Bayesian statistics, Occam razor, and model-independent learning of continuous probability densities

Ilya Nemenman
ITP, UCSB

Joint work with:

William Bialek, Princeton University

Bayesian statistics . . .

This is how it compares to other ultimate answers:

Bayesian statistics . . .

. . . claims that the the answer to the Great Question of Life, The Universe and Everything is not 42, but

This is how it compares to other ultimate answers:

Bayesian statistics ...

... claims that the the answer to the Great Question of Life, The Universe and Everything is not 42, but

$$P(\mathrm{Model}|\mathrm{Data}) = \frac{P(\mathrm{Data}|\mathrm{Model})\, \mathcal{P}(\mathrm{Model})}{P(\mathrm{Data})}$$

This is how it compares to other ultimate answers:

Bayesian statistics . . .

. . . claims that the the answer to the Great Question of Life, The Universe and Everything is not 42, but

$$P(\mathrm{Model}|\mathrm{Data}) = \frac{P(\mathrm{Data}|\mathrm{Model})\,\mathcal{P}(\mathrm{Model})}{P(\mathrm{Data})}$$

This is how it compares to other ultimate answers:

Bayes Best others

	Bayes	Best others
consistency	✓	✓
convergence rates	optimal	optimal
model selection	?	✓ (disagreement remains)
use of prior knowledge	✓	?

	Bayes	Best others
model selection	✓ (disagreement remains)	
use of prior knowledge	?	later

	Bayes	Best others	
model selection	?		today

Bayesian model selection for finitely parameterizable distributions

Bayesian model selection for finitely parameterizable distributions

$$P(x)$$

unknown

Bayesian model selection for finitely parameterizable distributions

$$P(x) \xrightarrow{\text{i.i.d.}} X = \{x_1 \cdots x_N\}$$

unknown

Bayesian model selection for finitely parameterizable distributions

$$P(x) \xrightarrow{\text{i.i.d.}} X = \{x_1 \cdots x_N\}$$

unknown

Model family A
$Q_A(x|\boldsymbol{\alpha})$
$\dim \boldsymbol{\alpha} = K_A$
$\mathcal{P}_A(\boldsymbol{\alpha})$, $Pr(A)$

Bayesian model selection for finitely parameterizable distributions

$$P(x) \xrightarrow{\text{i.i.d.}} X = \{x_1 \cdots x_N\}$$

unknown

Model family A
$Q_A(x|\boldsymbol{\alpha})$
$\dim \boldsymbol{\alpha} = K_A$
$\mathcal{P}_A(\boldsymbol{\alpha})$, $Pr(A)$

Model family B
$Q_B(x|\boldsymbol{\beta})$
$\dim \boldsymbol{\beta} = K_B$
$\mathcal{P}_B(\boldsymbol{\beta})$, $Pr(B)$

Bayesian model selection for finitely parameterizable distributions

$$P(x) \xrightarrow{\text{i.i.d.}} X = \{x_1 \cdots x_N\}$$

unknown

A or B?

Model family A
$Q_A(x|\boldsymbol{\alpha})$
$\dim \boldsymbol{\alpha} = K_A$
$\mathcal{P}_A(\boldsymbol{\alpha})$, $Pr(A)$

Model family B
$Q_B(x|\boldsymbol{\beta})$
$\dim \boldsymbol{\beta} = K_B$
$\mathcal{P}_B(\boldsymbol{\beta})$, $Pr(B)$

Solution

Find the model with maximum posterior probability!

Solution

Find the model with maximum posterior probability!

For example, for model A:

$$P(A|X) = \frac{P(X|A)Pr(A)}{P(X)} \longleftarrow P(X|A)Pr(A) + P(X|B)Pr(B) \equiv Z$$

$$P(X|A) = \int d\boldsymbol{\alpha} \mathcal{P}_A(\boldsymbol{\alpha})\, P(X|\boldsymbol{\alpha}) \sim P(X|\boldsymbol{\alpha}_{\mathrm{ML}})\ ||\delta\boldsymbol{\alpha}_{\mathrm{ML}}||$$

Solution

Find the model with maximum posterior probability!

For example, for model A:

$$P(A|X) = \frac{P(X|A)Pr(A)}{P(X)} \longleftarrow P(X|A)Pr(A) + P(X|B)Pr(B) \equiv Z$$

$$P(X|A) = \int d\boldsymbol{\alpha} \mathcal{P}_A(\boldsymbol{\alpha})\, P(X|\boldsymbol{\alpha}) \sim P(X|\boldsymbol{\alpha}_{\mathrm{ML}})\ ||\delta\boldsymbol{\alpha}_{\mathrm{ML}}||$$

For large K_A, $\delta\boldsymbol{\alpha}_{\mathrm{ML}}$ (region of "good" $\boldsymbol{\alpha}$) decreases.

More complicated models are penalized!

Solution

Find the model with maximum posterior probability!

For example, for model A:

$$P(A|X) = \frac{P(X|A)Pr(A)}{P(X)} \longleftarrow P(X|A)Pr(A) + P(X|B)Pr(B) \equiv Z$$

$$P(X|A) = \int d\boldsymbol{\alpha}\mathcal{P}_A(\boldsymbol{\alpha})\, P(X|\boldsymbol{\alpha}) \sim P(X|\boldsymbol{\alpha}_{\rm ML})\ ||\delta\boldsymbol{\alpha}_{\rm ML}||$$

For large K_A, $\delta\boldsymbol{\alpha}_{\rm ML}$ (region of "good" $\boldsymbol{\alpha}$) decreases.

More complicated models are penalized!

(See: Bayes factors, Occam factors; Jaynes 1968, 1979)

Large N expansion

Saddle point (large N) expansion is almost always valid.

(See: Schwartz 1978, MacKay 1992, Balasubramanian 1996)

Large N expansion

Saddle point (large N) expansion is almost always valid.

$$\log P(A|X) \to \sum_i \underbrace{\log Q_A(x_i|\boldsymbol{\alpha}_{\mathrm{ML}})}$$

(See: Schwartz 1978, MacKay 1992, Balasubramanian 1996)

Large N expansion

Saddle point (large N) expansion is almost always valid.

$$\log P(A|X) \to \sum_i \underbrace{\log Q_A(x_i|\boldsymbol{\alpha}_{\rm ML})}$$

$$\underbrace{- \frac{K_A}{2}\log N - \log\det \partial^2{}_{\boldsymbol{\alpha}_{\rm ML}} \frac{\sum_i \log Q(x_i|\boldsymbol{\alpha}_{\rm ML})}{N}}$$

(See: Schwartz 1978, MacKay 1992, Balasubramanian 1996)

Large N expansion

Saddle point (large N) expansion is almost always valid.

$$\log P(A|X) \to \sum_i \underbrace{\log Q_A(x_i|\boldsymbol{\alpha}_{\mathrm{ML}})}$$

$$- \underbrace{\frac{K_A}{2}\log N - \log\det\partial^2{}_{\boldsymbol{\alpha}_{\mathrm{ML}}}\frac{\sum_i \log Q(x_i|\boldsymbol{\alpha}_{\mathrm{ML}})}{N}}$$

$$+ \log\mathcal{P}(\boldsymbol{\alpha}_{\mathrm{ML}}) + o(N^0)$$

(See: Schwartz 1978, MacKay 1992, Balasubramanian 1996)

Large N expansion

Saddle point (large N) expansion is almost always valid.

$$\log P(A|X) \rightarrow \sum_i \underbrace{\log Q_A(x_i|\boldsymbol{\alpha}_{\rm ML})}_{\text{goodness of fit}}$$

$$\underbrace{- \frac{K_A}{2} \log N - \log \det \partial^2{}_{\boldsymbol{\alpha}_{\rm ML}} \frac{\sum_i \log Q(x_i|\boldsymbol{\alpha}_{\rm ML})}{N}}$$

$$+ \log \mathcal{P}(\boldsymbol{\alpha}_{\rm ML}) + o(N^0)$$

(See: Schwartz 1978, MacKay 1992, Balasubramanian 1996)

Large N expansion

Saddle point (large N) expansion is almost always valid.

$$\log P(A|X) \to \sum_i \underbrace{\log Q_A(x_i|\boldsymbol{\alpha}_{\mathrm{ML}})}_{\text{goodness of fit}}$$

$$\underbrace{-\frac{K_A}{2}\log N - \log\det\partial^2{}_{\boldsymbol{\alpha}_{\mathrm{ML}}}\frac{\sum_i \log Q(x_i|\boldsymbol{\alpha}_{\mathrm{ML}})}{N}}_{\text{generalization error, fluctuations, complexity; weak dependence on priors}}$$

$$+\log\mathcal{P}(\boldsymbol{\alpha}_{\mathrm{ML}}) + o(N^0)$$

(See: Schwartz 1978, MacKay 1992, Balasubramanian 1996)

Large N expansion

Saddle point (large N) expansion is almost always valid.

$$\log P(A|X) \to \sum_i \underbrace{\log Q_A(x_i|\boldsymbol{\alpha}_{\mathrm{ML}})}_{\text{goodness of fit}}$$

$$\underbrace{- \frac{K_A}{2} \log N - \log \det \partial^2{}_{\boldsymbol{\alpha}_{\mathrm{ML}}} \frac{\sum_i \log Q(x_i|\boldsymbol{\alpha}_{\mathrm{ML}})}{N}}$$

generalization error, fluctuations, complexity; weak dependence on priors

$$+ \log \mathcal{P}(\boldsymbol{\alpha}_{\mathrm{ML}}) + o(N^0)$$

(See: Schwartz 1978, MacKay 1992, Balasubramanian 1996)

Conclusions

Conclusions

- Bayesian inference penalizes for complexity (large K)

Conclusions

- Bayesian inference penalizes for complexity (large K)
- Fight between the goodness of fit and the complexity selects an optimal model family.

Conclusions

- Bayesian inference penalizes for complexity (large K)
- Fight between the goodness of fit and the complexity selects an optimal model family.
- This is a Bayesian analogue of the MDL principle.

Conclusions

- Bayesian inference penalizes for complexity (large K)
- Fight between the goodness of fit and the complexity selects an optimal model family.
- This is a Bayesian analogue of the MDL principle.

Does this generalize to infinite–dimensional models?

Estimating density

Estimating density

Standard setting
(solving IE)

Fisher–Wald setting
(minimizing risk)

Estimating density

Standard setting
(solving IE)
$F(t) = \int_{-\infty}^{t} Q(x) dx$

Fisher–Wald setting
(minimizing risk)
$R[Q] = -\int_{-\infty}^{+\infty} \log Q(x) dF(x)$

Estimating density

Standard setting
(solving IE)
$F(t) = \int_{-\infty}^{t} Q(x)dx$
$\frac{1}{N}\sum_{x_i} \Theta(x_i - t) = \int_{-\infty}^{t} Q(x)dx$

Fisher–Wald setting
(minimizing risk)
$R[Q] = -\int_{-\infty}^{+\infty} \log Q(x)dF(x)$
$R_{\mathrm{emp}}[Q] = -\sum_{x_i} \log Q(x_i)$

Estimating density

Standard setting
(solving IE)
$F(t) = \int_{-\infty}^{t} Q(x) dx$
$\frac{1}{N} \sum_{x_i} \Theta(x_i - t) = \int_{-\infty}^{t} Q(x) dx$

Fisher–Wald setting
(minimizing risk)
$R[Q] = -\int_{-\infty}^{+\infty} \log Q(x) dF(x)$
$R_{\mathrm{emp}}[Q] = -\sum_{x_i} \log Q(x_i)$

Both settings hypersensitive to fluctuations in $F(t)$.
Smoothing is *required*.

Bayesian learning for $K \to \infty$

Finite	Infinite

Bayesian learning for $K \to \infty$

Finite	Infinite
$\boldsymbol{\alpha}$	$\phi(x) = -\log \ell_0 Q(x)$

Bayesian learning for $K \to \infty$

Finite	Infinite
$\boldsymbol{\alpha}$	$\phi(x) = -\log \ell_0 Q(x)$
$\mathcal{P}(\boldsymbol{\alpha})$	$\mathcal{P}[Q] \propto \exp\Big[-\frac{\ell^{2\eta-1}}{2} \underbrace{\int dx (\partial_x^\eta \phi)^2}_{}$

Bayesian learning for $K \to \infty$

Finite	Infinite
$\boldsymbol{\alpha}$	$\phi(x) = -\log \ell_0 Q(x)$
$\mathcal{P}(\boldsymbol{\alpha})$	$\underbrace{\mathcal{P}[Q] \propto \exp\left[-\frac{\ell^{2\eta-1}}{2} \underbrace{\int dx (\partial_x^\eta \phi)^2}_{\text{smoothness penalty}} \right]}$

Bayesian learning for $K \to \infty$

Finite	Infinite
$\boldsymbol{\alpha}$	$\phi(x) = -\log \ell_0 Q(x)$
$\mathcal{P}(\boldsymbol{\alpha})$	$\underbrace{\mathcal{P}[Q] \propto \exp\left[-\frac{\ell^{2\eta-1}}{2} \underbrace{\int dx (\partial_x^\eta \phi)^2}_{\text{smoothness penalty}} \right]}_{\text{spline prior of order } 2\eta - 1}$

Bayesian learning for $K \to \infty$

Finite	Infinite
$\boldsymbol{\alpha}$	$\phi(x) = -\log \ell_0 Q(x)$
$\mathcal{P}(\boldsymbol{\alpha})$	$\mathcal{P}[Q] \propto \exp \underbrace{\left[-\frac{\ell^{2\eta-1}}{2} \underbrace{\int dx (\partial_x^\eta \phi)^2}_{\text{smoothness penalty}} \right]}_{\text{spline prior of order } 2\eta-1}$
$\{A, K_A\}$	$\{\ell, \eta(?)\}$ – index continuum of families

Bayesian learning for $K \to \infty$

Finite	Infinite
$\boldsymbol{\alpha}$	$\phi(x) = -\log \ell_0 Q(x)$
$\mathcal{P}(\boldsymbol{\alpha})$	$\underbrace{\mathcal{P}[Q] \propto \exp\left[-\frac{\ell^{2\eta-1}}{2} \underbrace{\int dx (\partial_x^\eta \phi)^2}_{\text{smoothness penalty}}\right]}_{\text{spline prior of order } 2\eta-1}$
$\{A, K_A\}$	$\{\ell, \eta(?)\}$ – index continuum of families
$Pr(A)$	$Pr(\ell, \eta(?))$

Bayesian learning for $K \to \infty$

Finite	Infinite
$\boldsymbol{\alpha}$	$\phi(x) = -\log \ell_0 Q(x)$
$\mathcal{P}(\boldsymbol{\alpha})$	$\mathcal{P}[Q] \propto \exp \underbrace{\left[-\frac{\ell^{2\eta-1}}{2} \underbrace{\int dx (\partial_x^\eta \phi)^2}_{\text{smoothness penalty}} \right]}_{\text{spline prior of order } 2\eta-1}$
$\{A, K_A\}$	$\{\ell, \eta(?)\}$ – index continuum of families
$Pr(A)$	$Pr(\ell, \eta(?))$

(See: Bialek, Callan, Strong, 1996)

Quantum Field Theory analogy

Fix ℓ and η:

$$= \frac{\langle Q(x)Q(x_1)\cdots Q(x_N)\rangle^0}{\underbrace{\langle Q(x_1)\cdots Q(x_N)\rangle^0}}$$

Correlation function in a QFT

defined by $\mathcal{P}[Q]$

Quantum Field Theory analogy

Fix ℓ and η:

$$P[Q|X] = \frac{P(X|Q)\mathcal{P}[Q]}{P(X)}$$

$$= \frac{\langle Q(x)Q(x_1)\cdots Q(x_N)\rangle^0}{\underbrace{\langle Q(x_1)\cdots Q(x_N)\rangle^0}}$$

Correlation function in a QFT

defined by $\mathcal{P}[Q]$

Quantum Field Theory analogy

Fix ℓ and η:

$$
\begin{aligned}
P[Q|X] &= \frac{P(X|Q)\mathcal{P}[Q]}{P(X)} \\
\langle Q \rangle &= \frac{\int [dQ]\, \mathcal{P}[Q]\, Q(x) \prod_{i=1}^{N} Q(x_i)}{\int [dQ]\, P[Q]\, \prod_{i=1}^{N} Q(x_i)} \\
&= \frac{\langle Q(x) Q(x_1) \cdots Q(x_N) \rangle^0}{\underbrace{\langle Q(x_1) \cdots Q(x_N) \rangle^0}}
\end{aligned}
$$

Correlation function in a QFT

defined by $\mathcal{P}[Q]$

Explicit form of correlation functions

$$\text{C. F.} \equiv \int [dQ]\mathcal{P}[Q]\prod_{i=1}^{N} Q(x_i)$$

$$= \int [d\phi]\,\frac{1}{\ell_0^N}\,\mathrm{e}^{-S[\phi]}\,\delta\left[\int dx \frac{1}{\ell_0}\mathrm{e}^{-\phi} - 1\right]$$

$$\underbrace{S[\phi]}_{\text{action}} = \underbrace{\frac{\ell^{2\eta-1}}{2}\int dx(\partial_x^\eta\phi)^2}_{\text{kinetic term}} + \underbrace{\sum_i \phi(x_i)}_{\text{random potential}}$$

Large N approximation for $\eta = 1$

ML (classical, saddle point) solution dominates

Large N approximation for $\eta = 1$

ML (classical, saddle point) solution dominates

$$\ell \partial_x^2 \phi_{\rm cl}(x) + \frac{N}{\ell_0} {\rm e}^{-\phi_{\rm cl}(x)} = \sum_j \delta(x - x_j)$$

Large N approximation for $\eta = 1$

ML (classical, saddle point) solution dominates

converges to
$-\log \ell_0 P(x)$

changes on scale
$\delta x \sim \sqrt{\ell / N P(x)}$

$$\ell \partial_x^2 \phi_{\rm cl}(x) + \frac{N}{\ell_0} {\rm e}^{-\phi_{\rm cl}(x)} = \sum_j \delta(x - x_j)$$

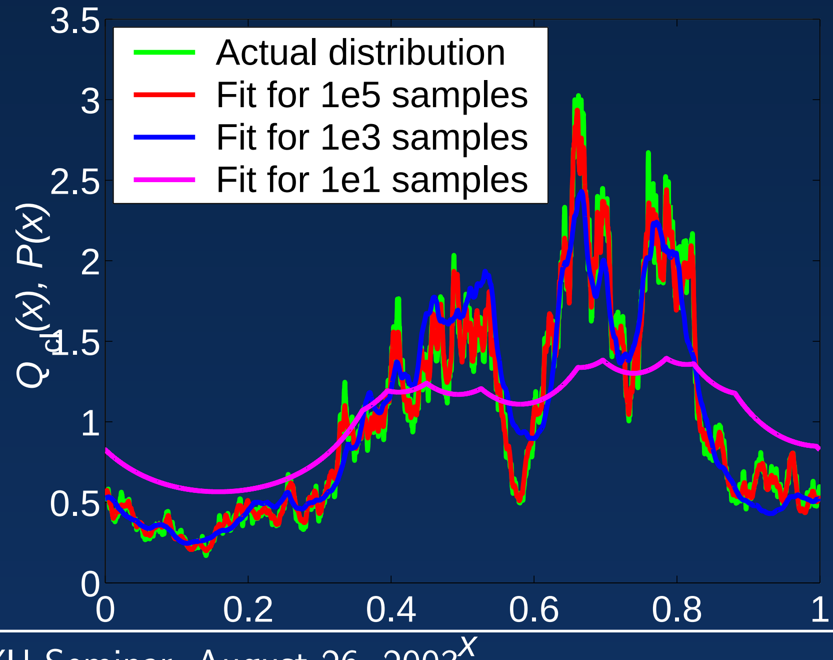
Large N approximation for $\eta = 1$

ML (classical, saddle point) solution dominates

converges to
$-\log \ell_0 P(x)$

changes on scale
$\delta x \sim \sqrt{\ell / N P(x)}$

$$\ell \partial_x^2 \phi_{\rm cl}(x) + \frac{N}{\ell_0} \mathrm{e}^{-\phi_{\rm cl}(x)} = \sum_j \delta(x - x_j)$$



Large N approximation for $\eta = 1$, continued

Van Vleck calculation of functional determinant:

Large N approximation for $\eta = 1$, continued

Van Vleck calculation of functional determinant:

$$\mathrm{C.\ F.} \;\approx\; (1/\ell_0)^N \mathrm{e}^{-S_{\mathrm{eff}}[\phi_{\mathrm{cl}}(x)]}$$

Large N approximation for $\eta = 1$, continued

Van Vleck calculation of functional determinant:

$$
\begin{aligned}
\mathrm{C.\,F.} &\approx (1/\ell_0)^N \mathrm{e}^{-S_{\mathrm{eff}}[\phi_{\mathrm{cl}}(x)]} \\
S_{\mathrm{eff}}[\phi_{\mathrm{cl}}] &= \underbrace{\frac{\ell}{2}\int dx (\partial\phi_{\mathrm{cl}})^2} + \underbrace{\sum \phi_{\mathrm{cl}}(x_i)} \\
&+ \underbrace{\frac{1}{2}\sqrt{\frac{N}{\ell\ell_0}}\int dx \mathrm{e}^{-\phi_{\mathrm{cl}}(x)/2}}
\end{aligned}
$$

Large N approximation for $\eta = 1$, continued

Van Vleck calculation of functional determinant:

$$\begin{aligned}
\text{C. F.} &\approx (1/\ell_0)^N \mathrm{e}^{-S_{\mathrm{eff}}[\phi_{\mathrm{cl}}(x)]} \\
S_{\mathrm{eff}}[\phi_{\mathrm{cl}}] &= \underbrace{\frac{\ell}{2}\int dx (\partial\phi_{\mathrm{cl}})^2}_{\text{prior, smoothness}} + \underbrace{\sum \phi_{\mathrm{cl}}(x_i)}_{\text{goodness of fit}} \\
&+ \underbrace{\frac{1}{2}\sqrt{\frac{N}{\ell\ell_0}}\int dx \mathrm{e}^{-\phi_{\mathrm{cl}}(x)/2}}_{\text{fluctuations, complexity, } error}
\end{aligned}$$

How do we measure performance?

How do we measure performance?

For $x \in [0, L)$ the *universal* learning curve is

$$\Lambda(N) \to \langle D_{\rm KL}(P||Q_{\rm cl})\rangle^0_{\{x_i\}} \sim \sqrt{\frac{L}{\ell N}}$$

How do we measure performance?

For $x \in [0, L)$ the *universal* learning curve is

$$\Lambda(N) \to \langle D_{\mathrm{KL}}(P||Q_{\mathrm{cl}})\rangle^{0}_{\{x_i\}} \sim \sqrt{\frac{L}{\ell N}}$$

For a different η:

$$\Lambda(N) \sim \left(\frac{L}{\ell}\right)^{1/2\eta} N^{1/2\eta-1}$$

Learning curves for fixed ℓ, $\eta = 1$

Learning curves for fixed ℓ, $\eta = 1$

Learner's assumptions $\mathcal{P}_{\ell,\eta=1}[Q]$

Learning curves for fixed ℓ, $\eta = 1$

Learner's assumptions	$\mathcal{P}_{\ell,\eta=1}[Q]$
Actual target distribution	$\mathcal{P}'_{\ell_a,\eta_a}[Q]$

Learning curves for fixed ℓ, $\eta = 1$

Learner's assumptions $\mathcal{P}_{\ell,\eta=1}[Q]$

Actual target distribution $\mathcal{P}'_{\ell_a,\eta_a}[Q]$

$\eta = \eta_a$, $\ell = \ell_a$ learning typical cases, $\mathcal{P} = \mathcal{P}'$

Learning curves for fixed ℓ, $\eta = 1$

Learner's assumptions	$\mathcal{P}_{\ell,\eta=1}[Q]$
Actual target distribution	$\mathcal{P}'_{\ell_a,\eta_a}[Q]$

$\eta = \eta_a$, $\ell = \ell_a$ learning typical cases, $\mathcal{P} = \mathcal{P}'$

$\eta = \eta_a$, $\ell \neq \ell_a$ marginal outliers of $\mathcal{P}$

Learning curves for fixed ℓ, $\eta = 1$

Learner's assumptions	$\mathcal{P}_{\ell,\eta=1}[Q]$
Actual target distribution	$\mathcal{P}'_{\ell_a,\eta_a}[Q]$

$\eta = \eta_a$, $\ell = \ell_a$	learning typical cases, $\mathcal{P} = \mathcal{P}'$
$\eta = \eta_a$, $\ell \neq \ell_a$	marginal outliers of $\mathcal{P}$
$\eta > \eta_a$	extremely rough outliers

Learning curves for fixed ℓ, $\eta = 1$

Learner's assumptions	$\mathcal{P}_{\ell,\eta=1}[Q]$
Actual target distribution	$\mathcal{P}'_{\ell_a,\eta_a}[Q]$

$\eta = \eta_a$, $\ell = \ell_a$	learning typical cases, $\mathcal{P} = \mathcal{P}'$
$\eta = \eta_a$, $\ell \neq \ell_a$	marginal outliers of $\mathcal{P}$
$\eta > \eta_a$	extremely rough outliers
$\eta < \eta_a$	extremely smooth outliers

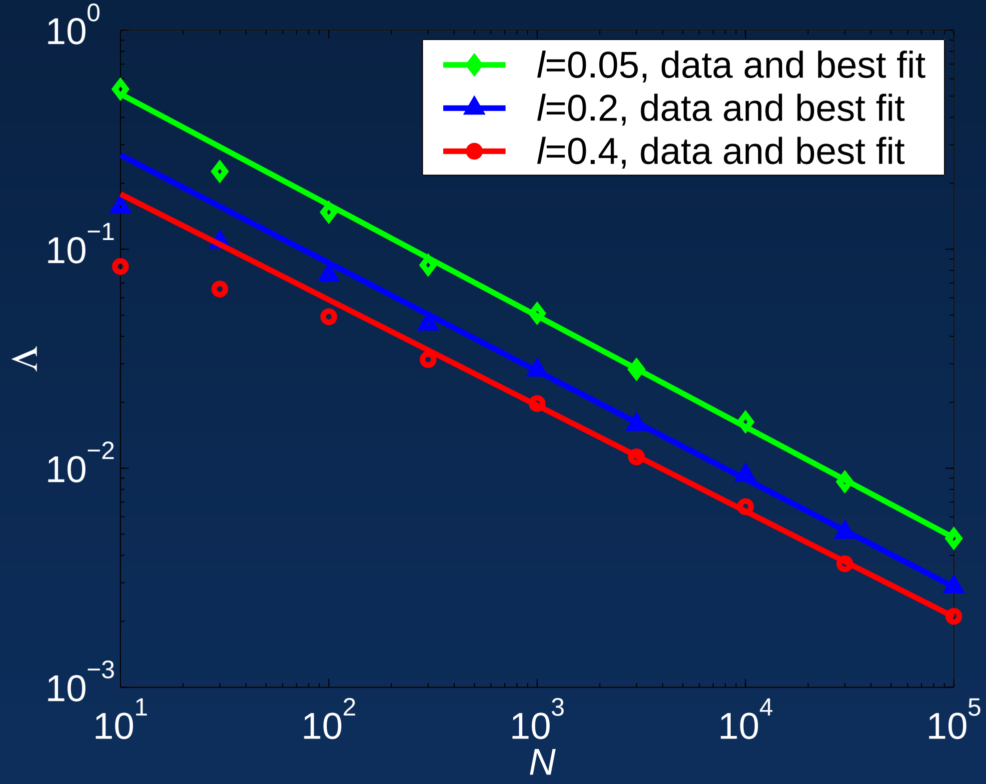
Learning curves for fixed ℓ, $\eta = 1$

Learner's assumptions	$\mathcal{P}_{\ell,\eta=1}[Q]$
Actual target distribution	$\mathcal{P}'_{\ell_a,\eta_a}[Q]$

$\eta = \eta_a$, $\ell = \ell_a$	learning typical cases, $\mathcal{P} = \mathcal{P}'$
$\eta = \eta_a$, $\ell \neq \ell_a$	marginal outliers of $\mathcal{P}$
$\eta > \eta_a$	extremely rough outliers
$\eta < \eta_a$	extremely smooth outliers

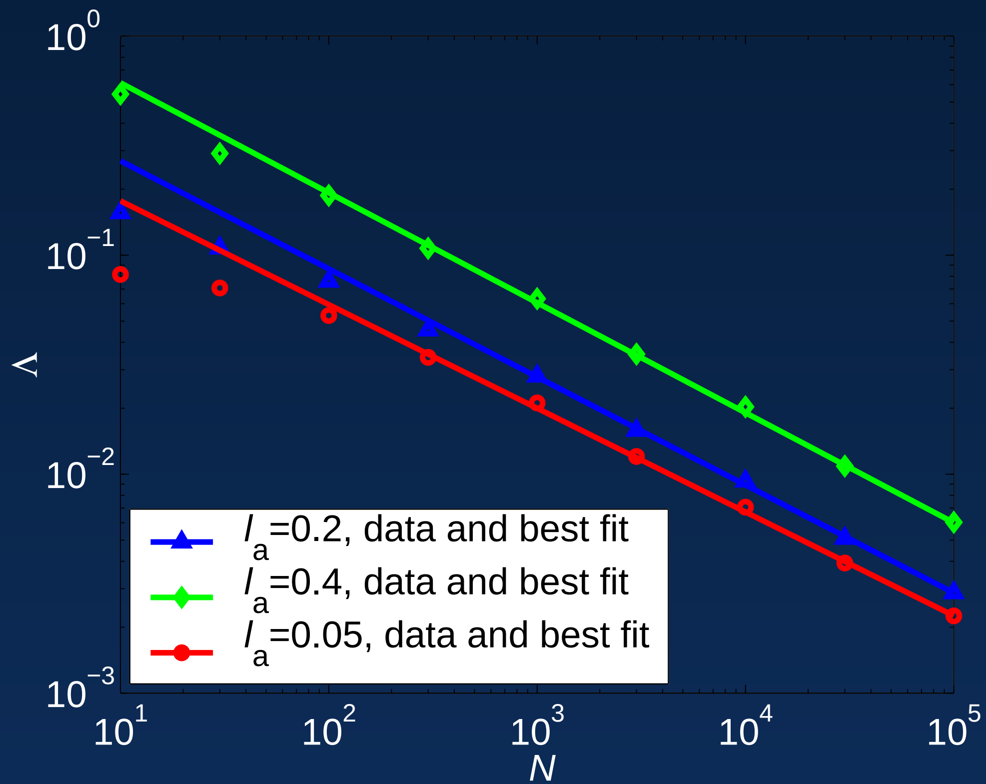
Note: we must have $\eta > 1/2$ for convergence of the integrals.

Learning typical cases



$$
\begin{aligned}
\ell = 0.4, &\quad \Lambda = (0.54 \pm 0.07) N^{-0.483 \pm 0.014} \\
\ell = 0.2, &\quad \Lambda = (0.83 \pm 0.08) N^{-0.493 \pm 0.09} \\
\ell = 0.05, &\quad \Lambda = (1.64 \pm 0.16) N^{-0.507 \pm 0.09}
\end{aligned}
$$

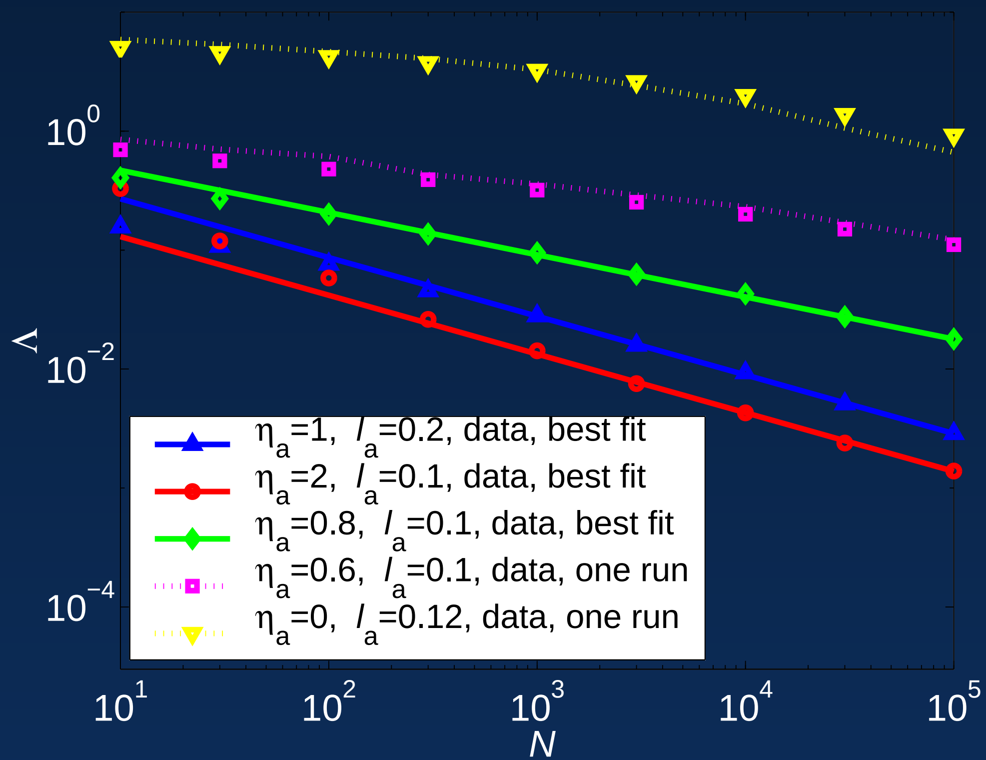
Learning marginal outliers



$$\ell_a = 0.4, \quad \Lambda = (0.56 \pm 0.08) N^{-0.477 \pm 0.015}$$
$$\ell_a = 0.05, \quad \Lambda = (1.90 \pm 0.16) N^{-0.502 \pm 0.008}$$

Learning at $\ell = 0.2$.

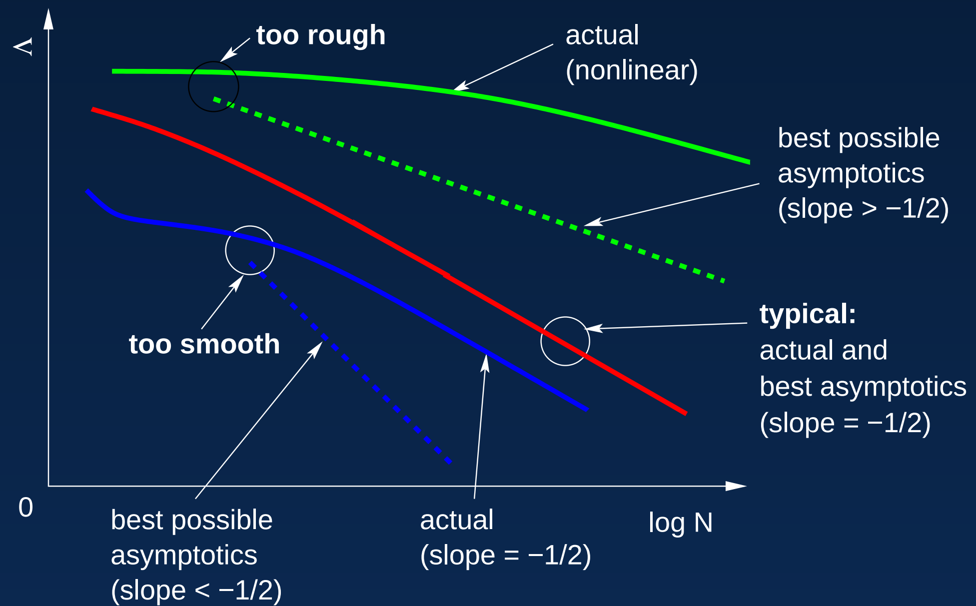
Learning strong outliers



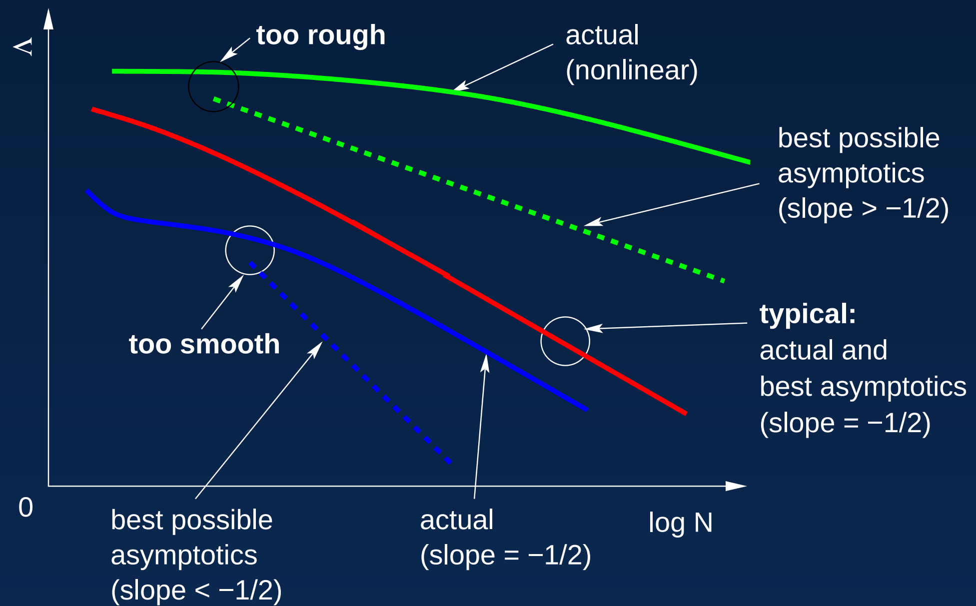
$$\eta_a = 2,\ \ell_a = 0.1, \qquad \Lambda = (0.40 \pm 0.05) N^{-0.493 \pm 0.013}$$
$$\eta_a = 0.8,\ \ell_a = 0.1, \quad \Lambda = (1.06 \pm 0.08) N^{-0.355 \pm 0.008}$$

$\ell = 0.1$ for $\eta_a = 0$ and $\ell = 0.2$ otherwise

Conclusions for fixed η and ℓ

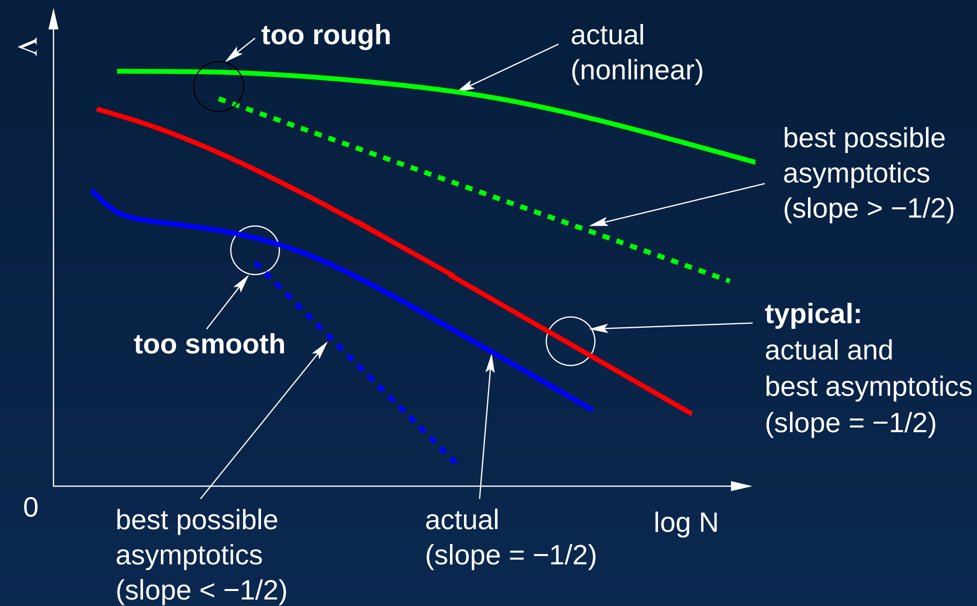


Conclusions for fixed η and ℓ



- No overfits!

Conclusions for fixed η and ℓ



- No overfits!
- but suboptimal performance for learning outliers

Smoothness scale selection

Smoothness scale selection

Allow a prior over ℓ, but keep $\eta = 1$

$$\mathrm{C.\ F.} \to \langle \mathrm{C.\ F.} \rangle_\ell$$

Smoothness scale selection

Allow a prior over ℓ, but keep $\eta = 1$

$$\mathrm{C.\ F.} \to \langle \mathrm{C.\ F.} \rangle_\ell = \int d\ell \; Pr(\ell) \; \mathrm{e}^{-S_{\mathrm{eff}}[\phi_{\mathrm{cl}}(\phi,\ell)]}$$

Smoothness scale selection

Allow a prior over ℓ, but keep $\eta = 1$

$$\mathrm{C.\ F.} \to \langle \mathrm{C.\ F.} \rangle_\ell = \int d\ell \, Pr(\ell) \, \mathrm{e}^{-S_{\mathrm{eff}}[\phi_{\mathrm{cl}}(\phi,\ell)]}$$

$$S_{\mathrm{eff}}[\phi_{\mathrm{cl}}] = \underbrace{\text{smoothing} + \text{data}} + \underbrace{\text{fluctuations}}$$

Smoothness scale selection

Allow a prior over ℓ, but keep $\eta = 1$

$$\mathrm{C.\ F.} \to \langle \mathrm{C.\ F.} \rangle_\ell = \int d\ell \; Pr(\ell) \; \mathrm{e}^{-S_{\mathrm{eff}}[\phi_{\mathrm{cl}}(\phi,\ell)]}$$

$$S_{\mathrm{eff}}[\phi_{\mathrm{cl}}] = \underbrace{\text{smoothing} + \text{data}}_{\text{grows with } \ell} + \underbrace{\text{fluctuations}}_{\text{grows with } 1/\ell}$$

Smoothness scale selection

Allow a prior over ℓ, but keep $\eta = 1$

$$\mathrm{C.\ F.} \to \langle \mathrm{C.\ F.} \rangle_\ell = \int d\ell \, Pr(\ell) \, \mathrm{e}^{-S_{\mathrm{eff}}[\phi_{\mathrm{cl}}(\phi,\ell)]}$$

$$S_{\mathrm{eff}}[\phi_{\mathrm{cl}}] = \underbrace{\text{smoothing} + \text{data}}_{\text{grows with } \ell} + \underbrace{\text{fluctuations}}_{\text{grows with } 1/\ell}$$

Some ℓ^* *always* dominates the C. F. and $\langle Q \rangle$!

Calculations: What is ℓ^* for η_a and ℓ_a?

Averaging over ℓ and allowing $\ell^* = \ell^*(N)$ deals with

Calculations: What is ℓ^* for η_a and ℓ_a?

If $\eta = \eta_a$, then $\ell^* = \ell_a$.

Averaging over ℓ and allowing $\ell^* = \ell^*(N)$ deals with

Calculations: What is ℓ^* for η_a and ℓ_a?

If $\eta = \eta_a$, then $\ell^* = \ell_a$. Otherwise:

$0.5 < \eta_a \le 1.5$	$1.5 < \eta_a$

Averaging over ℓ and allowing $\ell^* = \ell^*(N)$ deals with

Calculations: What is ℓ^* for η_a and ℓ_a?

If $\eta = \eta_a$, then $\ell^* = \ell_a$. Otherwise:

$0.5 < \eta_a \leq 1.5$	$1.5 < \eta_a$
data $>$ smoothing	smoothing $>$ data

Averaging over ℓ and allowing $\ell^* = \ell^*(N)$ deals with

Calculations: What is ℓ^* for η_a and ℓ_a?

If $\eta = \eta_a$, then $\ell^* = \ell_a$. Otherwise:

$0.5 < \eta_a \leq 1.5$	$1.5 < \eta_a$
data $>$ smoothing	smoothing $>$ data
$\ell^* \sim N^{(\eta_a - 1)/\eta_a}$	$\ell^* \sim N^{1/3}$

Averaging over ℓ and allowing $\ell^* = \ell^*(N)$ deals with

Calculations: What is ℓ^* for η_a and ℓ_a?

If $\eta = \eta_a$, then $\ell^* = \ell_a$. Otherwise:

$0.5 < \eta_a \leq 1.5$	$1.5 < \eta_a$
data $>$ smoothing	smoothing $>$ data
$\ell^* \sim N^{(\eta_a - 1)/\eta_a}$ $\Lambda \sim N^{1/2\eta_a - 1}$	$\ell^* \sim N^{1/3}$ $\Lambda \sim N^{-2/3}$

Averaging over ℓ and allowing $\ell^* = \ell^*(N)$ deals with

Calculations: What is ℓ^* for η_a and ℓ_a?

If $\eta = \eta_a$, then $\ell^* = \ell_a$. Otherwise:

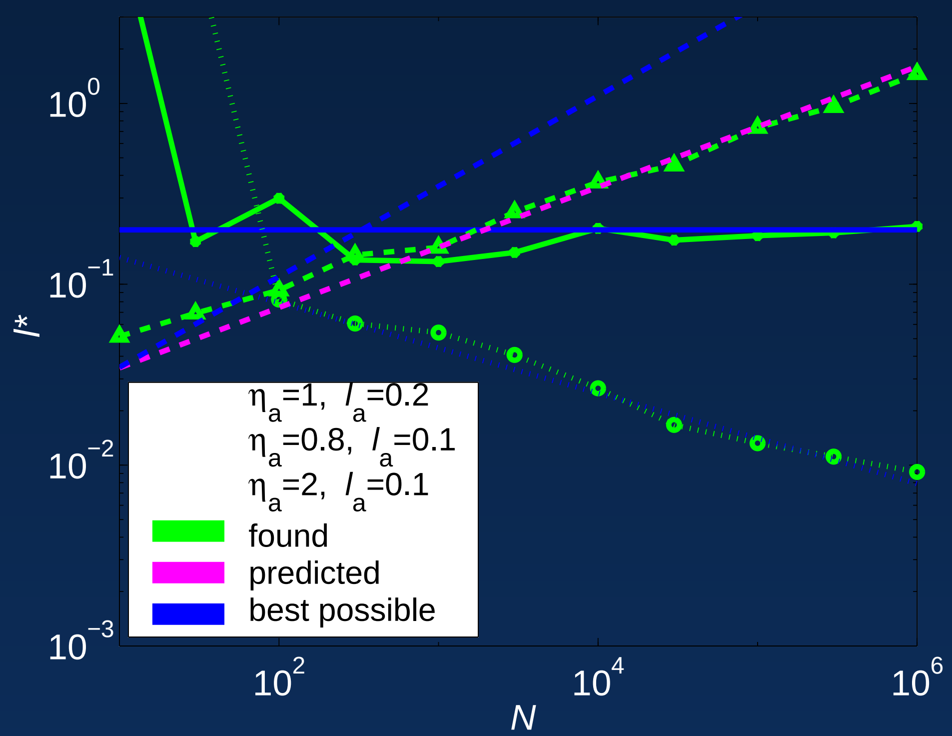
$0.5 < \eta_a \leq 1.5$	$1.5 < \eta_a$
data $>$ smoothing	smoothing $>$ data
$\ell^* \sim N^{(\eta_a - 1)/\eta_a}$ $\Lambda \sim N^{1/2\eta_a - 1}$	$\ell^* \sim N^{1/3}$ $\Lambda \sim N^{-2/3}$
best possible performance	better, but not best performance

Averaging over ℓ and allowing $\ell^* = \ell^*(N)$ deals with

qualitatively wrong smoothness $\eta_a \neq 1$!

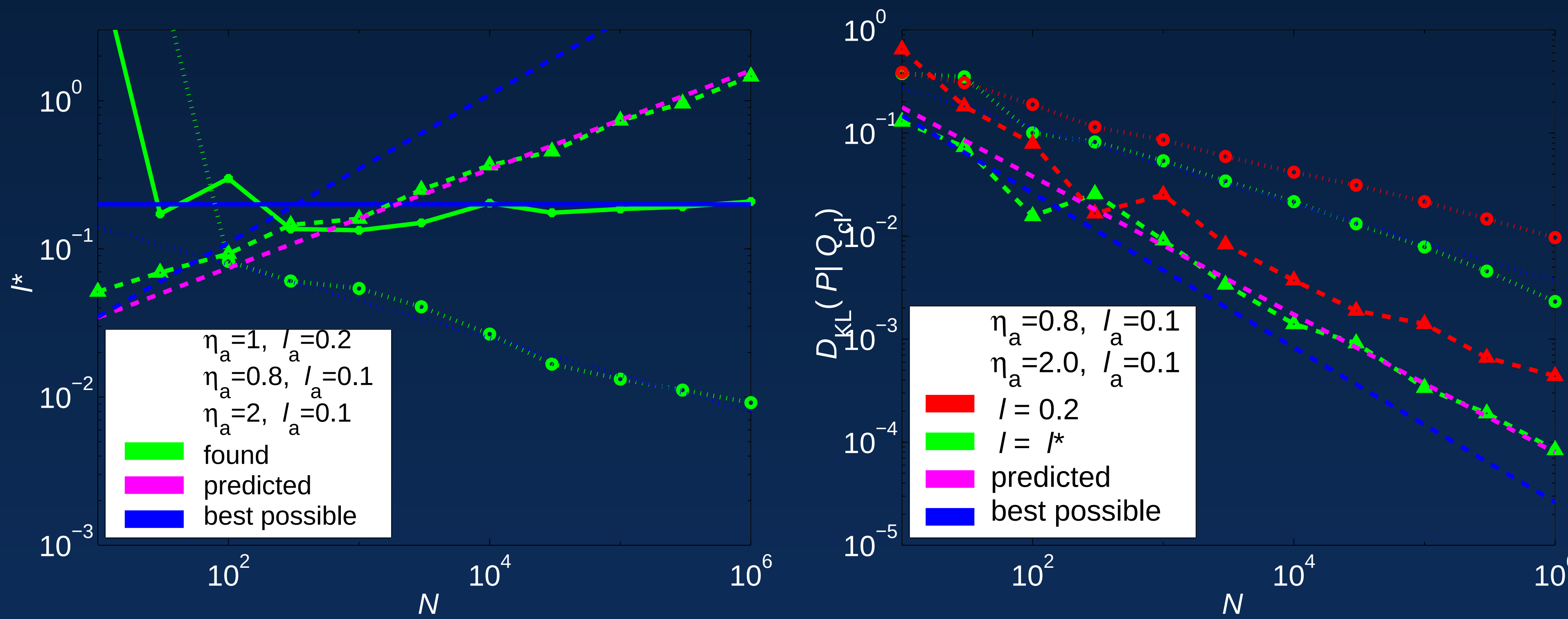
Numerics: What is ℓ^* for η_a and ℓ_a?

Numerics: What is ℓ^* for η_a and ℓ_a?



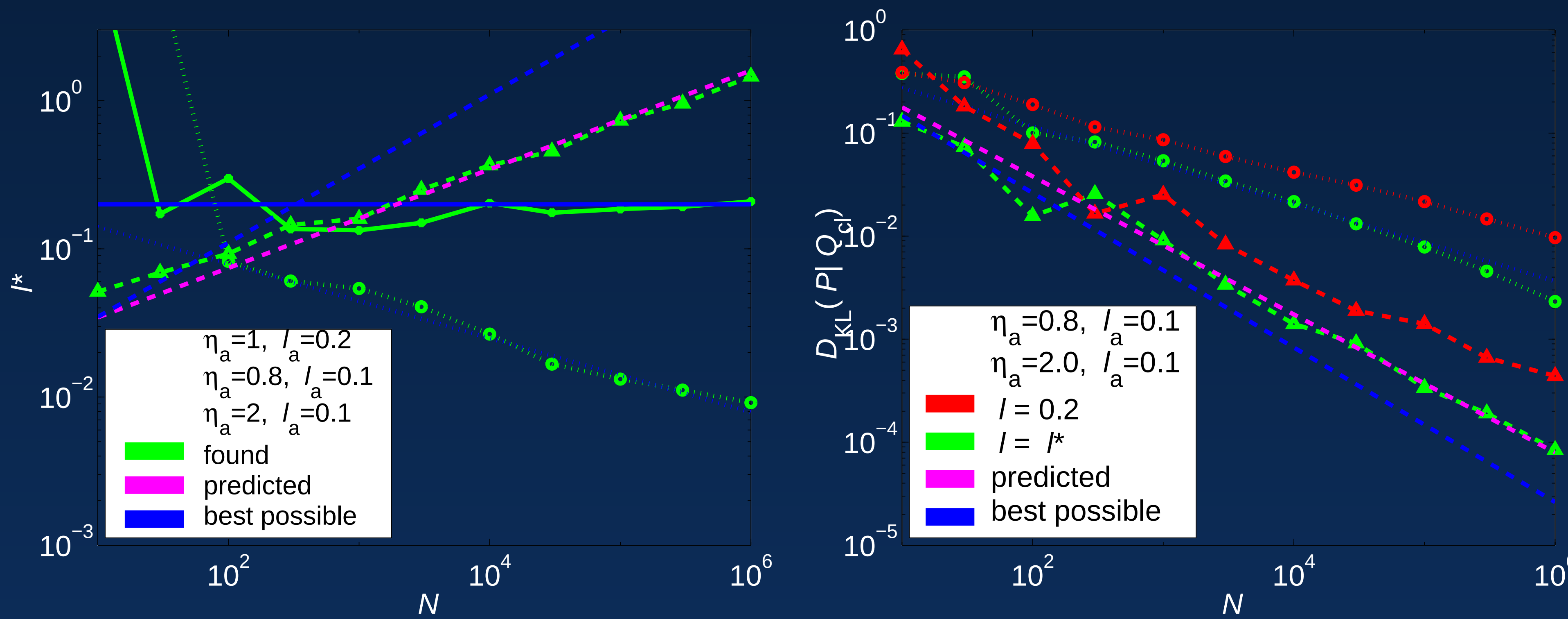
Note: just single runs shown.

Numerics: What is ℓ^* for η_a and ℓ_a?



Note: just single runs shown.

Numerics: What is ℓ^* for η_a and ℓ_a?



Note: just single runs shown.

Approaching model–independend optimal inference!

Analogies

Analogies

- choosing ℓ^* corresponds to selection of a structure element with $d_{\mathrm{VC}} = \sqrt{NL/\ell^*}$ in Vapnik's SRM theory

Analogies

- choosing ℓ^* corresponds to selection of a structure element with $d_{\mathrm{VC}} = \sqrt{NL/\ell^*}$ in Vapnik's SRM theory
- maximizing P over model families (ℓ's) asymptotically corresponds to searching for MDL

Analogies

- choosing ℓ^* corresponds to selection of a structure element with $d_{\mathrm{VC}} = \sqrt{NL/\ell^*}$ in Vapnik's SRM theory

- maximizing P over model families (ℓ's) asymptotically corresponds to searching for MDL

- a lot in common with the Gaussian Processes theory; however normalization constraint is important

Summary

Bayesian smoothness (model) selection works for nonparametric spline priors!

Open questions

Open questions

- constant factor or constant summand?

Open questions

- constant factor or constant summand?
- what to do with $\eta_a > 1.5$?

Open questions

- constant factor or constant summand?
- what to do with $\eta_a > 1.5$?
- reparameterization invariance

Open questions

- constant factor or constant summand?
- what to do with $\eta_a > 1.5$?
- reparameterization invariance
- information theoretic meaningful priors

Open questions

- constant factor or constant summand?
- what to do with $\eta_a > 1.5$?
- reparameterization invariance
- information theoretic meaningful priors
- higher dimensions

Open questions

- constant factor or constant summand?
- what to do with $\eta_a > 1.5$?
- reparameterization invariance
- information theoretic meaningful priors
- higher dimensions
- smooth transition from $K = \mathrm{const}$ to $K \to \infty$

Open questions

- constant factor or constant summand?
- what to do with $\eta_a > 1.5$?
- reparameterization invariance
- information theoretic meaningful priors
- higher dimensions
- smooth transition from $K = \mathrm{const}$ to $K \to \infty$
- which classes of priors are allowed?

Open questions

- constant factor or constant summand?
- what to do with $\eta_a > 1.5$?
- reparameterization invariance
- information theoretic meaningful priors
- higher dimensions
- smooth transition from $K = \text{const}$ to $K \to \infty$
- which classes of priors are allowed?

There is hope that all of this problems are resolvable in a single formulation.