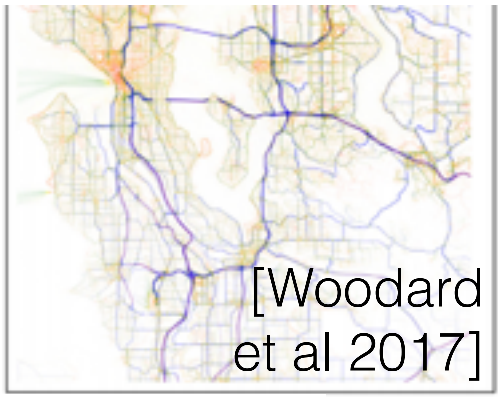
Automated Scalable Bayesian Inference via Data Summarization

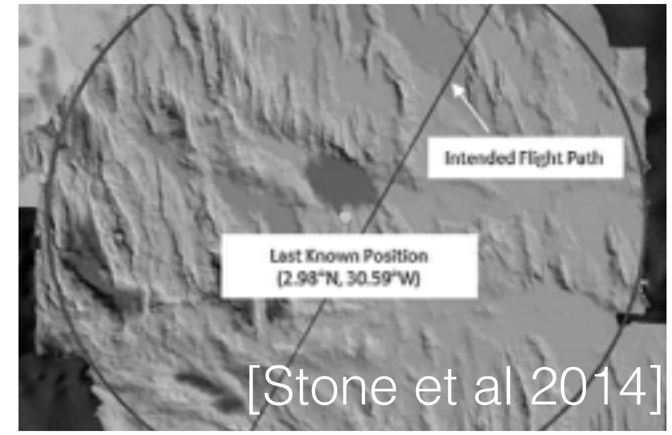
Tamara Broderick
ITT Career Development
Assistant Professor,
MIT

With: Trevor Campbell, Jonathan Huggins

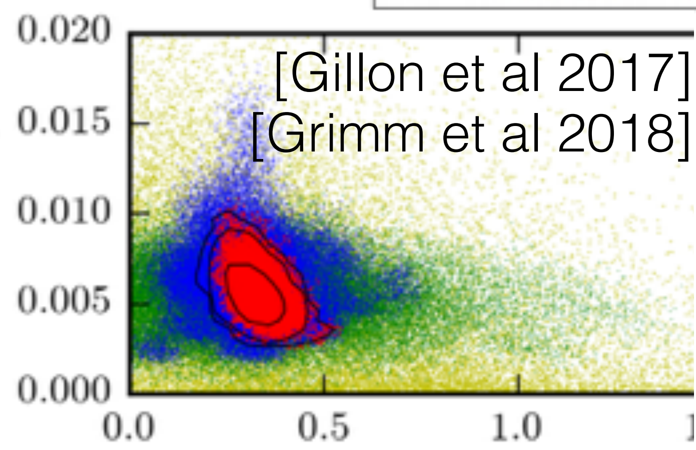
http://www.tamarabroderick.com/tutorials.html

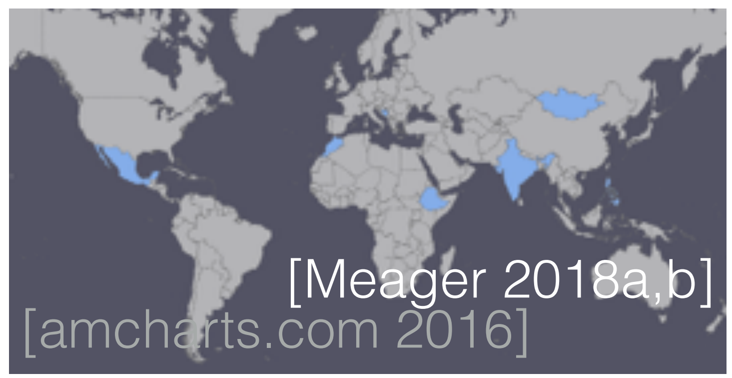
Bayesian inference





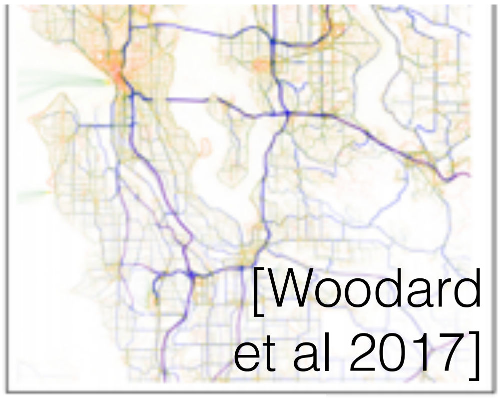
[mc-stan.org]

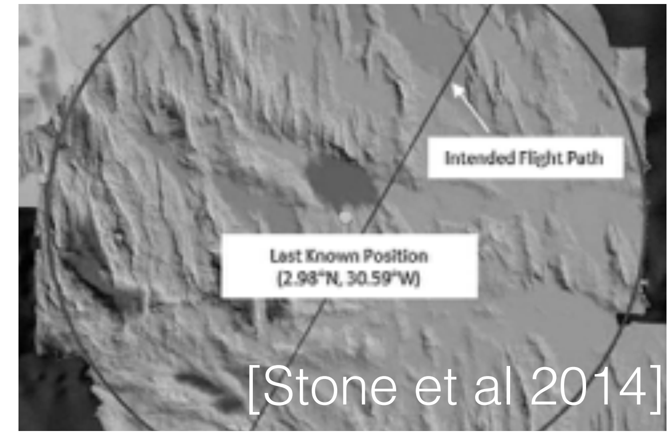




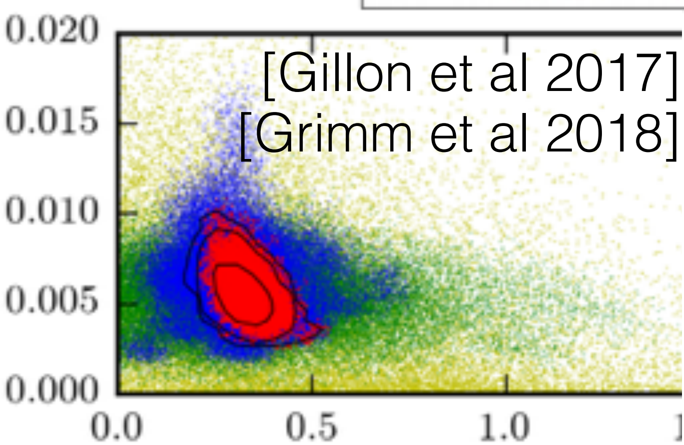
- Desiderata:
  - Point estimates, coherent uncertainties
  - Interpretable, complex, modular; prior information

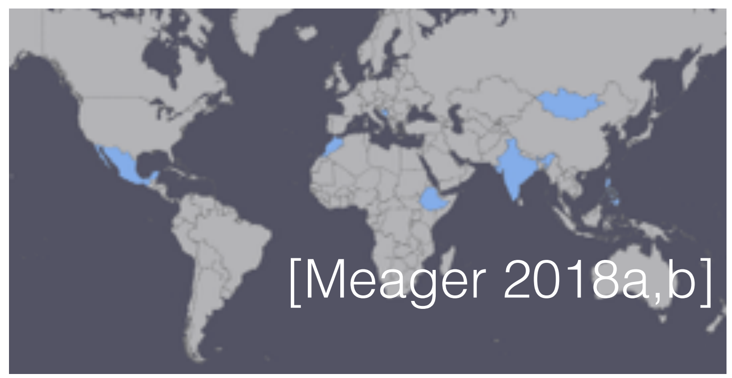
Bayesian inference





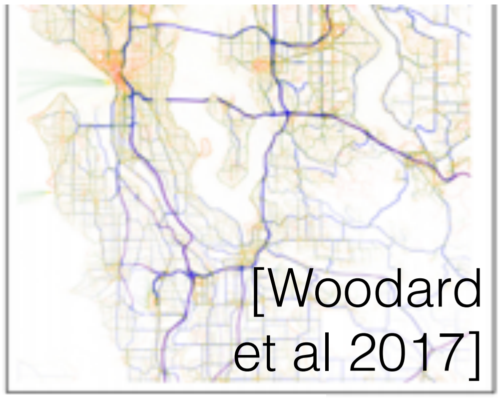
[mc-stan.org]

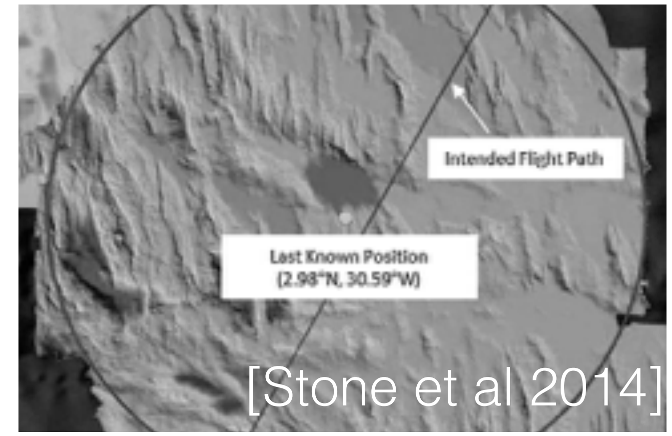




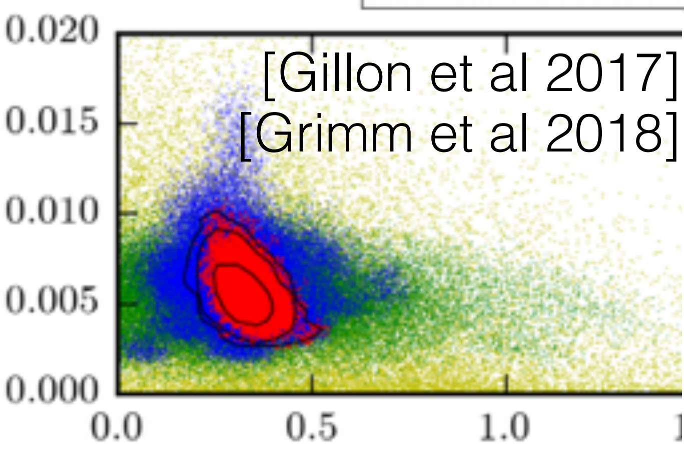
- Desiderata:
  - Point estimates, coherent uncertainties
  - Interpretable, complex, modular; prior information
- Challenge: existing methods can be slow, tedious, unreliable

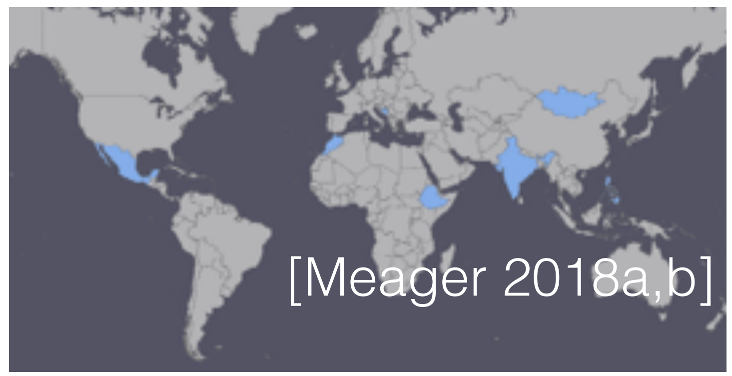
Bayesian inference





[mc-stan.org]





- Desiderata:
    - Point estimates, coherent uncertainties
    - Interpretable, complex, modular; prior information
- Challenge: existing methods can be slow, tedious, unreliable
- Our proposal: use *efficient data summaries* for **scalable**, **automated** algorithms with **error bounds for finite data**

Roadmap

- Approximate Bayes review
- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

Roadmap

- Approximate Bayes review
- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

Bayesian inference

Bayesian inference

$p(\theta)$

Bayesian inference

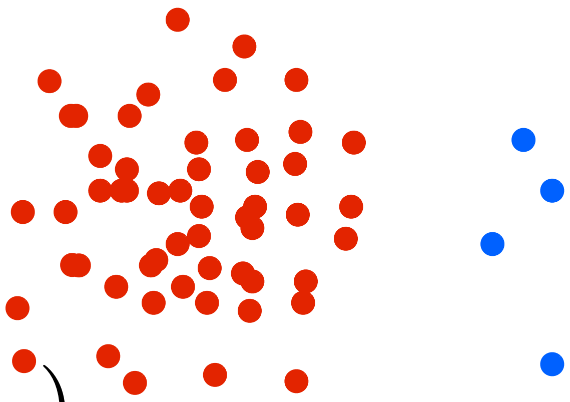
$p(y|\theta)p(\theta)$

Bayesian inference

$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$

Bayesian inference

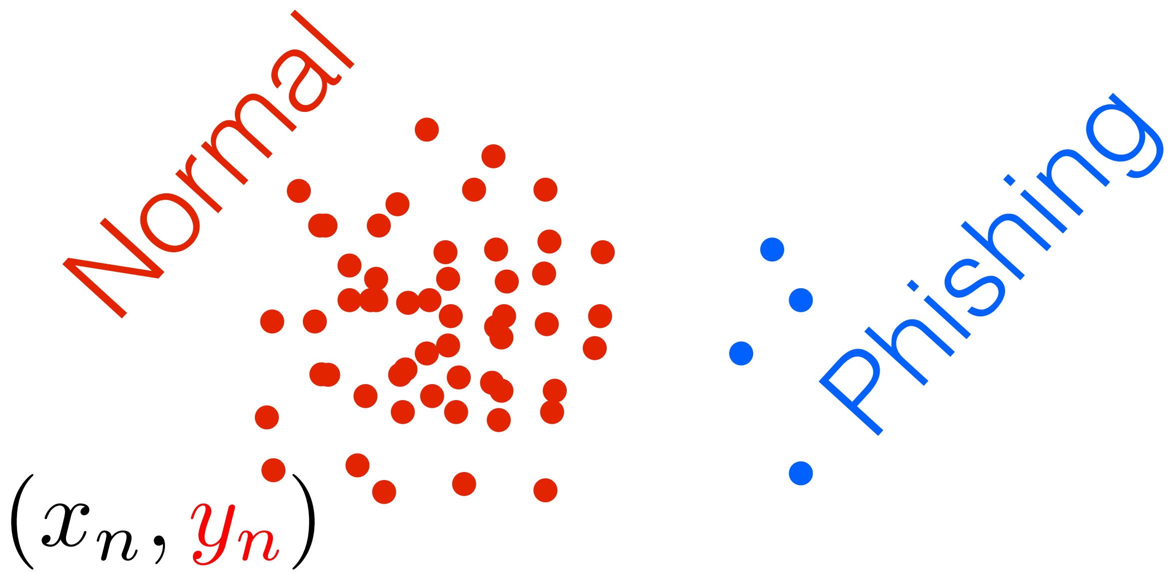
$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$



(x_n, y_n)

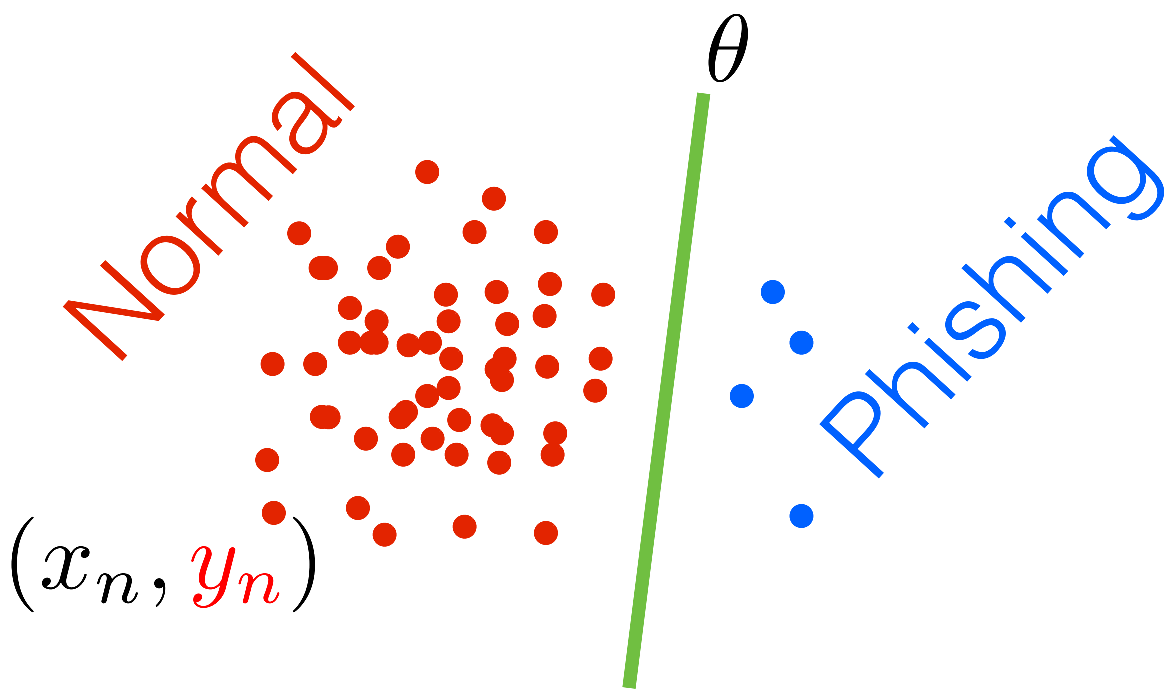
Bayesian inference

$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$



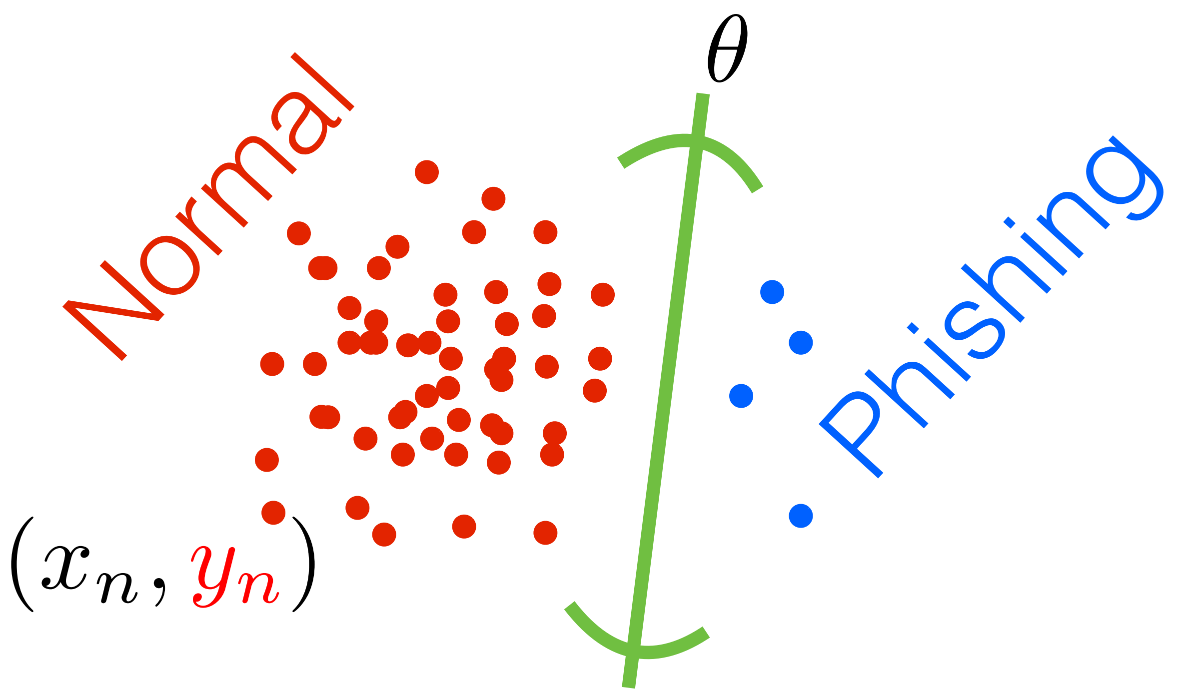
Bayesian inference

$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$



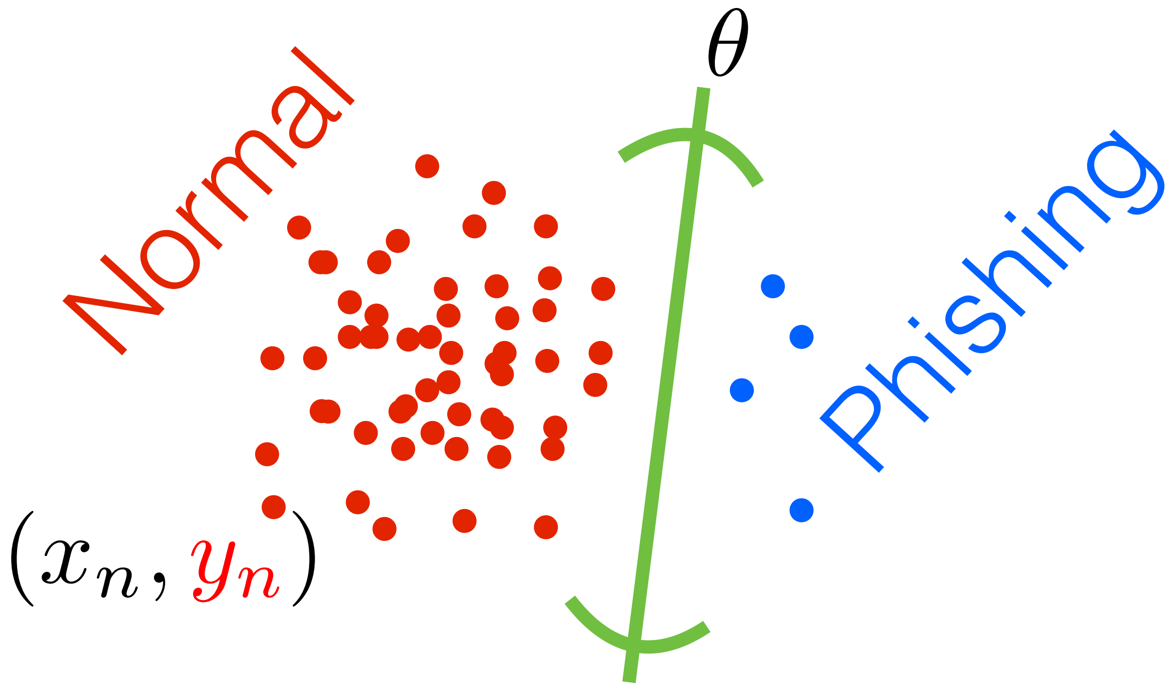
Bayesian inference

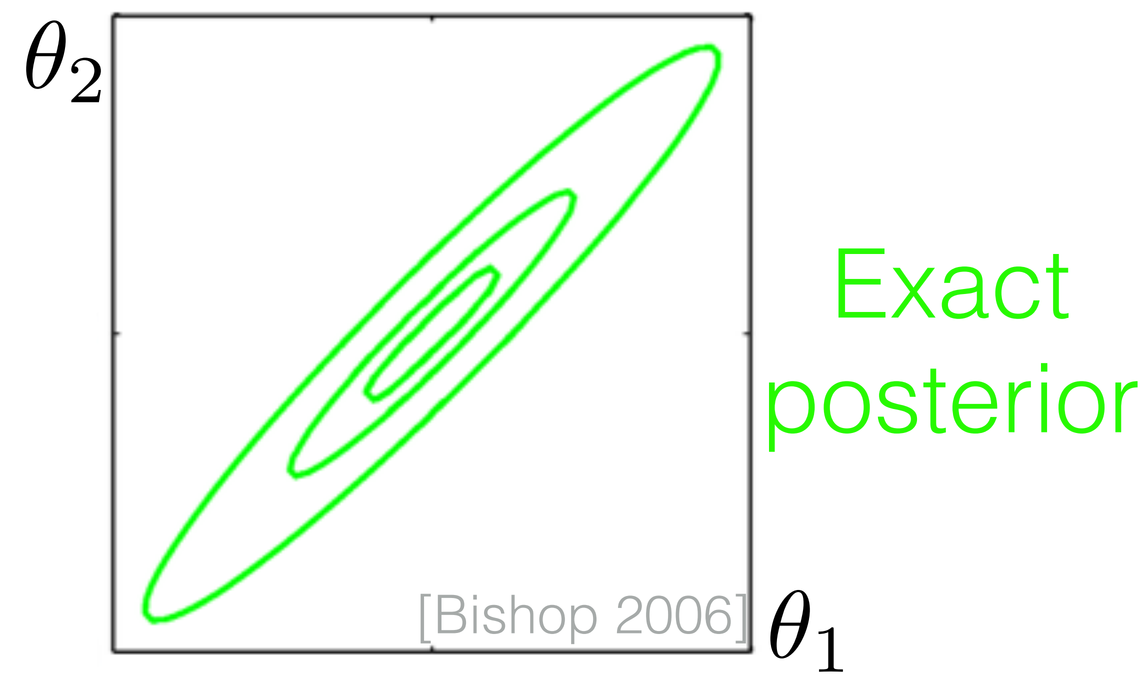
$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$



Bayesian inference

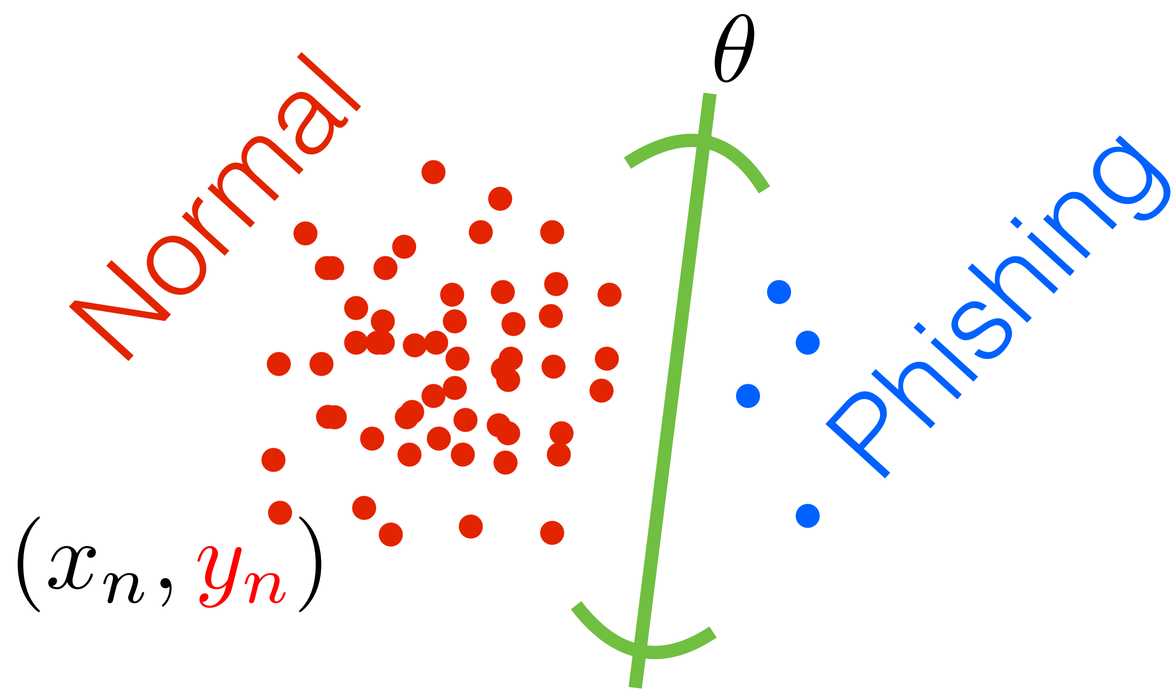
$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$

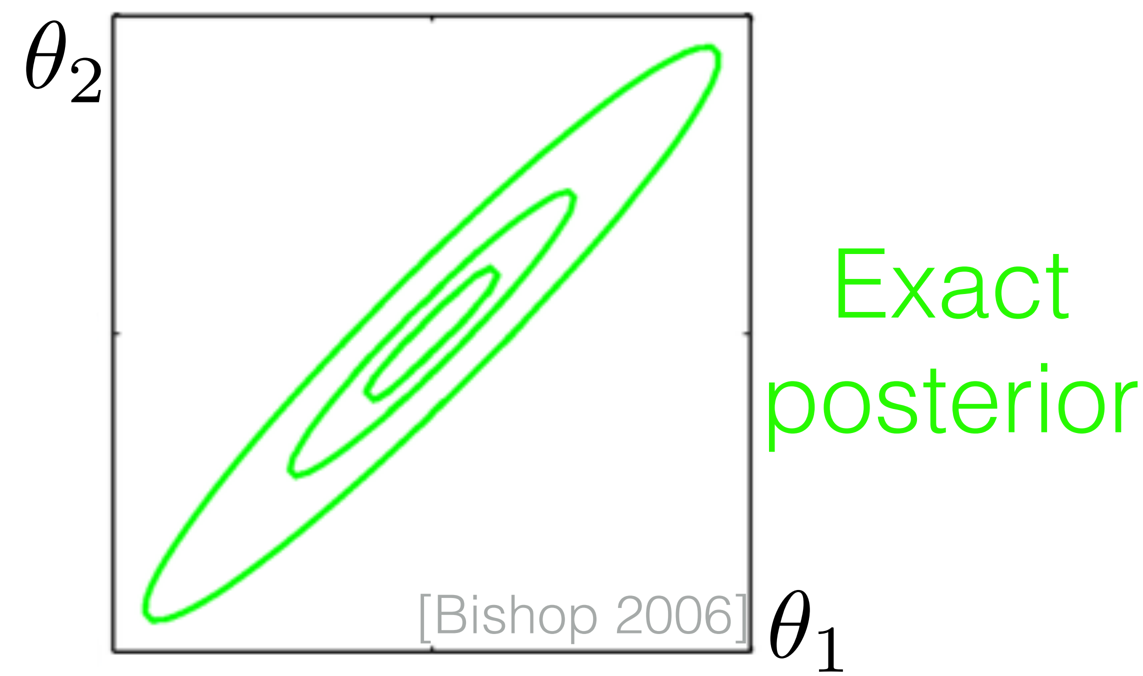




Bayesian inference

$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$

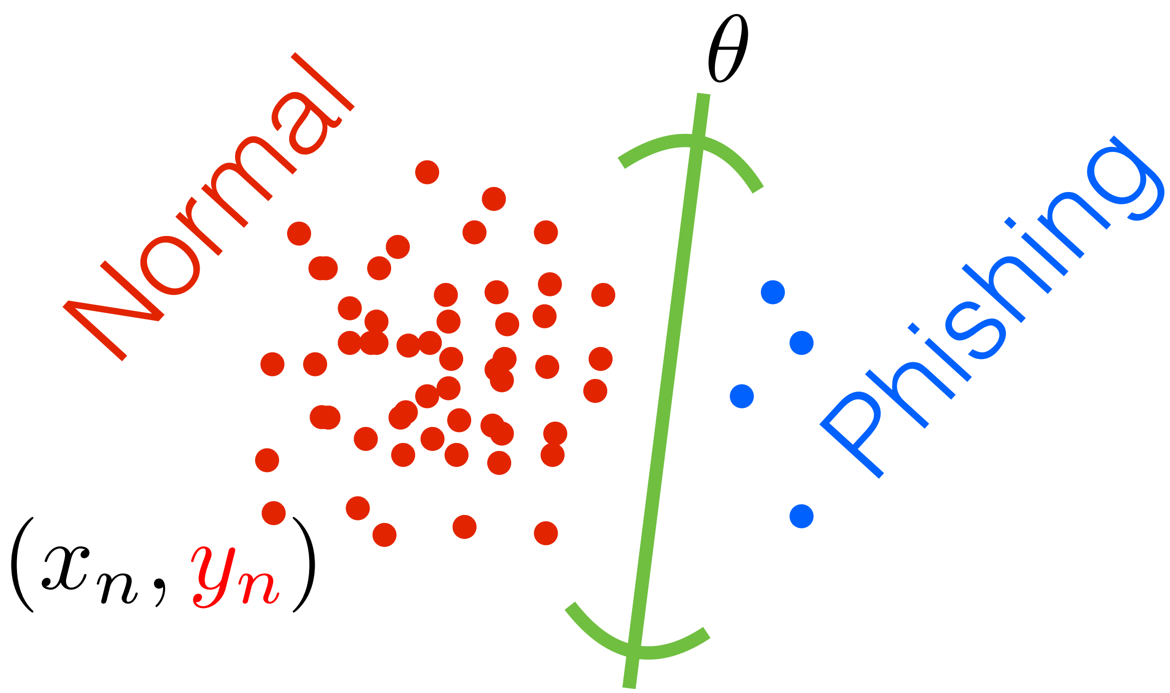


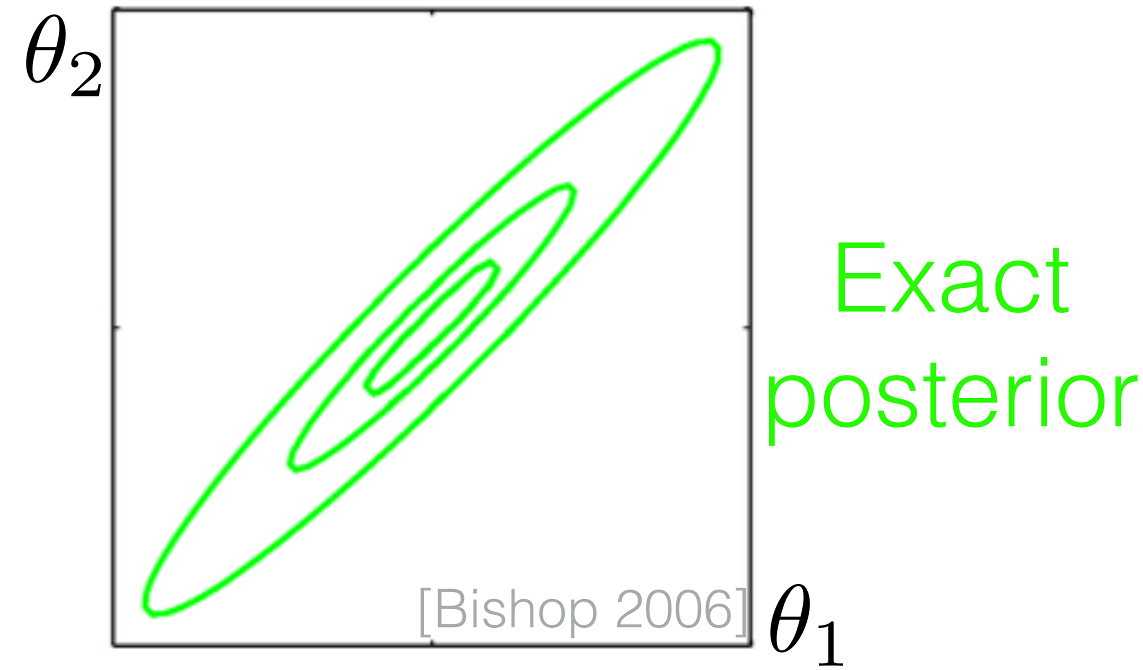


- MCMC: Eventually accurate but can be slow [Bardenet, Doucet, Holmes 2017]

Bayesian inference

$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$

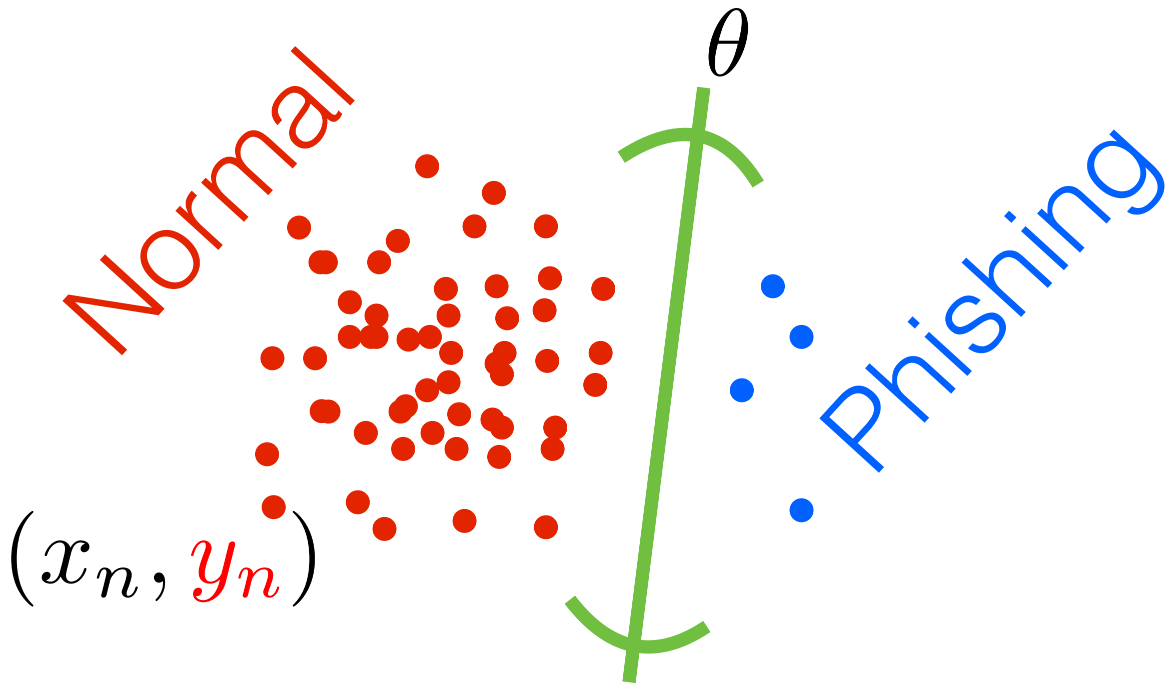


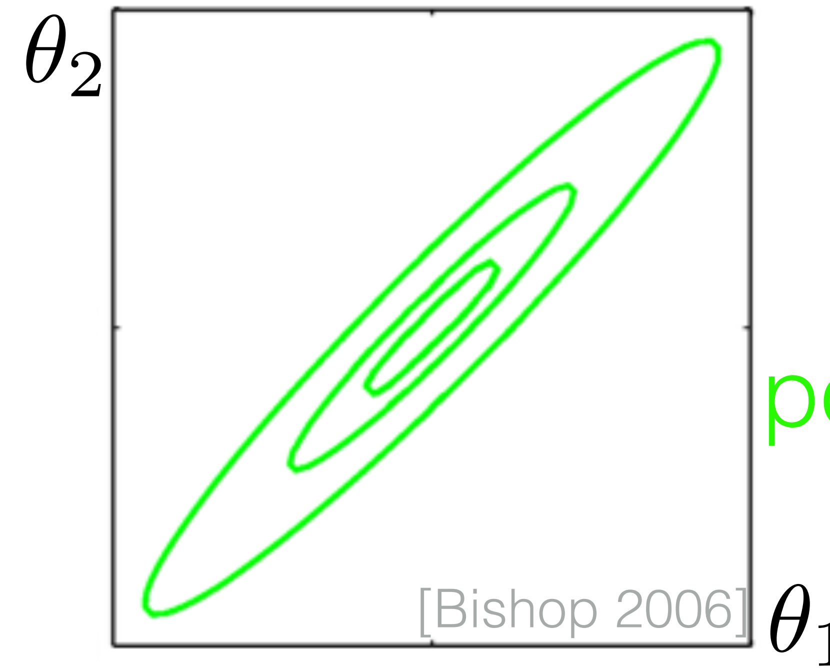


- MCMC: Eventually accurate but can be slow [Bardenet, Doucet, Holmes 2017]
- (Mean-field) variational Bayes: (MF)VB

Bayesian inference

$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$

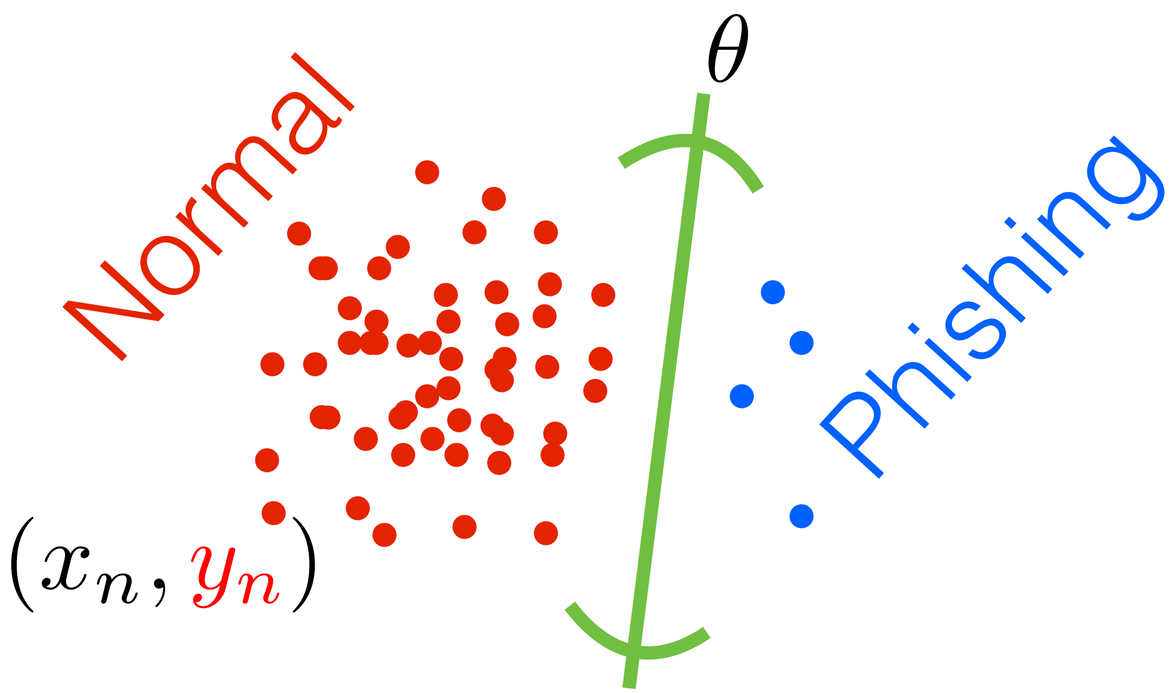


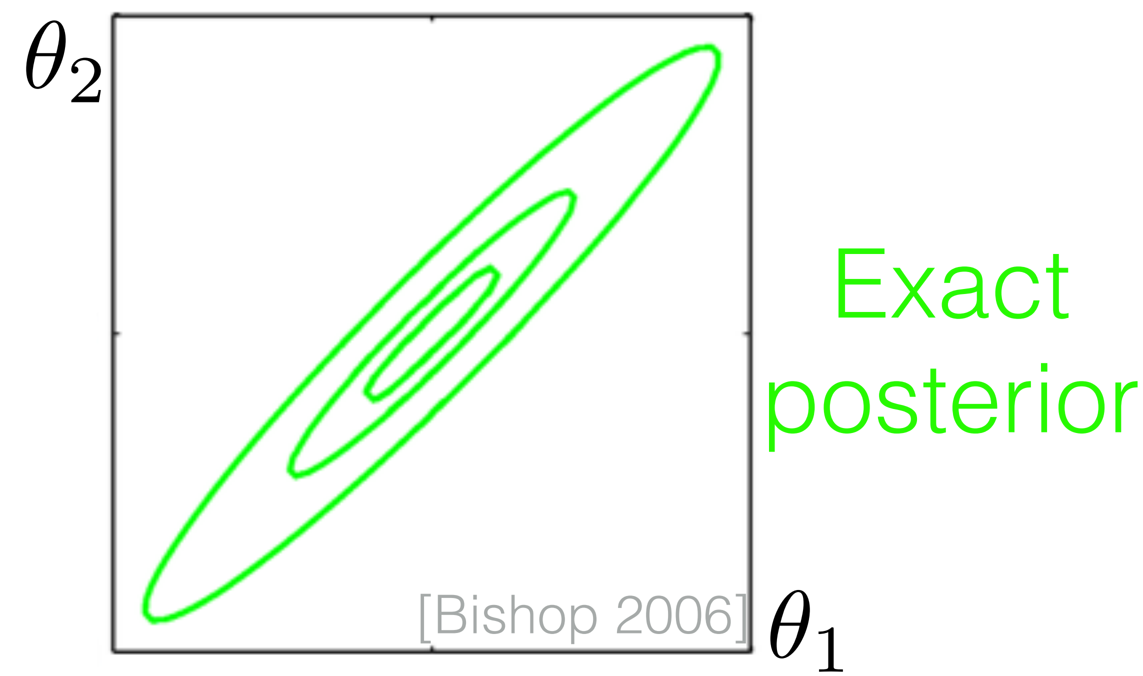


- MCMC: Eventually accurate but can be slow [Bardenet, Doucet, Holmes 2017]
- (Mean-field) variational Bayes: (MF)VB
  - Fast

Bayesian inference

$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$

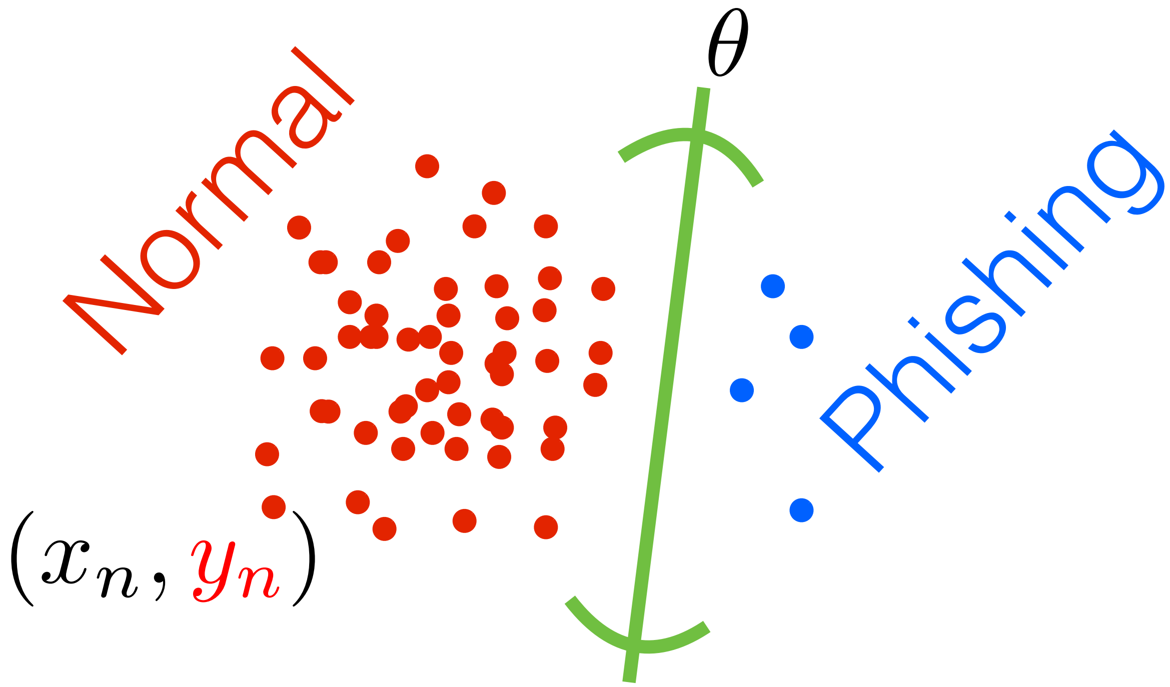


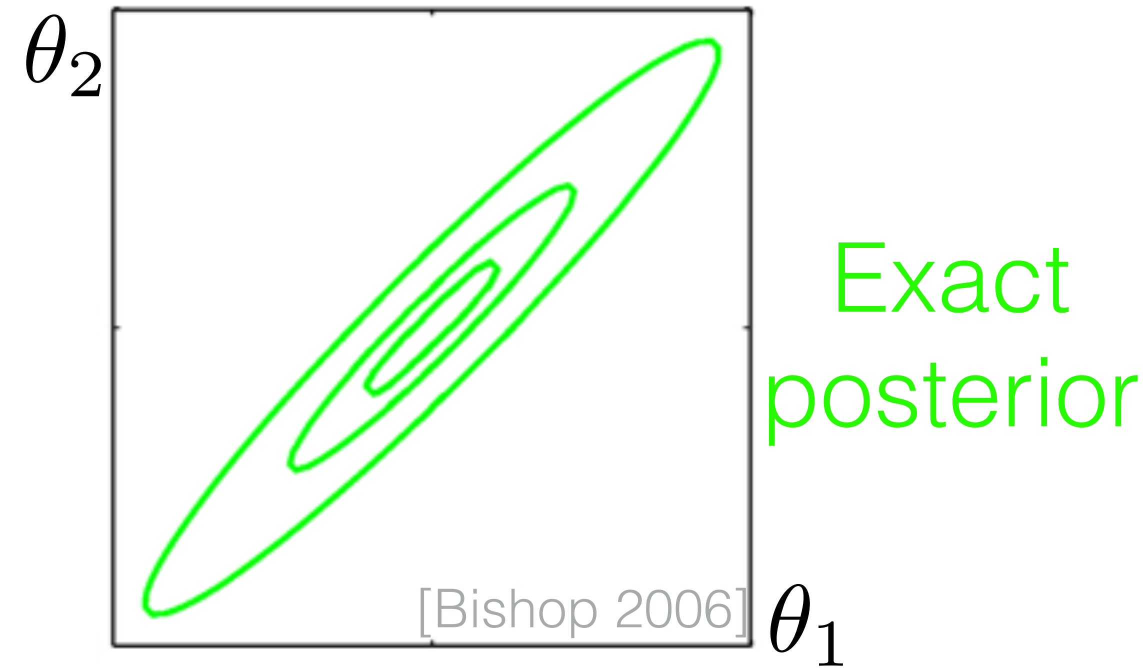


- MCMC: Eventually accurate but can be slow [Bardenet, Doucet, Holmes 2017]
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed [Broderick, Boyd, Wibisono, Wilson, Jordan 2013]

Bayesian inference

$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$

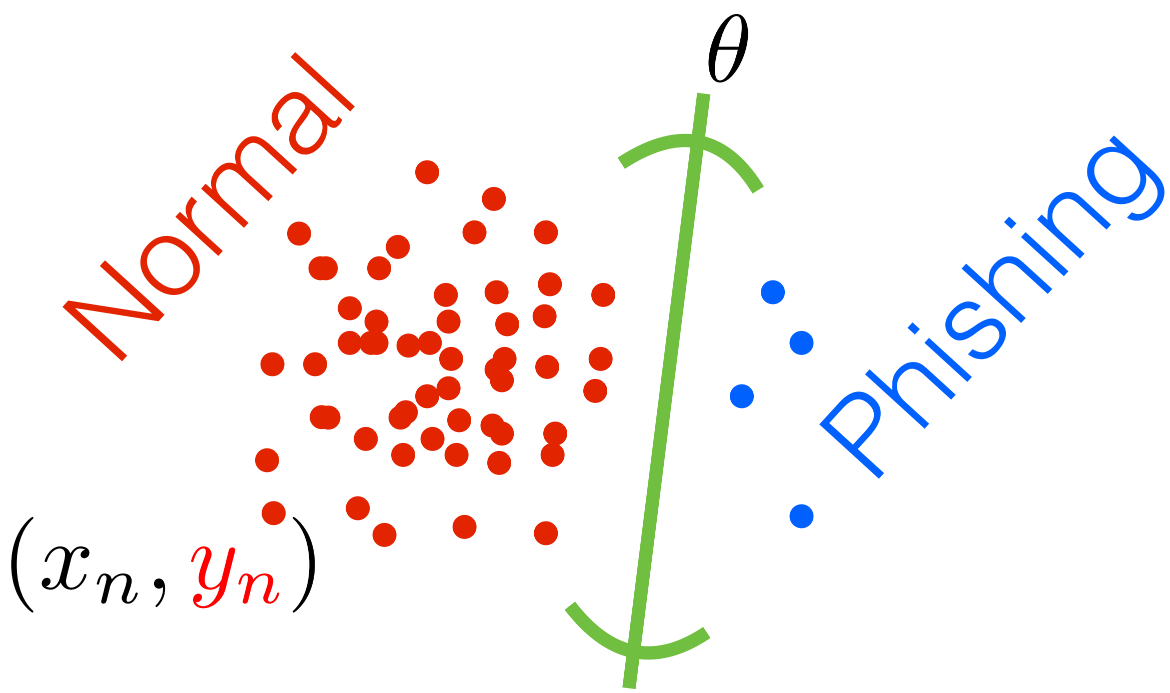


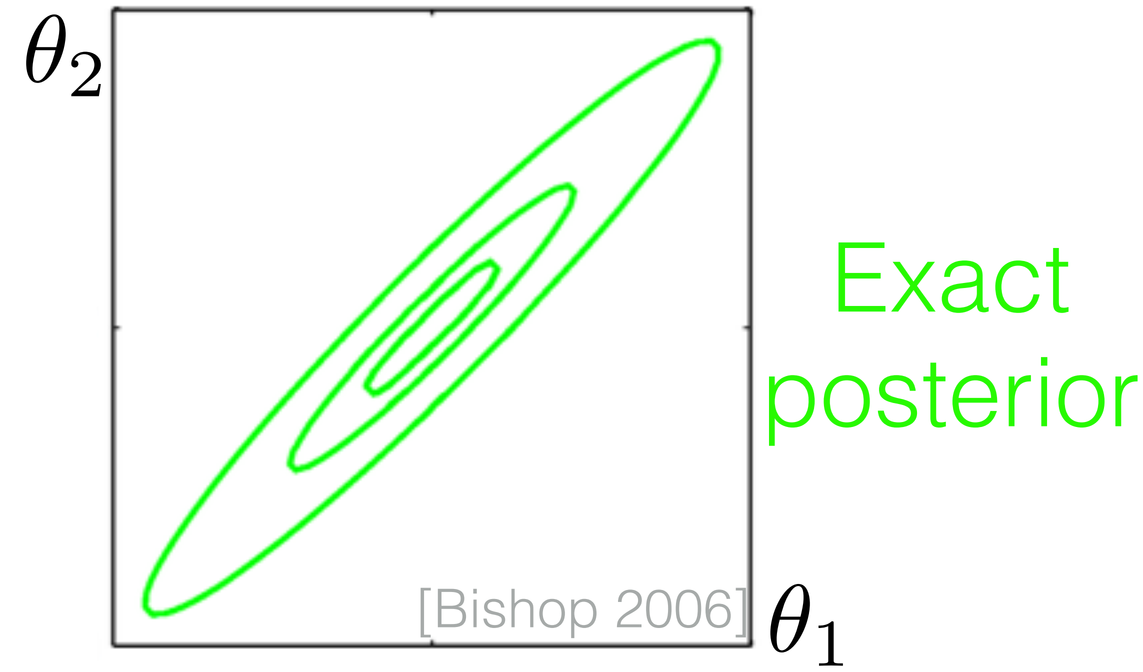


- MCMC: Eventually accurate but can be slow [Bardenet, Doucet, Holmes 2017]
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed [Broderick, Boyd, Wibisono, Wilson, Jordan 2013] (3.6M Wikipedia, 32 cores, ~hour)

Bayesian inference

$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$



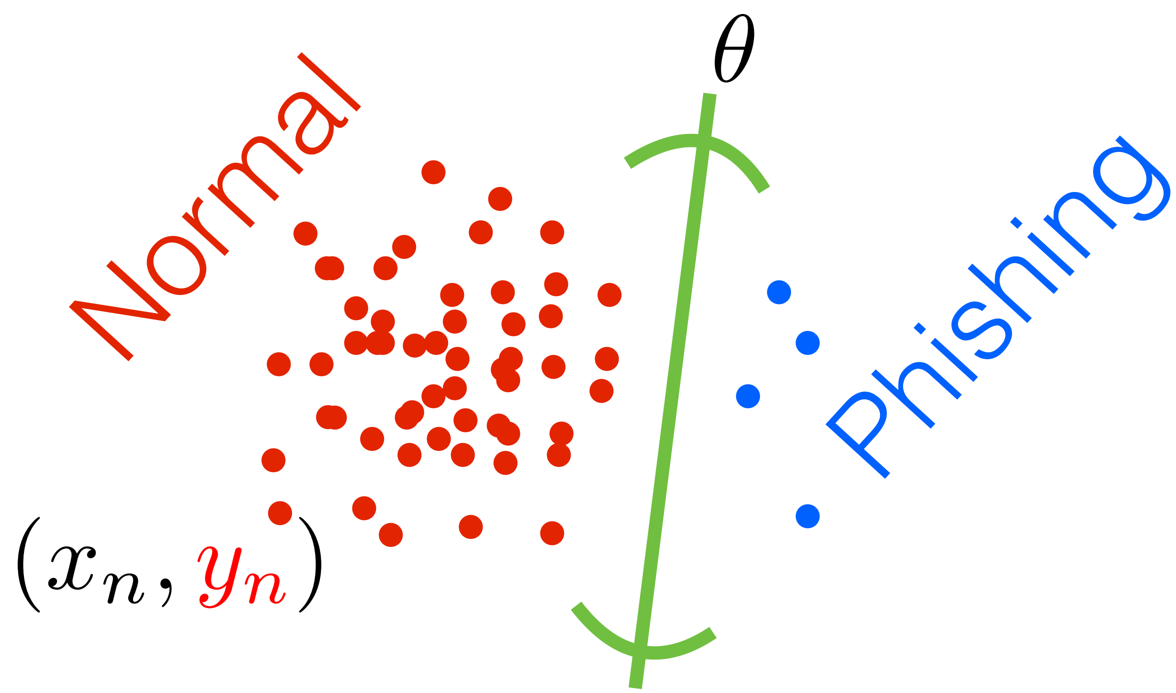


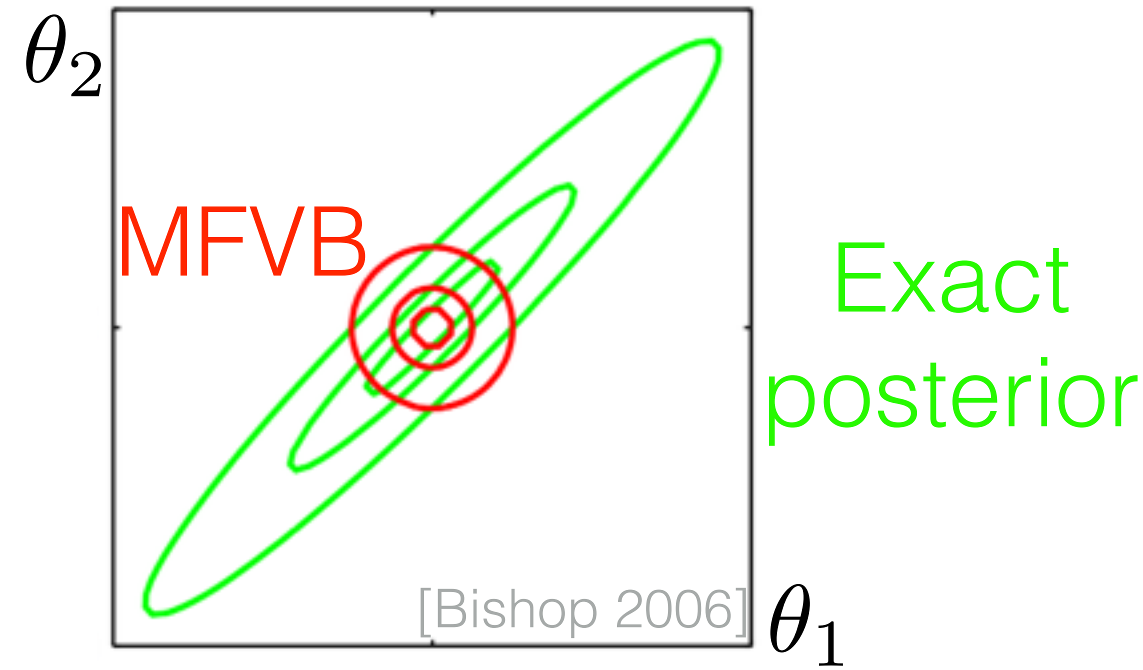
- MCMC: Eventually accurate but can be slow [Bardenet, Doucet, Holmes 2017]
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed [Broderick, Boyd, Wibisono, Wilson, Jordan 2013] (3.6M Wikipedia, 32 cores, ~hour)
  - Misestimation & lack of quality guarantees

[MacKay 2003; Bishop 2006; Wang, Titterington 2004; Turner, Sahani 2011; Fosdick 2013; Dunson 2014; Bardenet, Doucet, Holmes 2017; Opper, Winther 2003; Giordano, Broderick, Jordan 2015, 2017]

Bayesian inference

$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$



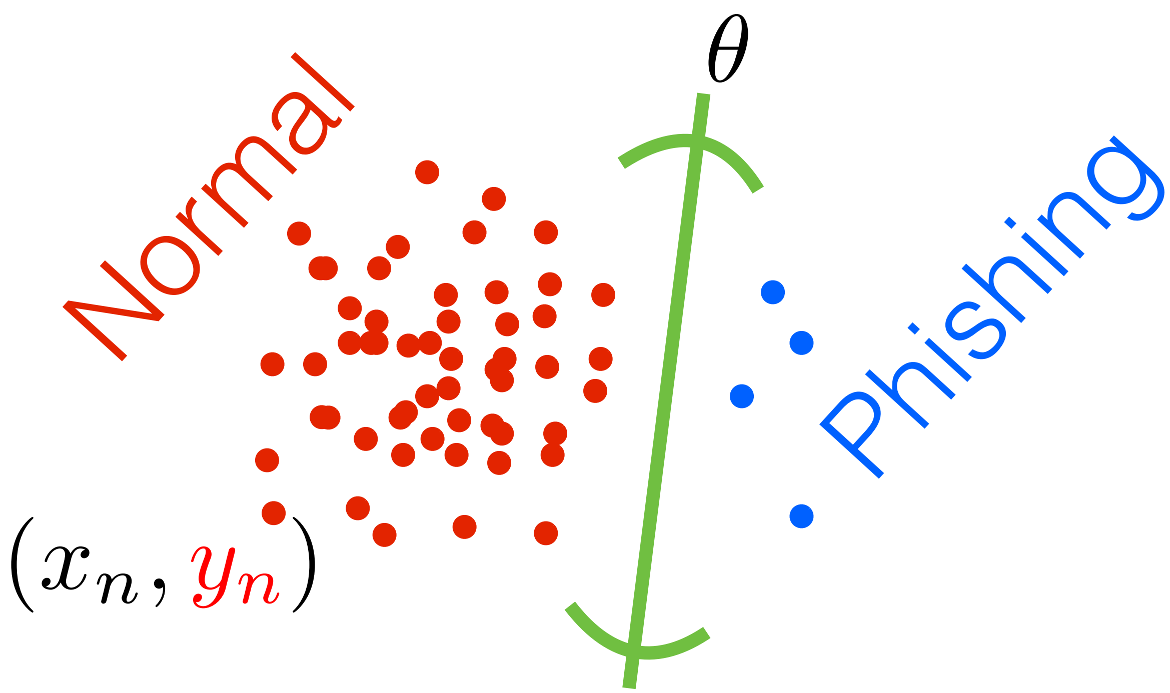


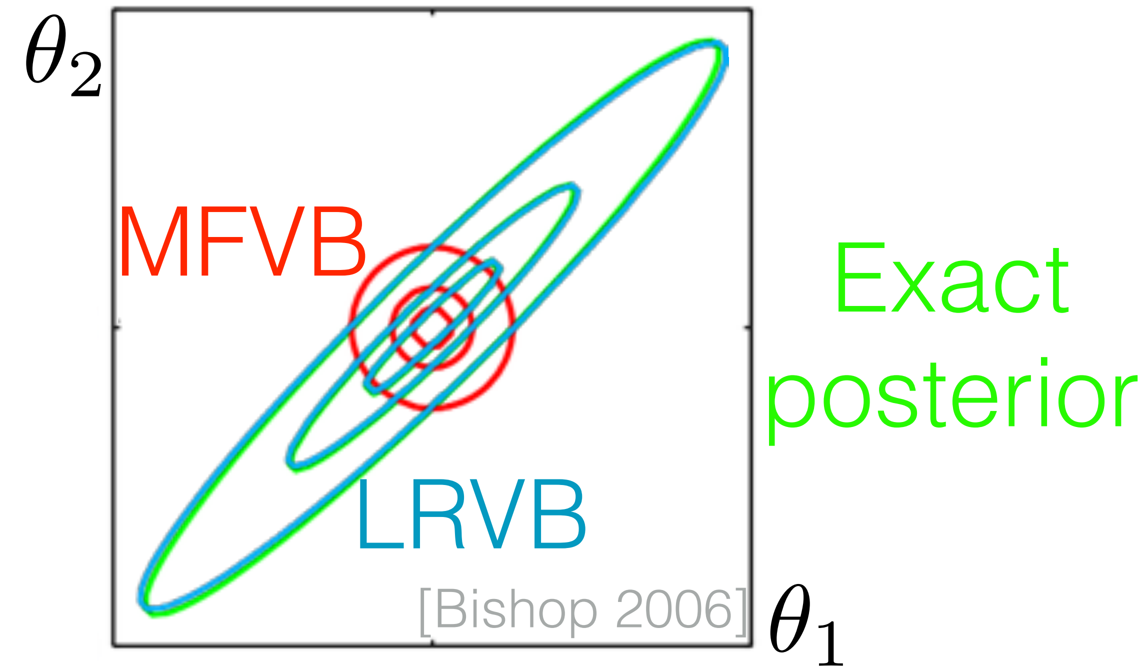
- MCMC: Eventually accurate but can be slow [Bardenet, Doucet, Holmes 2017]
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed [Broderick, Boyd, Wibisono, Wilson, Jordan 2013] (3.6M Wikipedia, 32 cores, ~hour)
  - Misestimation & lack of quality guarantees

[MacKay 2003; Bishop 2006; Wang, Titterington 2004; Turner, Sahani 2011; Fosdick 2013; Dunson 2014; Bardenet, Doucet, Holmes 2017; Opper, Winther 2003; Giordano, Broderick, Jordan 2015, 2017]

Bayesian inference

$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$



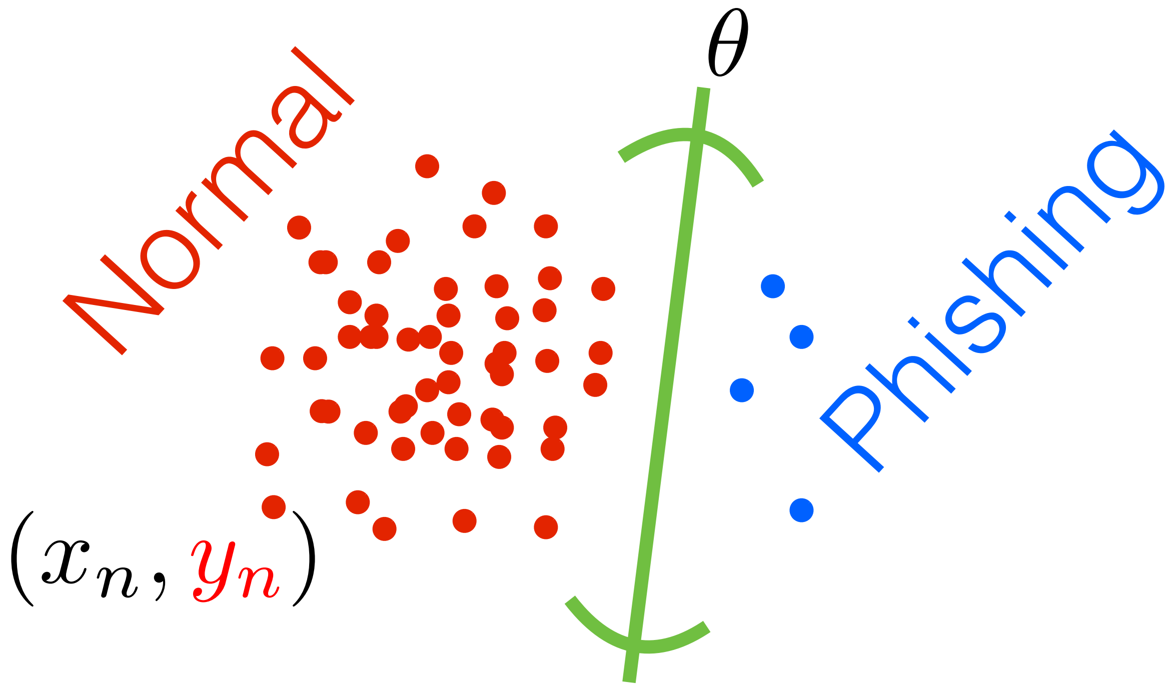


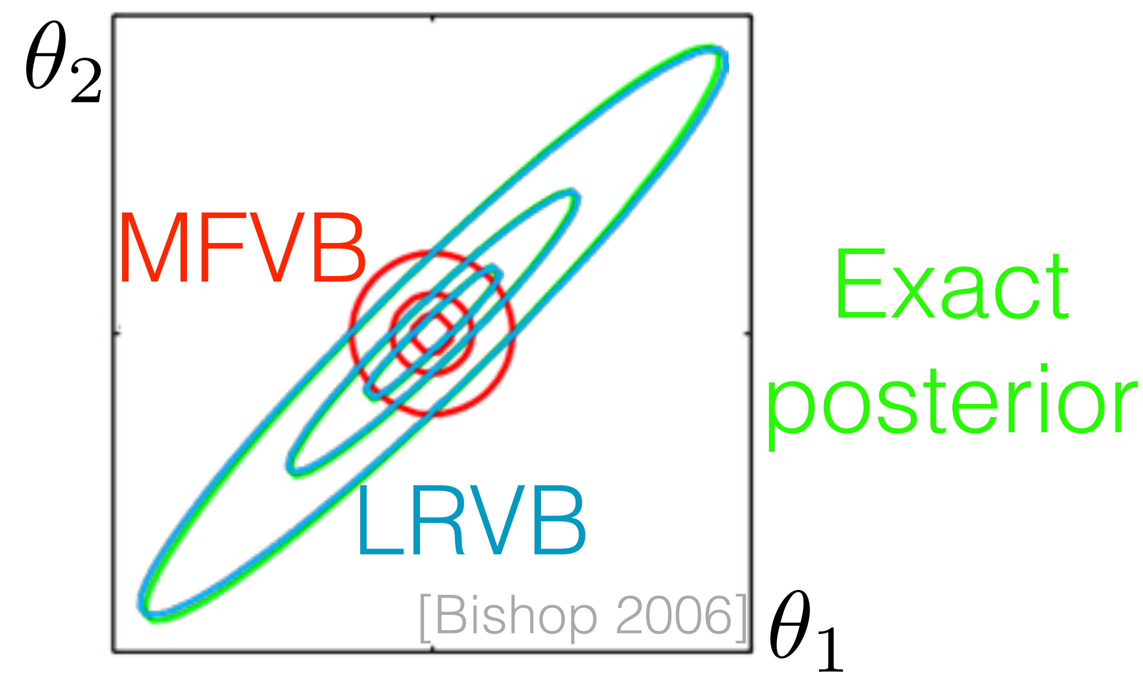
- MCMC: Eventually accurate but can be slow [Bardenet, Doucet, Holmes 2017]
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed [Broderick, Boyd, Wibisono, Wilson, Jordan 2013] (3.6M Wikipedia, 32 cores, ~hour)
  - Misestimation & lack of quality guarantees

[MacKay 2003; Bishop 2006; Wang, Titterington 2004; Turner, Sahani 2011; Fosdick 2013; Dunson 2014; Bardenet, Doucet, Holmes 2017; Opper, Winther 2003; Giordano, Broderick, Jordan 2015, 2017]

Bayesian inference

$$p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$$





- MCMC: Eventually accurate but can be slow [Bardenet, Doucet, Holmes 2017]
- (Mean-field) variational Bayes: (MF)VB
  - Fast, streaming, distributed [Broderick, Boyd, Wibisono, Wilson, Jordan 2013] (3.6M Wikipedia, 32 cores, ~hour)
  - Misestimation & lack of quality guarantees

[MacKay 2003; Bishop 2006; Wang, Titterington 2004; Turner, Sahani 2011; Fosdick 2013; Dunson 2014; Bardenet, Doucet, Holmes 2017; Opper, Winther 2003; Giordano, Broderick, Jordan 2015, 2017]

- Automation: e.g. Stan, NUTS, ADVI

[http://mc-stan.org/ ; Hoffman, Gelman 2014; Kucukelbir, Tran, Ranganath, Gelman, Blei 2017]

Roadmap

- Approximate Bayes review
- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

Roadmap

- Approximate Bayes review
- The "core" of the data set
- Uniform data subsampling isn't enough
- Importance sampling for "coresets"
- Optimization for "coresets"
- Approximate sufficient statistics

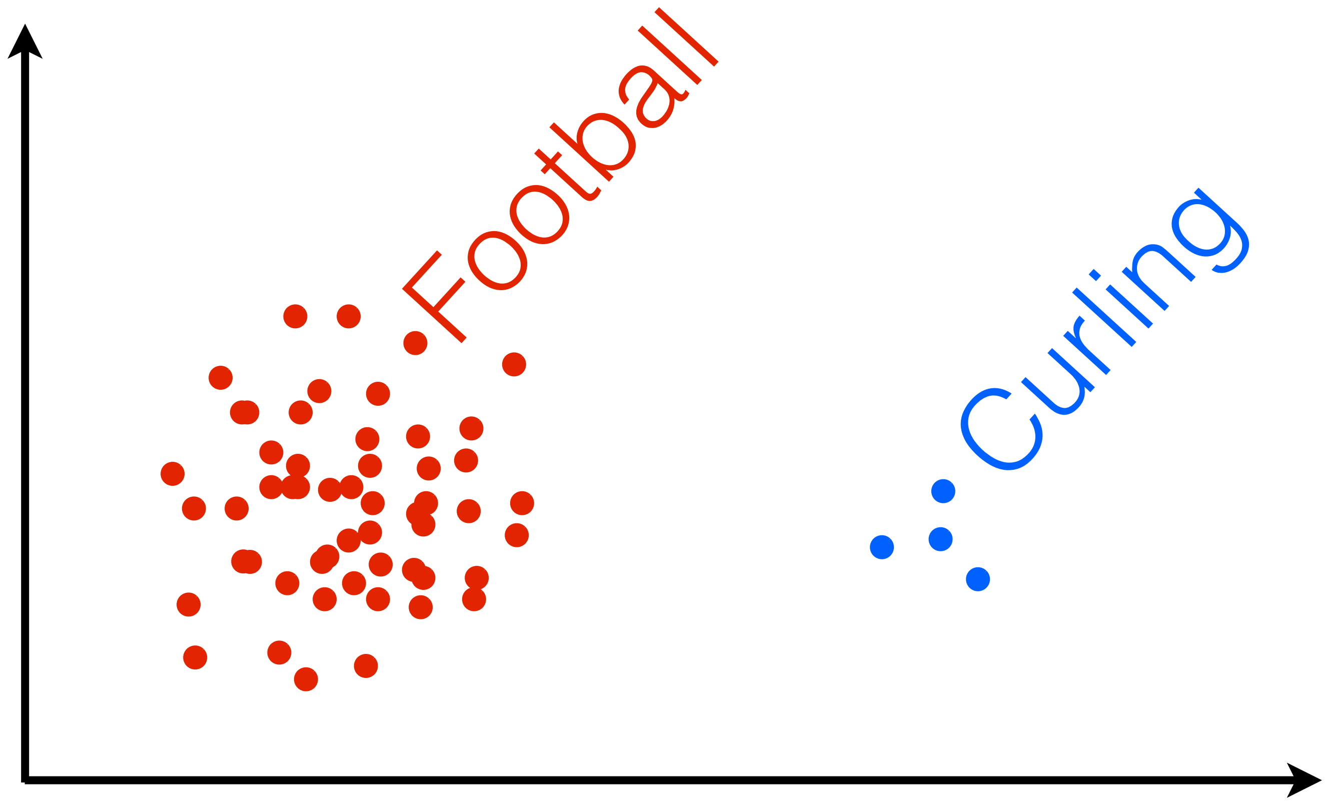
Bayesian coresets

Bayesian coresets

- Observe: redundancies can exist even if data isn't "tall"

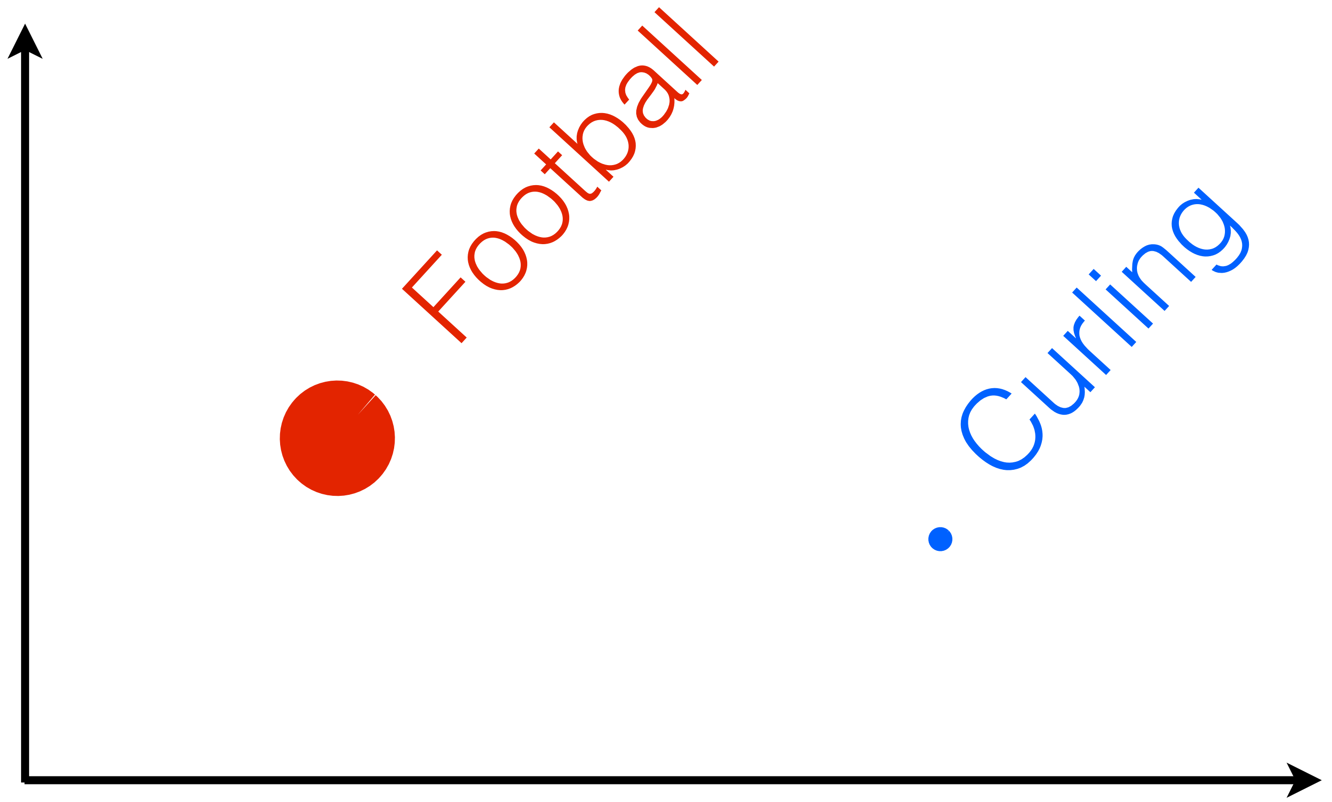
Bayesian coresets

- Observe: redundancies can exist even if data isn't "tall"



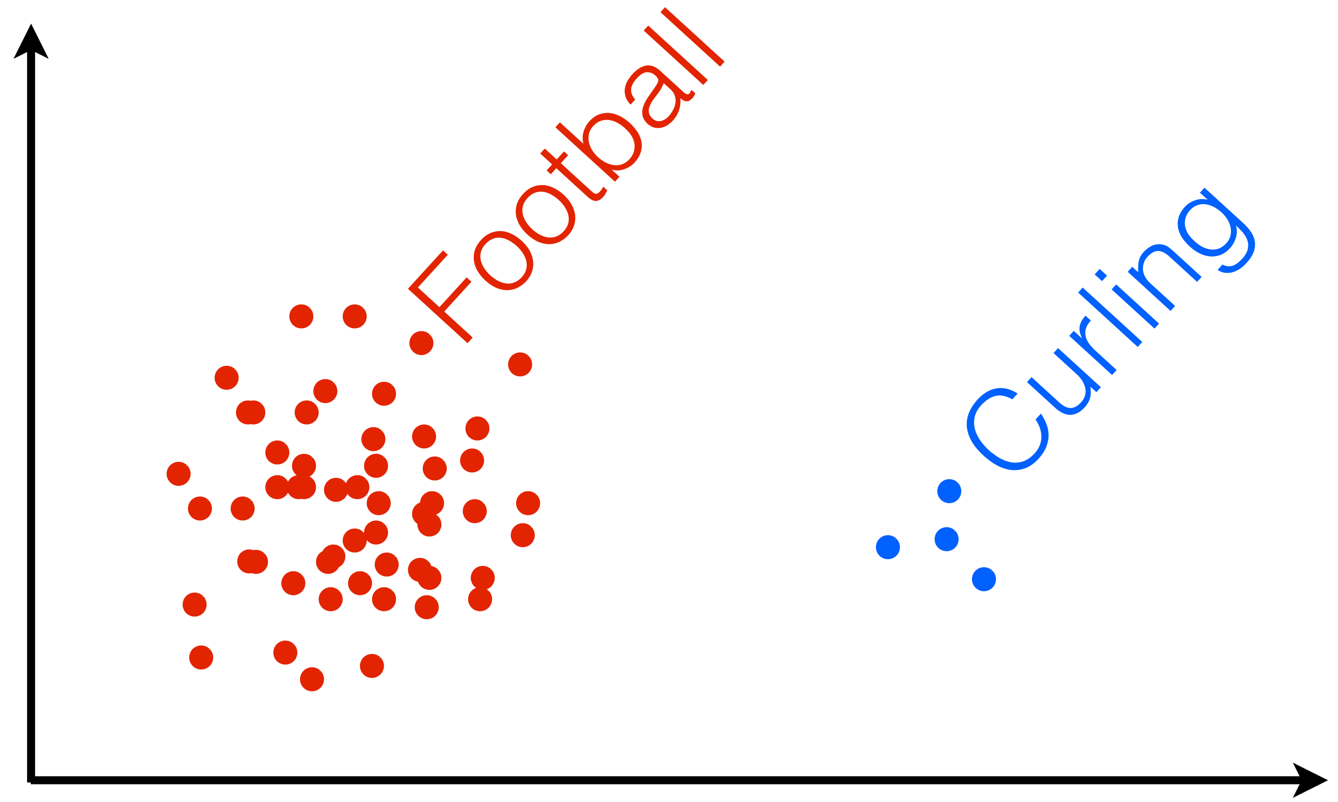
Bayesian coresets

- Observe: redundancies can exist even if data isn't "tall"



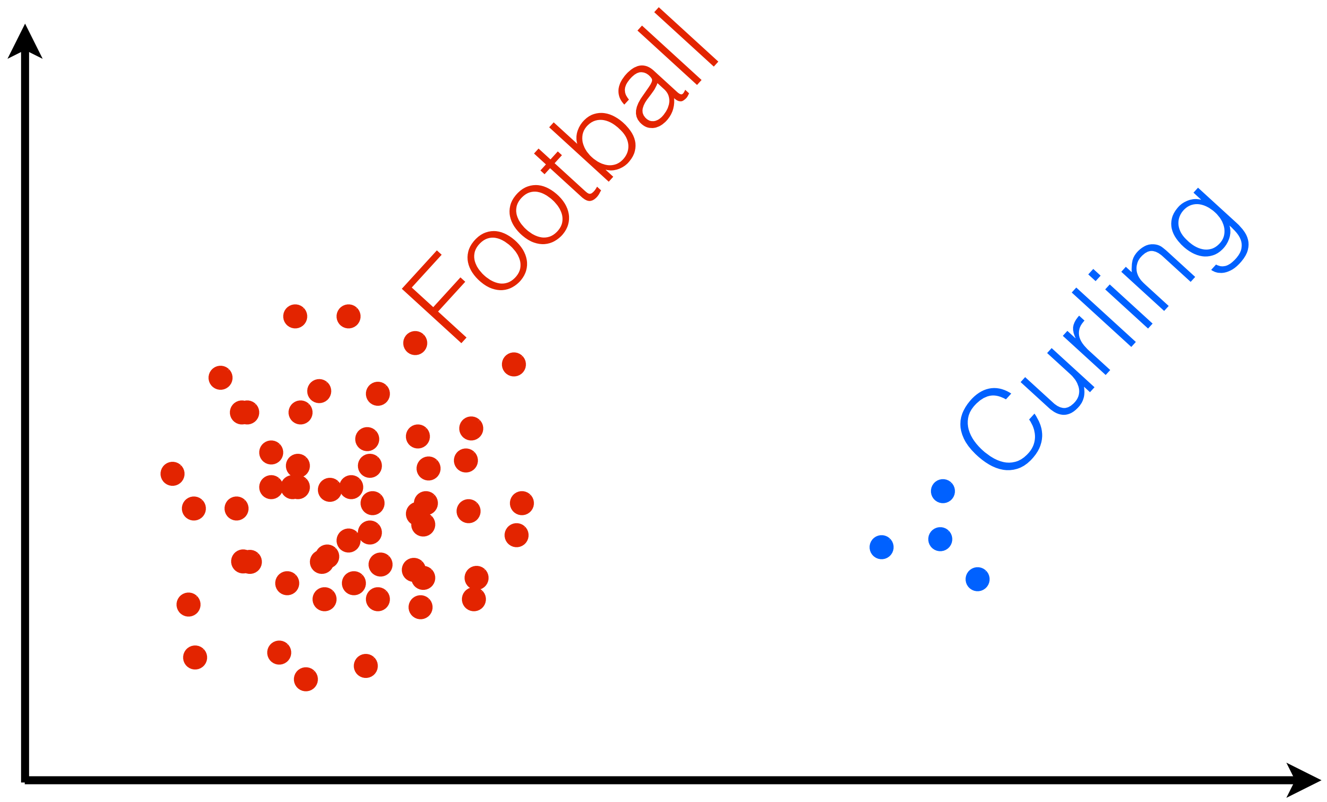
Bayesian coresets

- Observe: redundancies can exist even if data isn't "tall"



Bayesian coresets

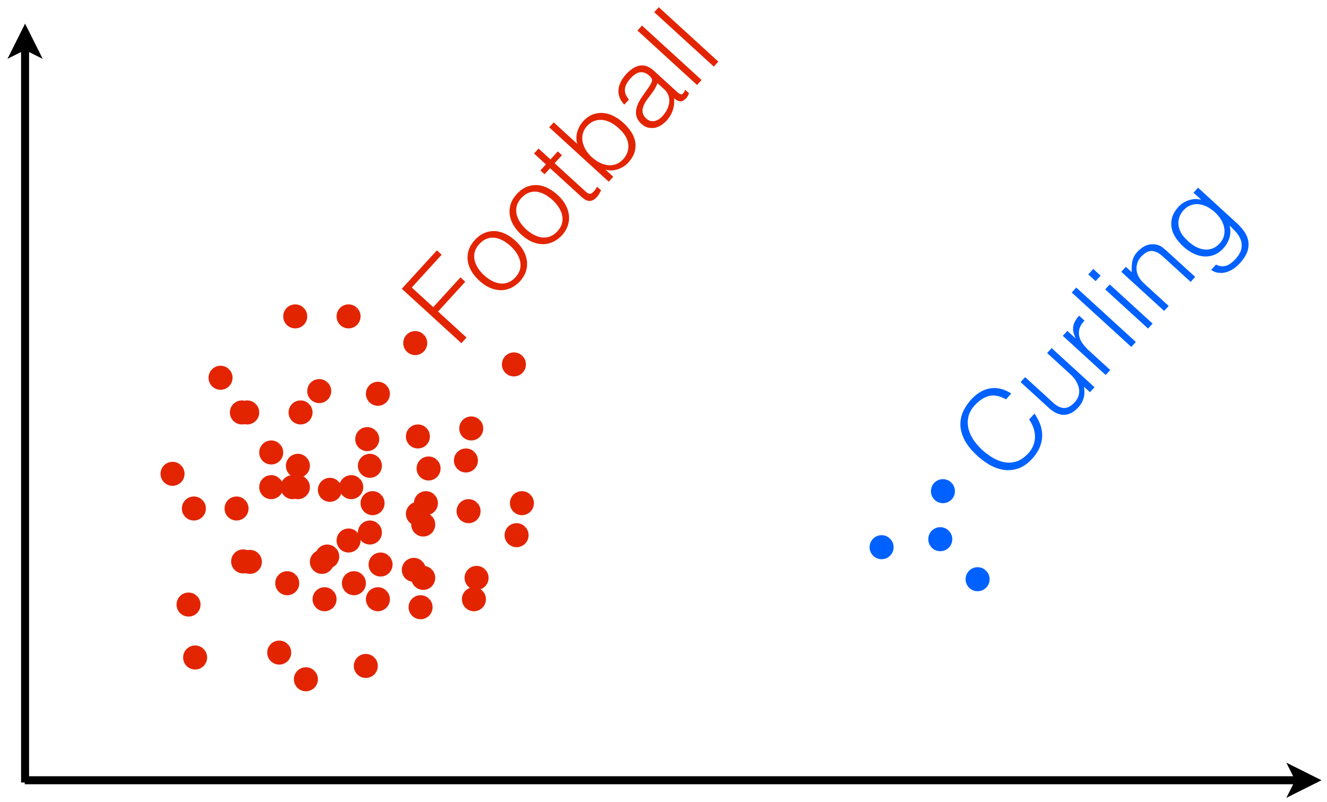
- Observe: redundancies can exist even if data isn't "tall"
- Coresets: pre-process data to get a smaller, weighted data set



[Agarwal et al 2005; Feldman & Langberg 2011]

Bayesian coresets

- Observe: redundancies can exist even if data isn't "tall"
- Coresets: pre-process data to get a smaller, weighted data set

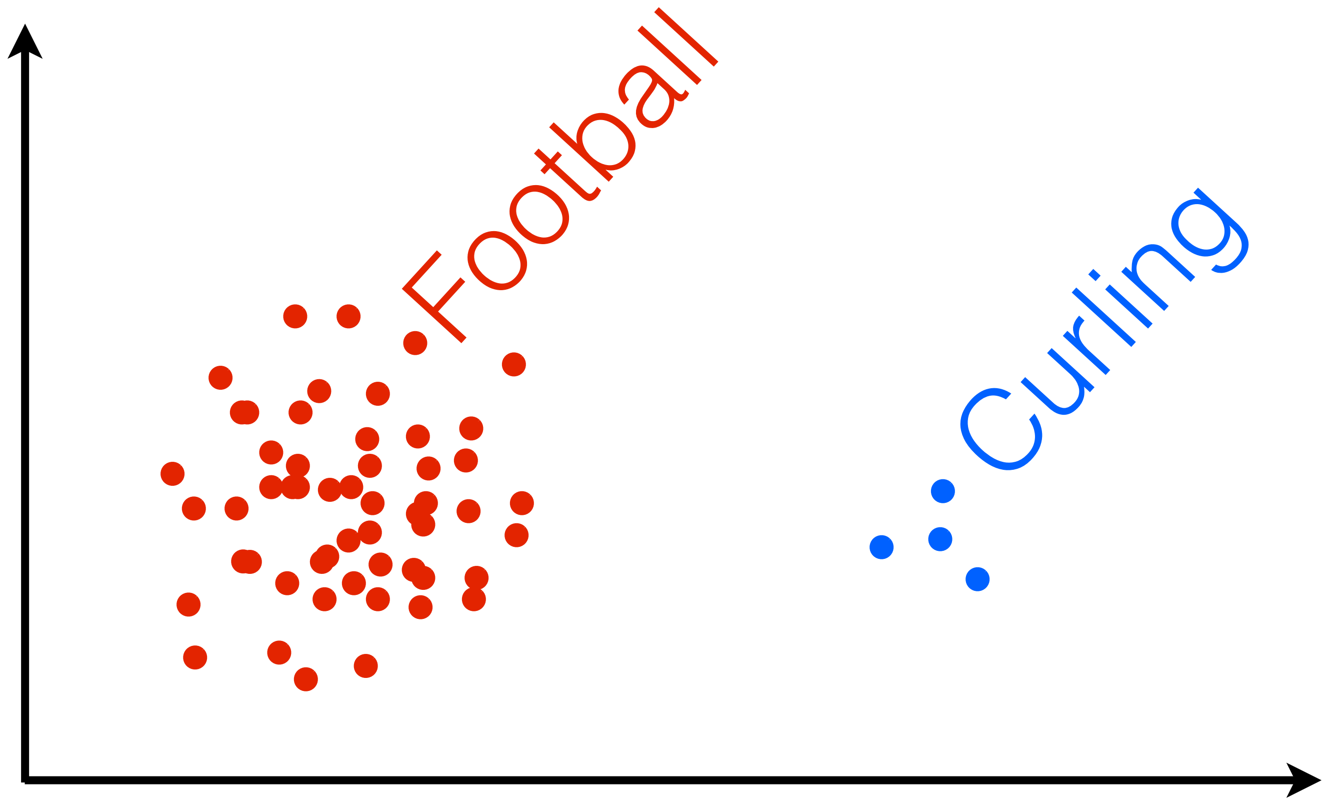


- Theoretical guarantees on quality

[Agarwal et al 2005; Feldman & Langberg 2011]

Bayesian coresets

- Observe: redundancies can exist even if data isn't "tall"
- Coresets: pre-process data to get a smaller, weighted data set

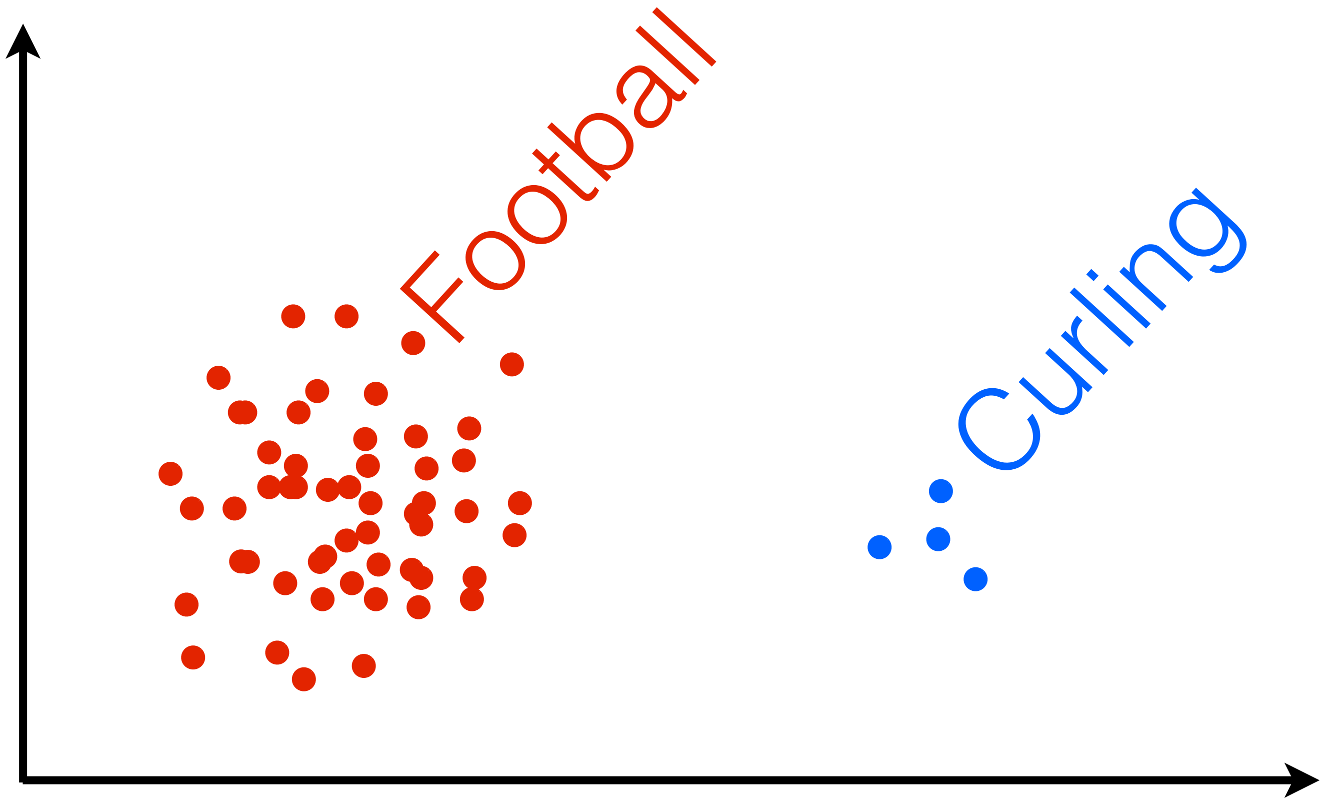


- Theoretical guarantees on quality
- Previous heuristics: data squashing, big data GPs

[Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002]

Bayesian coresets

- Observe: redundancies can exist even if data isn't "tall"
- Coresets: pre-process data to get a smaller, weighted data set

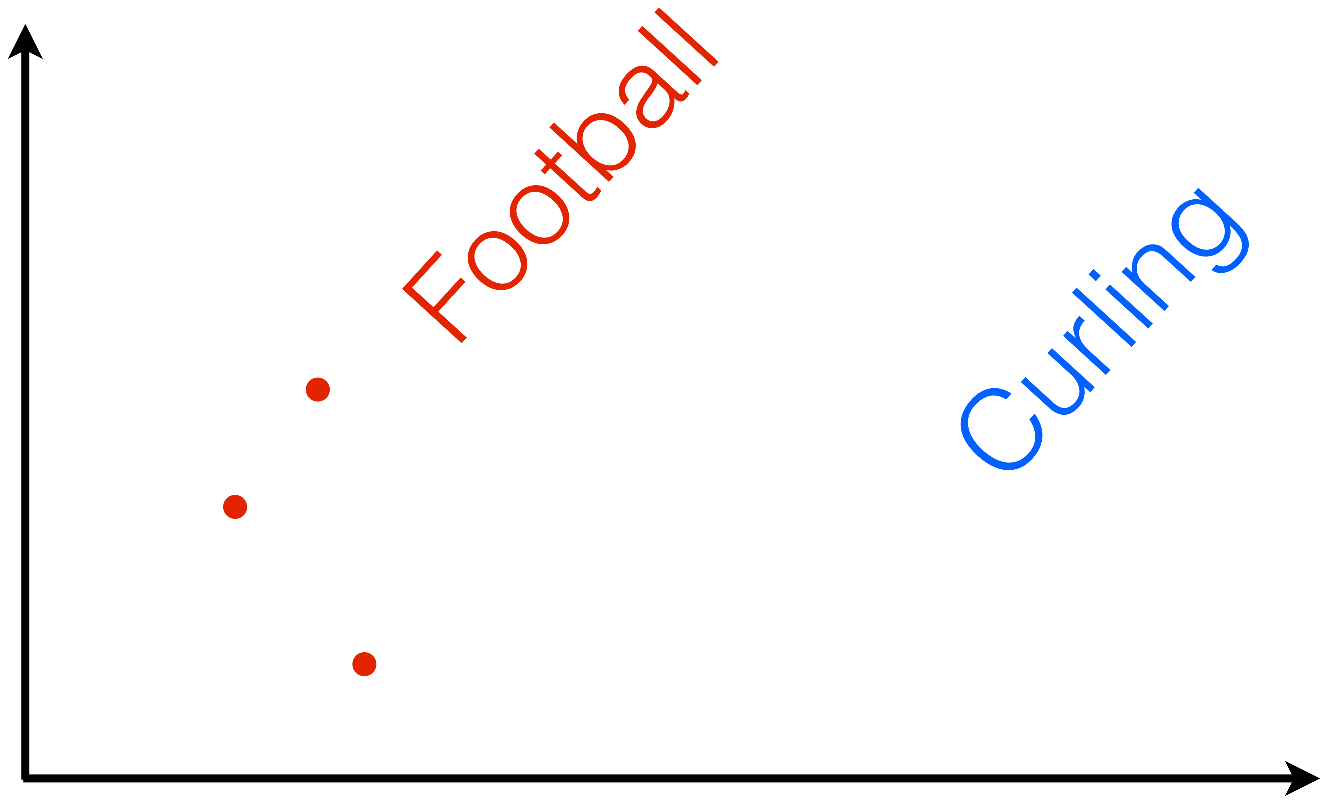


- Theoretical guarantees on quality
- Previous heuristics: data squashing, big data GPs
- Cf. subsampling

[Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002]

Bayesian coresets

- Observe: redundancies can exist even if data isn't "tall"
- Coresets: pre-process data to get a smaller, weighted data set

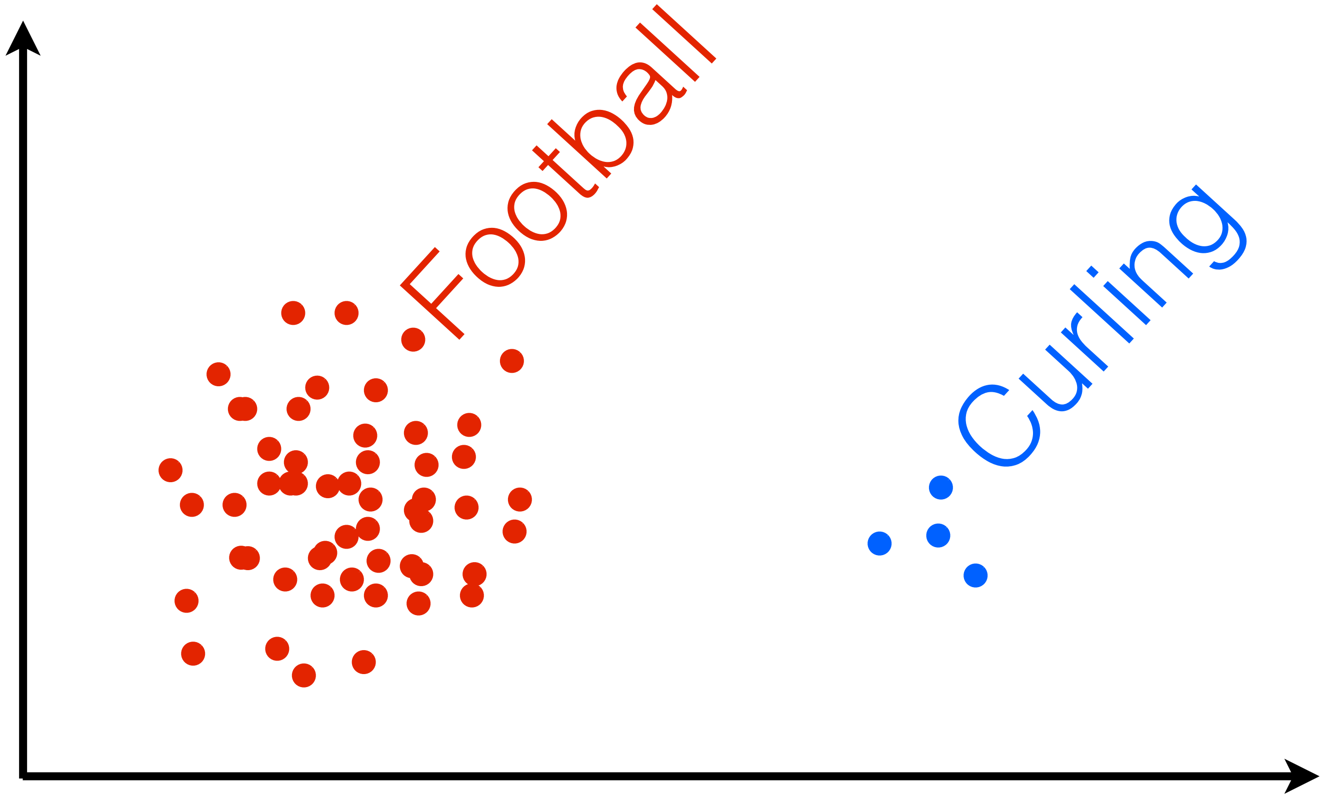


- Theoretical guarantees on quality
- Previous heuristics: data squashing, big data GPs
- Cf. subsampling

[Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002]

Bayesian coresets

- Observe: redundancies can exist even if data isn't "tall"
- Coresets: pre-process data to get a smaller, weighted data set

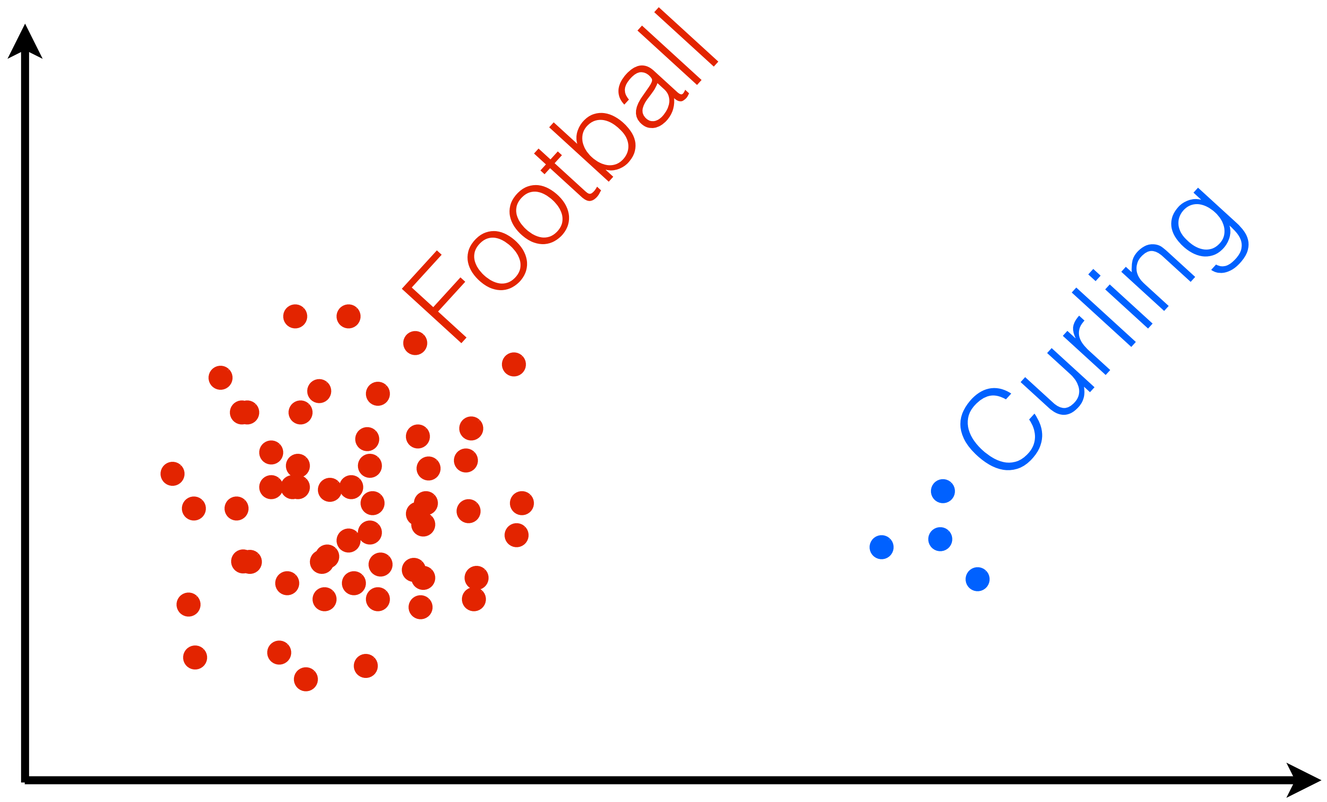


- Theoretical guarantees on quality
- Previous heuristics: data squashing, big data GPs
- Cf. subsampling

[Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002]

Bayesian coresets

- Observe: redundancies can exist even if data isn't "tall"
- Coresets: pre-process data to get a smaller, weighted data set

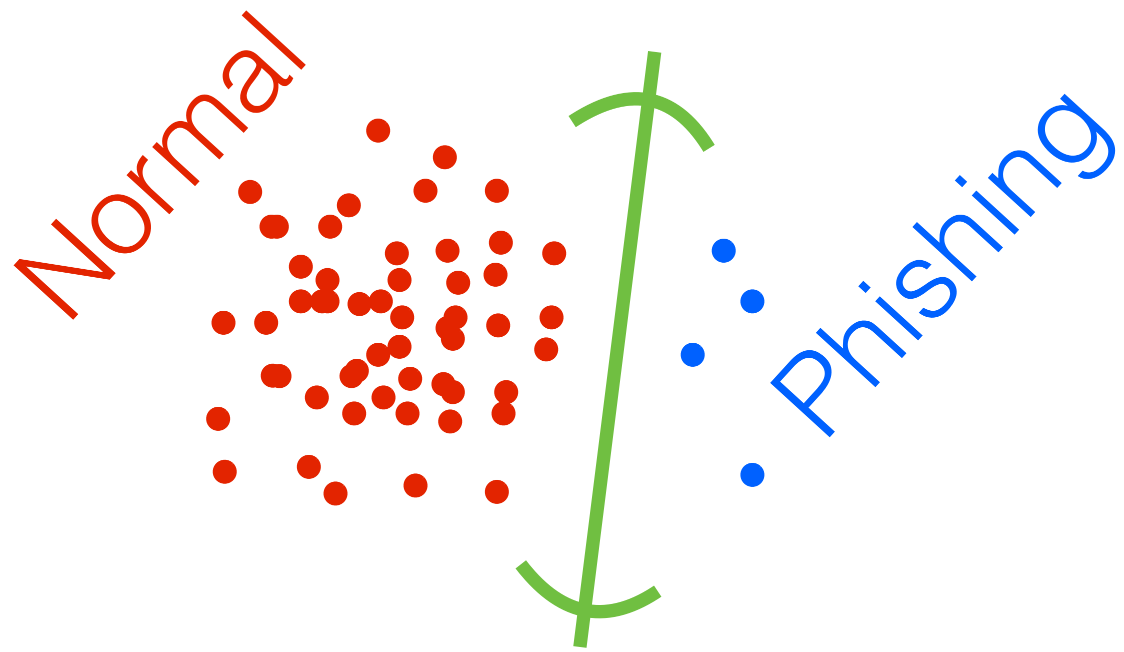


- Theoretical guarantees on quality
- Previous heuristics: data squashing, big data GPs
- Cf. subsampling
- How to develop coresets for Bayes?

[Agarwal et al 2005; Feldman & Langberg 2011; DuMouchel et al 1999; Madigan et al 2002; Huggins, Campbell, Broderick 2016; Campbell, Broderick 2017; Campbell, Broderick 2018]

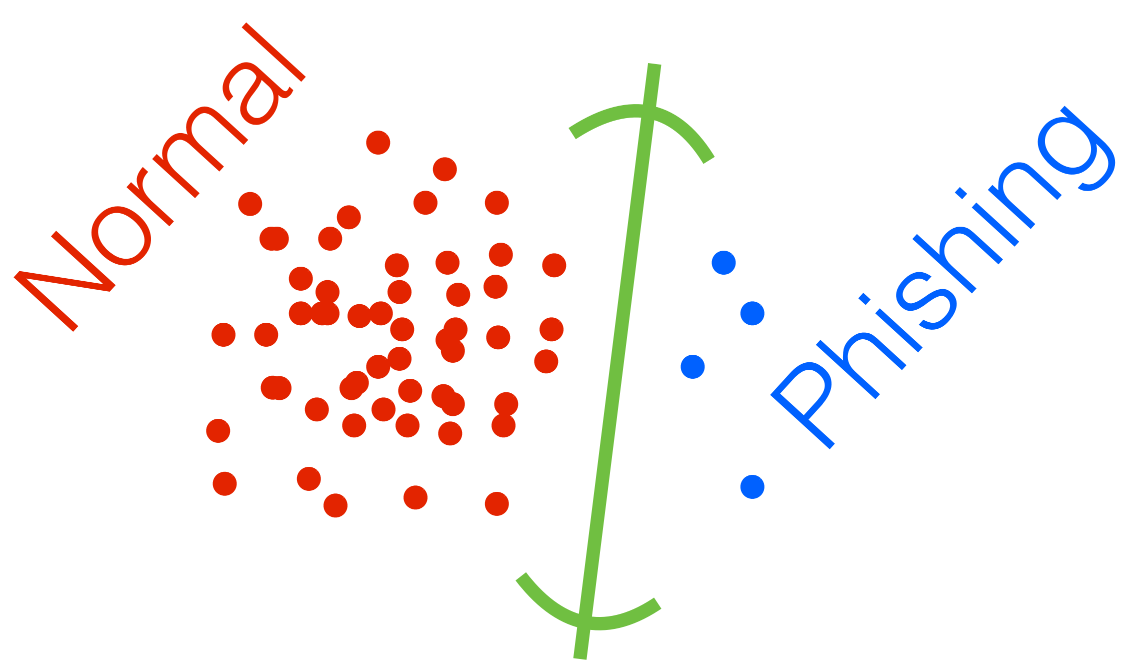
Bayesian coresets

- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$



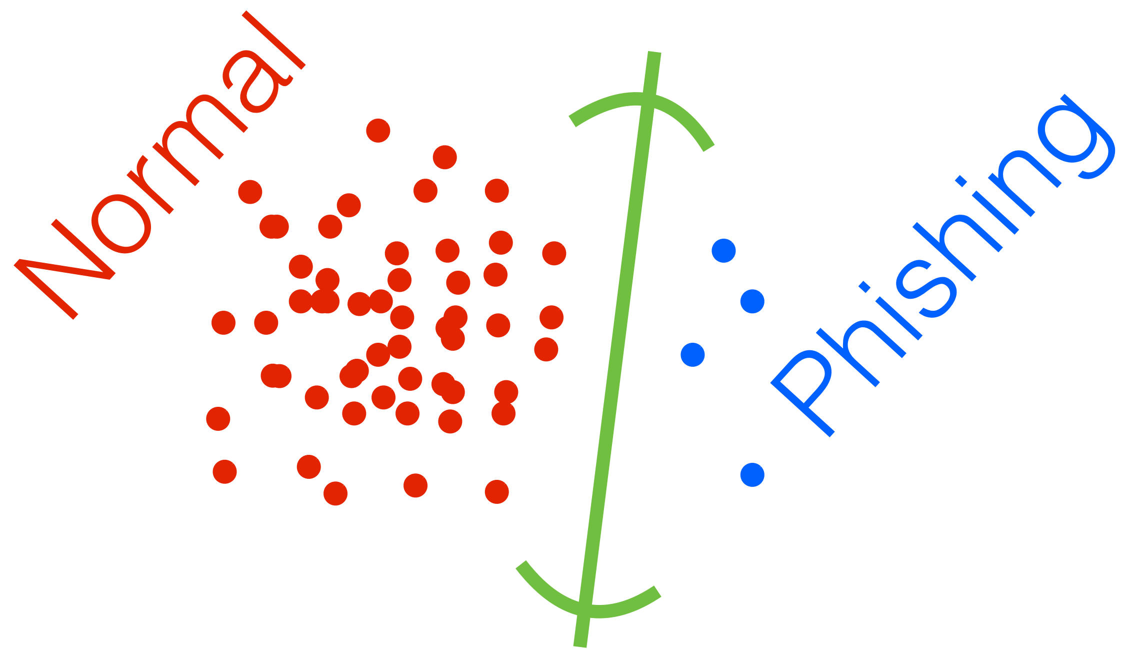
Bayesian coresets

- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,, \;\; \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$



Bayesian coresets

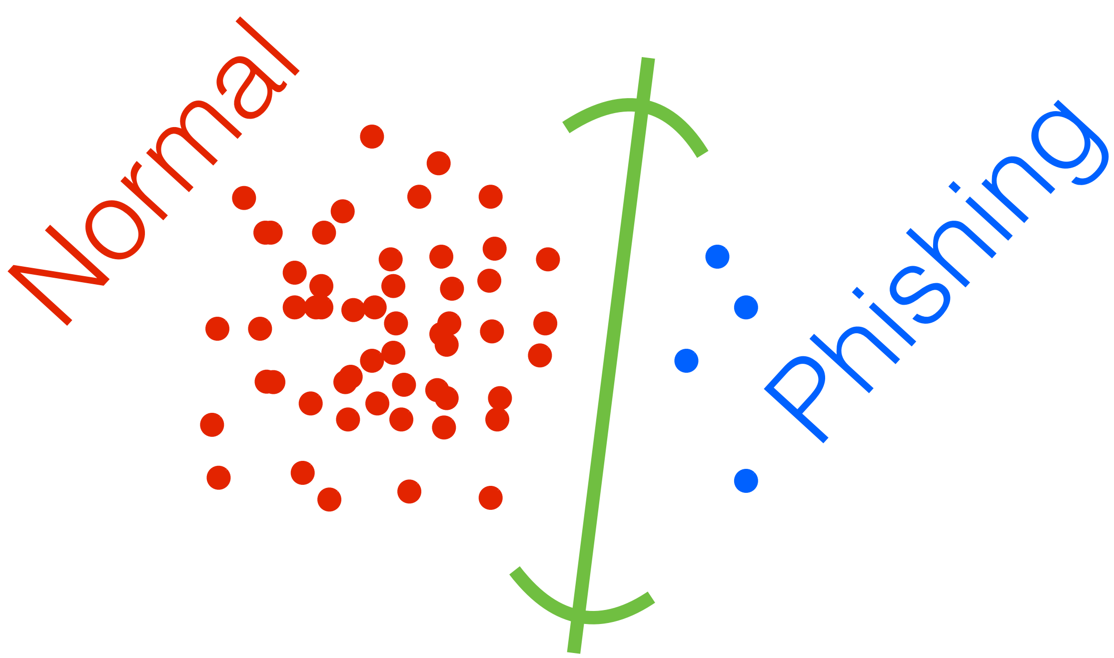
- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,, \;\; \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood



Bayesian coresets

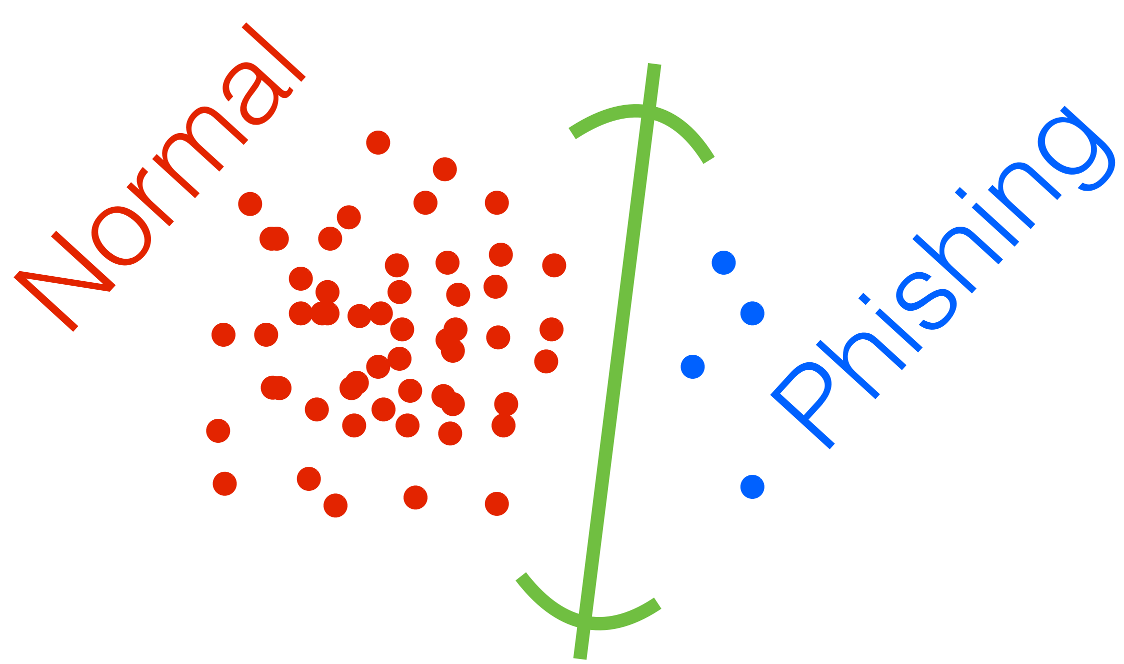
- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,, \;\; \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood

$$\|w\|_0 \ll N$$



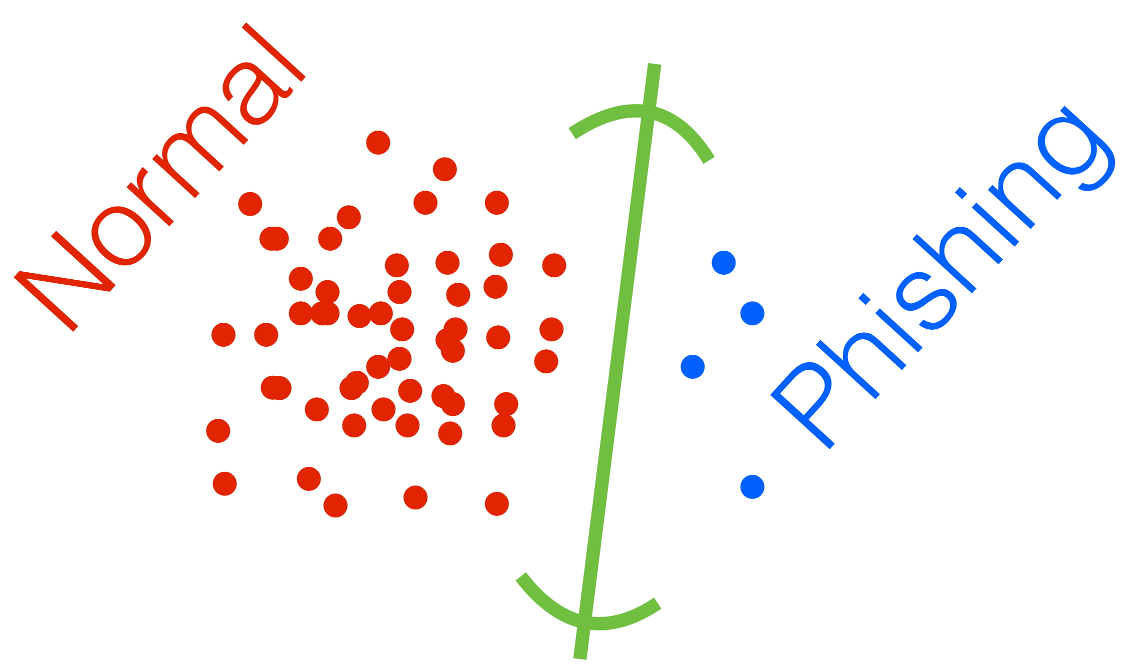
Bayesian coresets

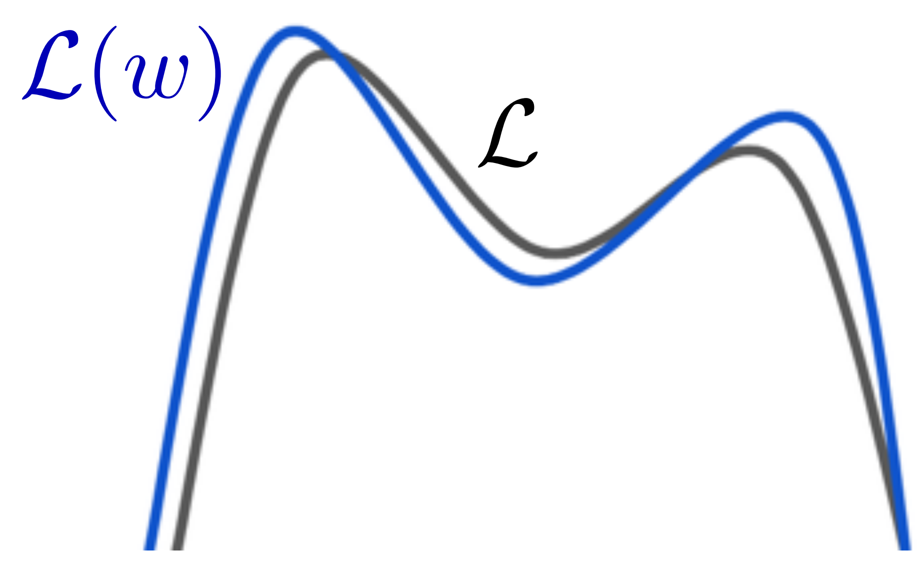
- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,,\;\; \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood $\mathcal{L}(w,\theta) := \sum_{n=1}^{N} w_n \mathcal{L}_n(\theta)$ s.t. $\|w\|_0 \ll N$



Bayesian coresets

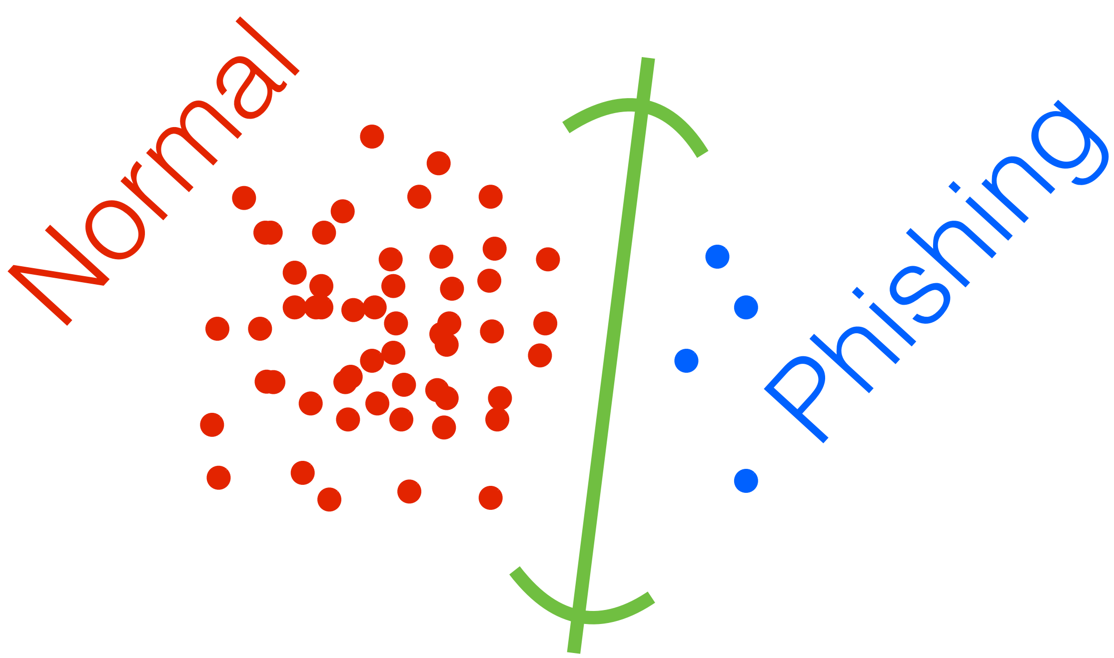
- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,, \quad \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood $\mathcal{L}(w,\theta) := \sum_{n=1}^{N} w_n \mathcal{L}_n(\theta)$ s.t. $\|w\|_0 \ll N$

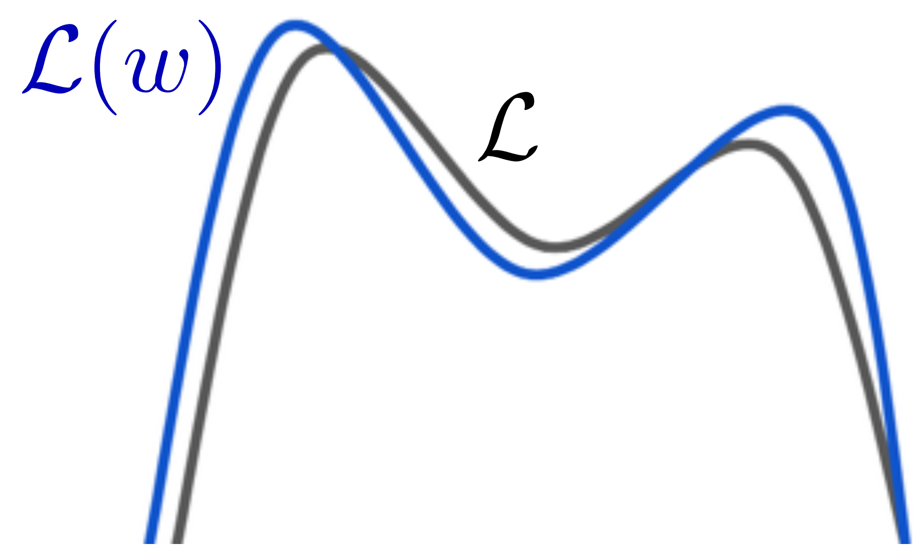




Bayesian coresets

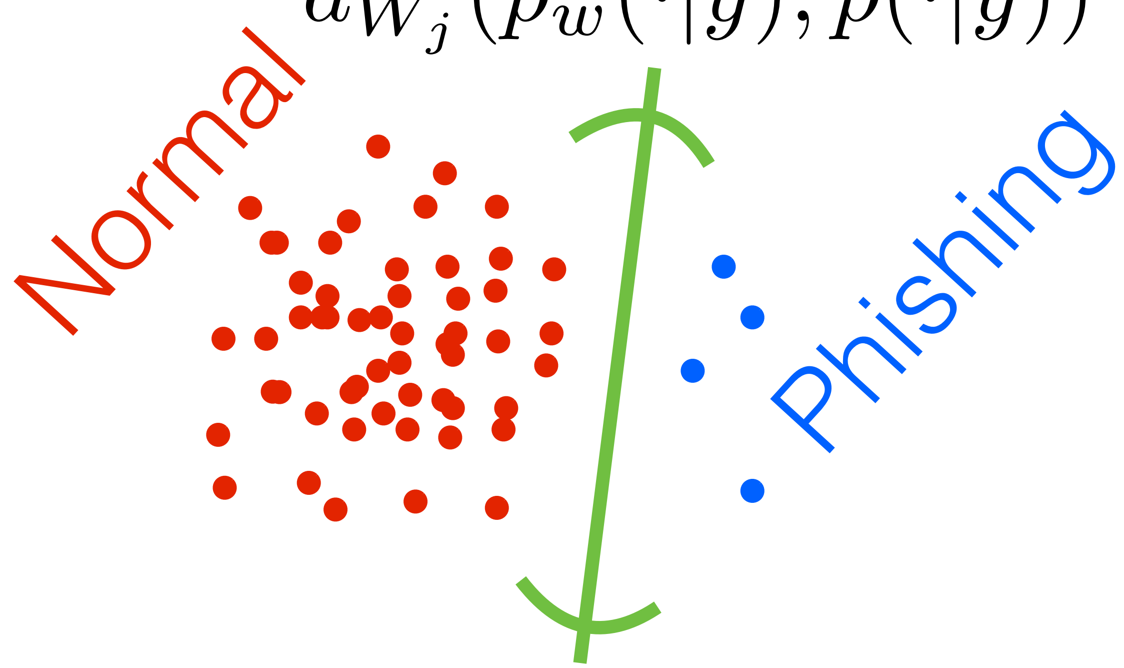
- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,, \quad \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood $\mathcal{L}(w,\theta) := \sum_{n=1}^{N} w_n \mathcal{L}_n(\theta)$ s.t. $\|w\|_0 \ll N$
- ε-coreset: $\|\mathcal{L}(w) - \mathcal{L}\| \leq \epsilon$

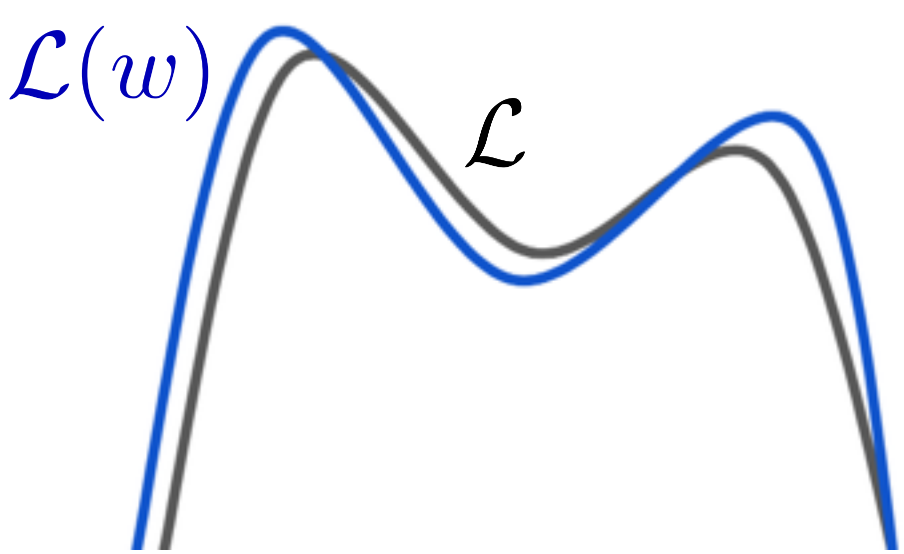




Bayesian coresets

- Posterior $p(\theta|y) \propto_\theta p(y|\theta)p(\theta)$
- Log likelihood $\mathcal{L}_n(\theta) := \log p(y_n|\theta)\,, \;\; \mathcal{L}(\theta) := \sum_{n=1}^{N} \mathcal{L}_n(\theta)$
- Coreset log likelihood $\mathcal{L}(w, \theta) := \sum_{n=1}^{N} w_n \mathcal{L}_n(\theta)$ s.t. $\|w\|_0 \ll N$
- ε-coreset: $\|\mathcal{L}(w) - \mathcal{L}\| \leq \epsilon$
    - Approximate posterior close in Wasserstein distance
      $d_{W_j}(p_w(\cdot|y), p(\cdot|y)) \leq C_j \|\mathcal{L}(w) - \mathcal{L}\|_{\mathrm{WFID}}, j \in \{1, 2\}$





[Huggins, Kasprzak, Campbell, Broderick, forthcoming]

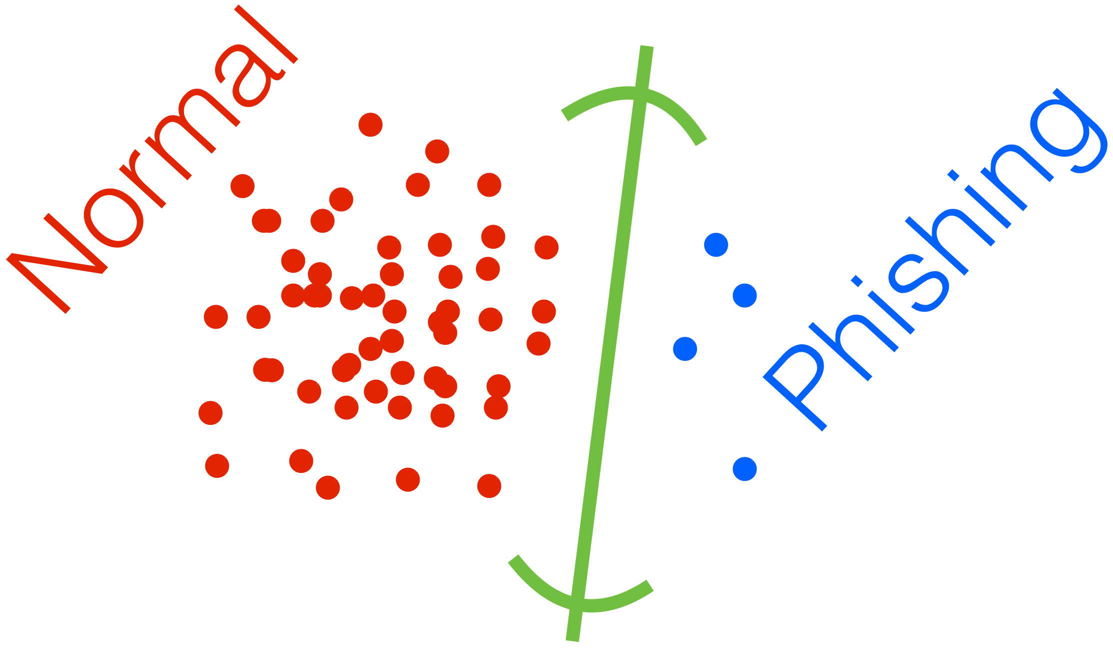
Roadmap

- Approximate Bayes review
- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

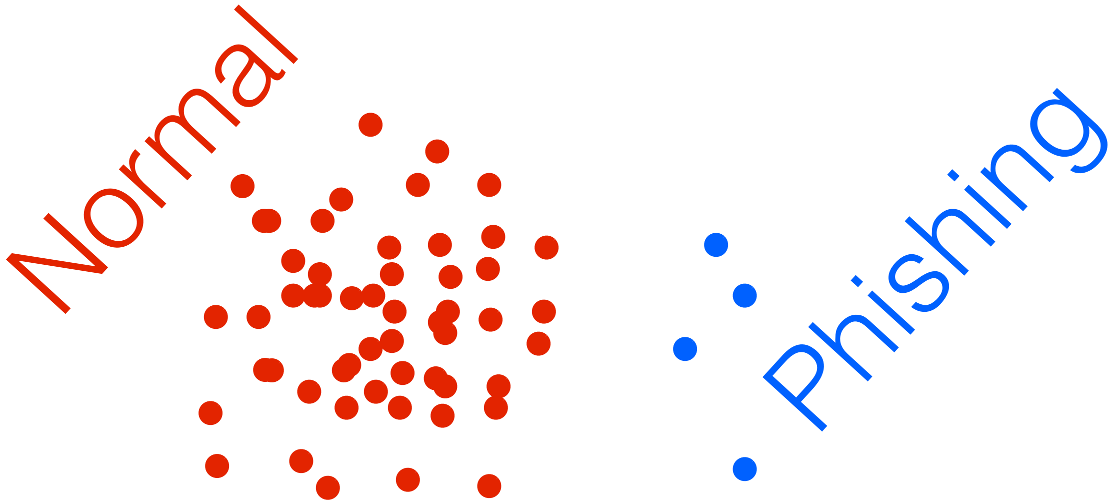
Roadmap

- Approximate Bayes review
- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

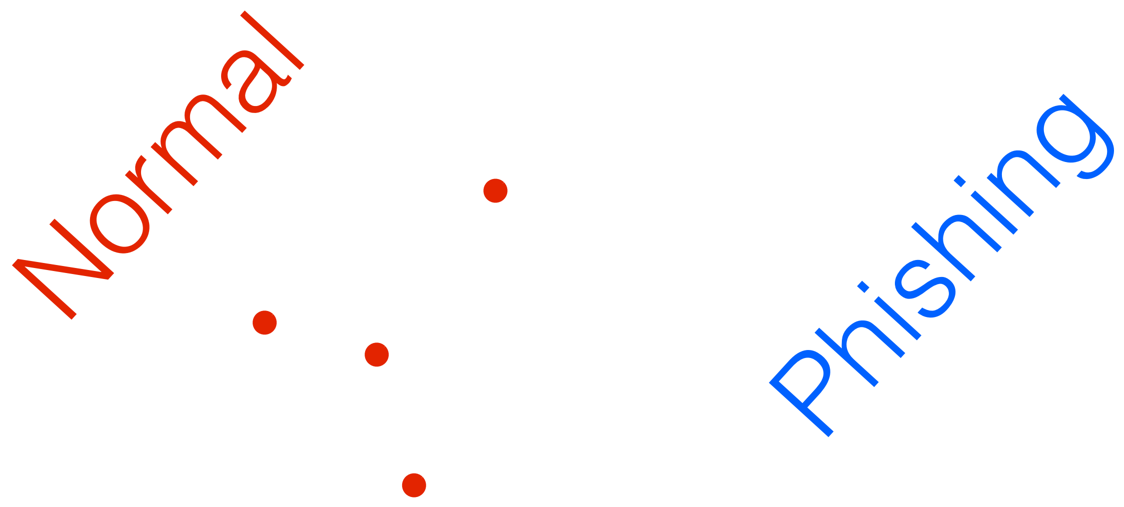
Uniform subsampling revisited



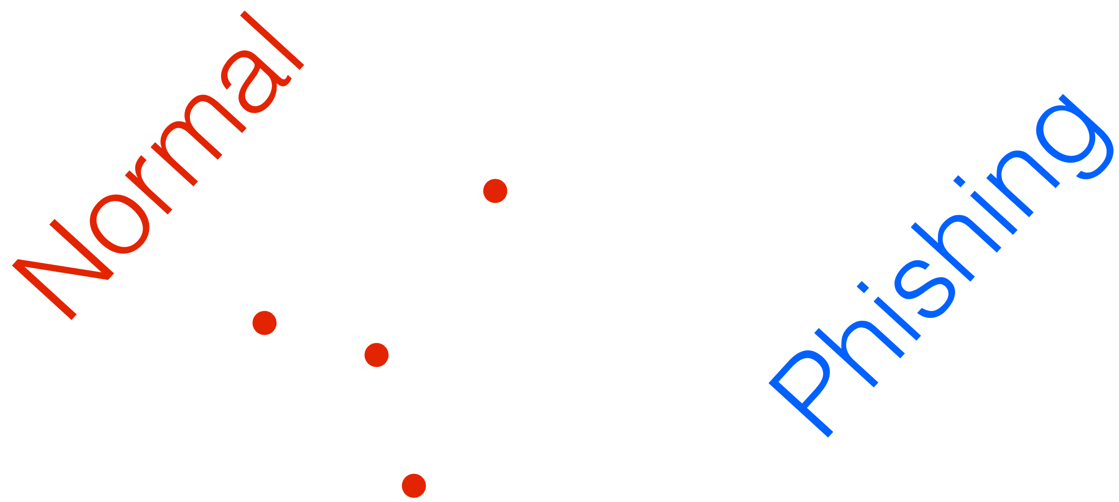
Uniform subsampling revisited



Uniform subsampling revisited

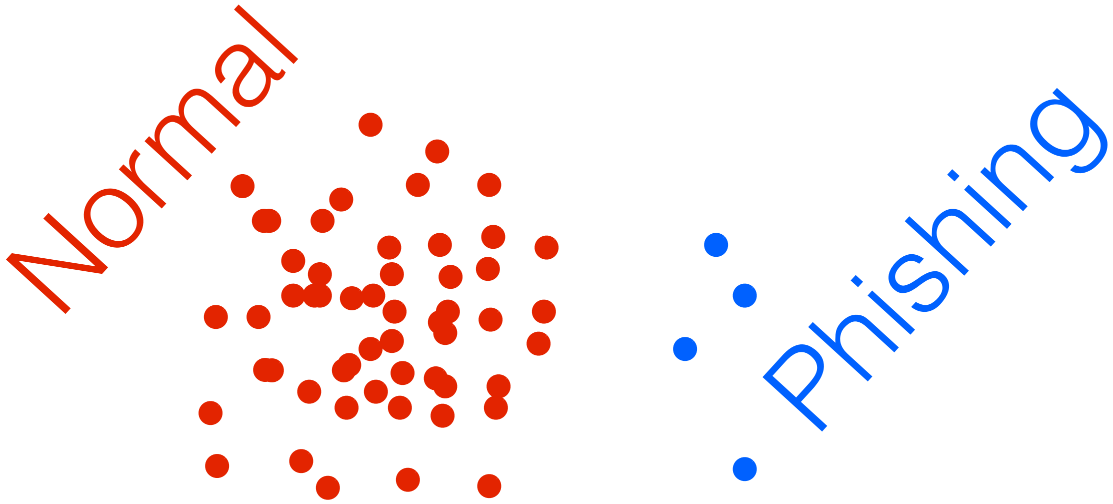


Uniform subsampling revisited



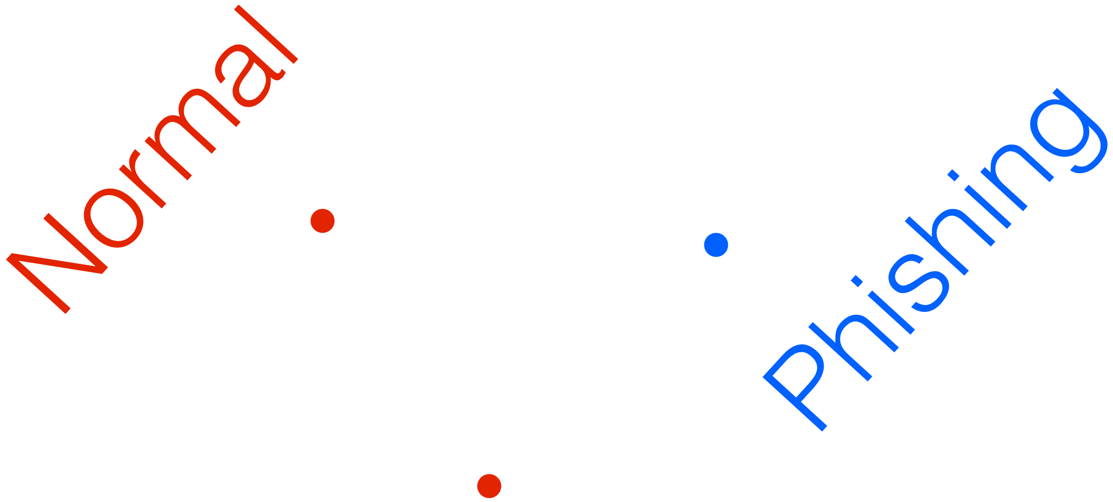
- Might miss important data

Uniform subsampling revisited



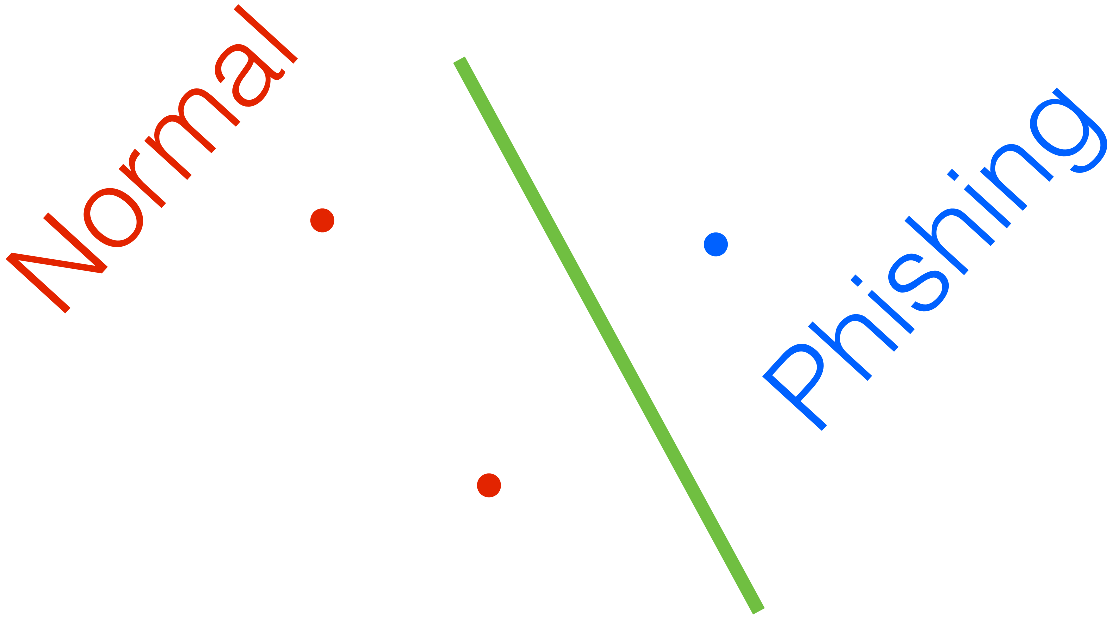
- Might miss important data

Uniform subsampling revisited



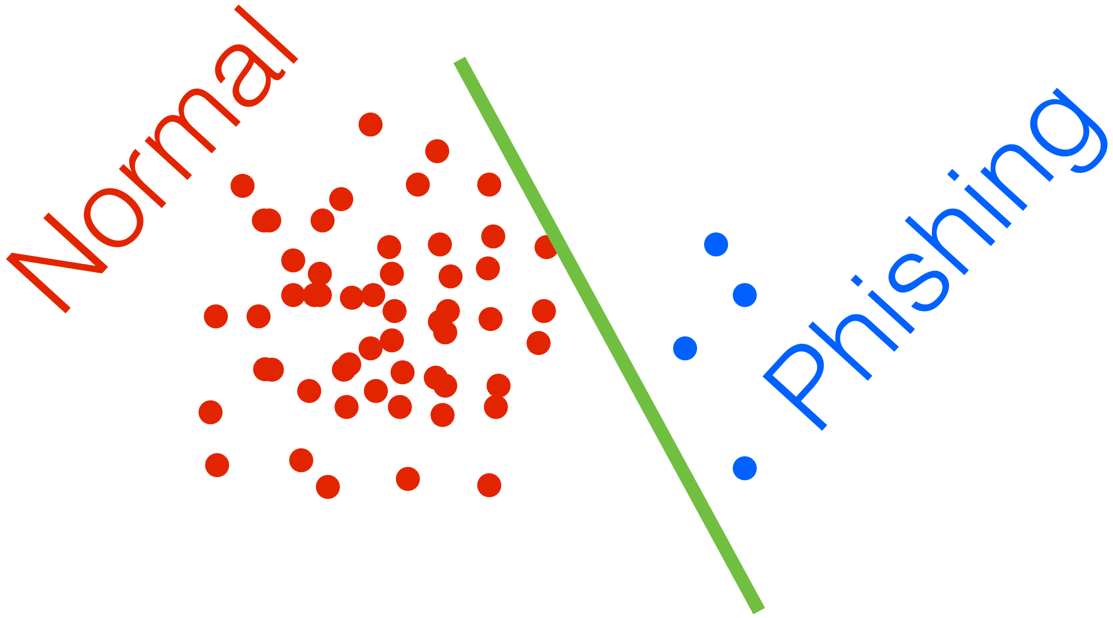
- Might miss important data

Uniform subsampling revisited



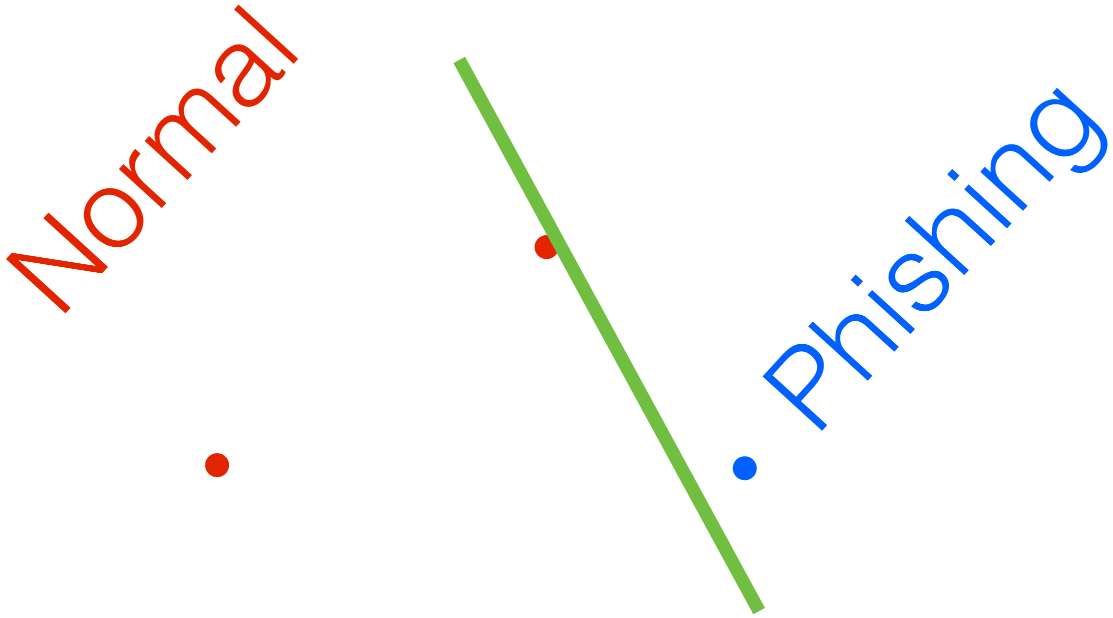
- Might miss important data

Uniform subsampling revisited



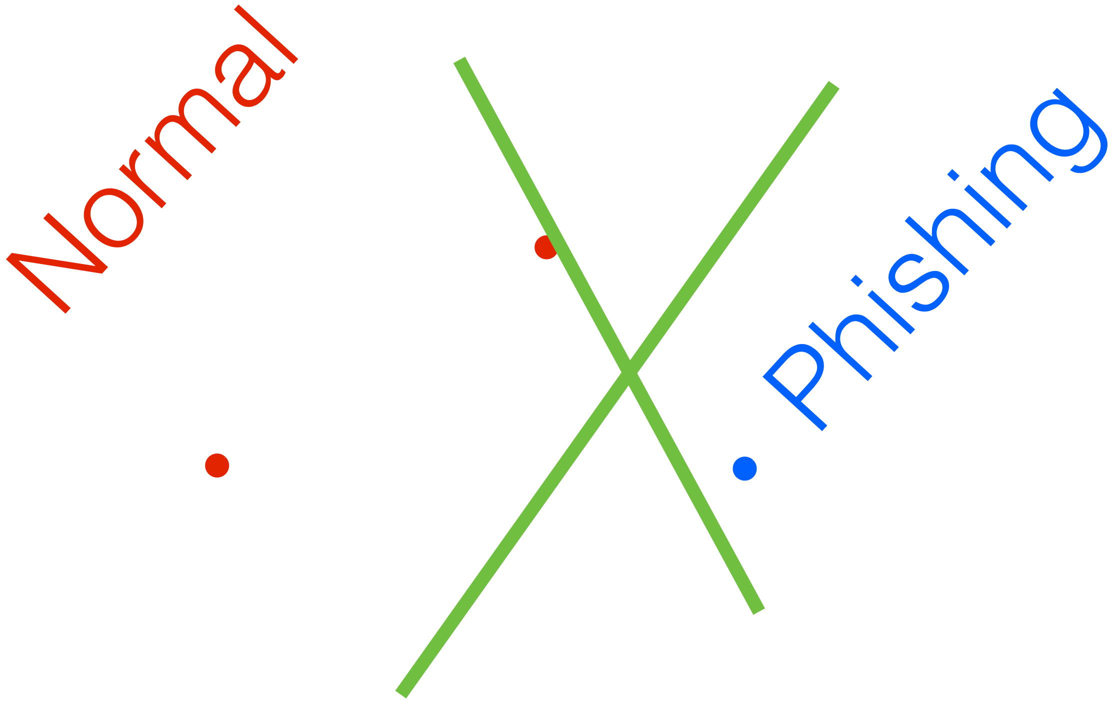
- Might miss important data

Uniform subsampling revisited



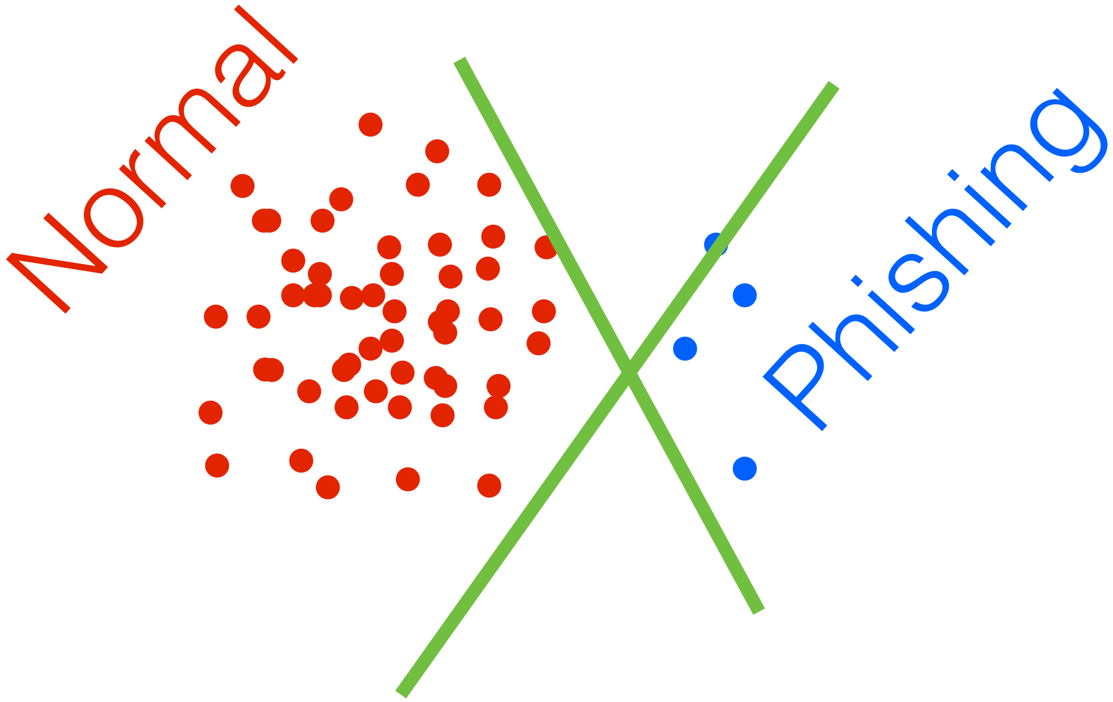
- Might miss important data

Uniform subsampling revisited



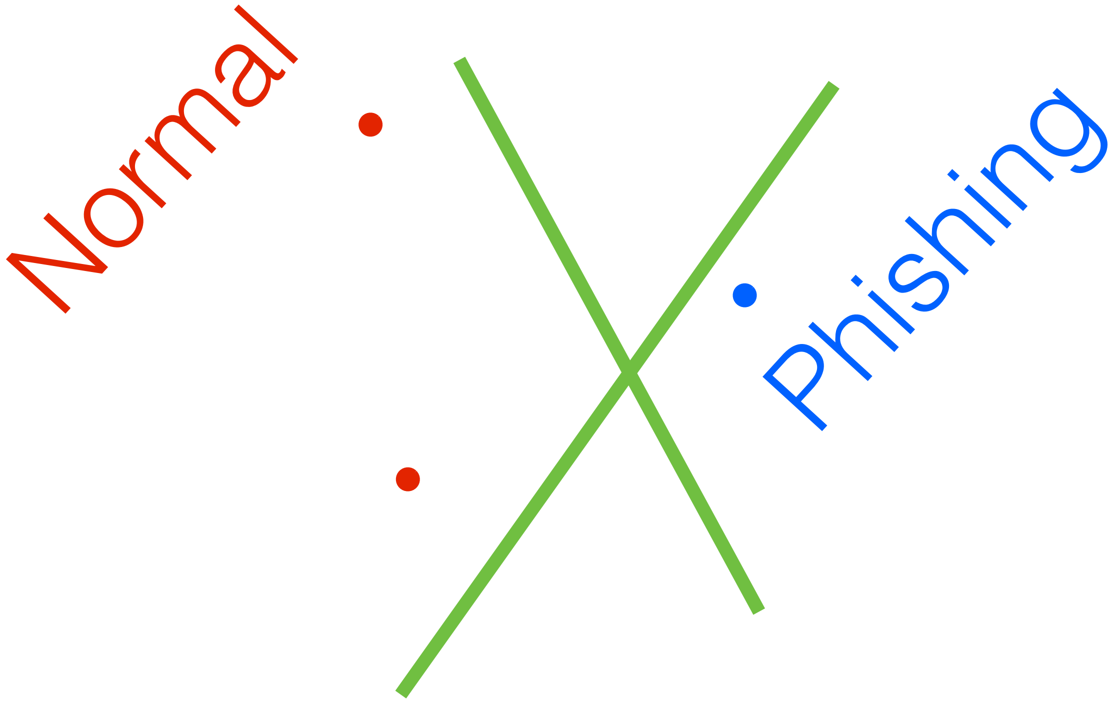
- Might miss important data

Uniform subsampling revisited



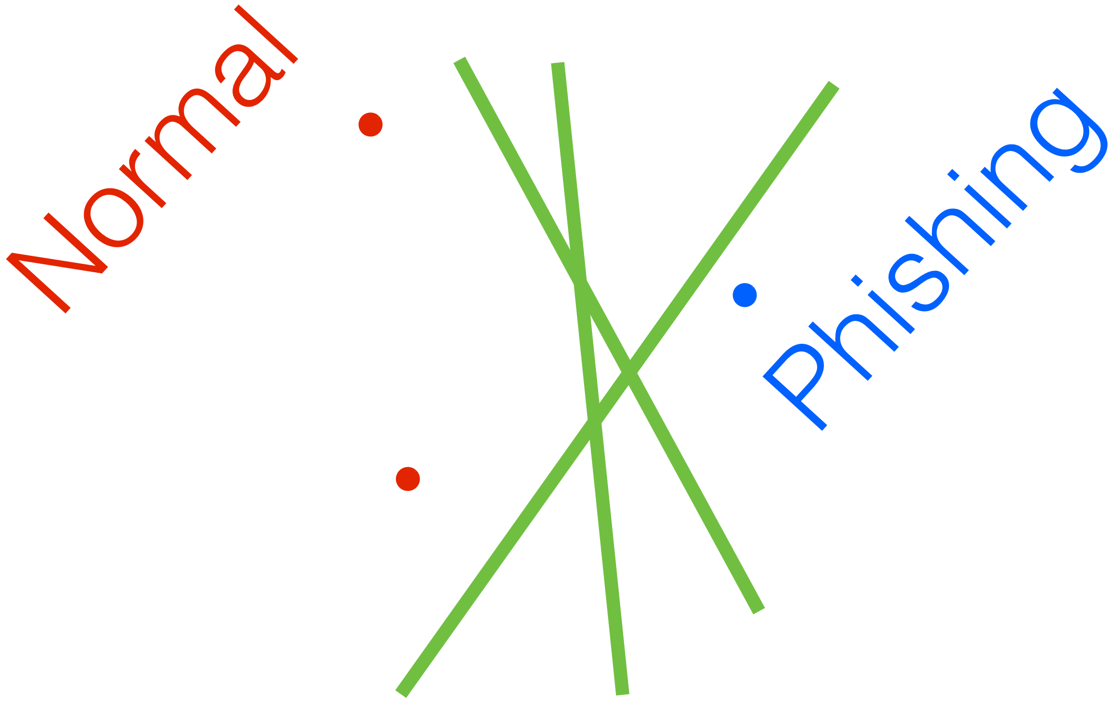
- Might miss important data

Uniform subsampling revisited



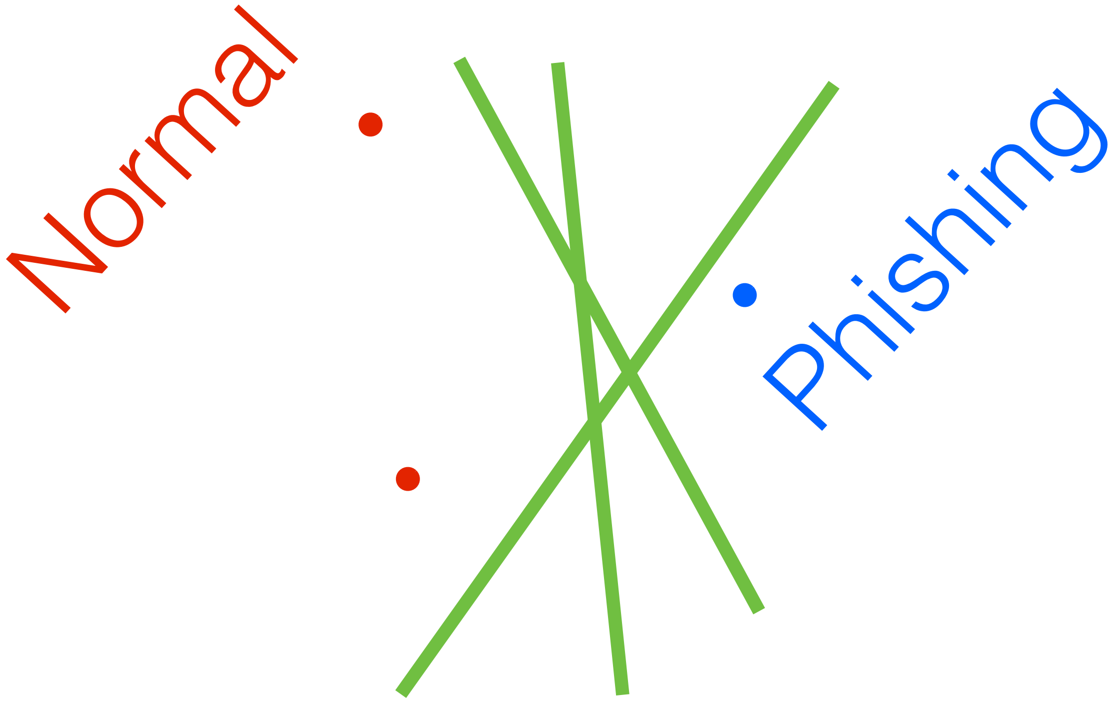
- Might miss important data

Uniform subsampling revisited



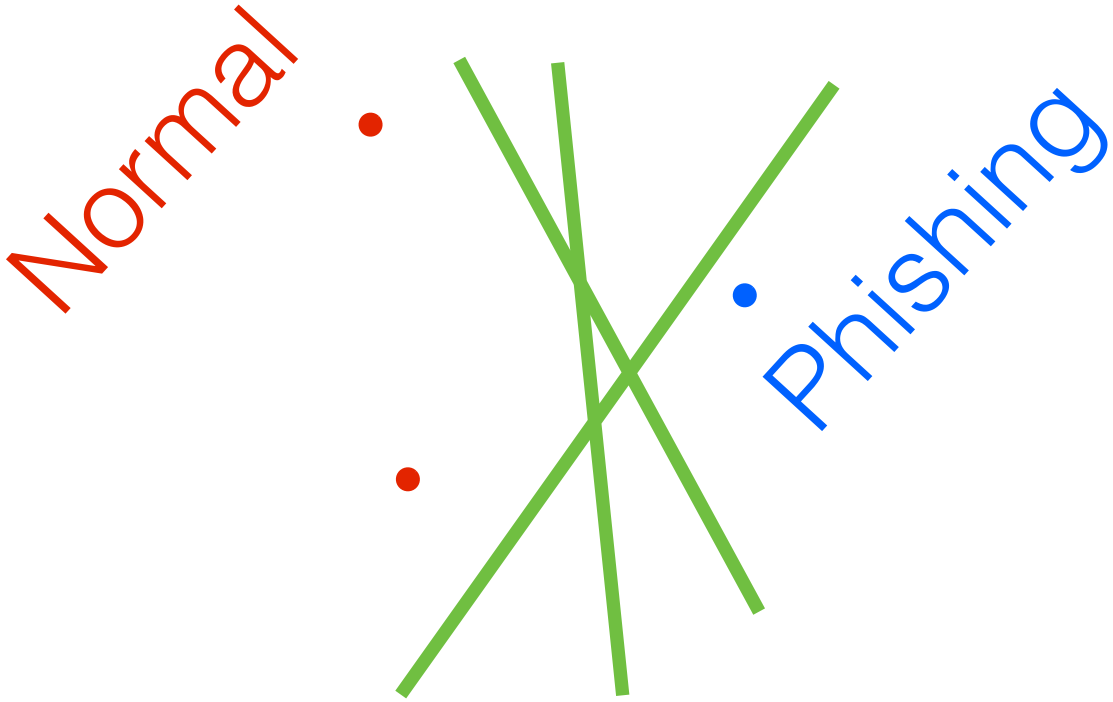
- Might miss important data

Uniform subsampling revisited

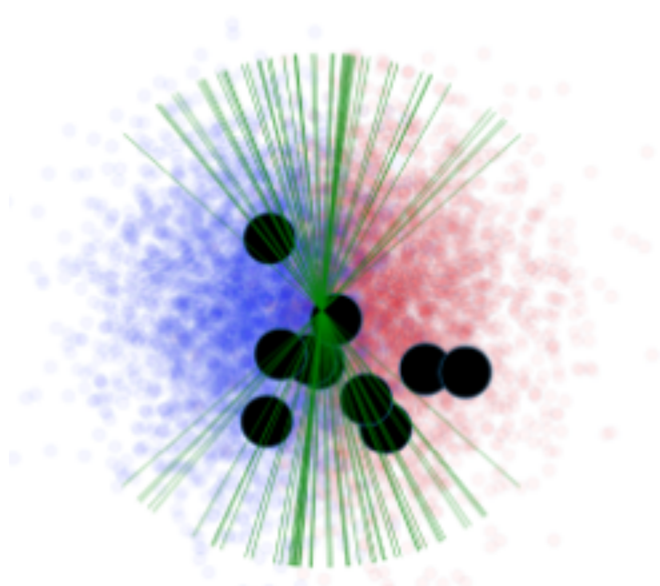


- Might miss important data
- Noisy estimates

Uniform subsampling revisited

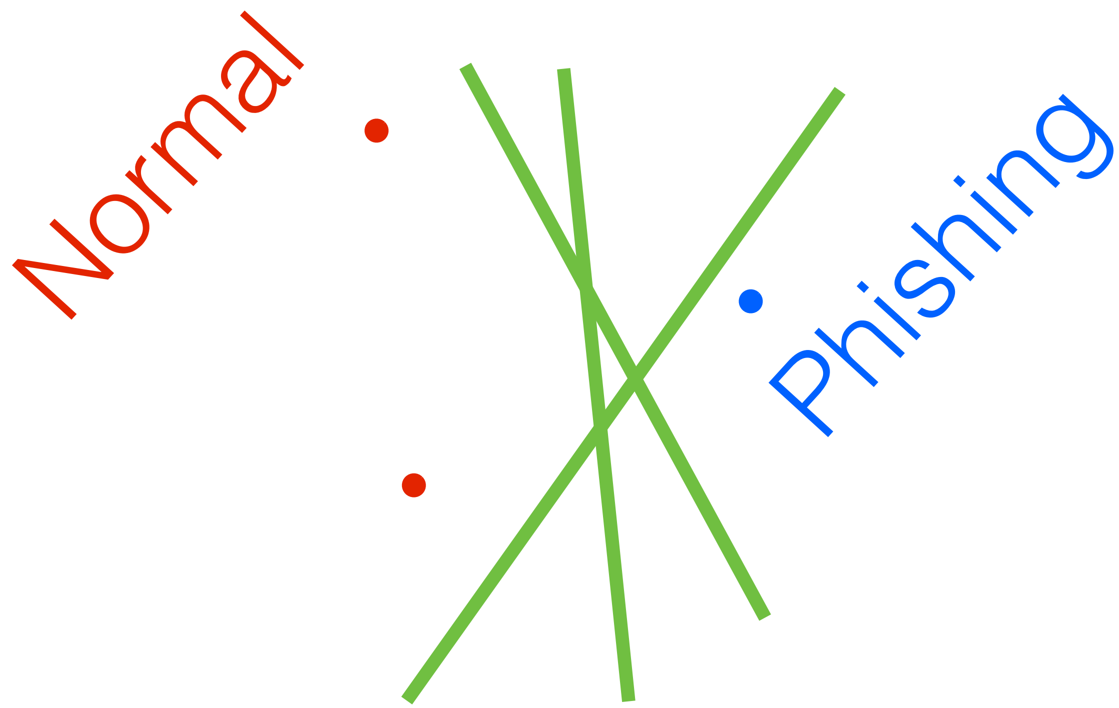


- Might miss important data
- Noisy estimates

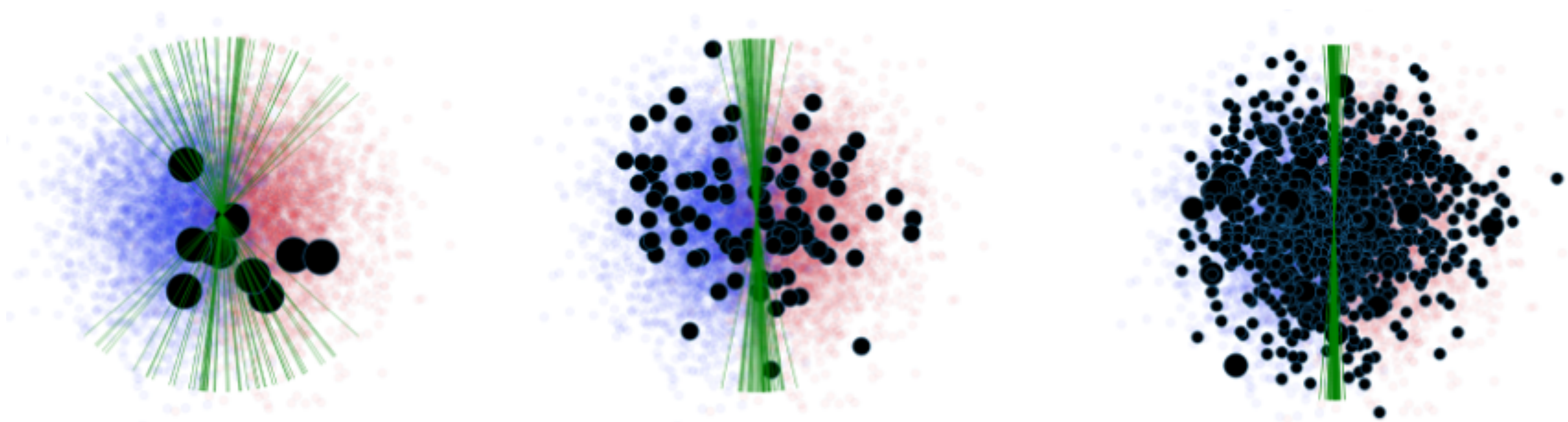


$M = 10$

Uniform subsampling revisited



- Might miss important data
- Noisy estimates



$M = 10$ $M = 100$ $M = 1000$

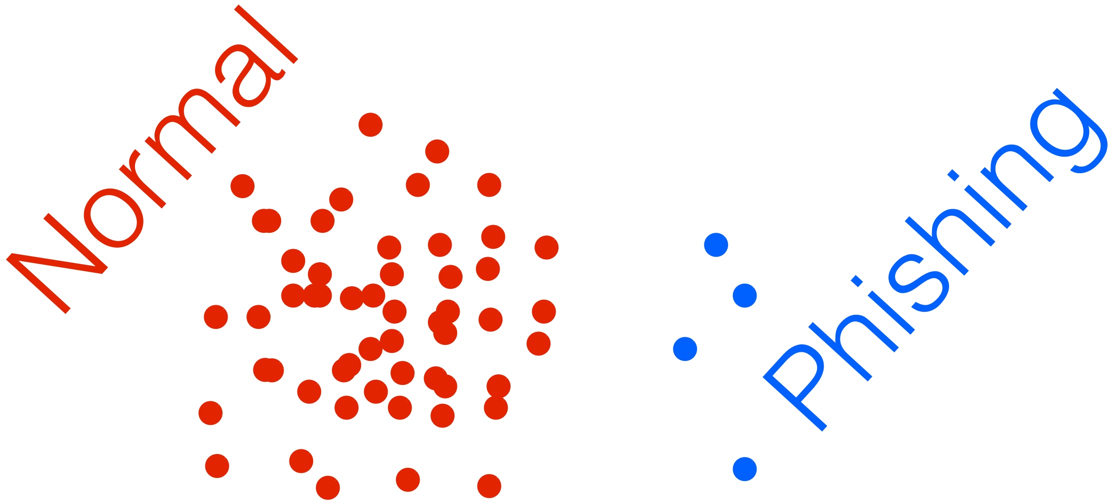
Roadmap

- Approximate Bayes review
- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

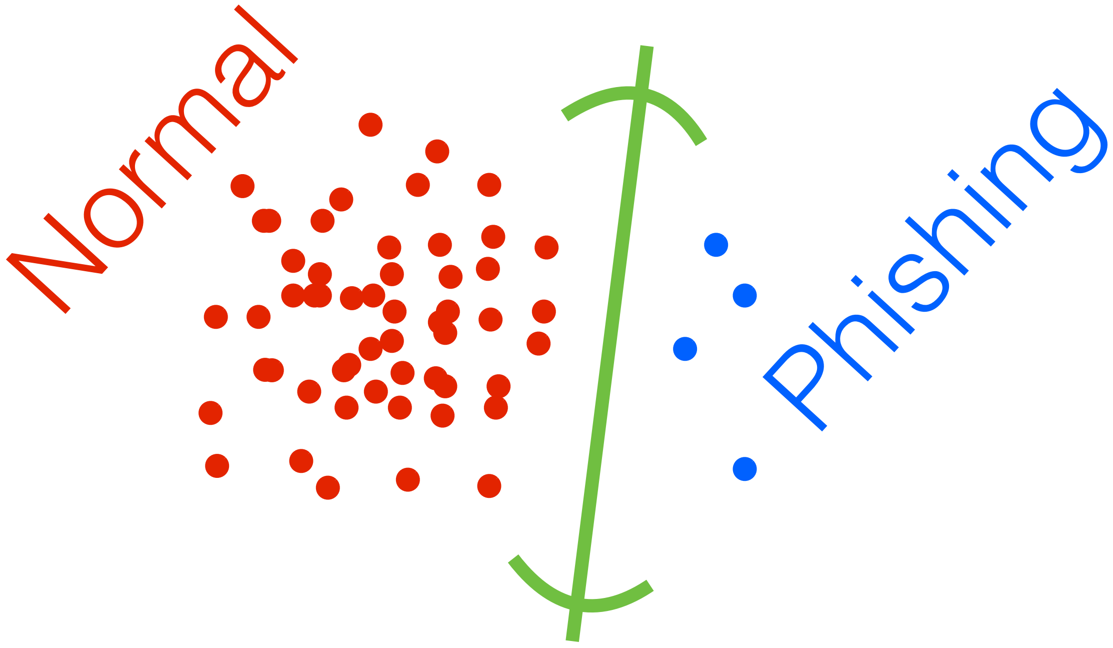
Roadmap

- Approximate Bayes review
- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

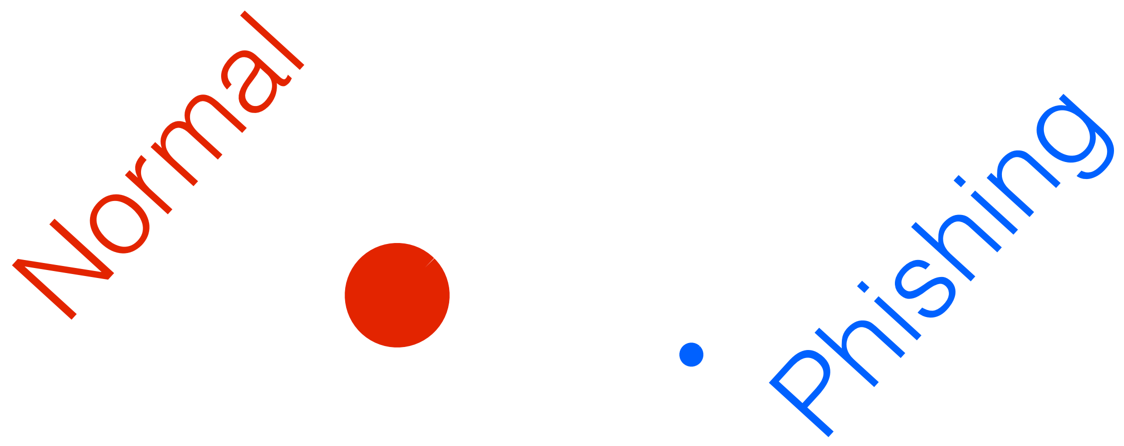
Importance sampling



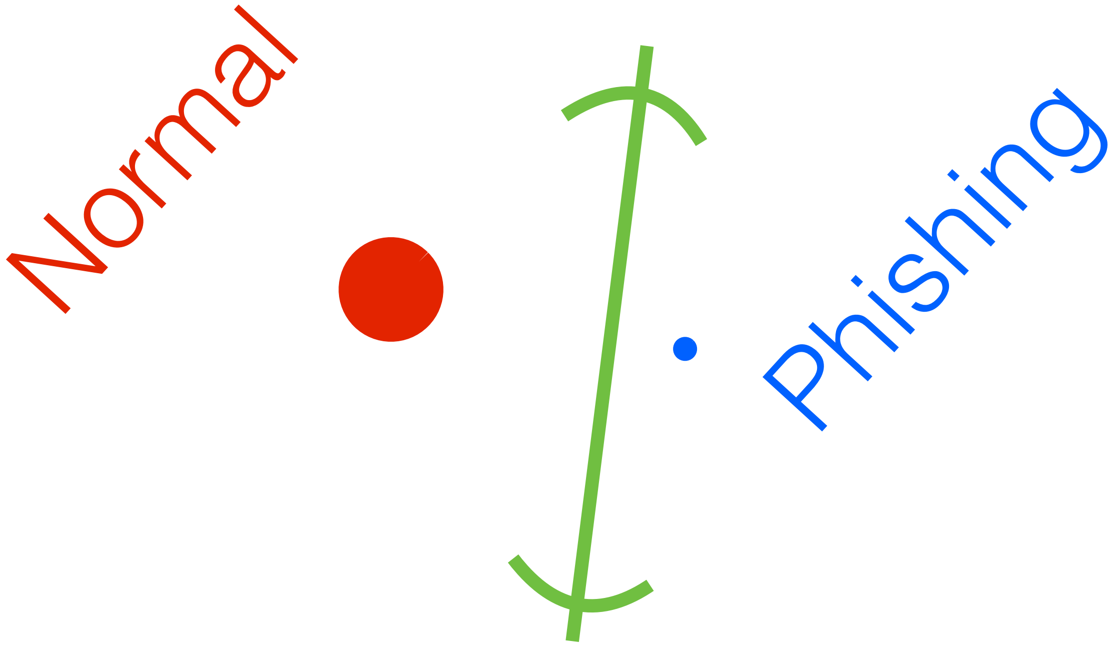
Importance sampling



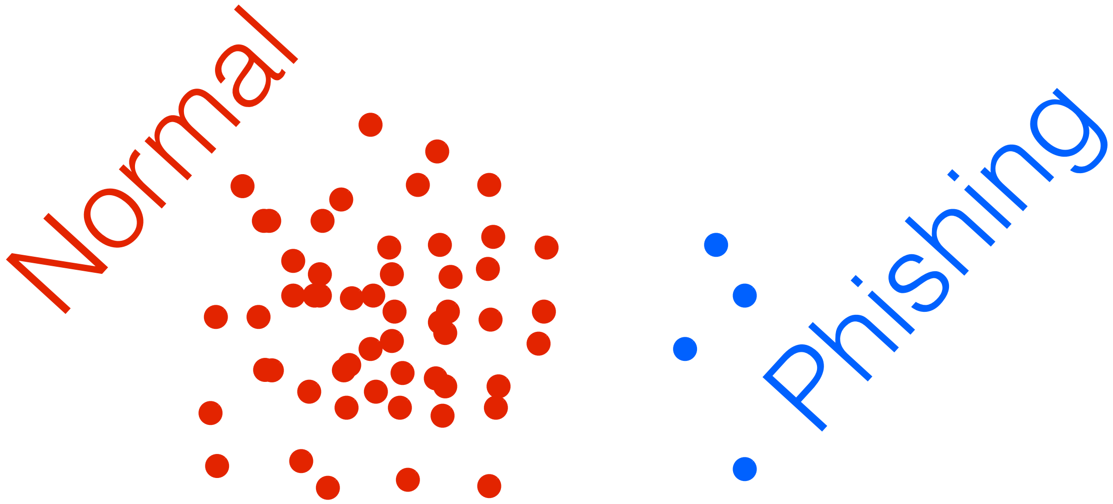
Importance sampling



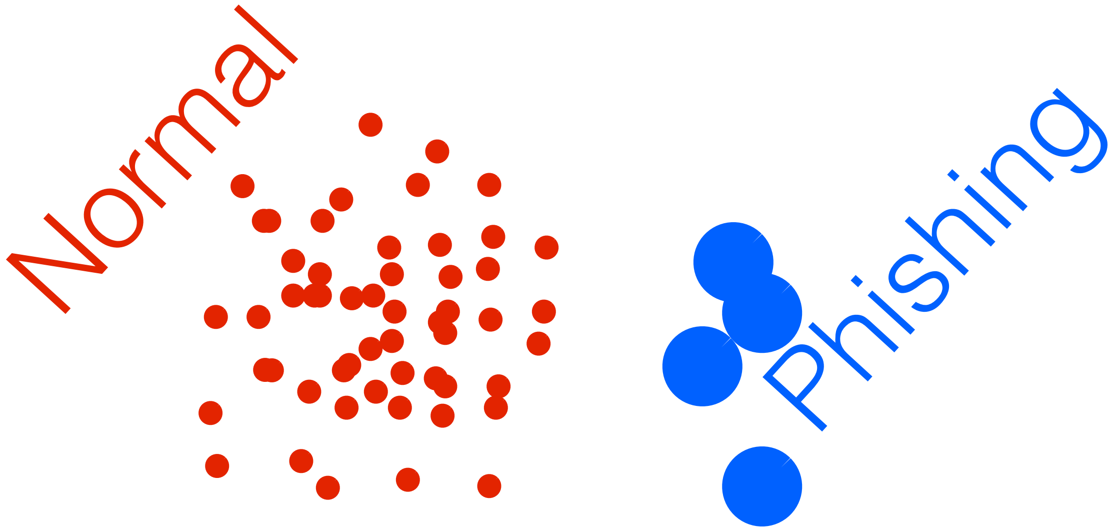
Importance sampling



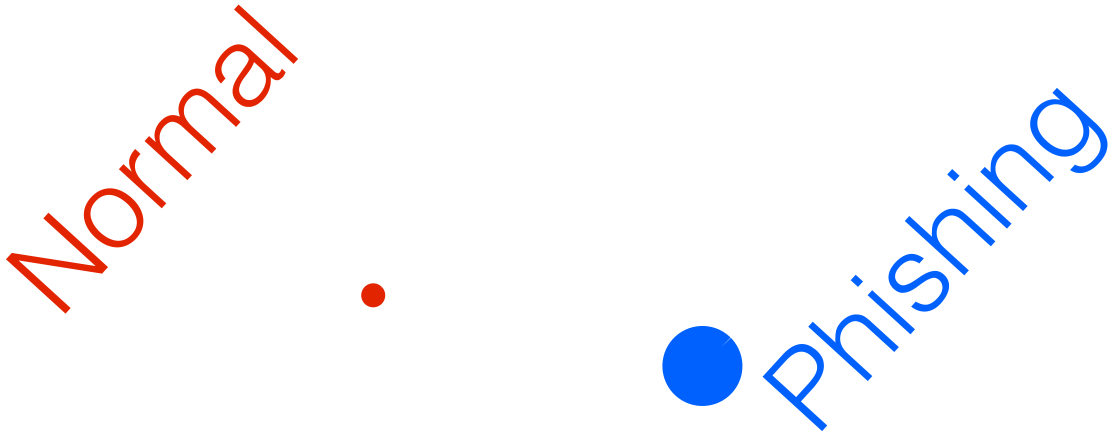
Importance sampling



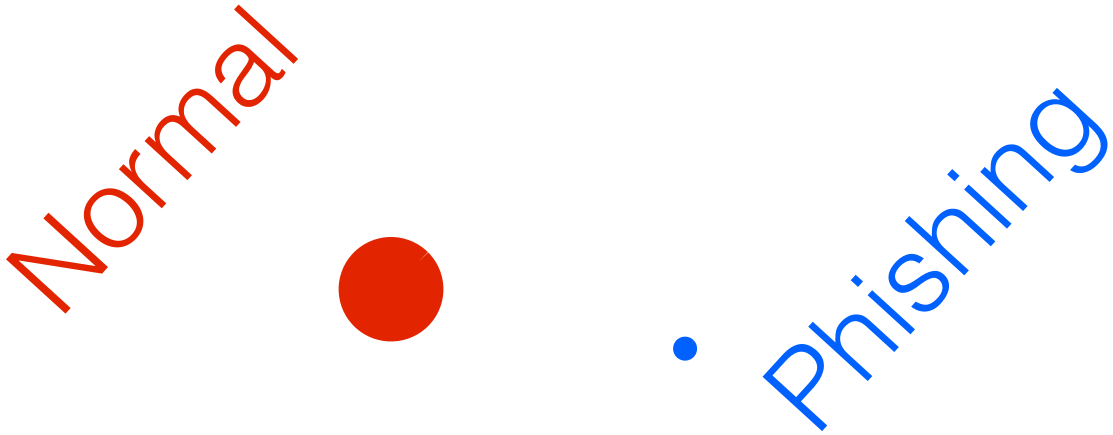
Importance sampling



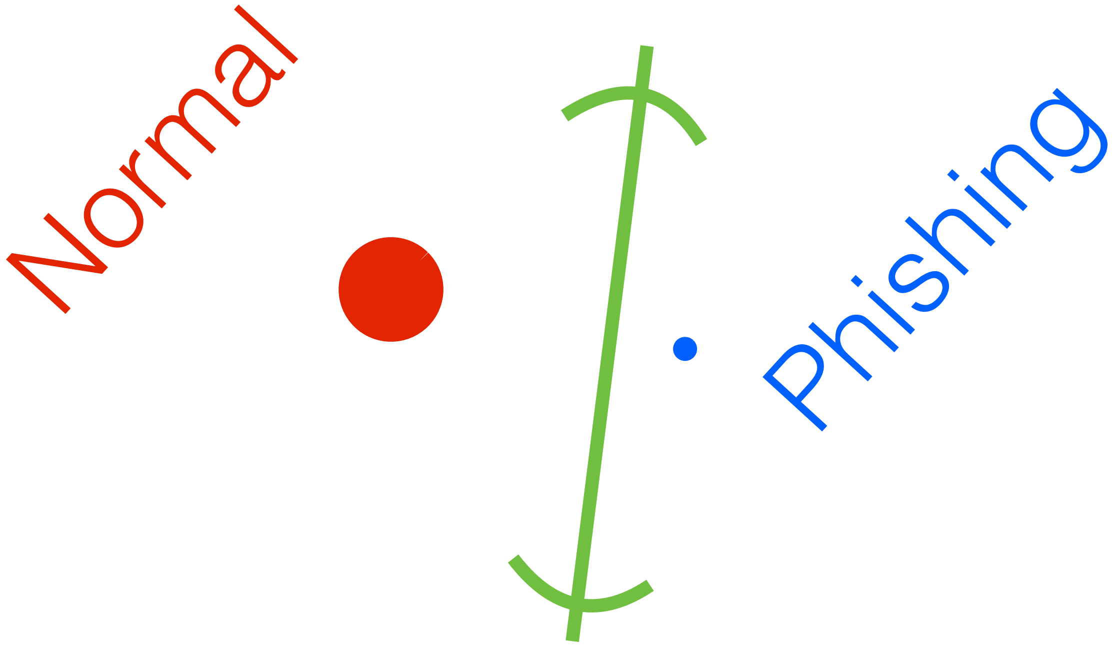
Importance sampling



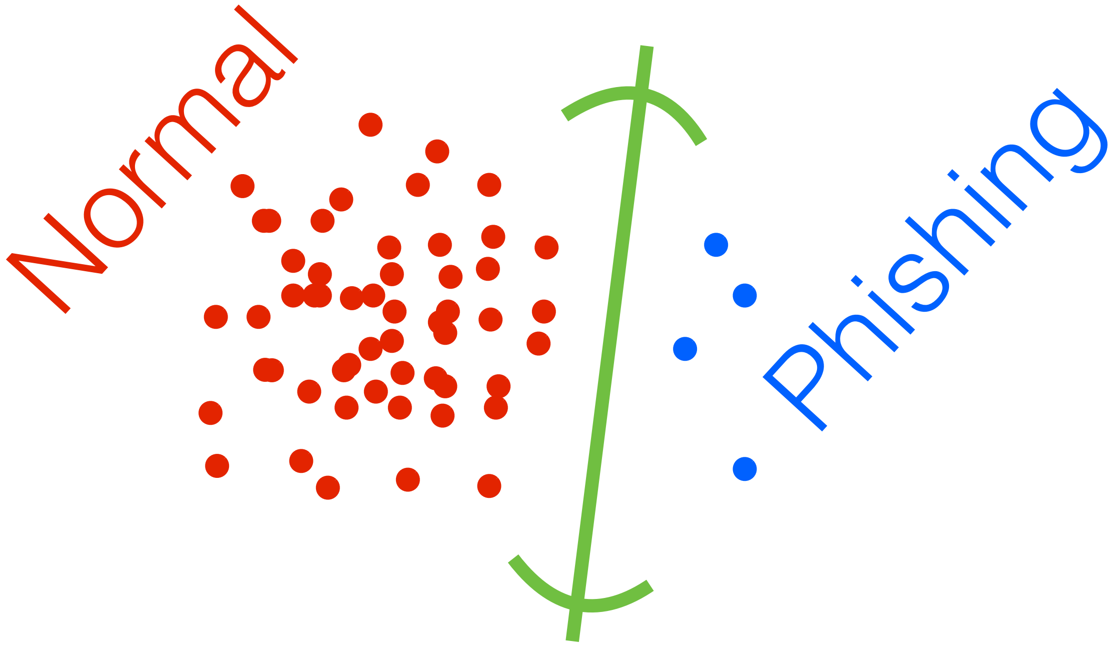
Importance sampling



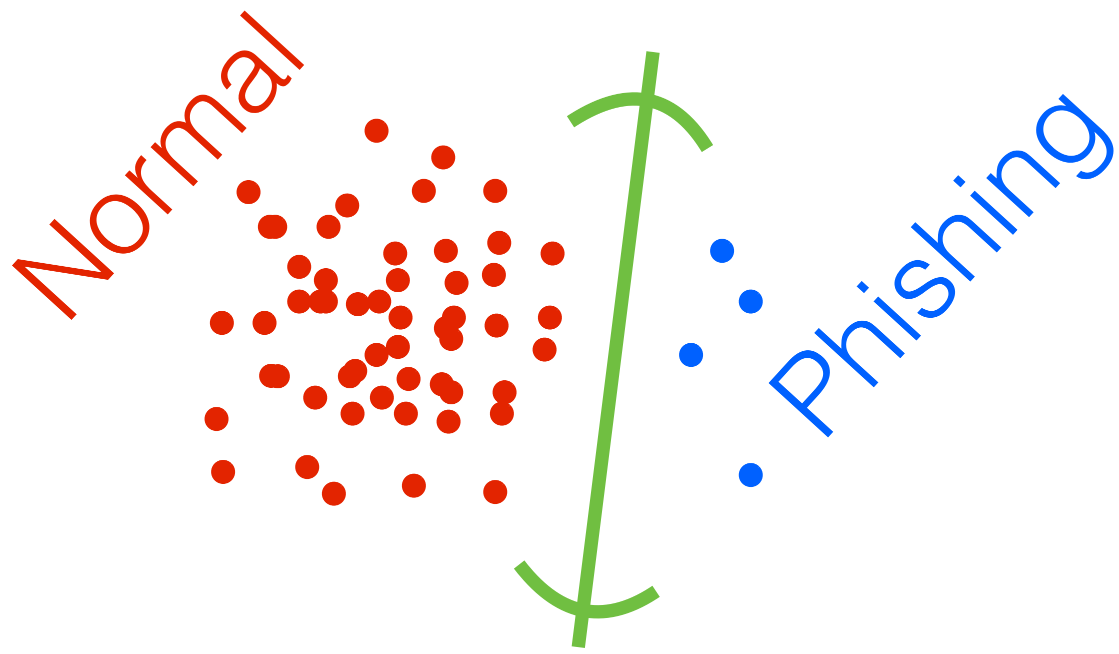
Importance sampling



Importance sampling

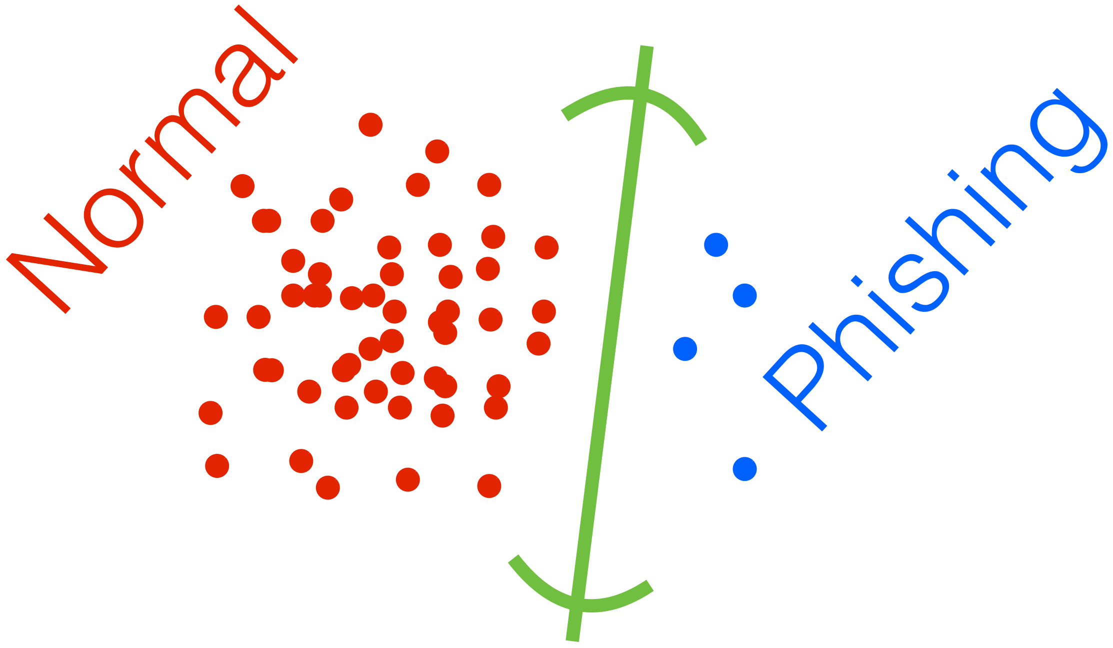


Importance sampling



$$\sigma_n \propto \|\mathcal{L}_n\|$$

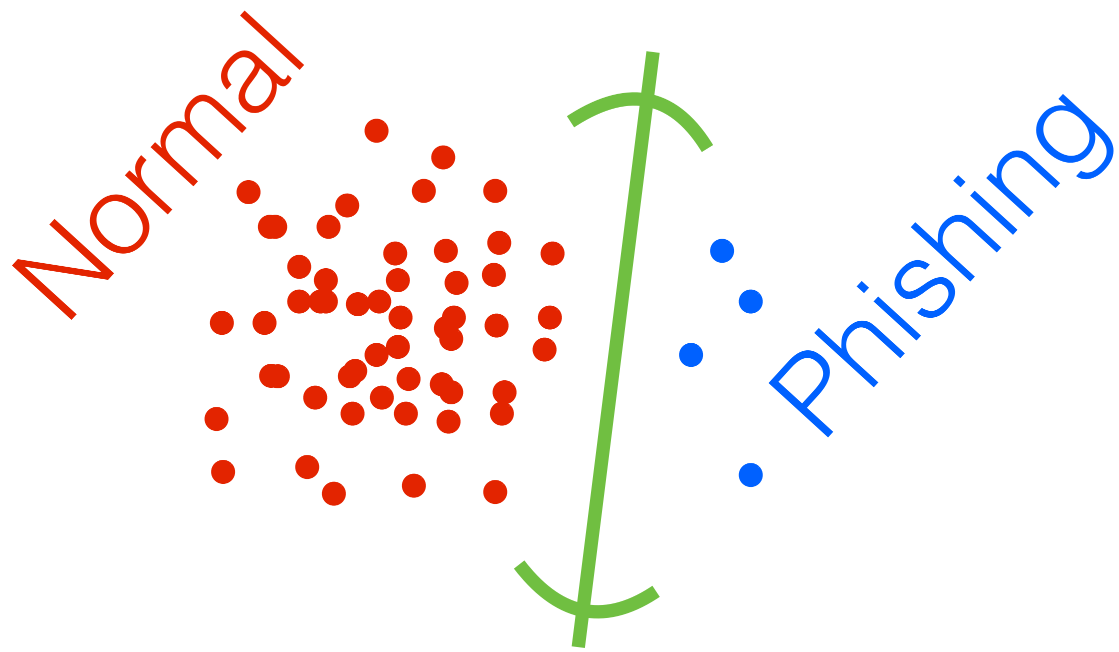
Importance sampling



$$\sigma := \sum_{n=1}^{N} \|\mathcal{L}_n\|$$

$$\sigma_n := \|\mathcal{L}_n\| / \sigma$$

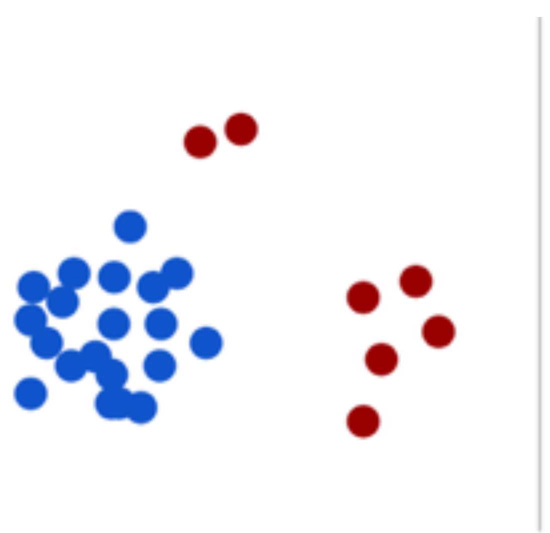
Importance sampling



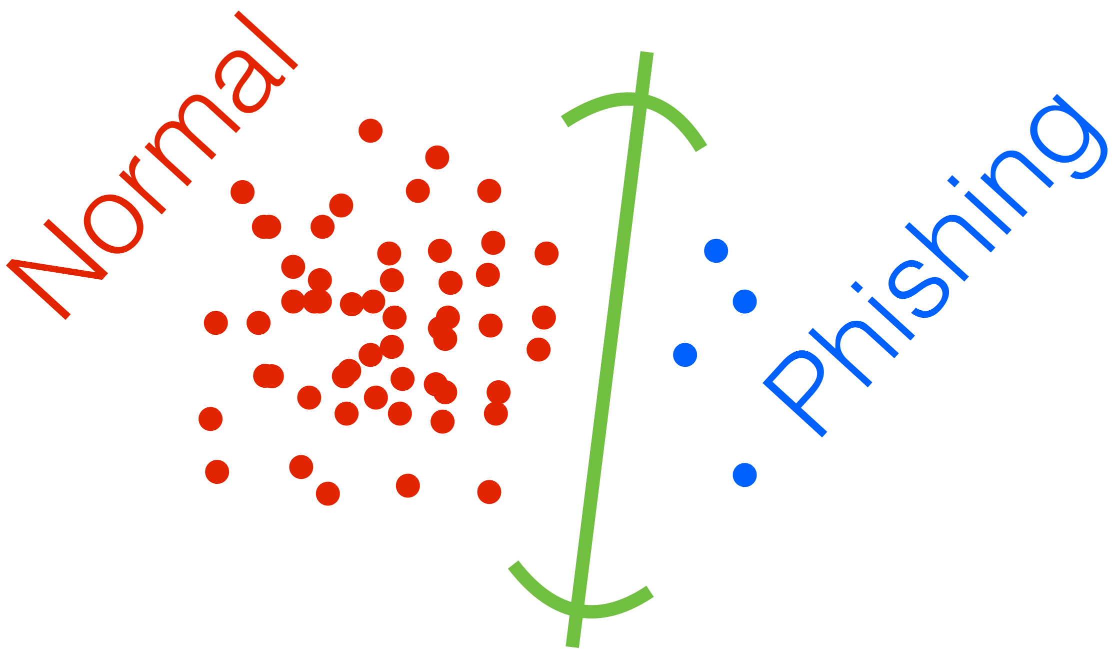
$$\sigma := \sum_{n=1}^{N} \|\mathcal{L}_n\|$$

$$\sigma_n := \|\mathcal{L}_n\| / \sigma$$

1. data

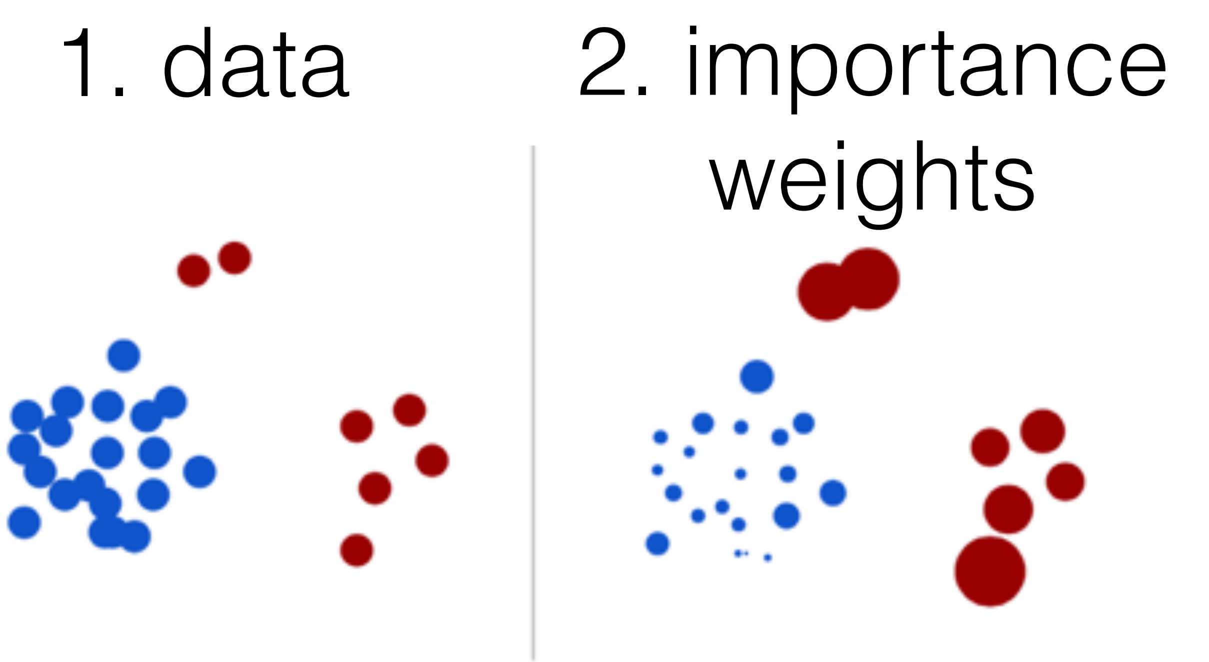


Importance sampling

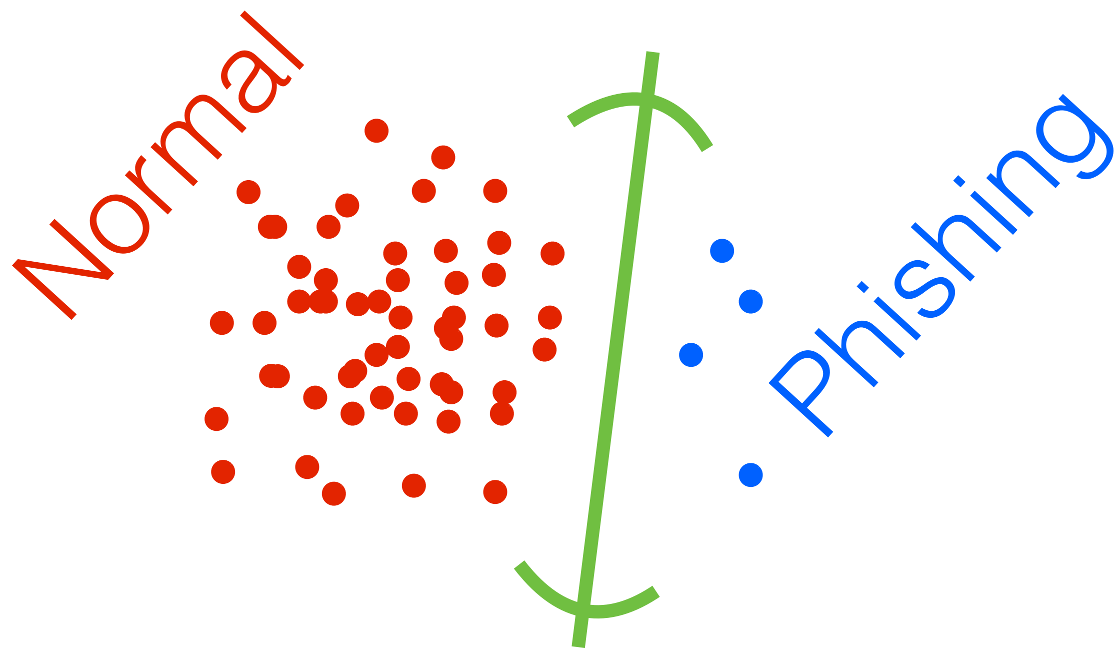


$$\sigma := \sum_{n=1}^{N} \|\mathcal{L}_n\|$$

$$\sigma_n := \|\mathcal{L}_n\|/\sigma$$



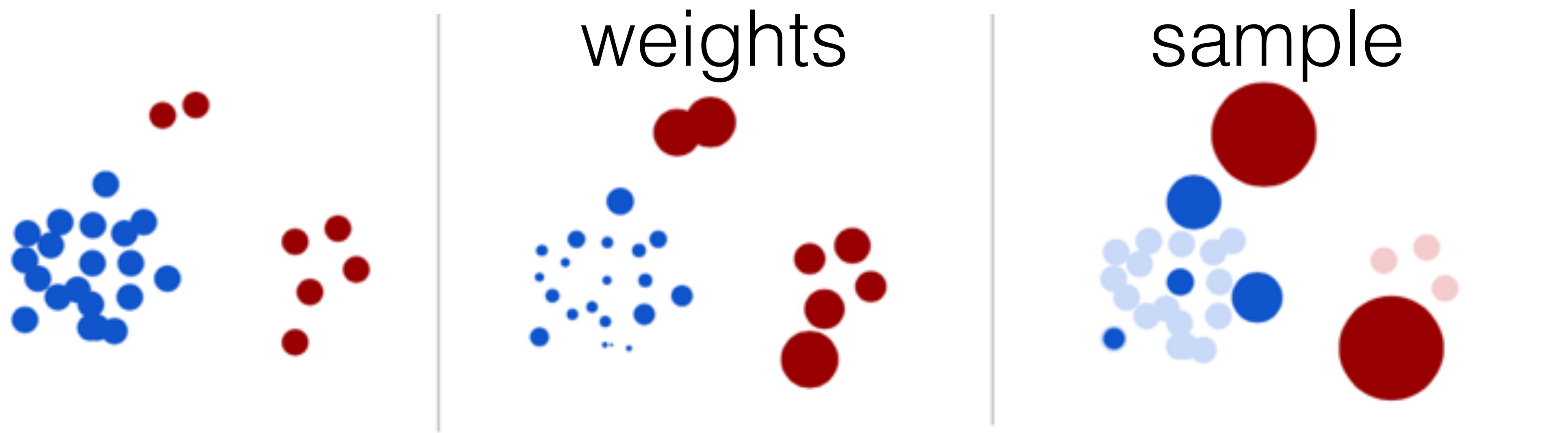
Importance sampling



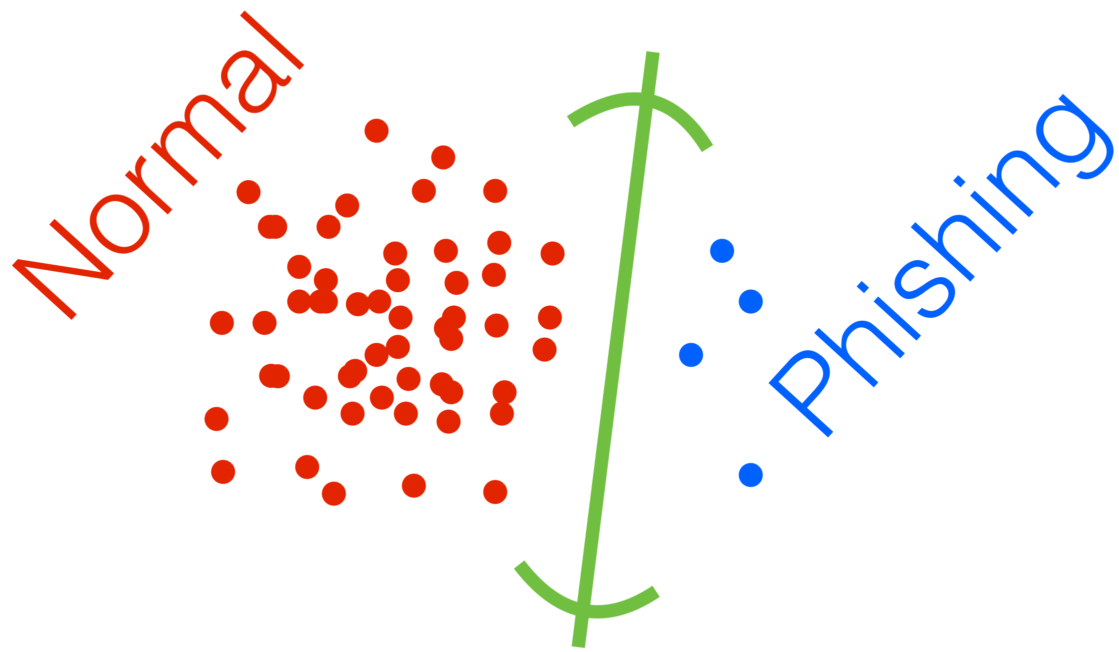
$$\sigma := \sum_{n=1}^{N} \|\mathcal{L}_n\|$$

$$\sigma_n := \|\mathcal{L}_n\| / \sigma$$

1. data
2. importance weights
3. importance sample

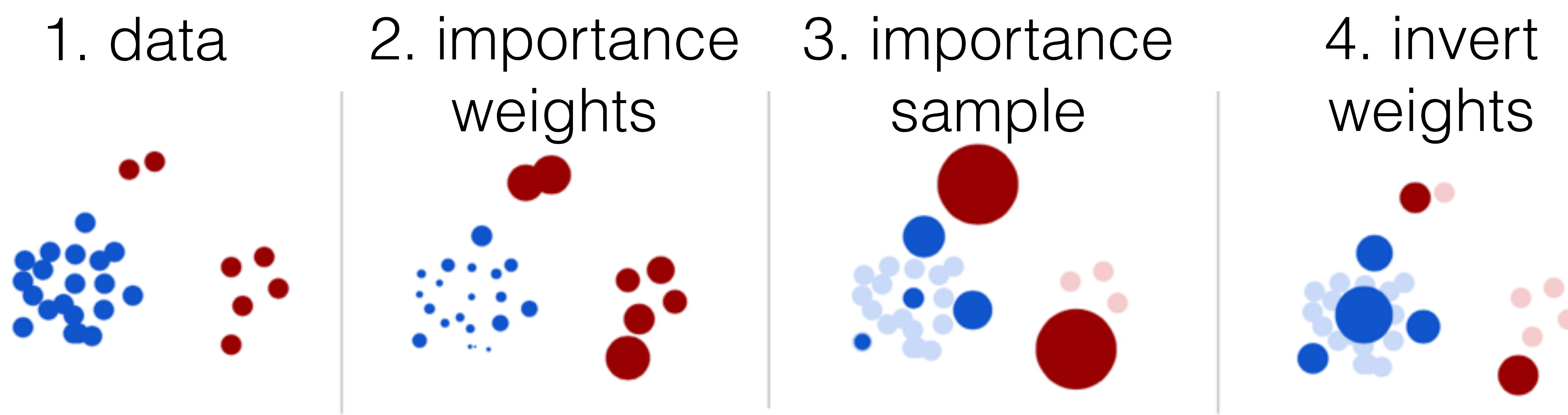


Importance sampling



$$\sigma := \sum_{n=1}^{N} \|\mathcal{L}_n\|$$

$$\sigma_n := \|\mathcal{L}_n\| / \sigma$$



Importance sampling

Thm sketch (CB). δ ∈ (0,1). W.p. ≥1 - δ, after *M* iterations,

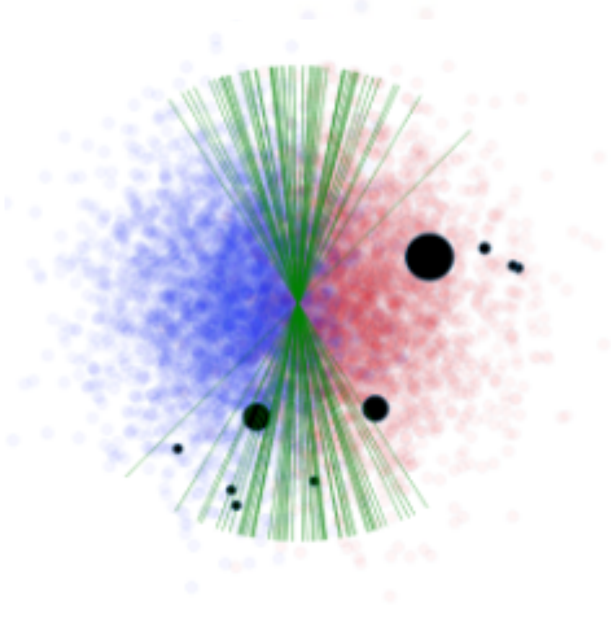
$$\|\mathcal{L}(w) - \mathcal{L}\| \leq \frac{\sigma\bar{\eta}}{\sqrt{M}}\left(1 + \sqrt{2\log\frac{1}{\delta}}\right)$$

Importance sampling

Thm sketch (CB). δ ∈ (0,1). W.p. ≥1 - δ, after *M* iterations,

$$\|\mathcal{L}(w) - \mathcal{L}\| \leq \frac{\sigma\bar{\eta}}{\sqrt{M}}\left(1 + \sqrt{2\log\frac{1}{\delta}}\right)$$

- Still noisy estimates



M = 10

Importance sampling

Thm sketch (CB). δ ∈ (0,1). W.p. ≥1 - δ, after *M* iterations,

$$\|\mathcal{L}(w) - \mathcal{L}\| \leq \frac{\sigma\bar{\eta}}{\sqrt{M}}\left(1 + \sqrt{2\log\frac{1}{\delta}}\right)$$

- Still noisy estimates

M = 10   *M* = 100   *M* = 1000

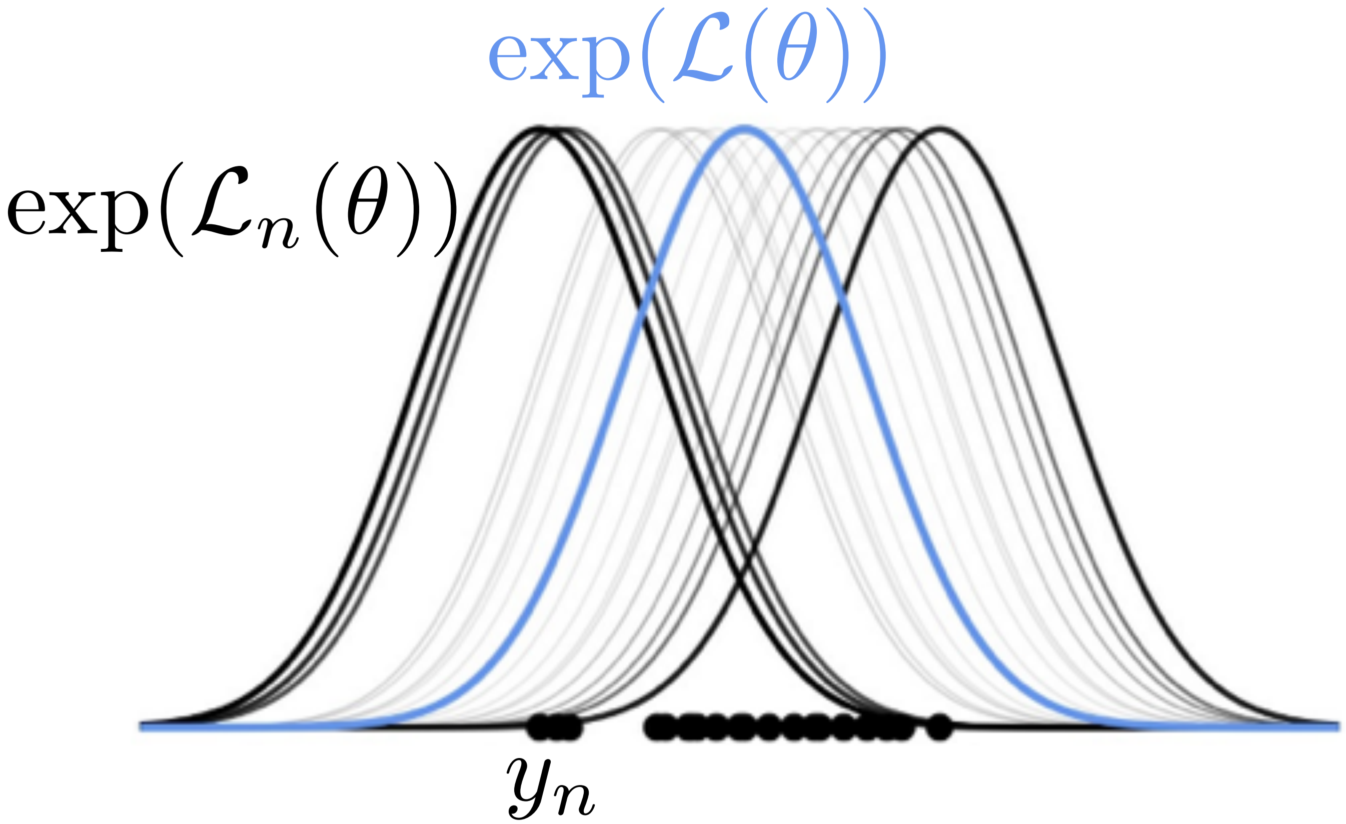
Hilbert coresets

- Want a good coreset: $$\min_{w\in\mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|$$ $$\text{s.t. } w \geq 0, \|w\|_0 \leq M$$

Hilbert coresets

- Want a good coreset: $\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|$

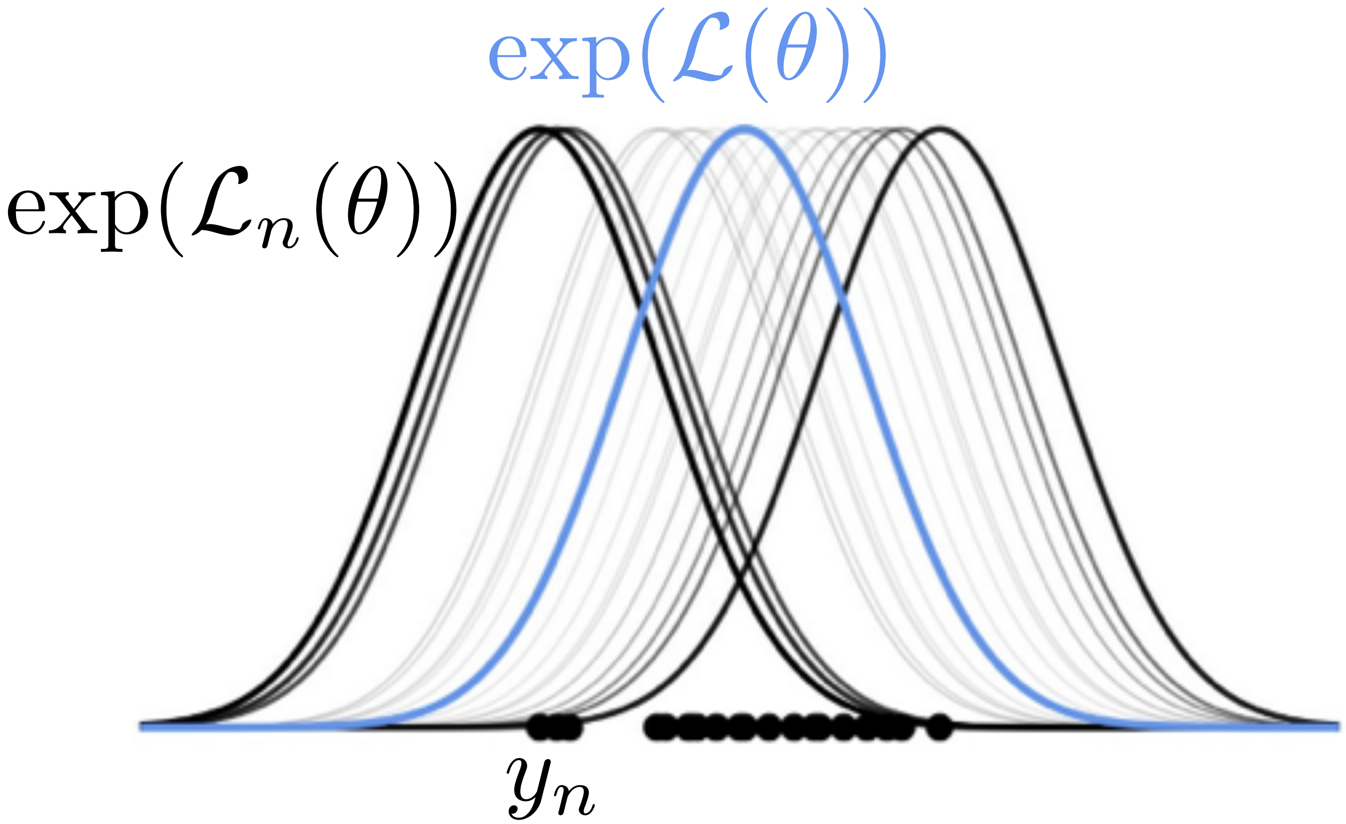
$$\text{s.t. } w \geq 0, \|w\|_0 \leq M$$

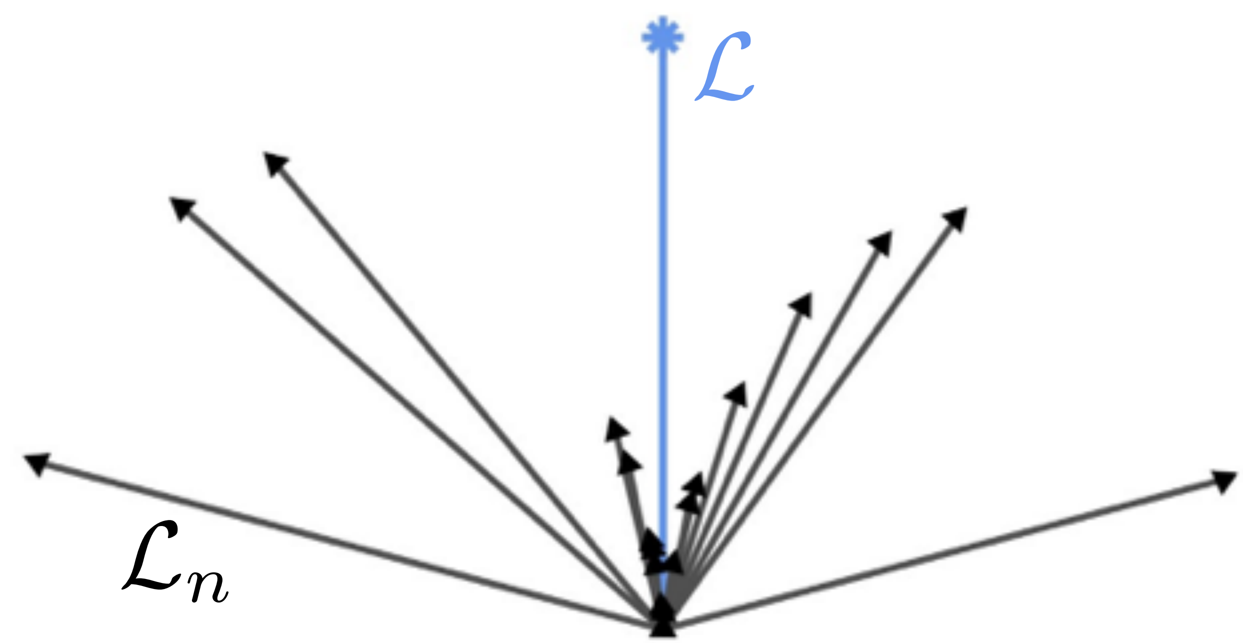


Hilbert coresets

- Want a good coreset: $\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|$

  s.t. $w \geq 0, \|w\|_0 \leq M$

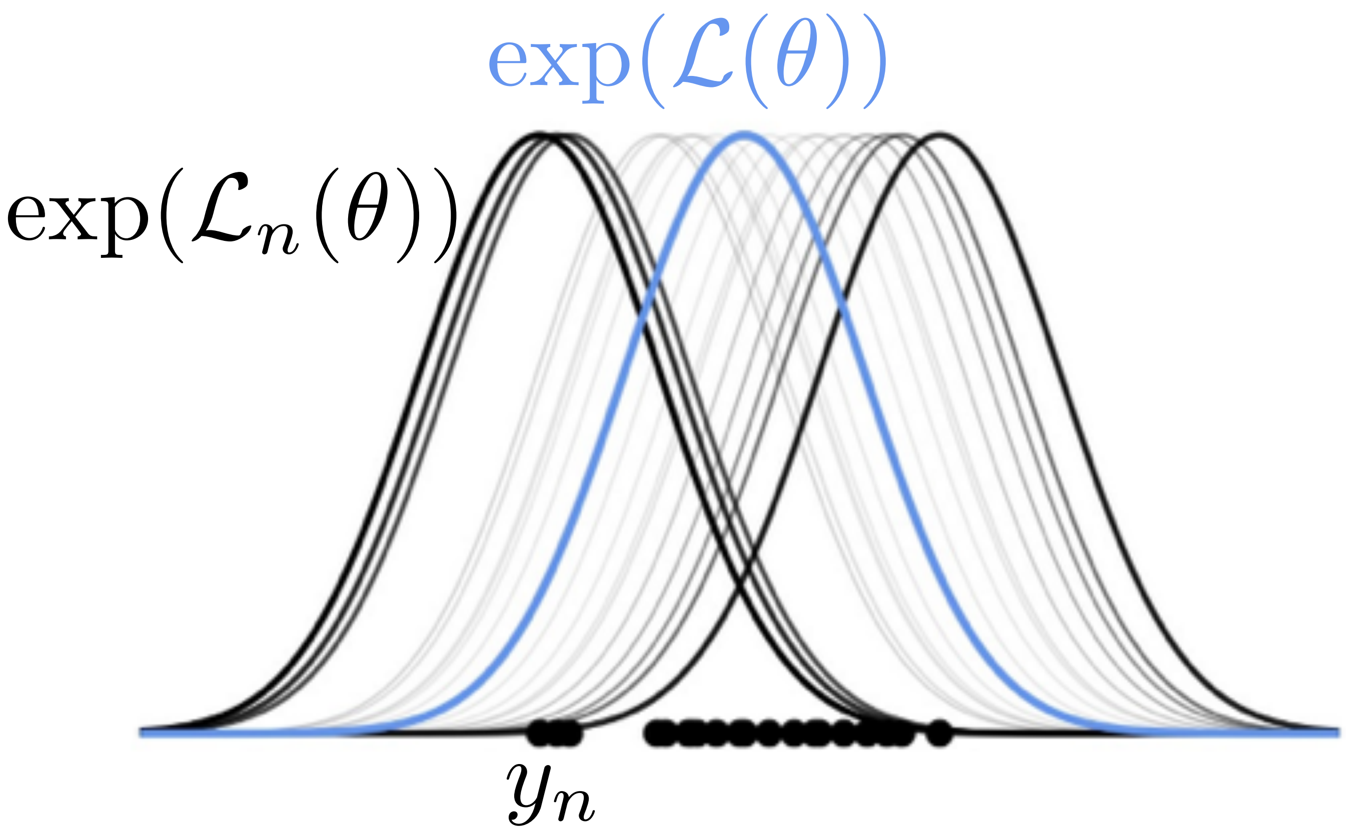


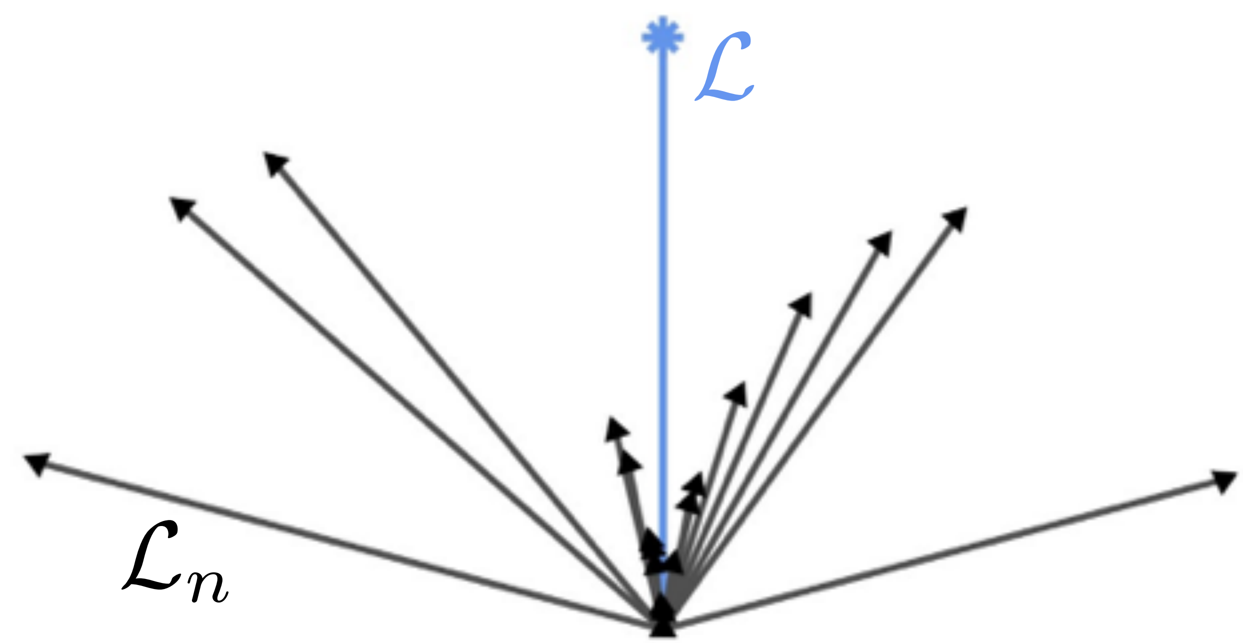


Hilbert coresets

- Want a good coreset: $\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|$

  s.t. $w \geq 0, \|w\|_0 \leq M$

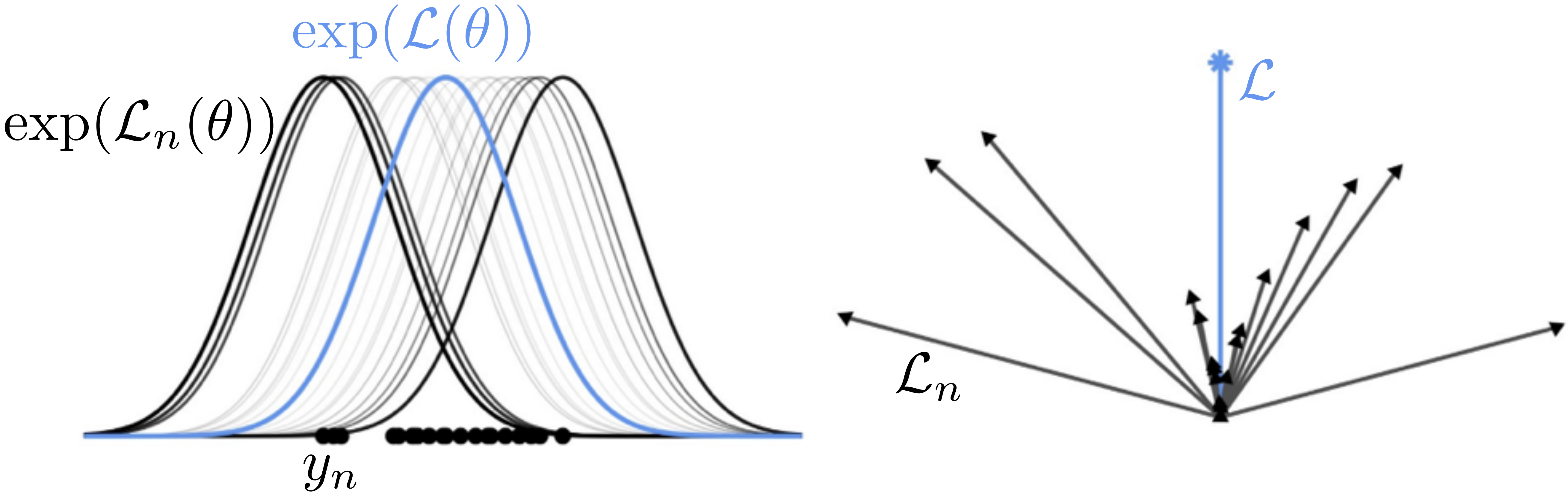




- need to consider (residual) error direction

Hilbert coresets

- Want a good coreset: $$\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\| \quad \text{s.t. } w \geq 0, \|w\|_0 \leq M$$



- need to consider (residual) error direction
- sparse optimization

Roadmap

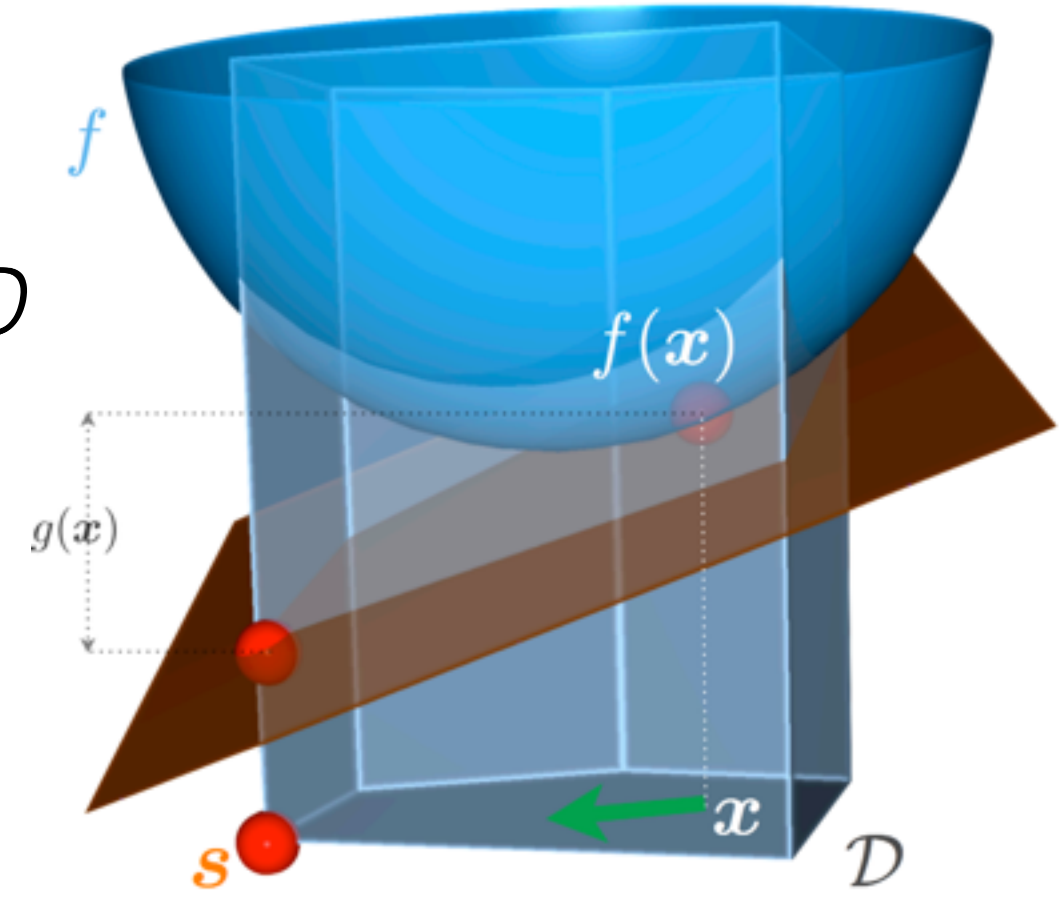
- Approximate Bayes review
- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

Roadmap

- Approximate Bayes review
- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

Frank-Wolfe

Convex optimization on a polytope *D*

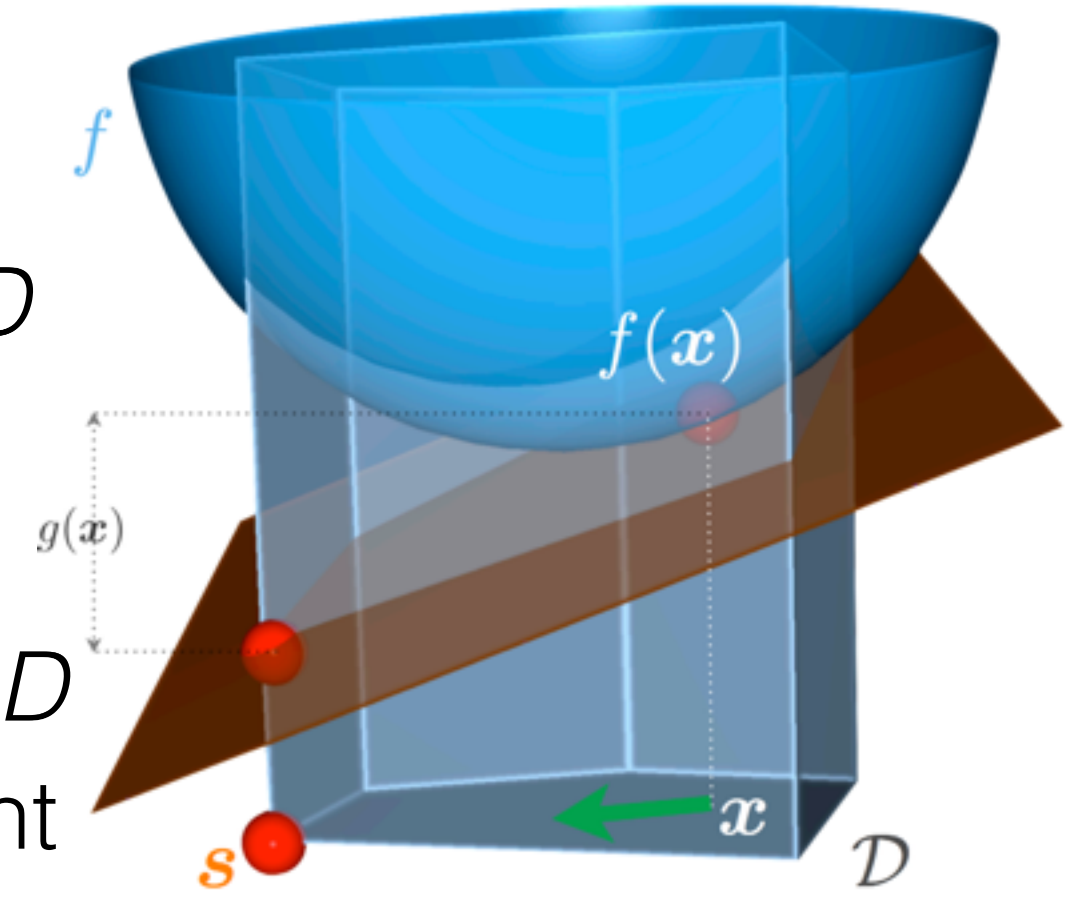


[Jaggi 2013]

Frank-Wolfe

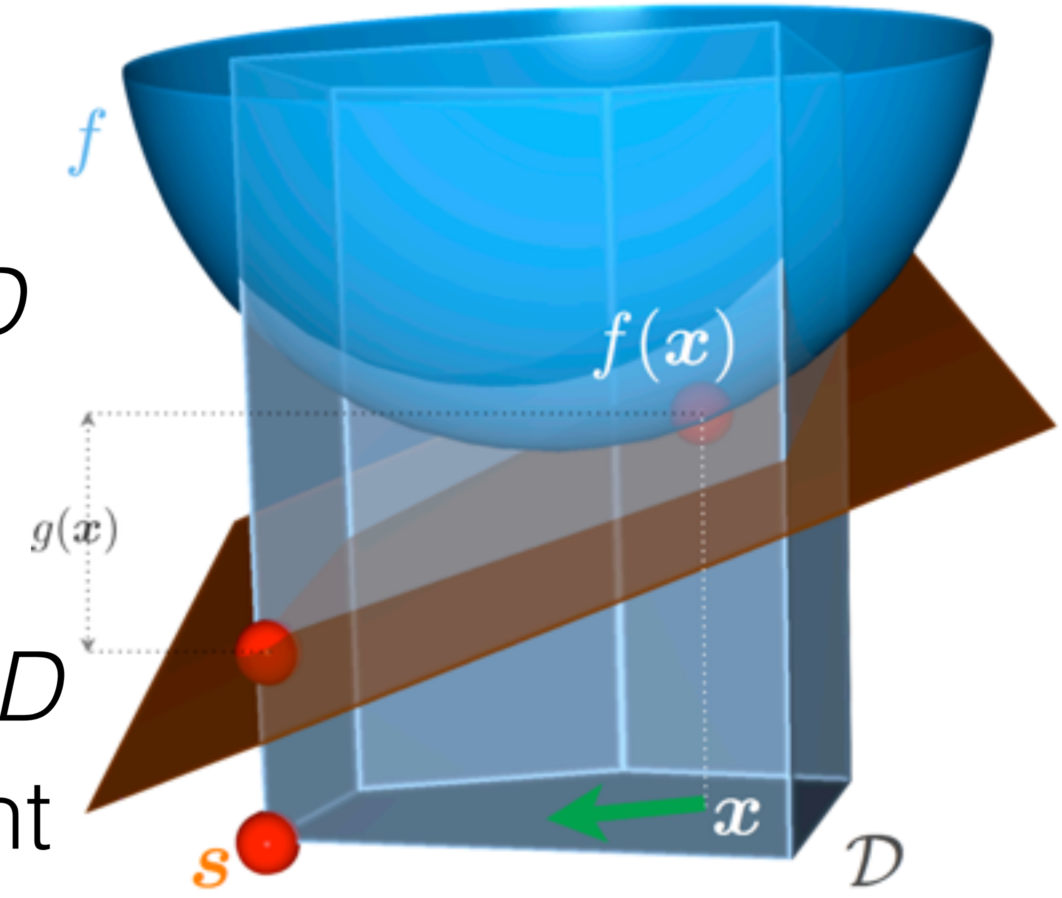
Convex optimization on a polytope *D*

- Repeat:
  1. Find gradient
  2. Find argmin point on plane in *D*
  3. Do line search between current point and argmin point



[Jaggi 2013]

Frank-Wolfe

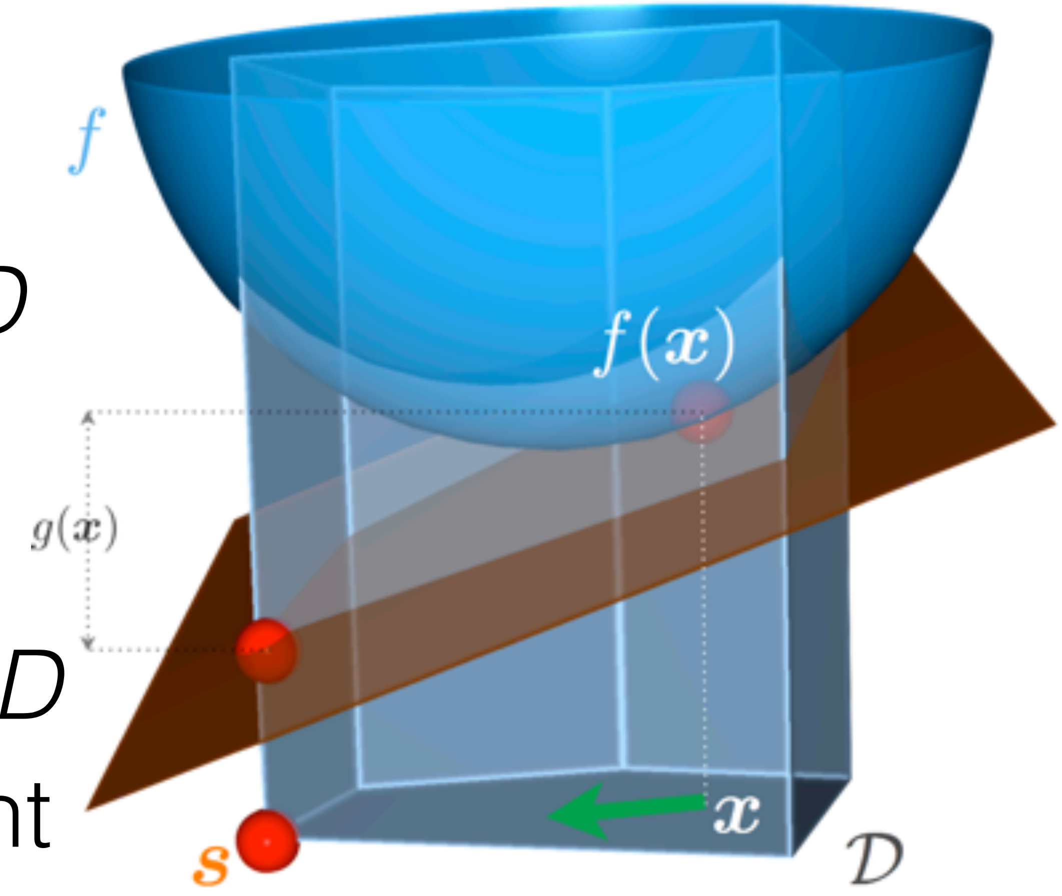


[Jaggi 2013]

Convex optimization on a polytope *D*

- Repeat:
  1. Find gradient
  2. Find argmin point on plane in *D*
  3. Do line search between current point and argmin point
- Convex combination of *M* vertices after *M* -1 steps

Frank-Wolfe

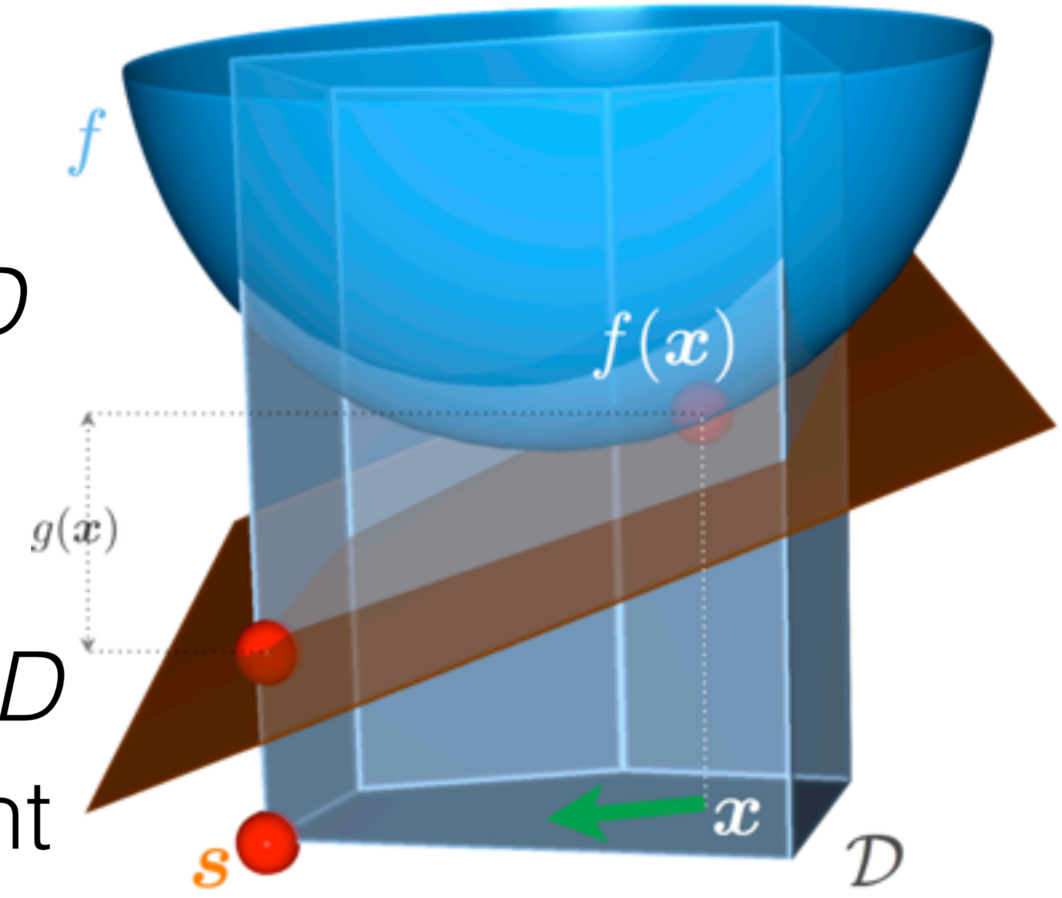


[Jaggi 2013]

Convex optimization on a polytope *D*

- Repeat:
  1. Find gradient
  2. Find argmin point on plane in *D*
  3. Do line search between current point and argmin point
- Convex combination of *M* vertices after *M* -1 steps
- Our problem: $\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|$

Frank-Wolfe

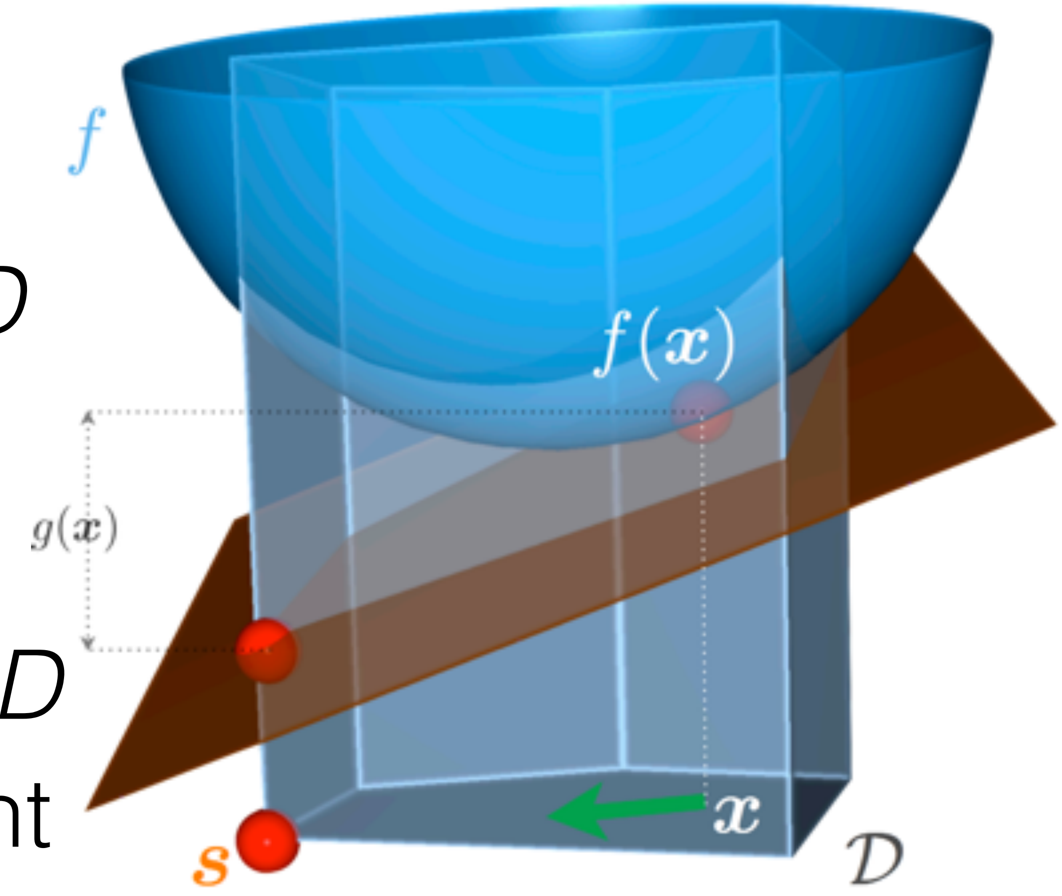


[Jaggi 2013]

Convex optimization on a polytope D

- Repeat:
  1. Find gradient
  2. Find argmin point on plane in D
  3. Do line search between current point and argmin point
- Convex combination of M vertices after M -1 steps
- Our problem: $\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|^2$

Frank-Wolfe



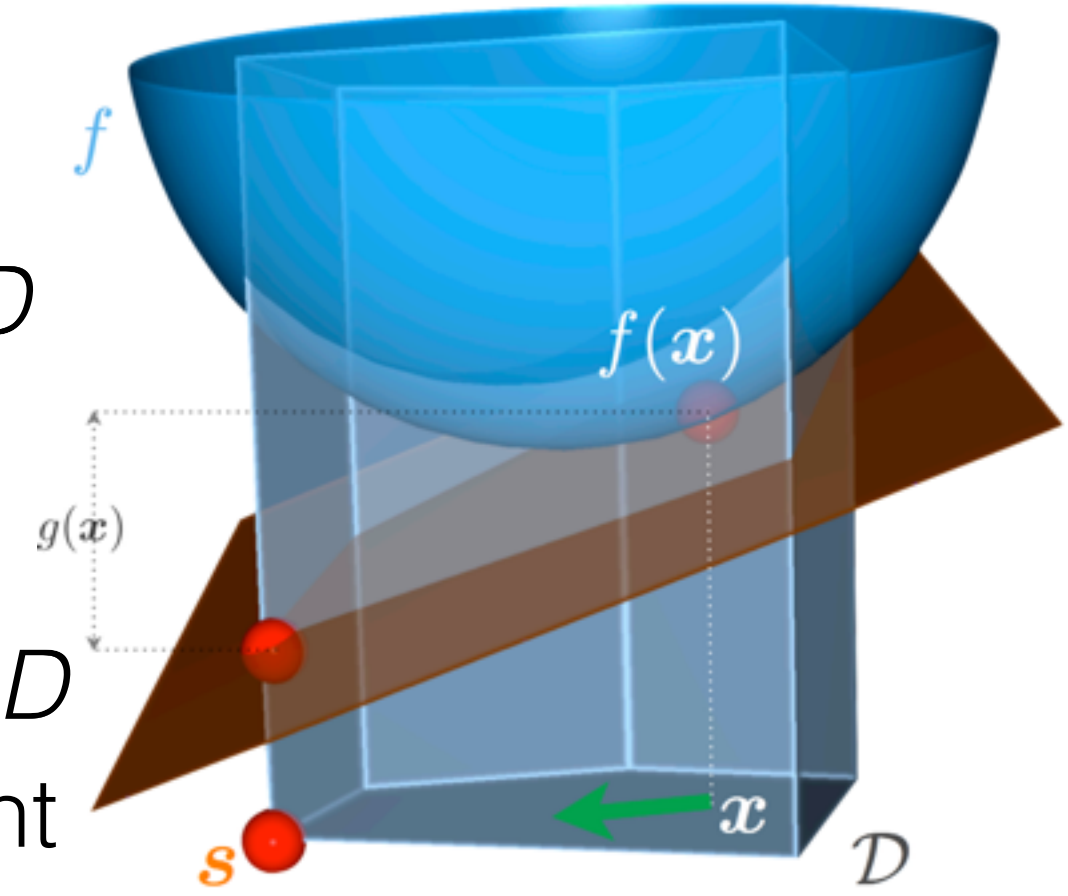
[Jaggi 2013]

Convex optimization on a polytope *D*

- Repeat:
  1. Find gradient
  2. Find argmin point on plane in *D*
  3. Do line search between current point and argmin point
- Convex combination of *M* vertices after *M* -1 steps
- Our problem: $$\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|^2$$

$$\text{s.t. } w \geq 0, \|w\|_0 \leq M$$

Frank-Wolfe



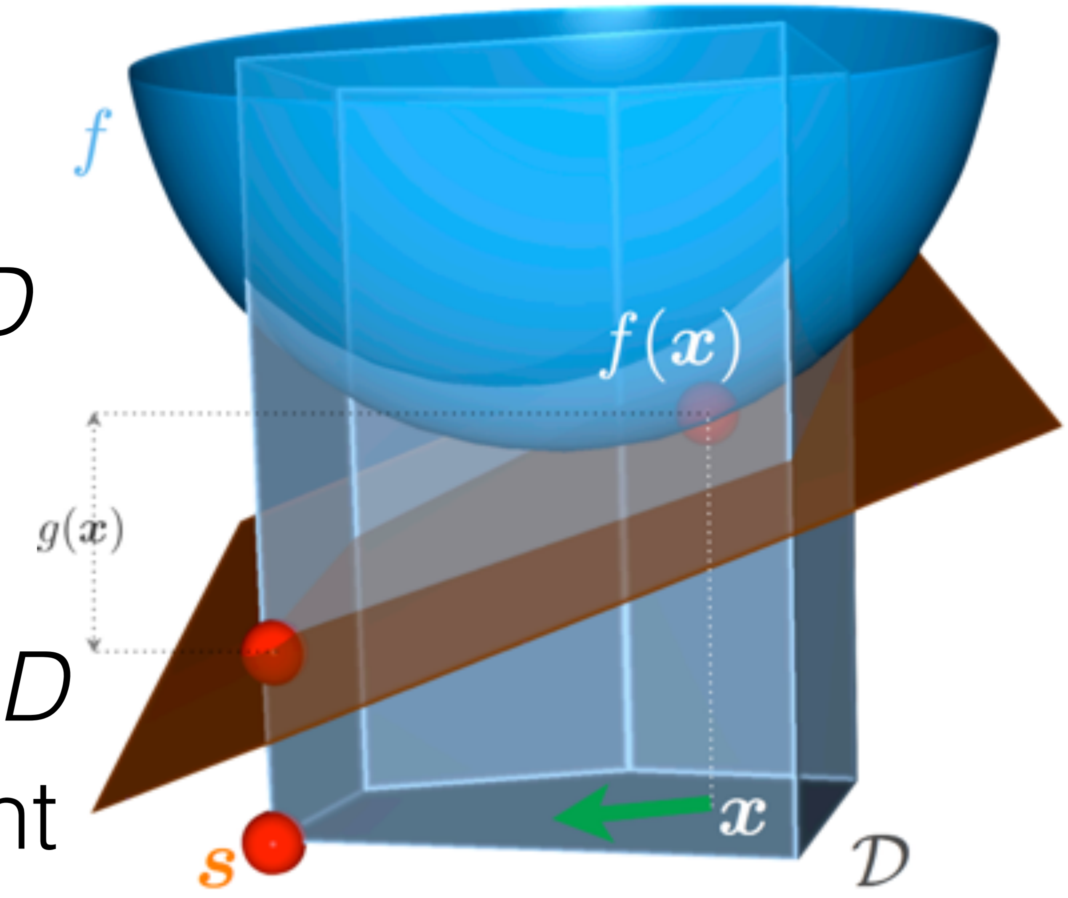
[Jaggi 2013]

Convex optimization on a polytope *D*

- Repeat:
  1. Find gradient
  2. Find argmin point on plane in *D*
  3. Do line search between current point and argmin point
- Convex combination of *M* vertices after *M* -1 steps
- Our problem: $\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|^2$

$$\Delta^{N-1} := \left\{ w \in \mathbb{R}^N : \sum_{n=1}^{N} \sigma_n w_n = \sigma, w \geq 0 \right\}$$

Frank-Wolfe



[Jaggi 2013]

Convex optimization on a polytope *D*

- Repeat:
  1. Find gradient
  2. Find argmin point on plane in *D*
  3. Do line search between current point and argmin point
- Convex combination of *M* vertices after *M* -1 steps
- Our problem: $\min_{w \in \mathbb{R}^N} \|\mathcal{L}(w) - \mathcal{L}\|^2$

$$\Delta^{N-1} := \left\{ w \in \mathbb{R}^N : \sum_{n=1}^{N} \sigma_n w_n = \sigma, w \geq 0 \right\}$$

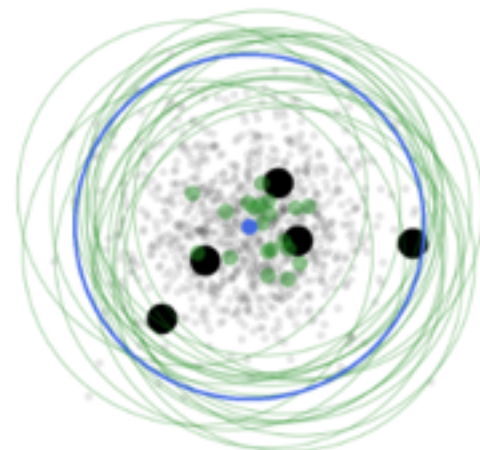
Thm sketch (CB). After *M* iterations,

$$\|\mathcal{L}(w) - \mathcal{L}\| \leq \frac{\sigma \bar{\eta}}{\sqrt{\alpha^{2M} + M}}$$

Gaussian model (simulated)

- 10K pts; norms, inference: closed-form

Uniform subsampling

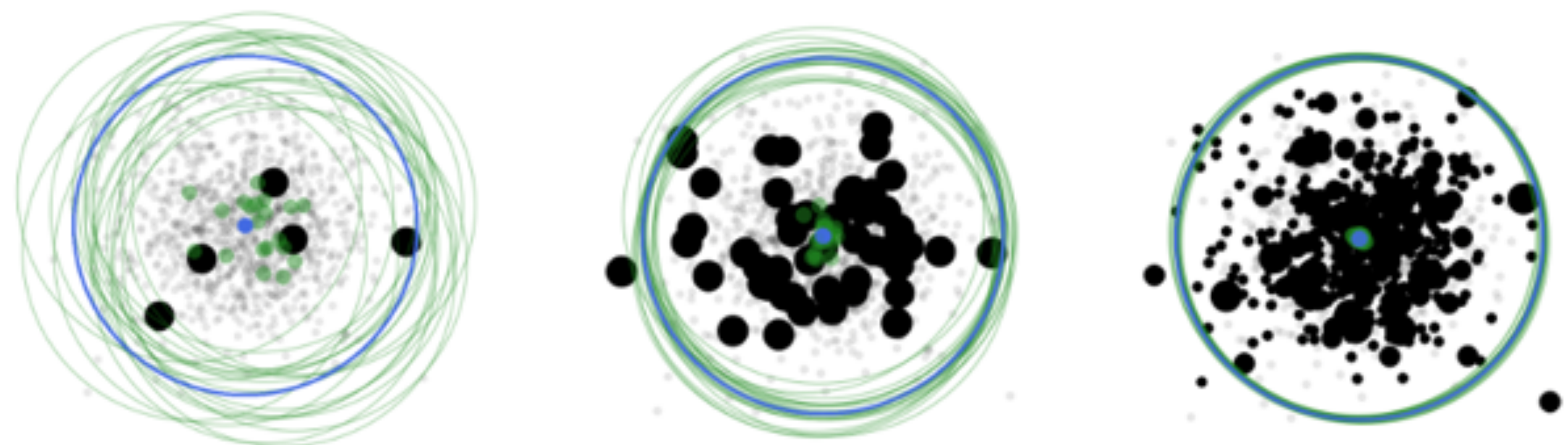


$M = 5$

Gaussian model (simulated)

- 10K pts; norms, inference: closed-form

Uniform subsampling



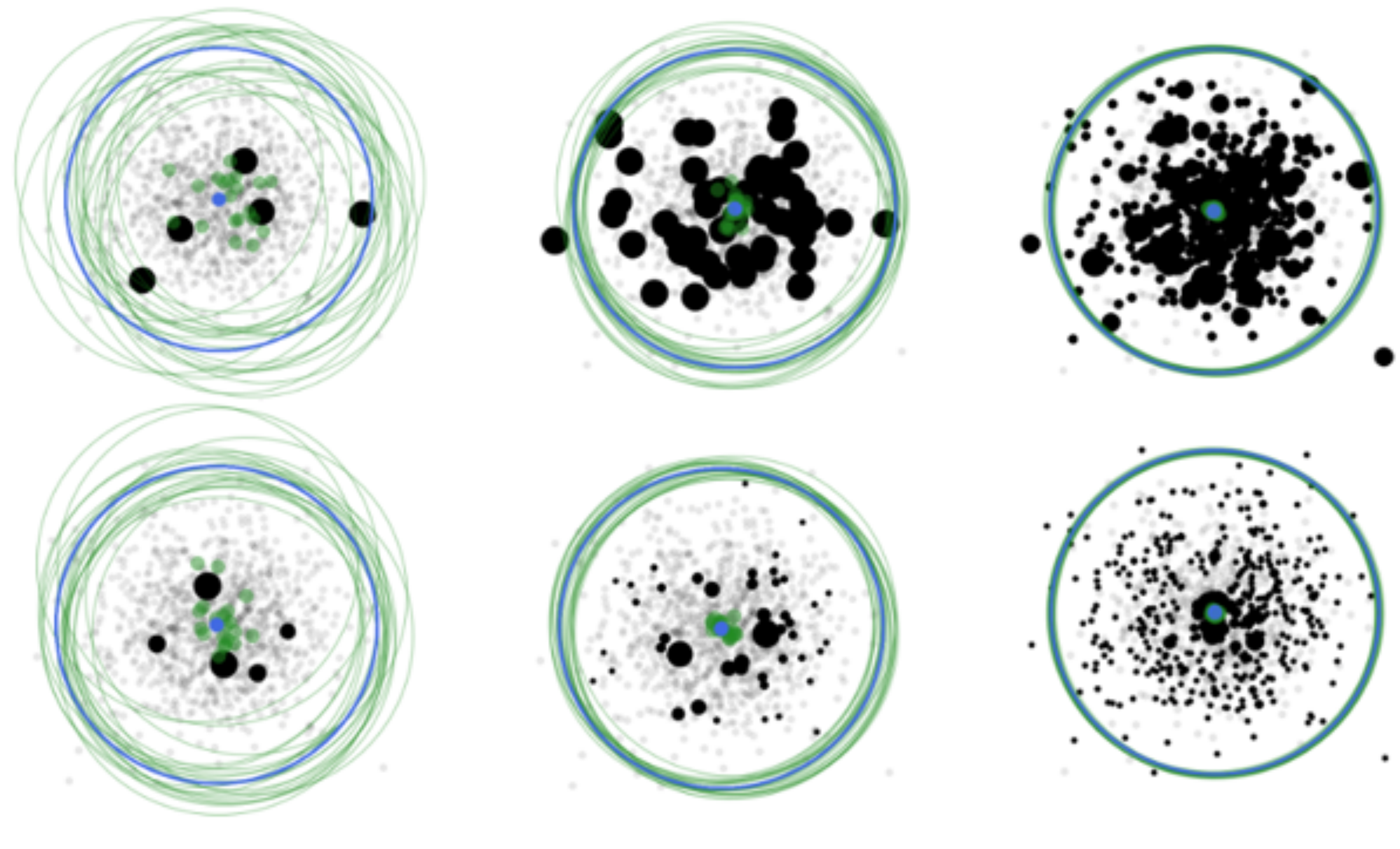
$M = 5$ $M = 50$ $M = 500$

Gaussian model (simulated)

- 10K pts; norms, inference: closed-form

Uniform subsampling

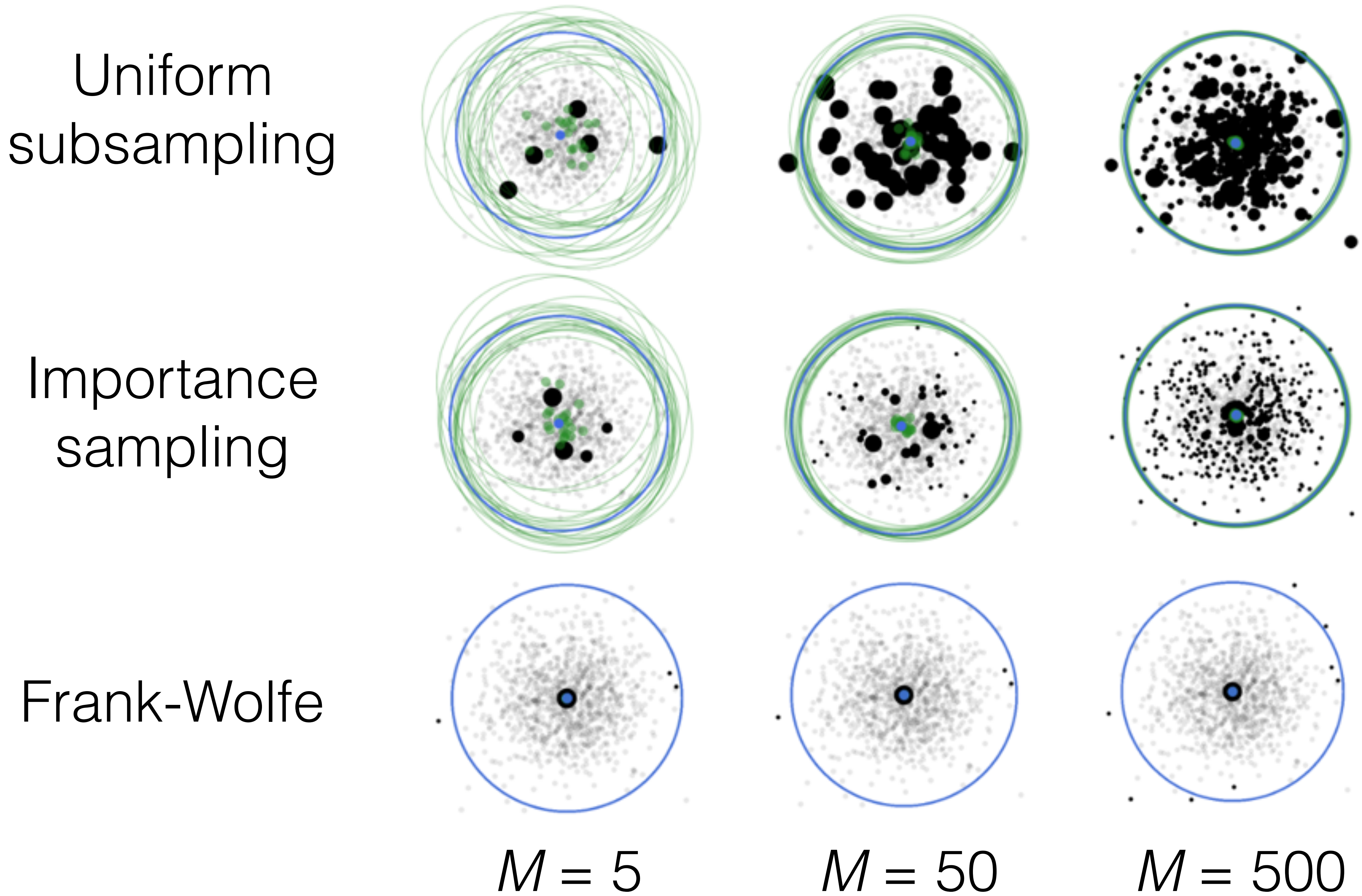
Importance sampling



$M = 5$ $M = 50$ $M = 500$

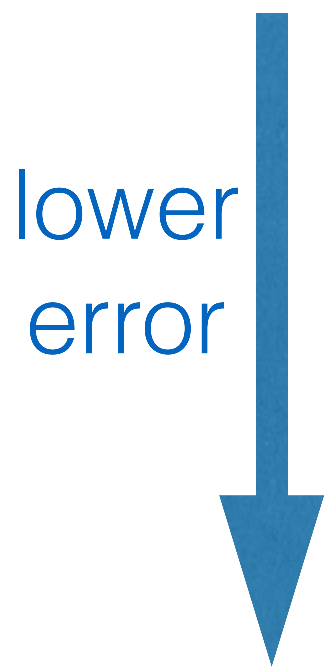
Gaussian model (simulated)

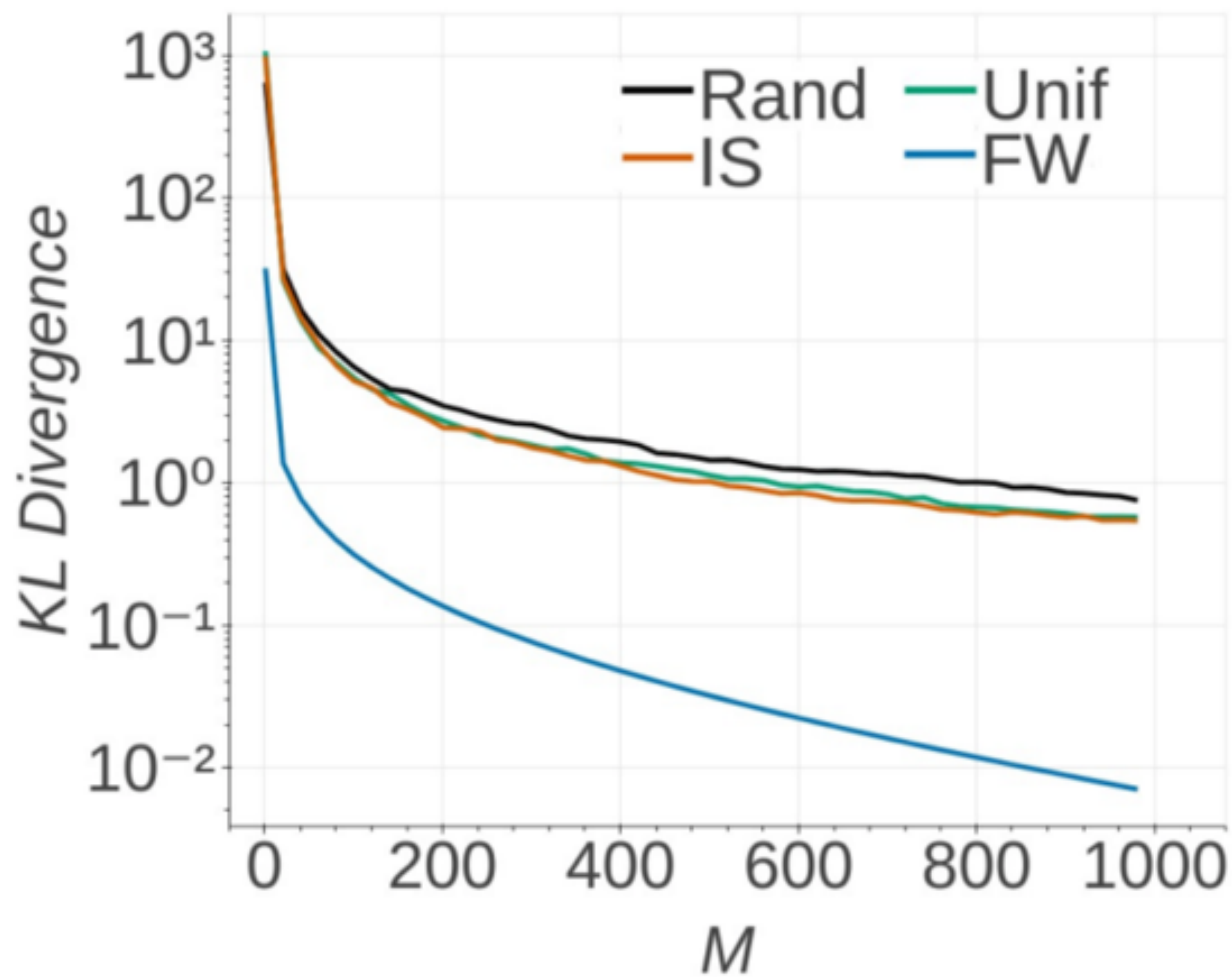
- 10K pts; norms, inference: closed-form



Gaussian model (simulated)

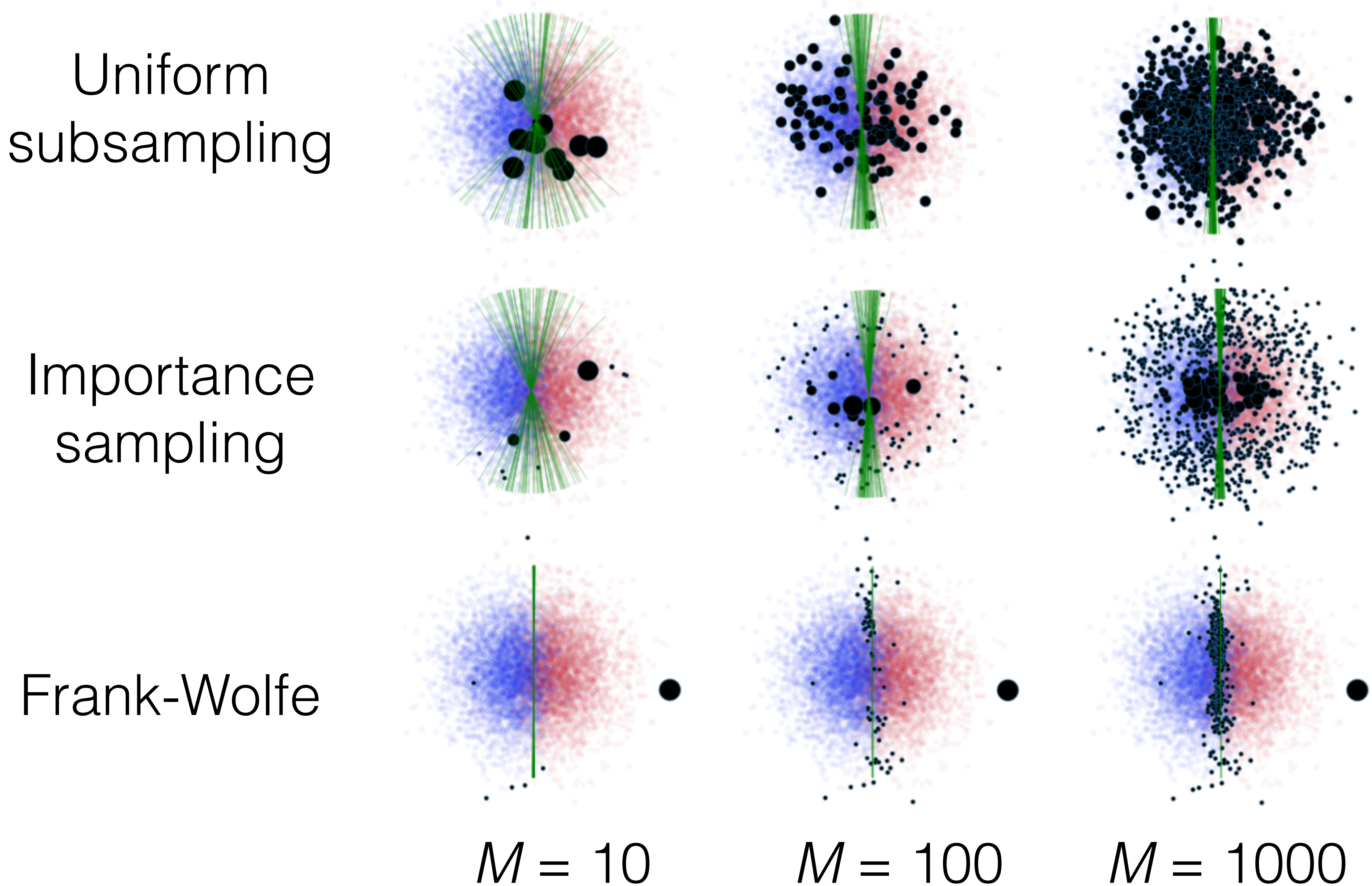
- 10K pts; norms, inference: closed-form





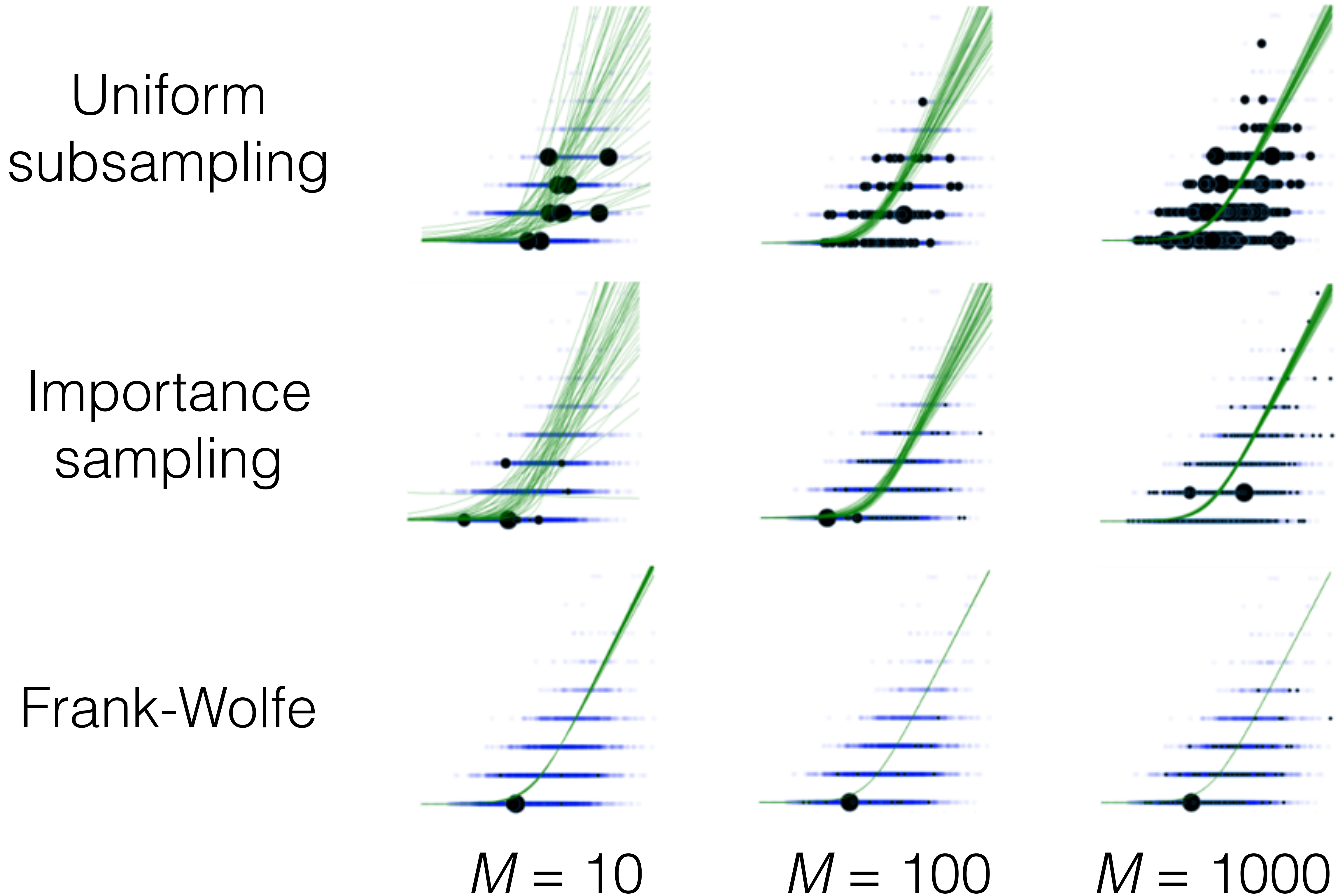
Logistic regression (simulated)

- 10K pts; general inference



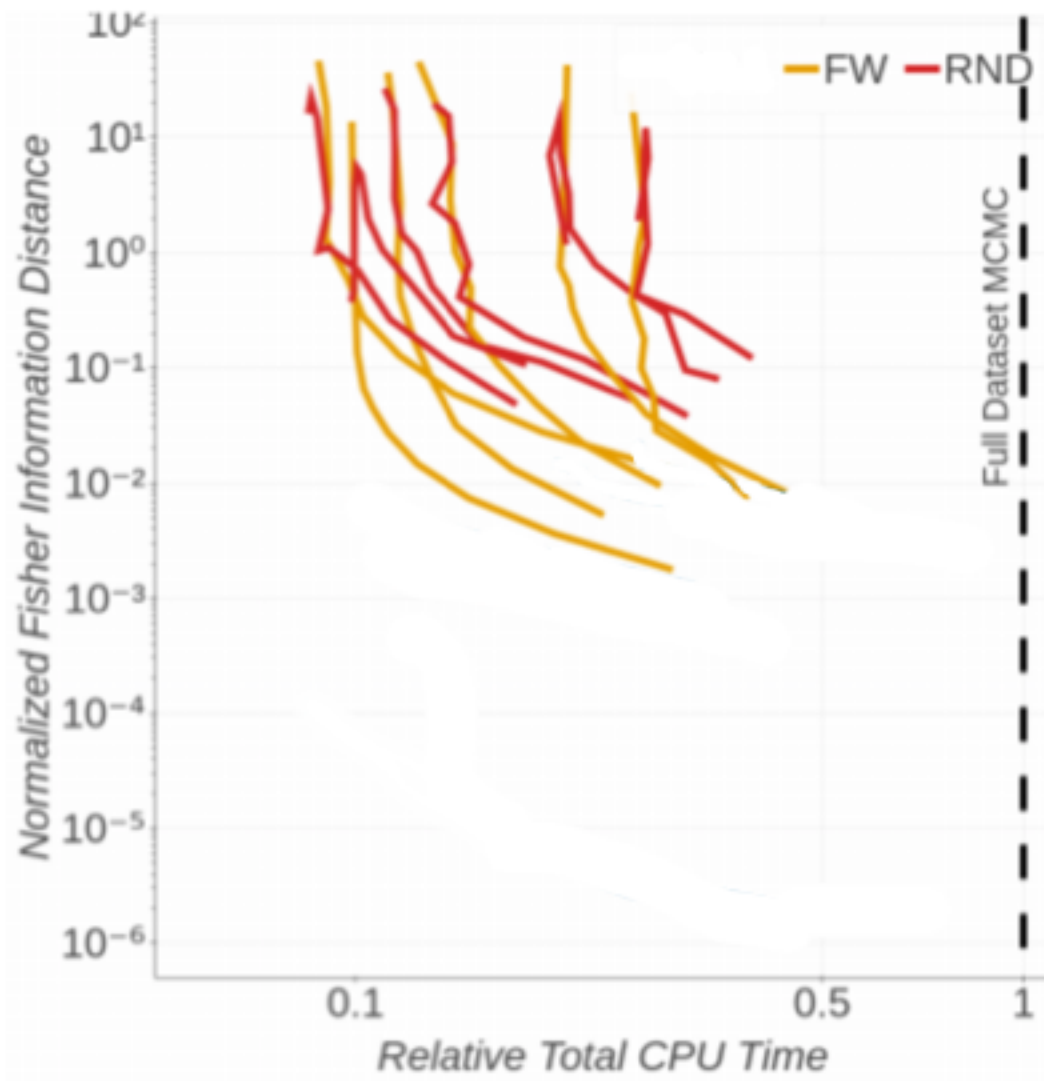
Poisson regression (simulated)

- 10K pts; general inference



Real data experiments

lower error



less total time

Uniform subsampling

Frank Wolfe coresets

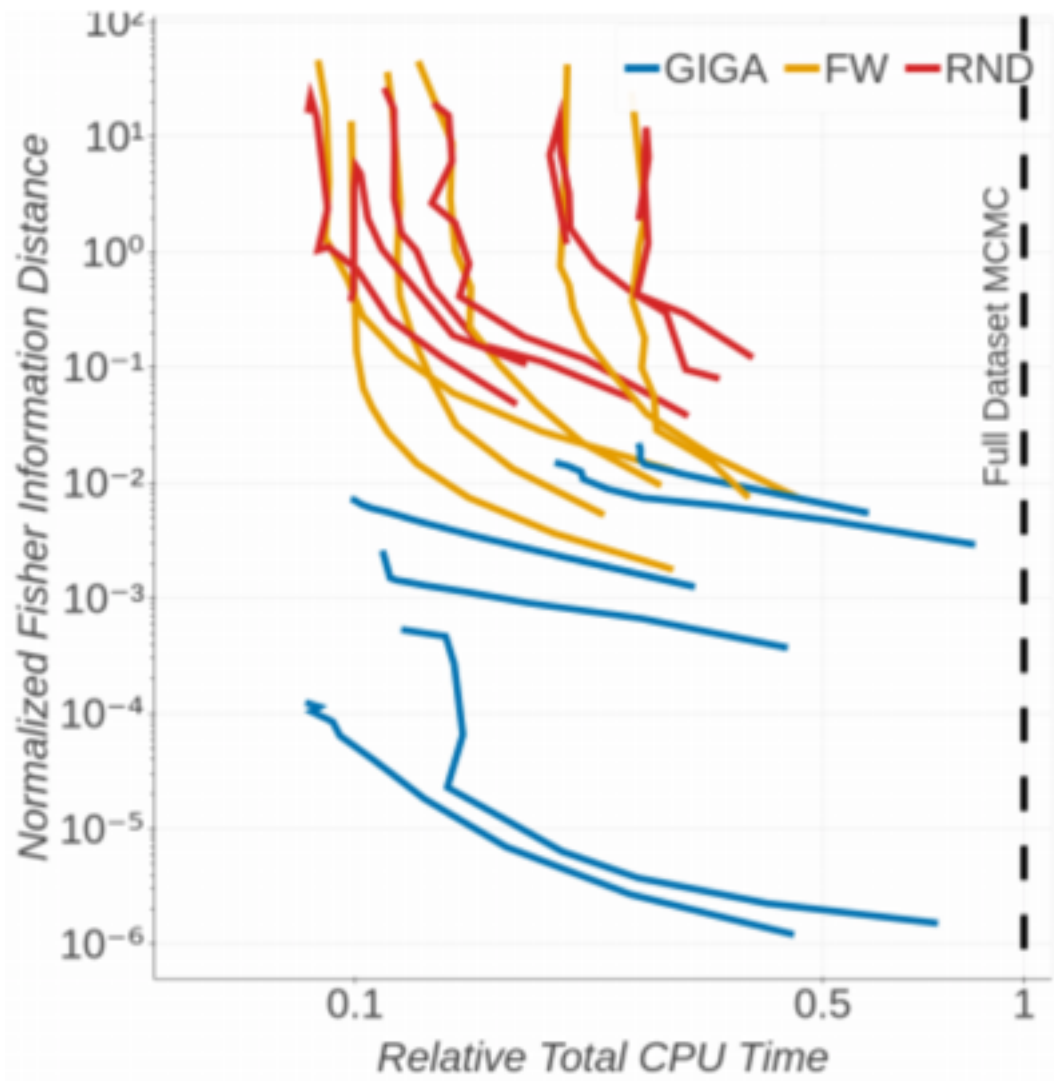
Data sets include:

- Phishing
- Chemical reactivity
- Bicycle trips
- Airport delays

[Campbell, Broderick 2017]

Real data experiments

lower error



less total time

Uniform subsampling

Frank Wolfe coresets

GIGA coresets

Data sets include:

- Phishing
- Chemical reactivity
- Bicycle trips
- Airport delays

[Campbell, Broderick 2017, 2018]

Roadmap

- Approximate Bayes review
- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

Roadmap

- Approximate Bayes review
- The “core” of the data set
- Uniform data subsampling isn’t enough
- Importance sampling for “coresets”
- Optimization for “coresets”
- Approximate sufficient statistics

Data summarization

Data summarization

- Exponential family likelihood

Data summarization

- Exponential family likelihood

Sufficient statistics

$$p(y_{1:N}|x_{1:N},\theta) = \prod_{n=1}^{N} \exp\left[T(y_n, x_n) \cdot \eta(\theta)\right]$$

Data summarization

- Exponential family likelihood

Sufficient statistics

$$p(y_{1:N}|x_{1:N}, \theta) = \prod_{n=1}^{N} \exp\left[T(y_n, x_n) \cdot \eta(\theta)\right]$$

$$= \exp\left[\left\{\sum_{n=1}^{N} T(y_n, x_n)\right\} \cdot \eta(\theta)\right]$$

Data summarization

- Exponential family likelihood

Sufficient statistics

$$p(y_{1:N}|x_{1:N},\theta) = \prod_{n=1}^{N} \exp\left[T(y_n,x_n)\cdot\eta(\theta)\right]$$

$$= \exp\left[\left\{\sum_{n=1}^{N} T(y_n,x_n)\right\}\cdot\eta(\theta)\right]$$

- Scalable, single-pass, streaming, distributed, complementary to MCMC

Data summarization

- Exponential family likelihood

Sufficient statistics

$$p(y_{1:N}|x_{1:N},\theta) = \prod_{n=1}^{N} \exp\left[T(y_n, x_n) \cdot \eta(\theta)\right]$$

$$= \exp\left[\left\{\sum_{n=1}^{N} T(y_n, x_n)\right\} \cdot \eta(\theta)\right]$$

- Scalable, single-pass, streaming, distributed, complementary to MCMC
- *But*: Often no simple sufficient statistics

Data summarization

- Exponential family likelihood

Sufficient statistics

$$p(y_{1:N}|x_{1:N},\theta) = \prod_{n=1}^{N} \exp\left[T(y_n, x_n) \cdot \eta(\theta)\right]$$

$$= \exp\left[\left\{\sum_{n=1}^{N} T(y_n, x_n)\right\} \cdot \eta(\theta)\right]$$

- Scalable, single-pass, streaming, distributed, complementary to MCMC
- *But*: Often no simple sufficient statistics
  - E.g. Bayesian logistic regression; GLMs; "deeper" models
    - Likelihood $p(y_{1:N}|x_{1:N},\theta) = \prod_{n=1}^{N} \frac{1}{1+\exp(-y_n x_n \cdot \theta)}$

Data summarization

- Exponential family likelihood

Sufficient statistics

$$p(y_{1:N}|x_{1:N},\theta) = \prod_{n=1}^{N} \exp\left[T(y_n, x_n) \cdot \eta(\theta)\right]$$

$$= \exp\left[\left\{\sum_{n=1}^{N} T(y_n, x_n)\right\} \cdot \eta(\theta)\right]$$

- Scalable, single-pass, streaming, distributed, complementary to MCMC
- *But*: Often no simple sufficient statistics
  - E.g. Bayesian logistic regression; GLMs; "deeper" models
    - Likelihood $p(y_{1:N}|x_{1:N},\theta) = \prod_{n=1}^{N} \frac{1}{1+\exp(-y_n x_n \cdot \theta)}$
- Our proposal: (polynomial) *approximate* sufficient statistics

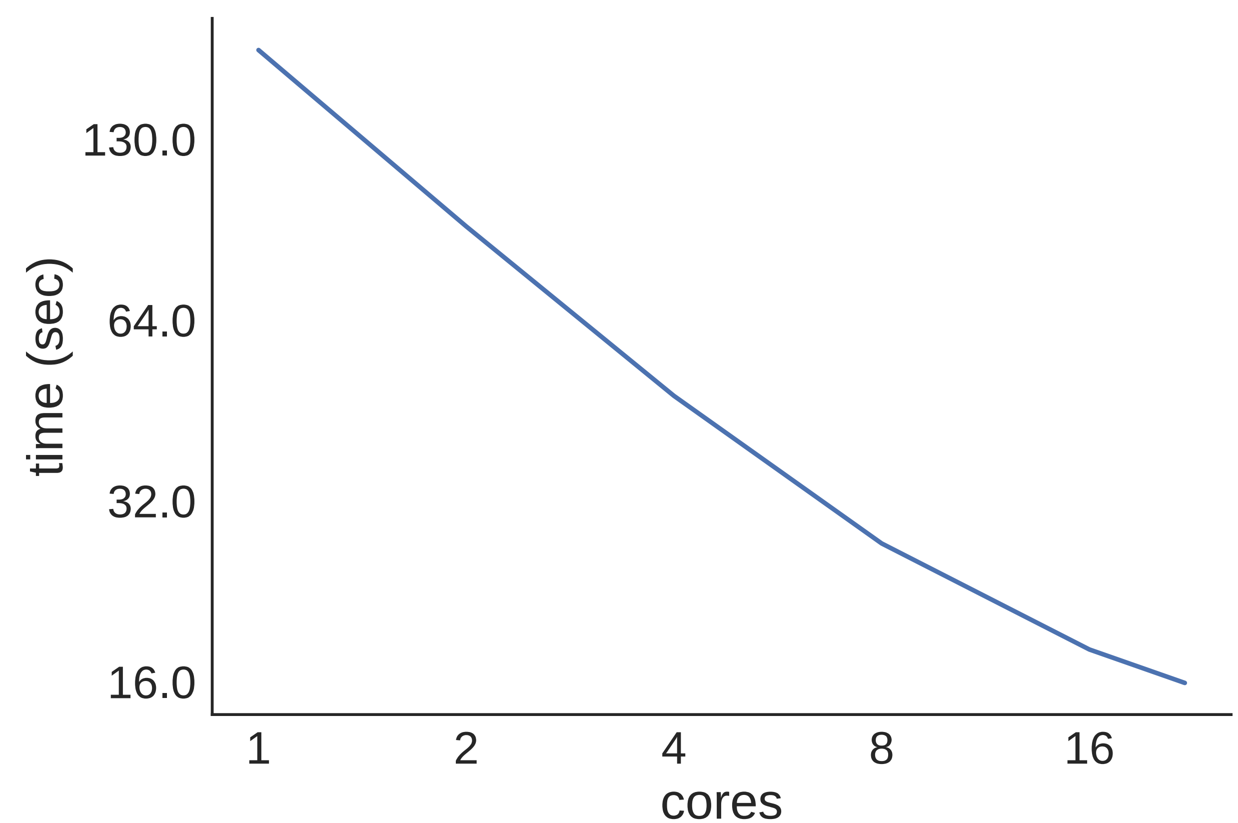
Data summarization

Criteo Labs > Algorithms > Criteo Releases its New Dataset

Criteo Releases its New Dataset

By: CriteoLabs / 31 Mar 2015

- 6M data points, 1000 features
- Streaming, distributed; minimal communication
- 22 cores, 16 sec
- Finite-data guarantees on Wasserstein distance to exact posterior



[Huggins, Adams, Broderick 2017]

Conclusions

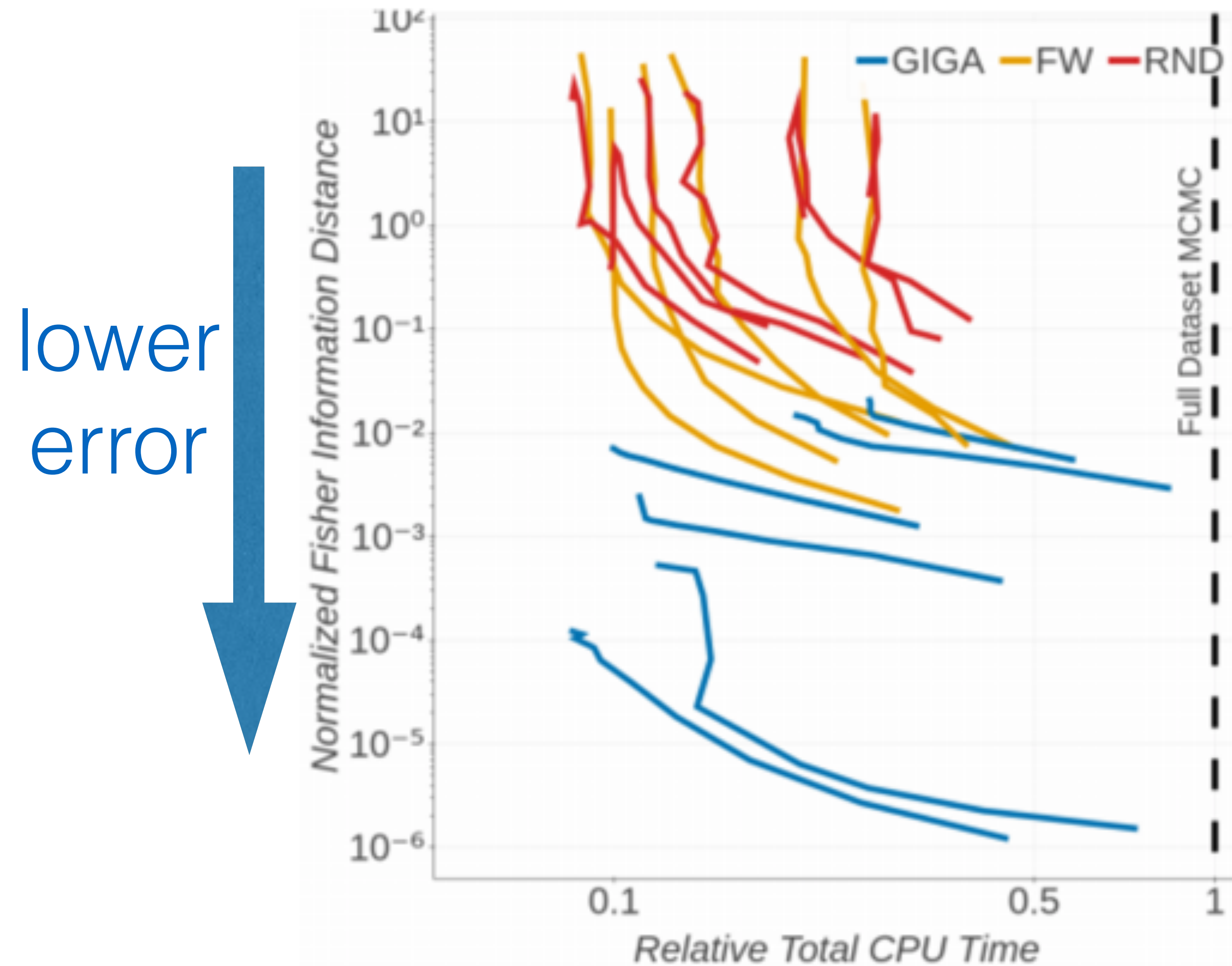
- *Data summarization* for **scalable**, **automated** approx. Bayes algorithms with **error bounds on quality for finite data**
  - Get more accurate with more computation investment
  - Coresets
  - Approx. suff. stats

Conclusions

- *Data summarization* for **scalable**, **automated** approx. Bayes algorithms with **error bounds on quality for finite data**
  - Get more accurate with more computation investment
  - Coresets
  - Approx. suff. stats
- A start
  - Lots of potential improvements/ directions

Conclusions

- *Data summarization* for **scalable**, **automated** approx. Bayes algorithms with **error bounds on quality for finite data**
    - Get more accurate with more computation investment
    - Coresets
    - Approx. suff. stats
- A start
    - Lots of potential improvements/directions



[Campbell, Broderick 2018]

References (1/5)

T Campbell and T Broderick. Automated scalable Bayesian inference via Hilbert coresets. Under review. ArXiv:1710.05053.

T Campbell and T Broderick. Bayesian Coreset Construction via Greedy Iterative Geodesic Ascent. *ICML* 2018.

JH Huggins, T Campbell, and T Broderick. Coresets for scalable Bayesian logistic regression. *NIPS* 2016.

JH Huggins, RP Adams, and T Broderick. PASS-GLM: Polynomial approximate sufficient statistics for scalable Bayesian GLM inference. *NIPS* 2017.

JH Huggins, T Campbell, M Kasprzak, and T Broderick. Scalable Gaussian Process inference with finite-data mean and variance guarantees. Under review. ArXiv:1806.10234.

JH Huggins, M Kasprzak, T Campbell, and T Broderick. Practical bounds on the error of Bayesian posterior approximations: A nonasymptotic approach, 2018. ArXiv:1809.09505.

References (2/5)

PK Agarwal, S Har-Peled, and KR Varadarajan. "Geometric approximation via coresets." *Combinatorial and Computational Geometry* 52 (2005): 1-30.

R Bardenet, A Doucet, and C Holmes. "On Markov chain Monte Carlo methods for tall data." *The Journal of Machine Learning Research* 18.1 (2017): 1515-1557.

T Broderick, N Boyd, A Wibisono, AC Wilson, and MI Jordan. Streaming variational Bayes. *NIPS* 2013.

CM Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006.

W DuMouchel, C Volinsky, T Johnson, C Cortes, and D Pregibon. "Squashing flat files flatter." In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 6-15. ACM, 1999.

D Dunson. Robust and scalable approach to Bayesian inference. Talk at *ISBA* 2014.

D Feldman, and M Langberg. "A unified framework for approximating and clustering data." In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 569-578. ACM, 2011.

B Fosdick. *Modeling Heterogeneity within and between Matrices and Arrays*, Chapter 4.7. PhD Thesis, University of Washington, 2013.

RJ Giordano, T Broderick, and MI Jordan. "Linear response methods for accurate covariance estimates from mean field variational Bayes." *NIPS* 2015.

R Giordano, T Broderick, R Meager, J Huggins, and MI Jordan. "Fast robustness quantification with variational Bayes." *ICML 2016 Workshop on #Data4Good: Machine Learning in Social Good Applications*, 2016.

References (3/5)

MD Hoffman, and A Gelman. "The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* 15, no. 1 (2014): 1593-1623.

M Jaggi. "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization." *ICML* 2013.

A Kucukelbir, R Ranganath, A Gelman, and D Blei. "Automatic variational inference in Stan." *NIPS* 2015.

A Kucukelbir, D Tran, R Ranganath, A Gelman, and DM Blei. "Automatic differentiation variational inference." *The Journal of Machine Learning Research* 18.1 (2017): 430-474.

DJC MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

D Madigan, N Raghavan, W Dumouchel, M Nason, C Posse, and G Ridgeway. "Likelihood-based data squashing: A modeling approach to instance construction." *Data Mining and Knowledge Discovery* 6, no. 2 (2002): 173-190.

M Opper and O Winther. Variational linear response. *NIPS* 2003.

Stan (open source software). http://mc-stan.org/ Accessed: 2018.

RE Turner and M Sahani. Two problems with variational expectation maximisation for time-series models. In D Barber, AT Cemgil, and S Chiappa, editors, *Bayesian Time Series Models*, 2011.

B Wang and M Titterington. Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *AISTATS*, 2004.

Application References (4/5)

Chati, Yashovardhan Sushil, and Hamsa Balakrishnan. "A Gaussian process regression approach to model aircraft engine fuel flow rate." *Cyber-Physical Systems (ICCPS), 2017 ACM/IEEE 8th International Conference on*. IEEE, 2017.

Meager, Rachael. “Understanding the impact of microcredit expansions: A Bayesian hierarchical analysis of 7 randomized experiments.” *AEJ: Applied*, to appear, 2018a.

Meager, Rachael. “Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature.” Working paper, 2018b.

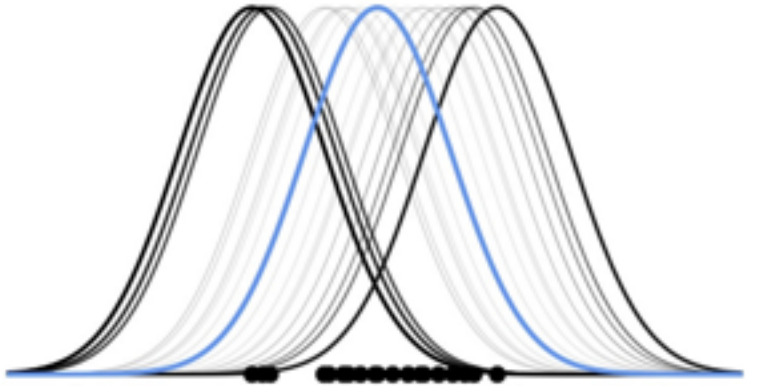
Webb, Steve, James Caverlee, and Calton Pu. "Introducing the Webb Spam Corpus: Using Email Spam to Identify Web Spam Automatically." In *CEAS*. 2006.

Additional image references (5/5)

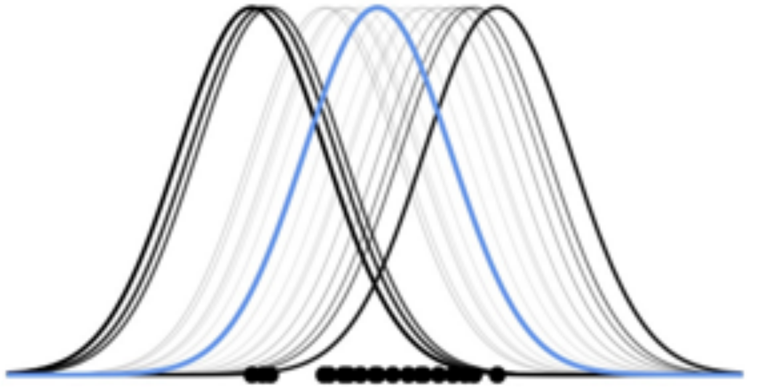
amCharts. Visited Countries Map. https://www.amcharts.com/visited_countries/ Accessed: 2016.

J. Herzog. 3 June 2016, 17:17:30. Obtained from: https://commons.wikimedia.org/wiki/File:Airbus_A350-941_F-WWCF_MSN002_ILA_Berlin_2016_17.jpg (Creative Commons Attribution 4.0 International License)

Practicalities

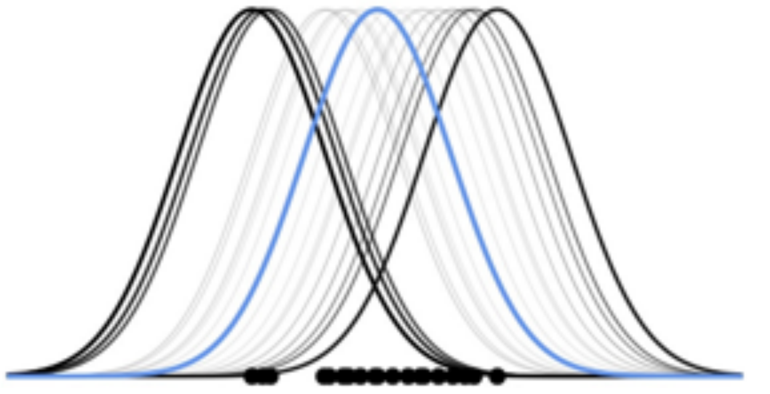


Practicalities



- Choice of norm

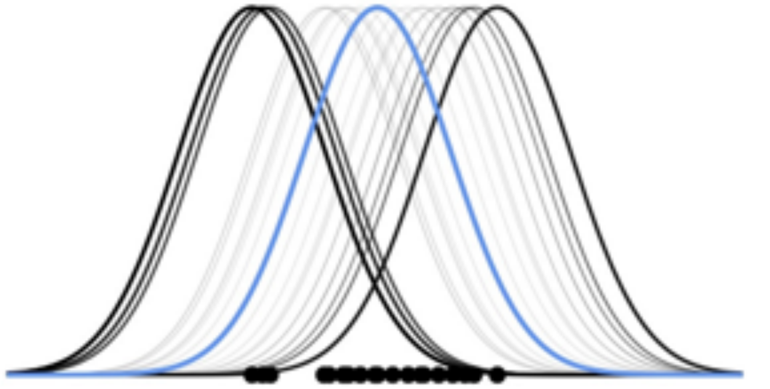
Practicalities



- Choice of norm
  - E.g. (weighted) Fisher information distance

$$\|\mathcal{L}(w) - \mathcal{L}\|^2_{\hat{\pi},F} := \mathbb{E}_{\hat{\pi}}\left[\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(w,\theta)\|_2^2\right]$$

Practicalities



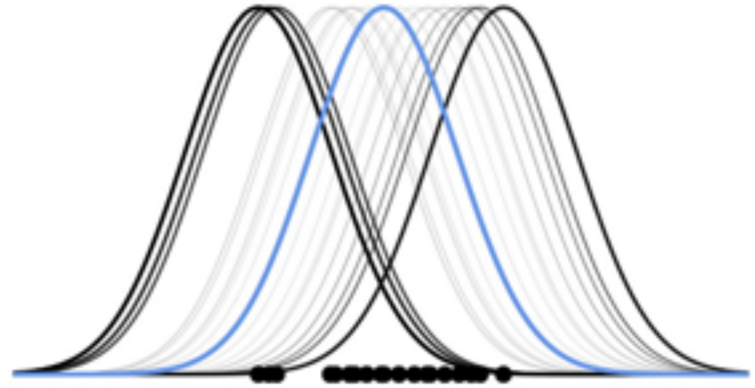
- Choice of norm
  - E.g. (weighted) Fisher information distance

$$\|\mathcal{L}(w) - \mathcal{L}\|^2_{\hat{\pi},F} := \mathbb{E}_{\hat{\pi}}\left[\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(w,\theta)\|_2^2\right]$$

- Associated inner product:

$$\langle\mathcal{L}_n, \mathcal{L}_m\rangle_{\hat{\pi},F} := \mathbb{E}_{\hat{\pi}}\left[\nabla\mathcal{L}_n(\theta)^T\nabla\mathcal{L}_m(\theta)\right]$$

Practicalities

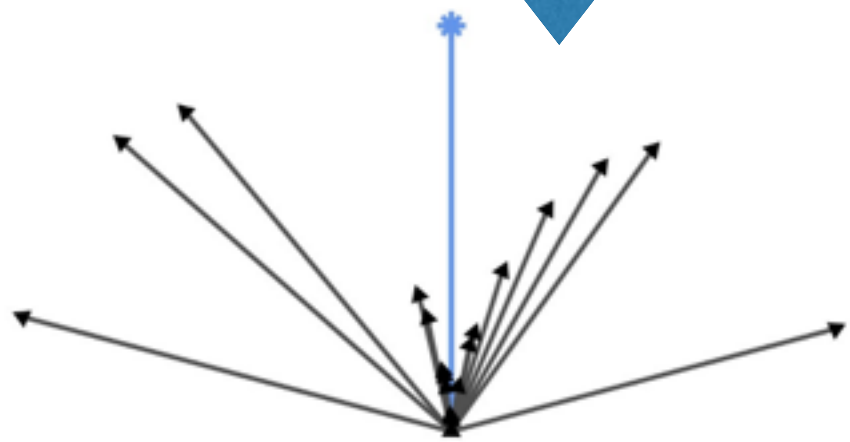


- Choice of norm
  - E.g. (weighted) Fisher information distance

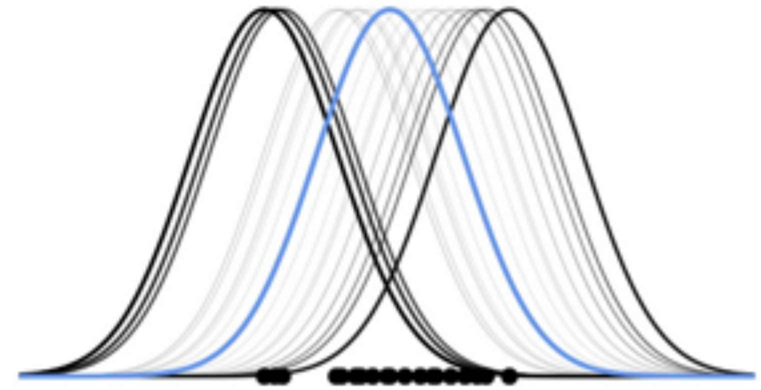
$$\|\mathcal{L}(w) - \mathcal{L}\|^2_{\hat{\pi},F} := \mathbb{E}_{\hat{\pi}}\left[\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(w,\theta)\|_2^2\right]$$

    - Associated inner product:

$$\langle\mathcal{L}_n, \mathcal{L}_m\rangle_{\hat{\pi},F} := \mathbb{E}_{\hat{\pi}}\left[\nabla\mathcal{L}_n(\theta)^T\nabla\mathcal{L}_m(\theta)\right]$$



Practicalities



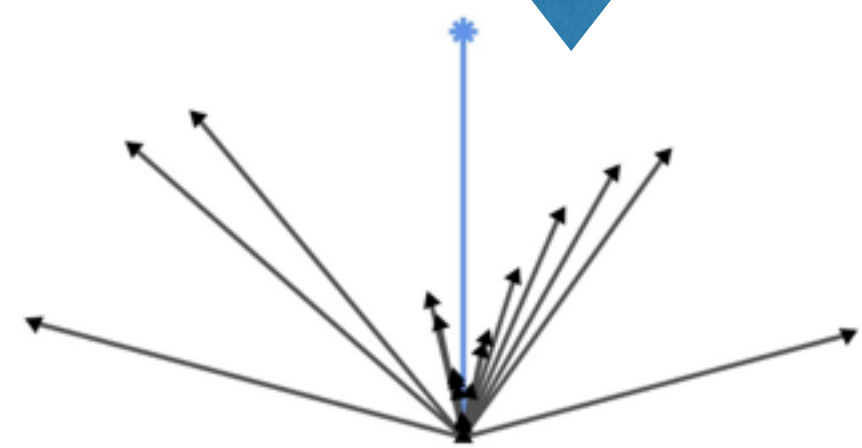
- Choice of norm
  - E.g. (weighted) Fisher information distance

$$\|\mathcal{L}(w) - \mathcal{L}\|^2_{\hat{\pi},F} := \mathbb{E}_{\hat{\pi}}\left[\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(w,\theta)\|^2_2\right]$$

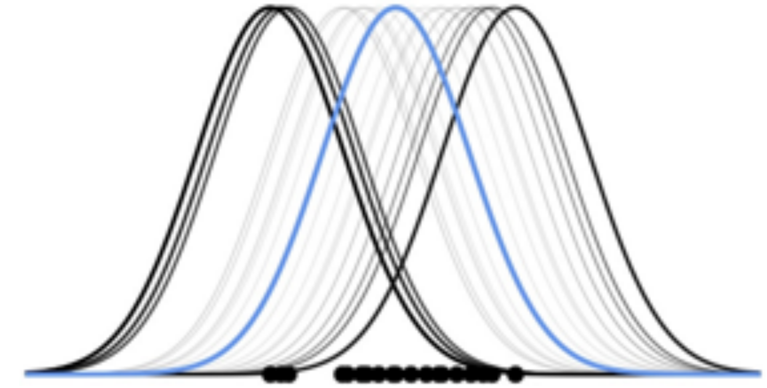
    - Associated inner product:

$$\langle\mathcal{L}_n, \mathcal{L}_m\rangle_{\hat{\pi},F} := \mathbb{E}_{\hat{\pi}}\left[\nabla\mathcal{L}_n(\theta)^T\nabla\mathcal{L}_m(\theta)\right]$$

- Random feature projection



Practicalities



- Choice of norm
  - E.g. (weighted) Fisher information distance

$$\|\mathcal{L}(w) - \mathcal{L}\|^2_{\hat{\pi},F} := \mathbb{E}_{\hat{\pi}}\left[\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(w,\theta)\|_2^2\right]$$

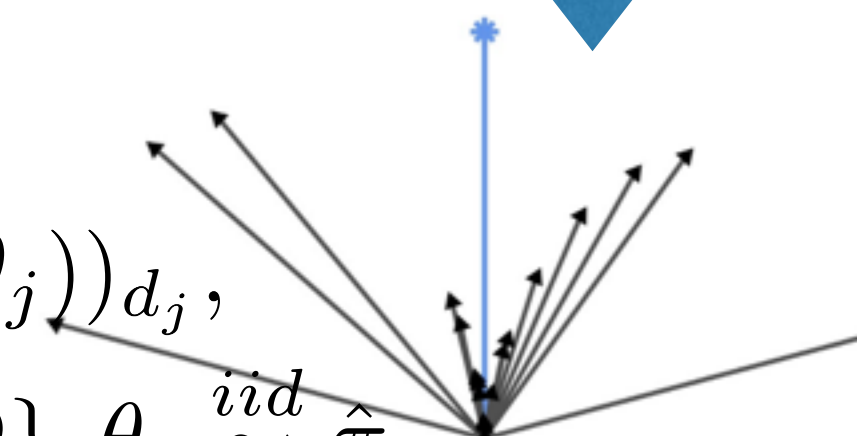
  - Associated inner product:

$$\langle\mathcal{L}_n, \mathcal{L}_m\rangle_{\hat{\pi},F} := \mathbb{E}_{\hat{\pi}}\left[\nabla\mathcal{L}_n(\theta)^T\nabla\mathcal{L}_m(\theta)\right]$$

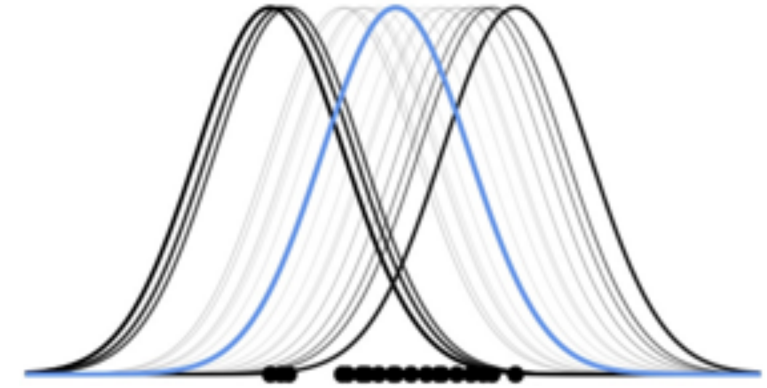
- Random feature projection

$$\langle\mathcal{L}_n, \mathcal{L}_m\rangle_{\hat{\pi},F} \approx \frac{D}{J}\sum_{j=1}^{J}(\nabla\mathcal{L}_n(\theta_j))_{d_j}(\nabla\mathcal{L}_m(\theta_j))_{d_j},$$

$$d_j \overset{iid}{\sim} \text{Unif}\{1,\ldots,D\}, \theta_j \overset{iid}{\sim} \hat{\pi}$$



Practicalities



- Choice of norm
  - E.g. (weighted) Fisher information distance

$$\|\mathcal{L}(w) - \mathcal{L}\|^2_{\hat{\pi},F} := \mathbb{E}_{\hat{\pi}}\left[\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(w,\theta)\|_2^2\right]$$

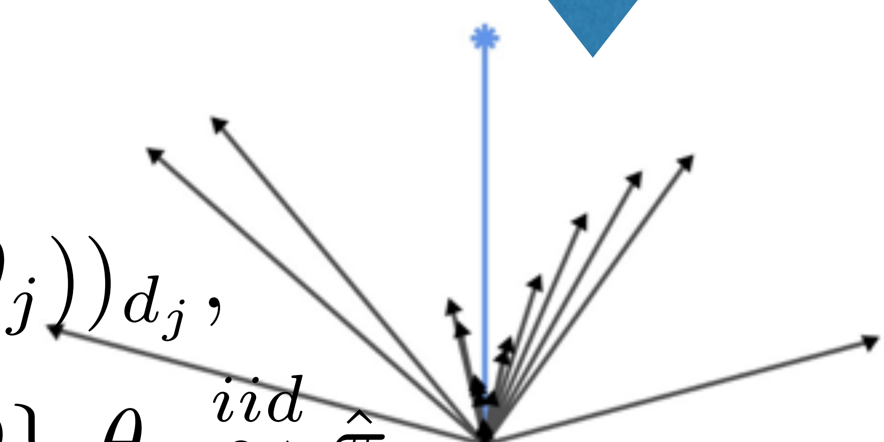
    - Associated inner product:

$$\langle\mathcal{L}_n, \mathcal{L}_m\rangle_{\hat{\pi},F} := \mathbb{E}_{\hat{\pi}}\left[\nabla\mathcal{L}_n(\theta)^T\nabla\mathcal{L}_m(\theta)\right]$$

- Random feature projection

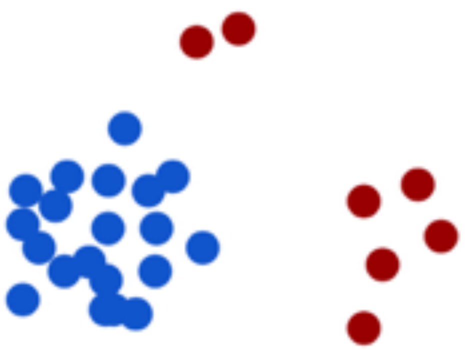
$$\langle\mathcal{L}_n, \mathcal{L}_m\rangle_{\hat{\pi},F} \approx \frac{D}{J}\sum_{j=1}^{J}(\nabla\mathcal{L}_n(\theta_j))_{d_j}(\nabla\mathcal{L}_m(\theta_j))_{d_j},$$

$$d_j \overset{iid}{\sim} \text{Unif}\{1,\ldots,D\}, \theta_j \overset{iid}{\sim} \hat{\pi}$$



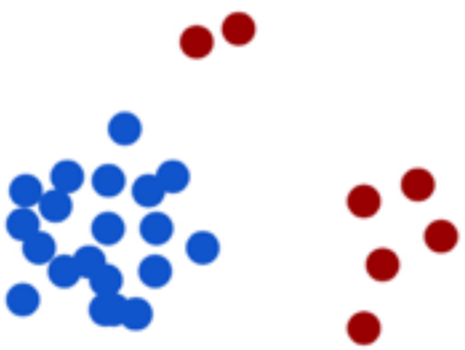
Thm sketch (CB). With high probability and large enough J, a good coreset after random feat. proj. is a good coreset for $(\mathcal{L}_n)_{n=1}^N$

Full pipeline



N
dataset size

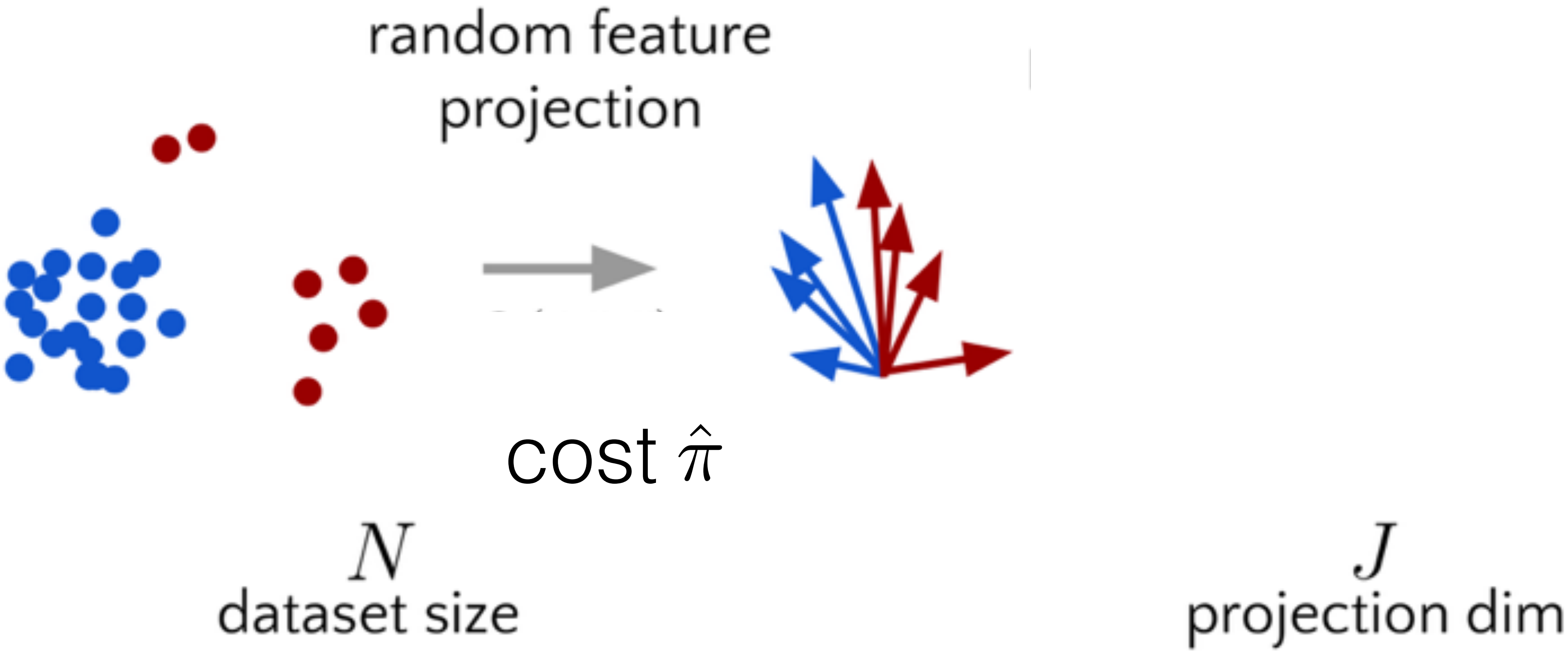
Full pipeline



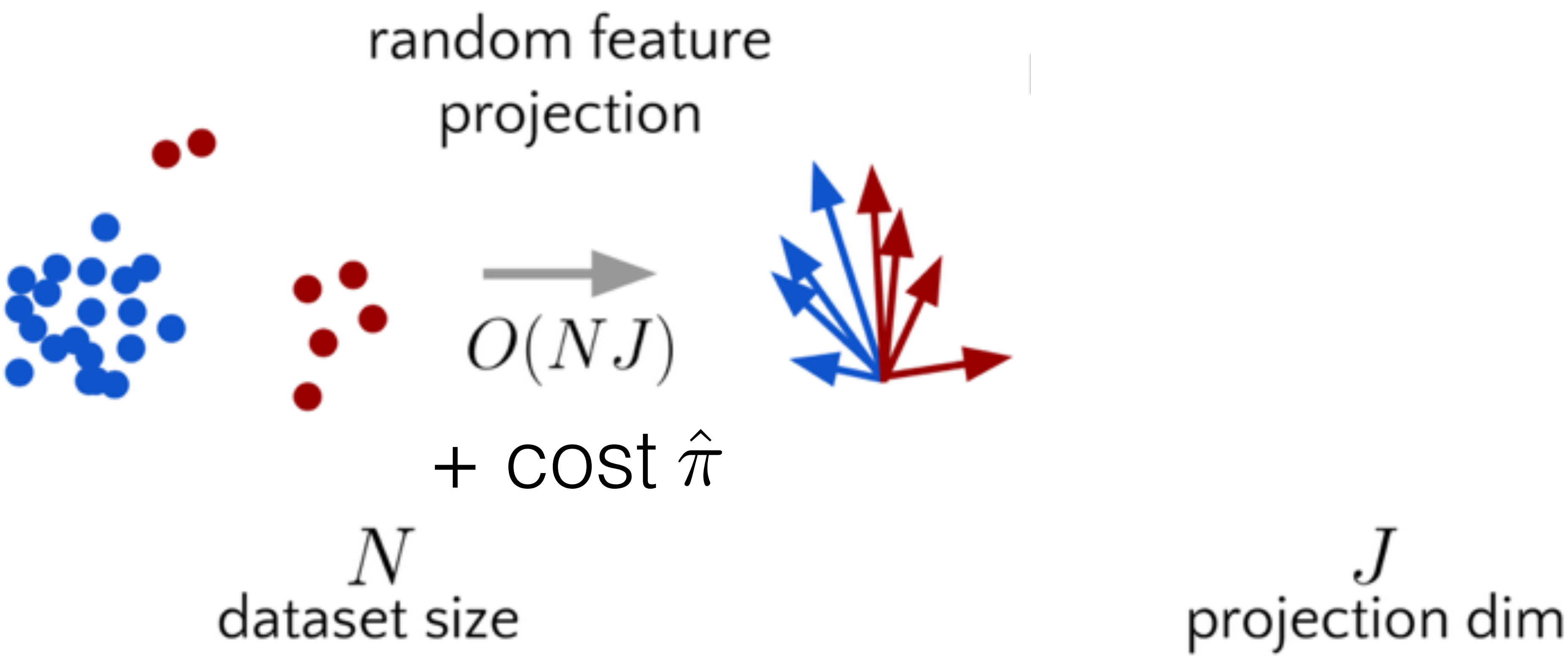
cost $\hat{\pi}$

N
dataset size

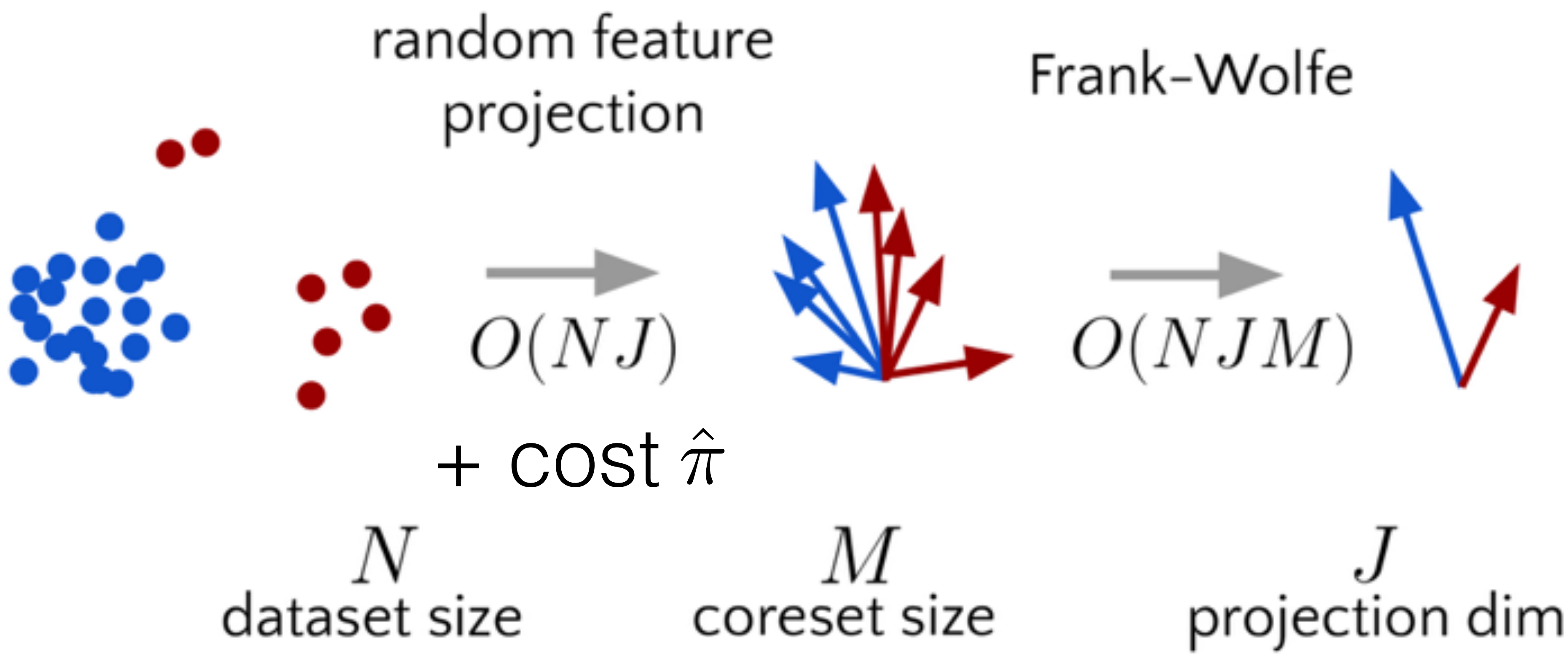
Full pipeline



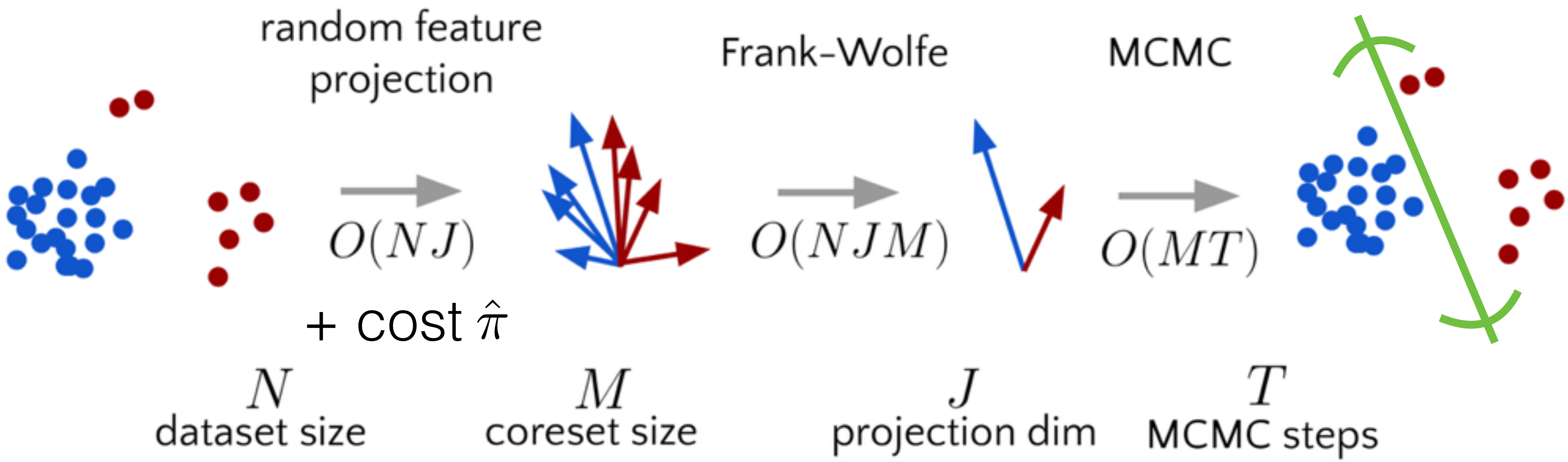
Full pipeline



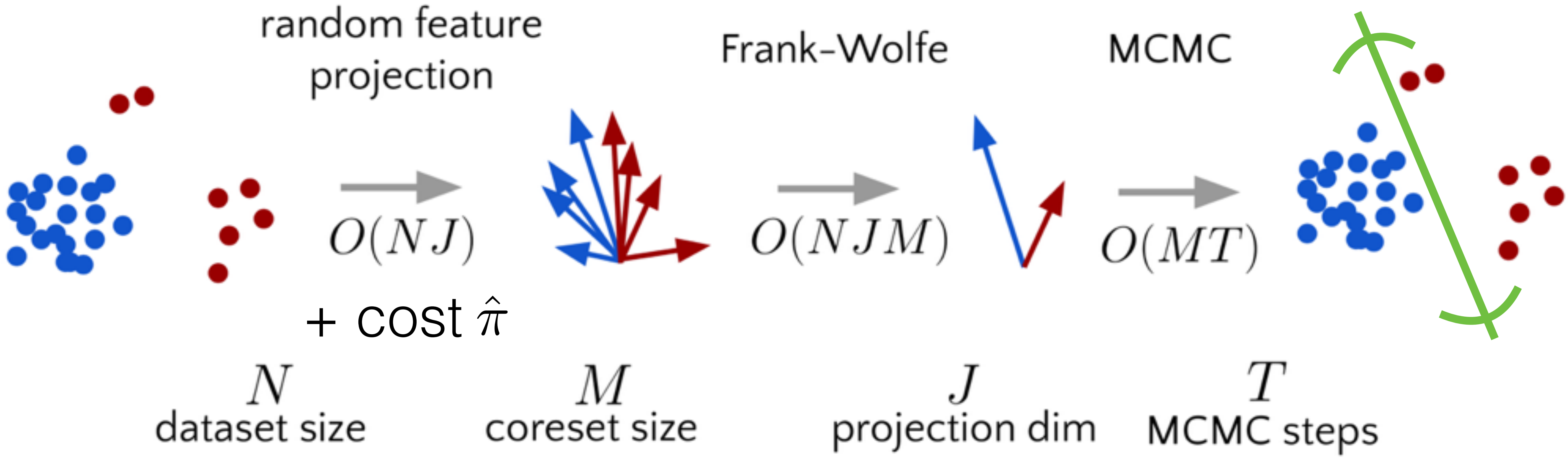
Full pipeline



Full pipeline

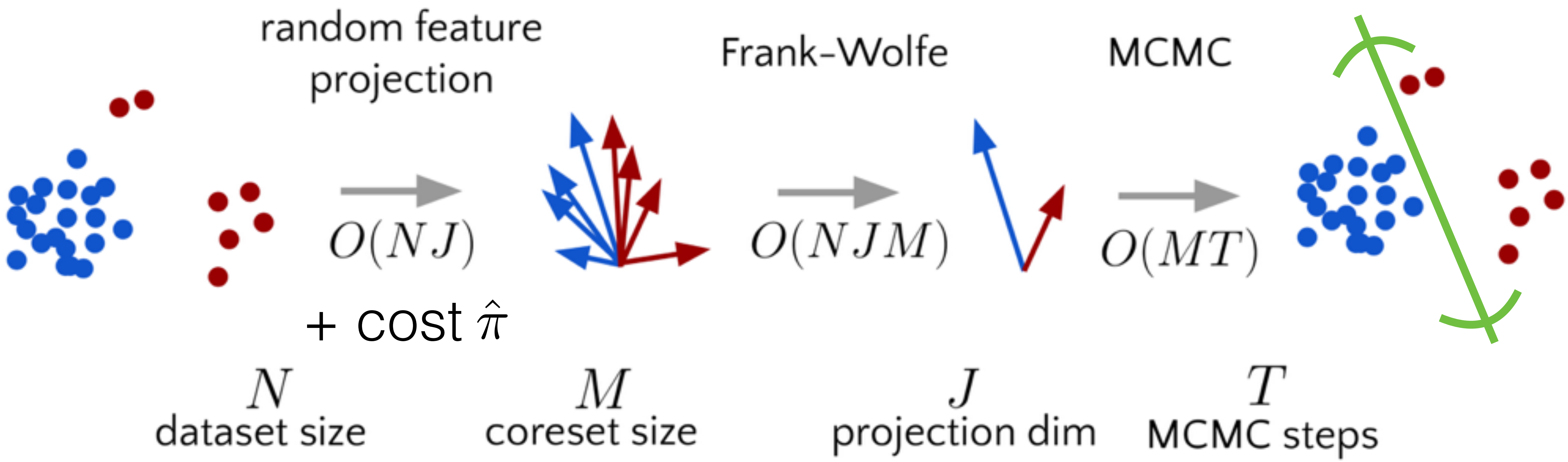


Full pipeline



- vs. $O(NT)$

Full pipeline



- vs. $O(NT)$
- Can make streaming, distributed