Generalization Error Rates in Kernel Regression: The Crossover from the Noiseless to Noisy Regime

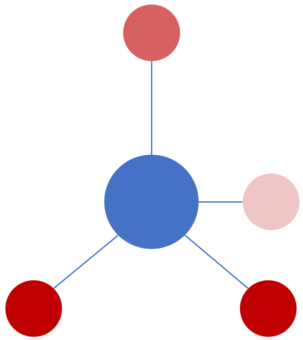
Hugo Cui,[1] Bruno Loureiro,[2] Florent Krzakala,[2] and Lenka Zdeborová[1]
[1] *SPOC, EPFL*
[2] *IDePHICS lab, EPFL*

arXiv: 2105.15004

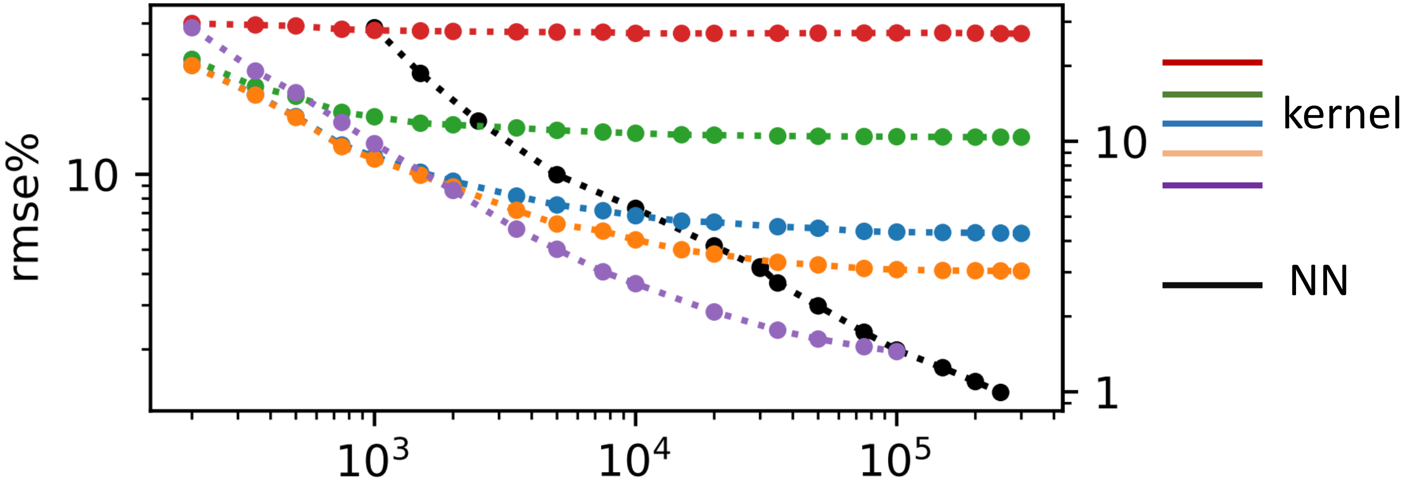
NEURAL INFORMATION PROCESSING SYSTEMS

EPFL
SPOC Statistical Physics Of Computation
IdEPHIX INFORMATION, LEARNING & PHYSICS LAB.

Why study kernels?

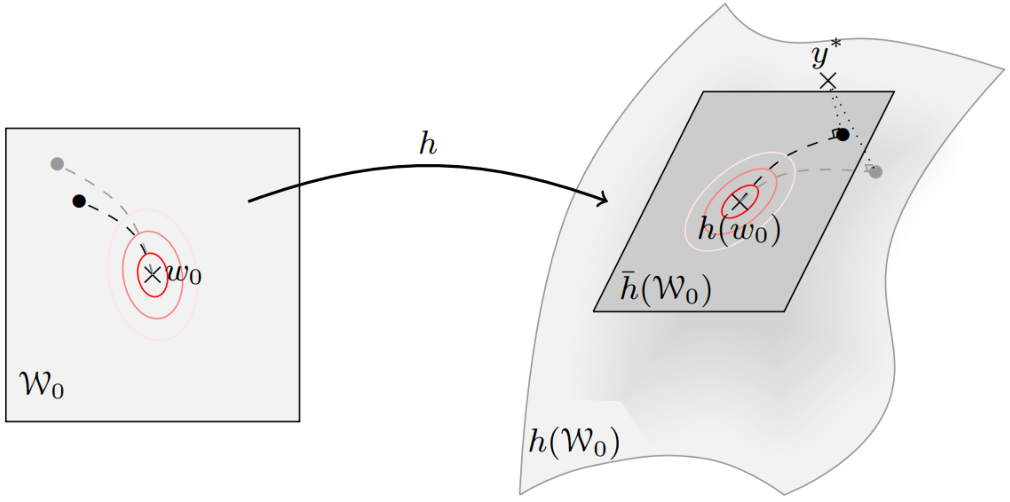


Kernels can outperform NN…



Nigam. Pozdnyakiv, Ceriotti, J. of Chem. Phys, 2020

… and be an interesting limit of NNs



Chizat, Oyallon, Bach, NeurIPS 2018
Jacot, Gabriel, Hongler, NeurIPS 2018

Quick appetizer

For a dataset characterized by

$\boldsymbol{\alpha}$: effective dimension of the dataset

$\boldsymbol{r}$: complexity of the label distribution

*At which rate does the excess error decay with the number of samples n for **kernel ridge**?*

Quick appetizer

For a dataset characterized by

α : effective dimension of the dataset
r : complexity of the label distribution

*At which rate does the excess error decay with the number of samples n for **kernel ridge**?*

$$n^{-2\alpha\min(1,r)}$$

$$n^{-\frac{2\alpha\min(1,r)}{1+2\alpha\min(1,r)}}$$

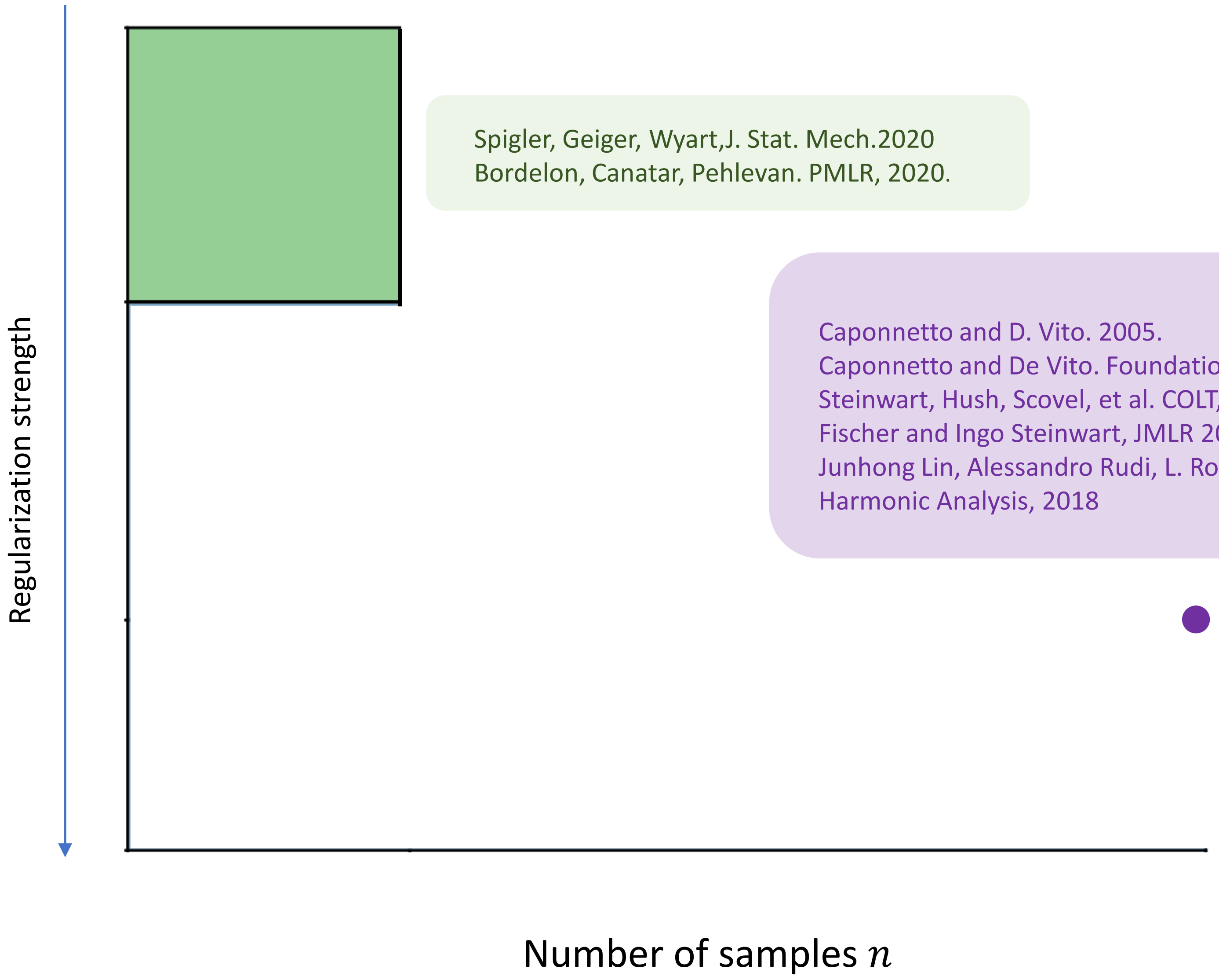
Spigler, Geiger, Wyart,J. Stat. Mech.2020
Bordelon, Canatar, Pehlevan. PMLR, 2020.

Caponnetto and D. Vito. 2005.
Caponnetto and De Vito. Foundations of Computational Mathematics, 2007.
Steinwart, Hush, Scovel, et al. COLT, 2009.
Fischer and Ingo Steinwart, JMLR 2020.
Junhong Lin, Alessandro Rudi, L. Rosasco, and V. Cevher Applied and Computational Harmonic Analysis, 2018

Why the discrepancy?

Quick appetizer

Actually these are two different regimes. We locate them on the (regularization, sample complexity plane)….

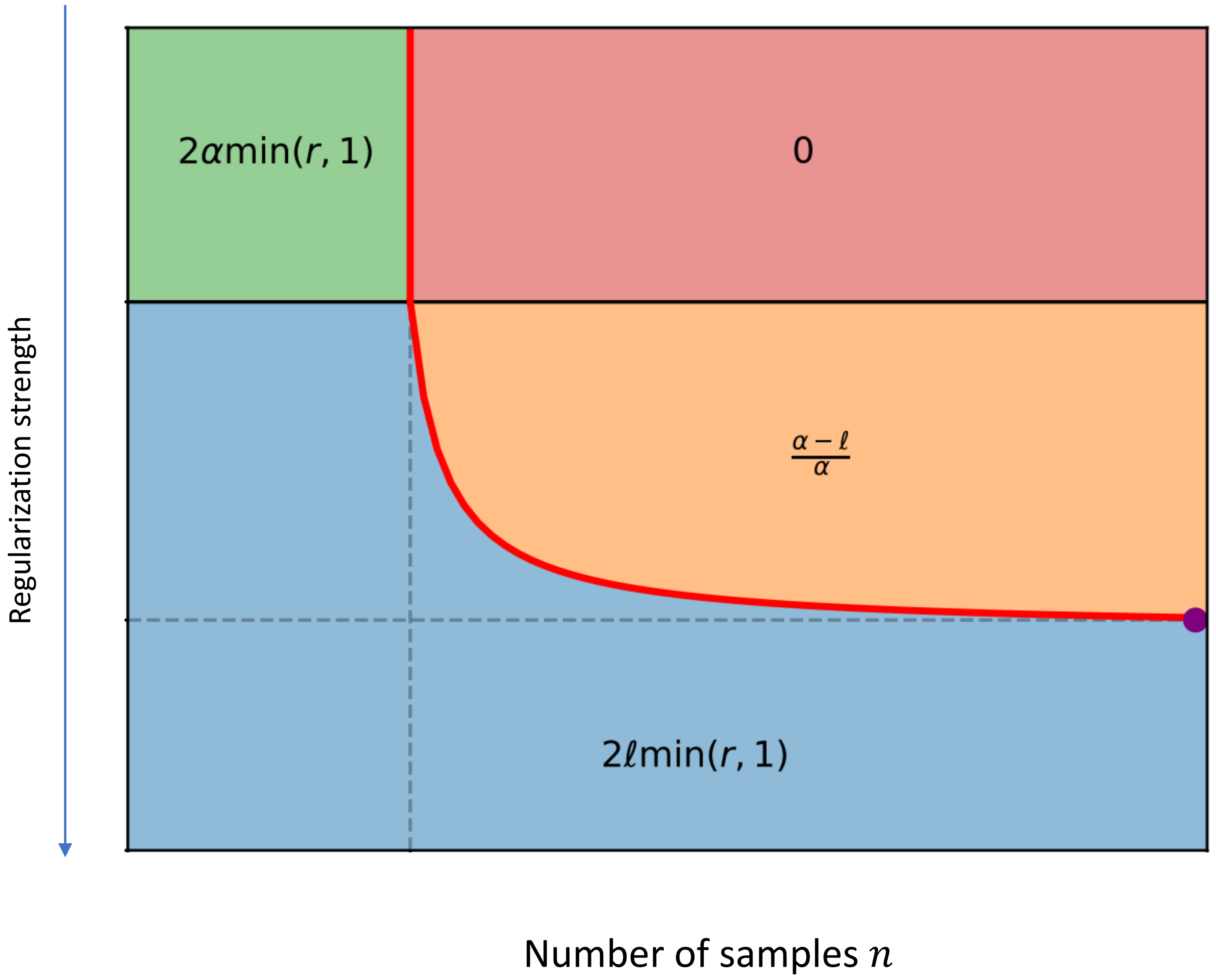


Spigler, Geiger, Wyart,J. Stat. Mech.2020
Bordelon, Canatar, Pehlevan. PMLR, 2020.

Caponnetto and D. Vito. 2005.
Caponnetto and De Vito. Foundations of Computational Mathematics, 2007.
Steinwart, Hush, Scovel, et al. COLT, 2009.
Fischer and Ingo Steinwart, JMLR 2020.
Junhong Lin, Alessandro Rudi, L. Rosasco, and V. Cevher Applied and Computational Harmonic Analysis, 2018

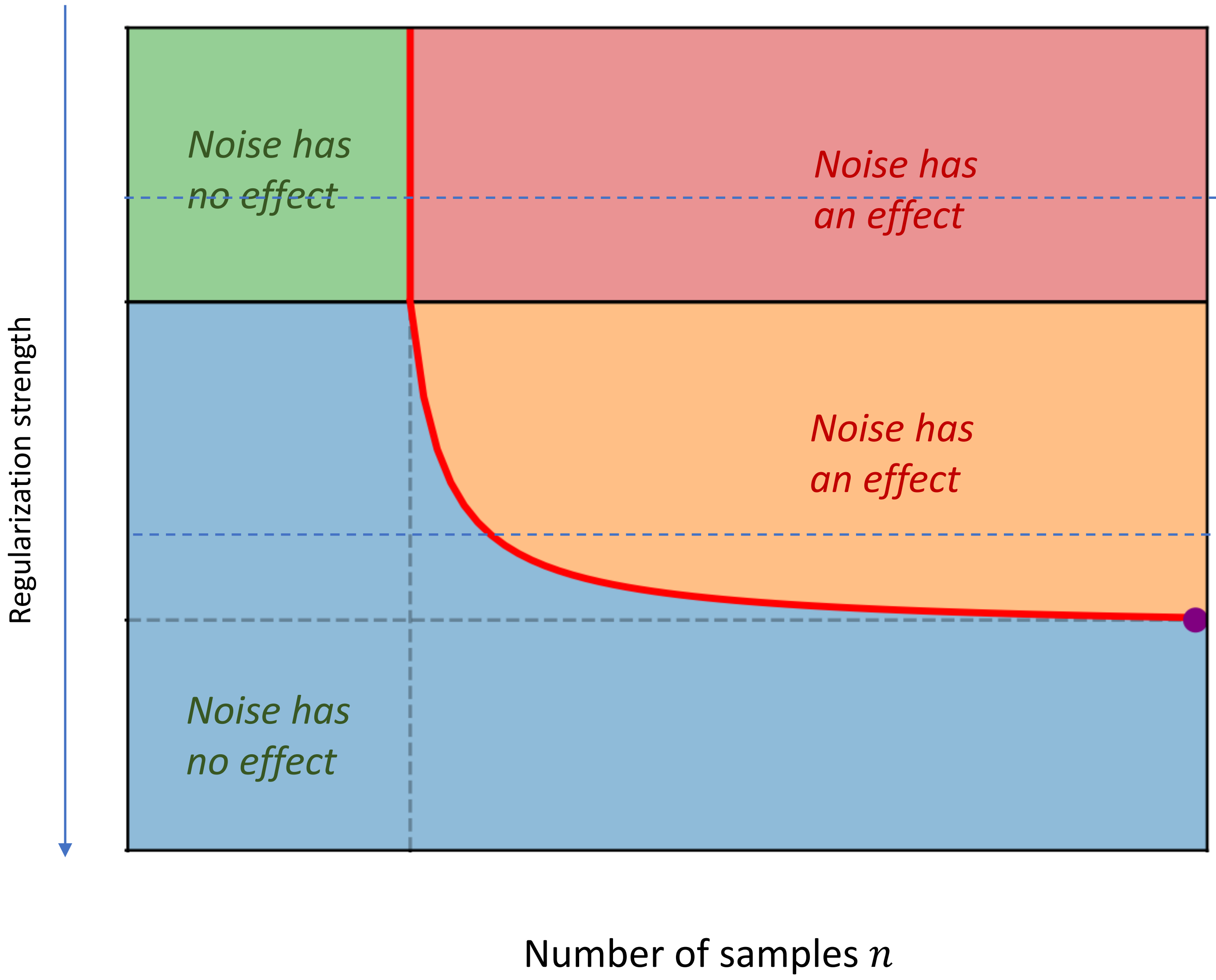
Quick appetizer

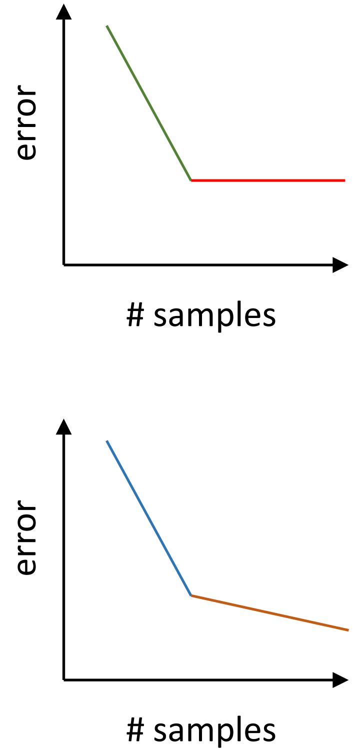
… and provide an unifying picture of the four regimes of KRR…



Quick appetizer

… and discuss when label noise affects the learning speed





A refresher on Kernels

Take a kernel K with Reproducing Kernel Hilbert Space $\mathcal{H}$

Kernel Ridge Regression (KRR)

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{\mu=1}^{n} (f(x^\mu) - y^\mu)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

samples (n), data (x^μ), label (y^μ), regularization (λ)

A refresher on Kernels

Take a kernel K with Reproducing Kernel Hilbert Space $\mathcal{H}$

Kernel Ridge Regression (KRR)

samples

label

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{\mu=1}^{n}(f(x^\mu) - y^\mu)^2 + \lambda\|f\|_{\mathcal{H}}^2$$

data

regularization

Using a feature map $\psi(x^\mu) \in \mathbb{R}^p$

$$\hat{\mathcal{R}}_n(w) = \frac{1}{n}\sum_{\mu=1}^{n}\left(w^\top\psi(x^\mu) - y^\mu\right)^2 + \lambda\, w^\top w.$$

A refresher on Kernels

Take a kernel K with Reproducing Kernel Hilbert Space $\mathcal{H}$

Kernel Ridge Regression (KRR)

samples

label

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{\mu=1}^{n}(f(x^{\mu}) - y^{\mu})^2 + \lambda \|f\|_{\mathcal{H}}^{2}$$

data

regularization

Using a feature map $\psi(x^{\mu}) \in \mathbb{R}^p$

$$\hat{\mathcal{R}}_n(w) = \frac{1}{n}\sum_{\mu=1}^{n}\left(w^{\top}\psi(x^{\mu}) - y^{\mu}\right)^2 + \lambda\, w^{\top}w.$$

Chosen so that the covariance is diagonal

$$\Sigma \equiv \mathbb{E}_{x\sim\rho_x}\left[\psi(x)\psi(x)^{\top}\right] = \mathrm{diag}(\eta_1, \eta_2, \ldots, \eta_p)$$

Working assumptions

Gaussian design

$$\psi(x) \stackrel{d}{=} \mathcal{N}(0, \Sigma)$$

Gaussian features

$$y^\mu = \theta^\star \psi(x^\mu) + \sigma \mathcal{N}(0,1)$$

teacher + additive gaussian noise

Regularization

$$\lambda = n^{-\ell}$$

Dicker et al. Bernoulli, 2016.
Hsu,Kakade, and Zhang.PMLR 2012.
Dobriban and Wager. The Annals of Statistics,2018.
Ledoit and Péché. Probability Theory and Related Fields, 2011

Working assumptions

$\exists \alpha > 1,\ r \geq 0$

$\operatorname{tr} \Sigma^{\frac{1}{\alpha}} < \infty$

$||\Sigma^{\frac{1}{2}-r}\theta^{\star}|| < \infty$

Working assumptions

$$\exists \alpha > 1,\ r \geq 0$$

$$\mathrm{tr}\, \Sigma^{\frac{1}{\alpha}} < \infty$$

$$||\Sigma^{\frac{1}{2}-r}\theta^{\star}|| < \infty$$

Capacity condition

Source condition

Eigenvalues of Σ

$$\eta_k = k^{-\alpha}$$

$$\theta^{\star}_k = k^{-\frac{1+\alpha(2r-1)}{2}}$$

Spigler, Geiger, Wyart,J. Stat. Mech.2020
Bordelon, Canatar, Pehlevan. PMLR, 2020.
Dobriban and Wager. Ann. of Stat.,2018.

Working assumptions

$$\exists \alpha > 1,\ r \geq 0$$

$$\mathrm{tr}\, \Sigma^{\frac{1}{\alpha}} < \infty$$

$$||\Sigma^{\frac{1}{2}-r}\theta^{\star}|| < \infty$$

Capacity condition

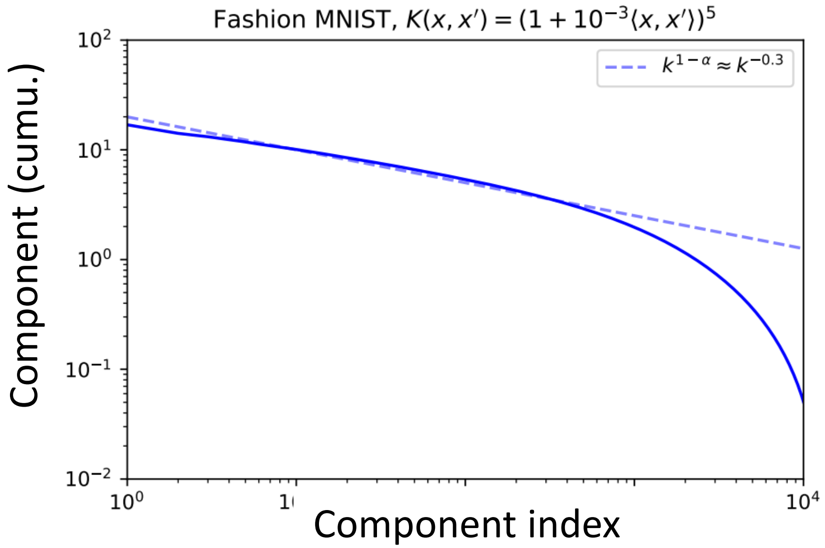
Source condition

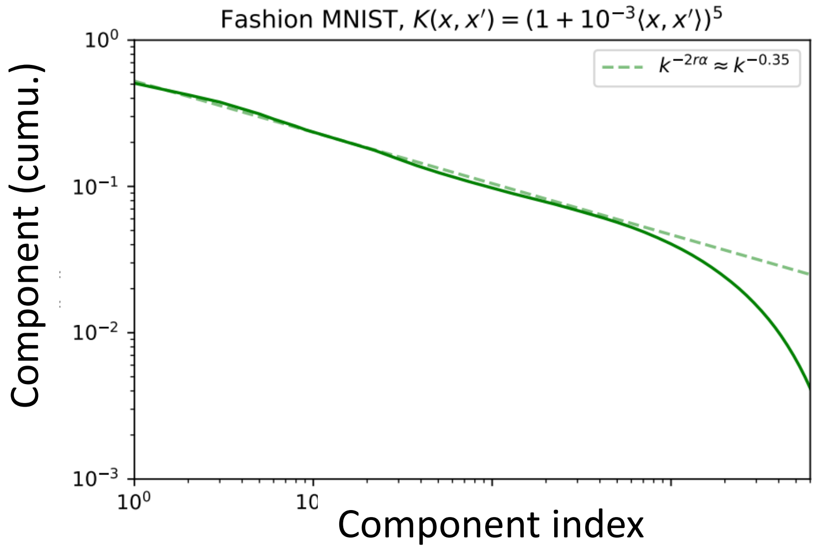
Eigenvalues of Σ

$$\eta_k = k^{-\alpha}$$

$$\theta^{\star}_k = k^{-\frac{1+\alpha(2r-1)}{2}}$$

Spigler, Geiger, Wyart,J. Stat. Mech.2020
Bordelon, Canatar, Pehlevan. PMLR, 2020.
Dobriban and Wager. Ann. of Stat.,2018.





Decay rates for the prediction error

$$\epsilon_g - \sigma^2 = \mathbb{E}_{u \sim \mathcal{N}(0,\Sigma)} \left(u^T \theta^\star - u^T \hat{w}\right)^2$$

Teacher *KRR Estimator*

At which rate does the excess error decay with the number of samples n?

Decay rates for the prediction error

$$\epsilon_g - \sigma^2 = \mathbb{E}_{u\sim\mathcal{N}(0,\Sigma)}(u^T\theta^\star - u^T\hat{w})^2$$

Teacher *KRR Estimator*

Spigler, Geiger, Wyart,J. Stat. Mech.2020
Bordelon, Canatar, Pehlevan. PMLR, 2020.

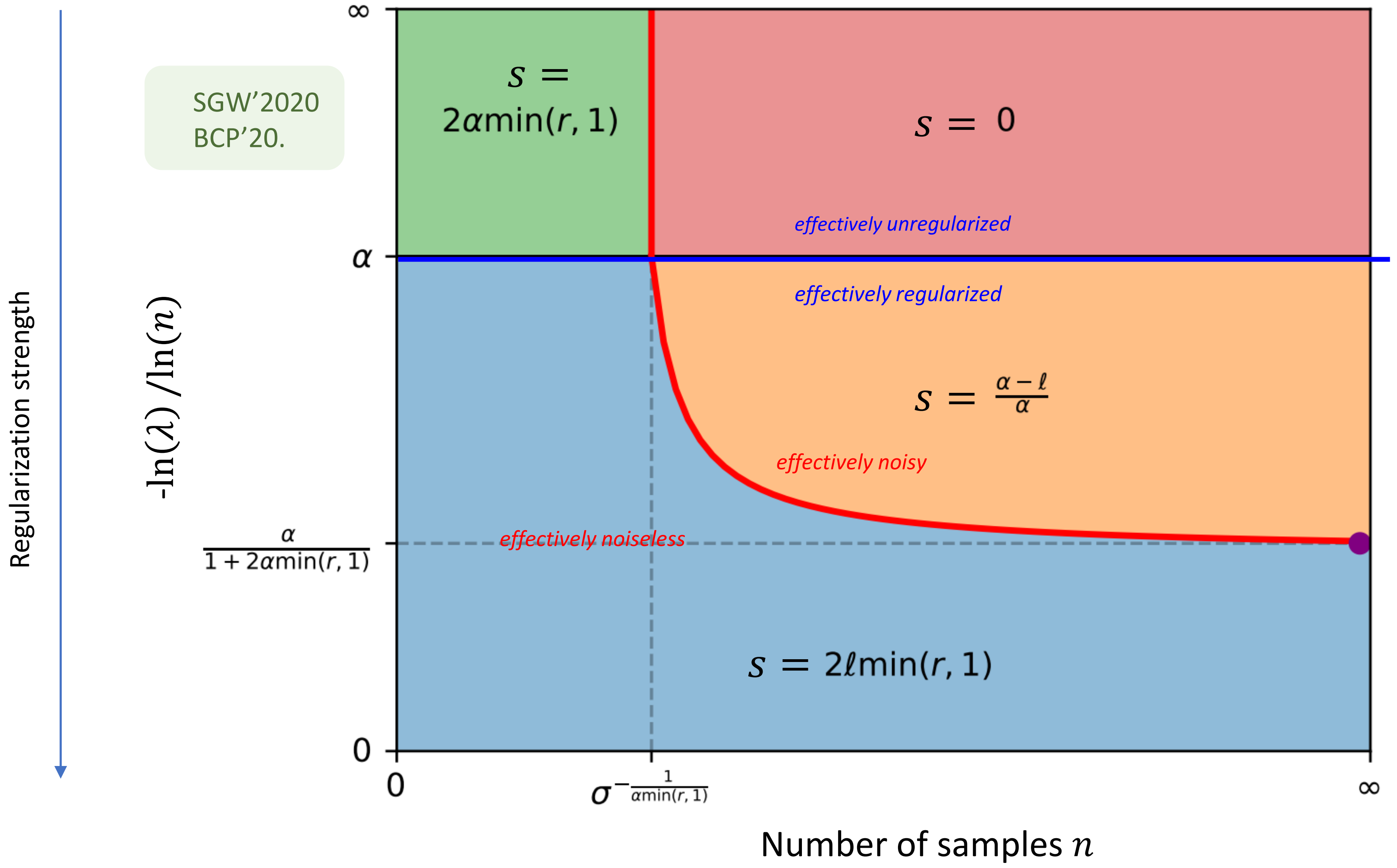
Typical case, **fast** decay

Caponnetto and D. Vito. 2005.
Caponnetto and De Vito. Foundations of Computational Mathematics, 2007.
Steinwart, Hush, Scovel, et al. COLT, 2009.
Fischer and Ingo Steinwart, JMLR 2020.
Junhong Lin, Alessandro Rudi, L. Rosasco, and V. Cevher Applied and Computational Harmonic Analysis, 2018

Worst case, **mild** decay

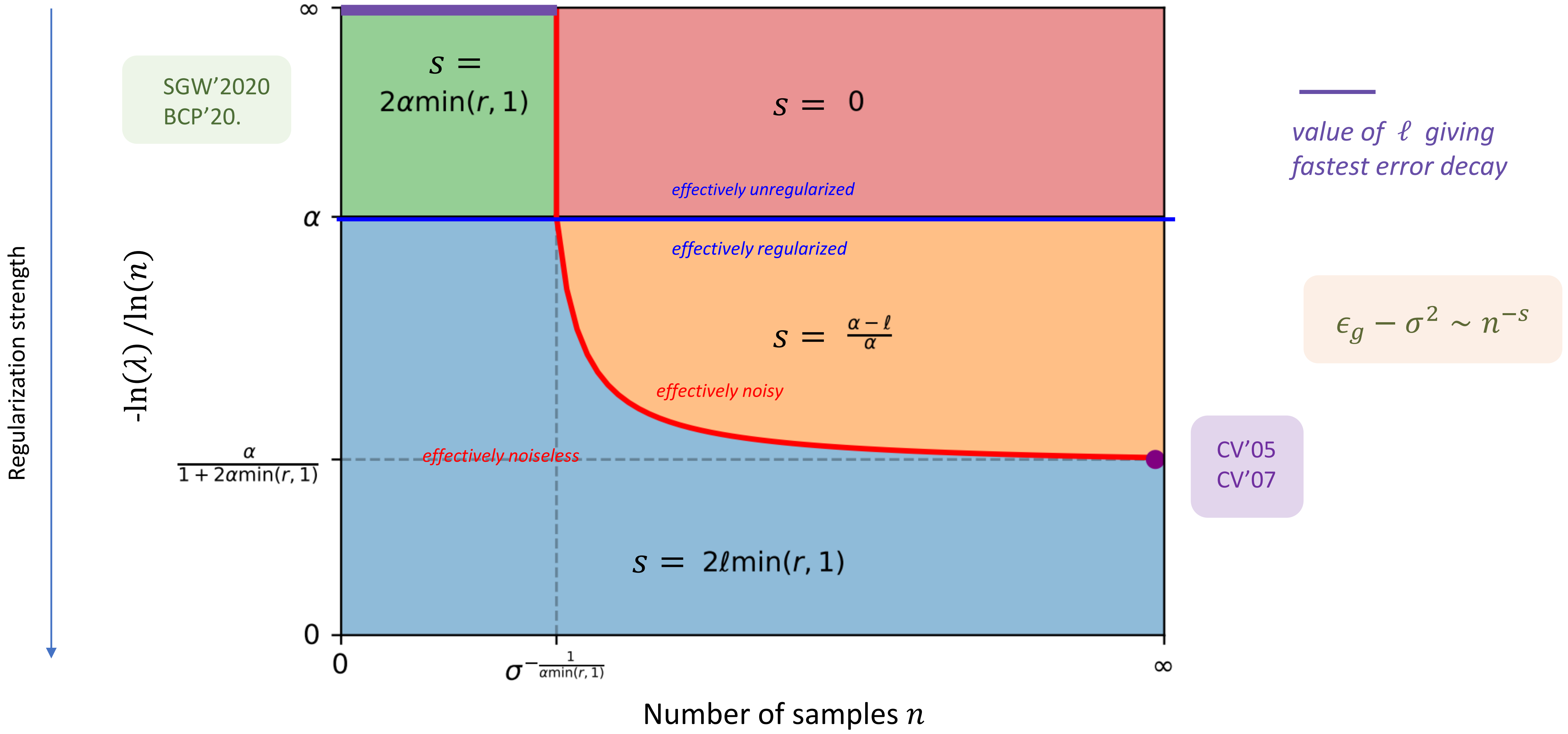
The four regimes

SGW'2020
BCP'20.

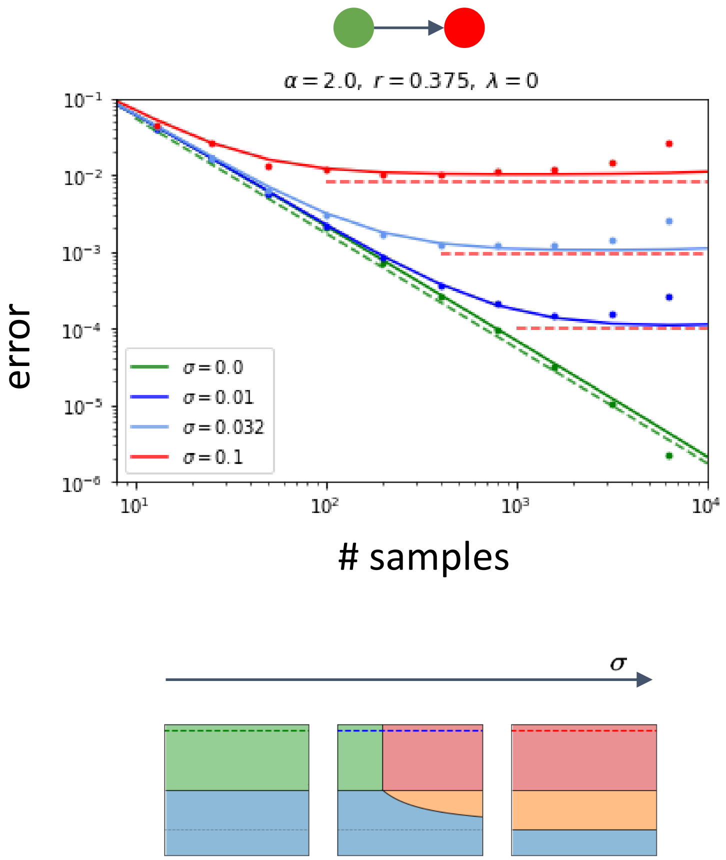


$\epsilon_g - \sigma^2 \sim n^{-s}$

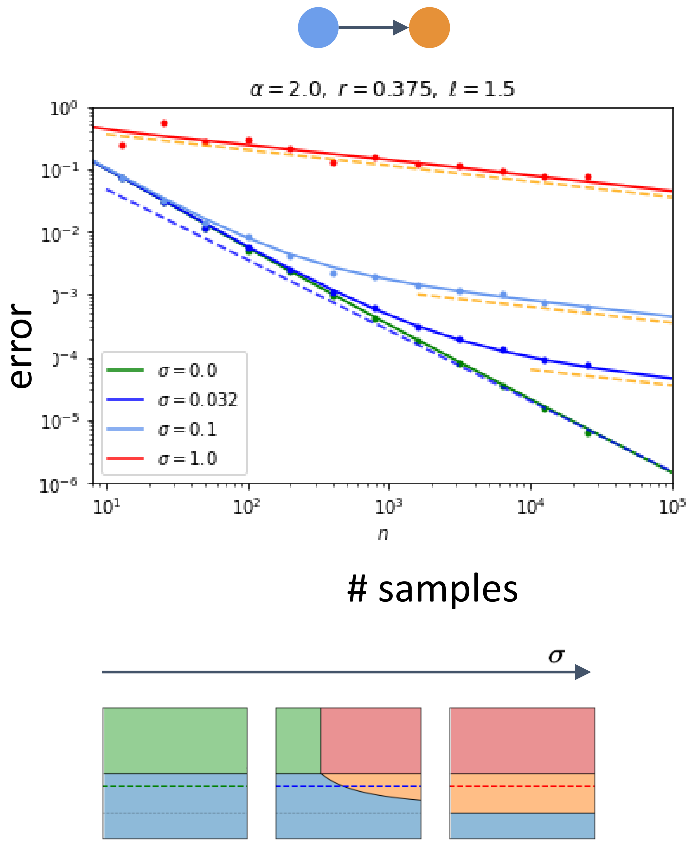
The four regimes



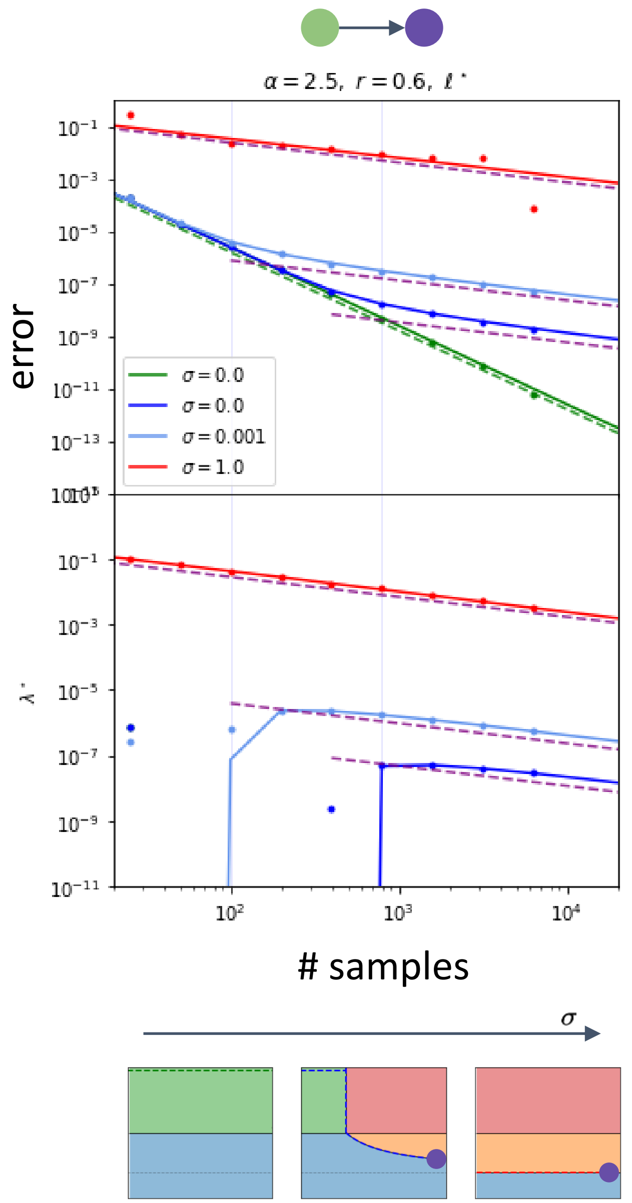
Crossovers



α = 2.0, r = 0.375, λ = 0
error
σ = 0.0
σ = 0.01
σ = 0.032
σ = 0.1
samples
σ

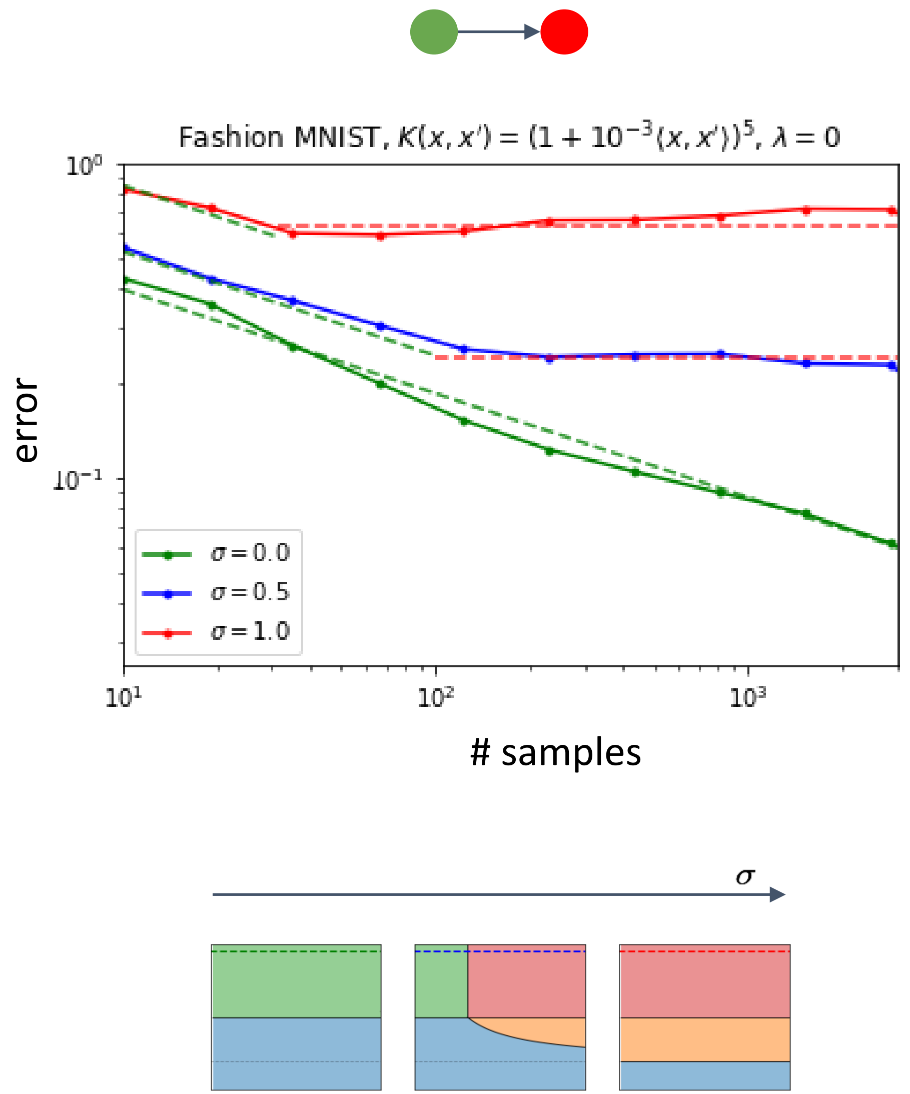


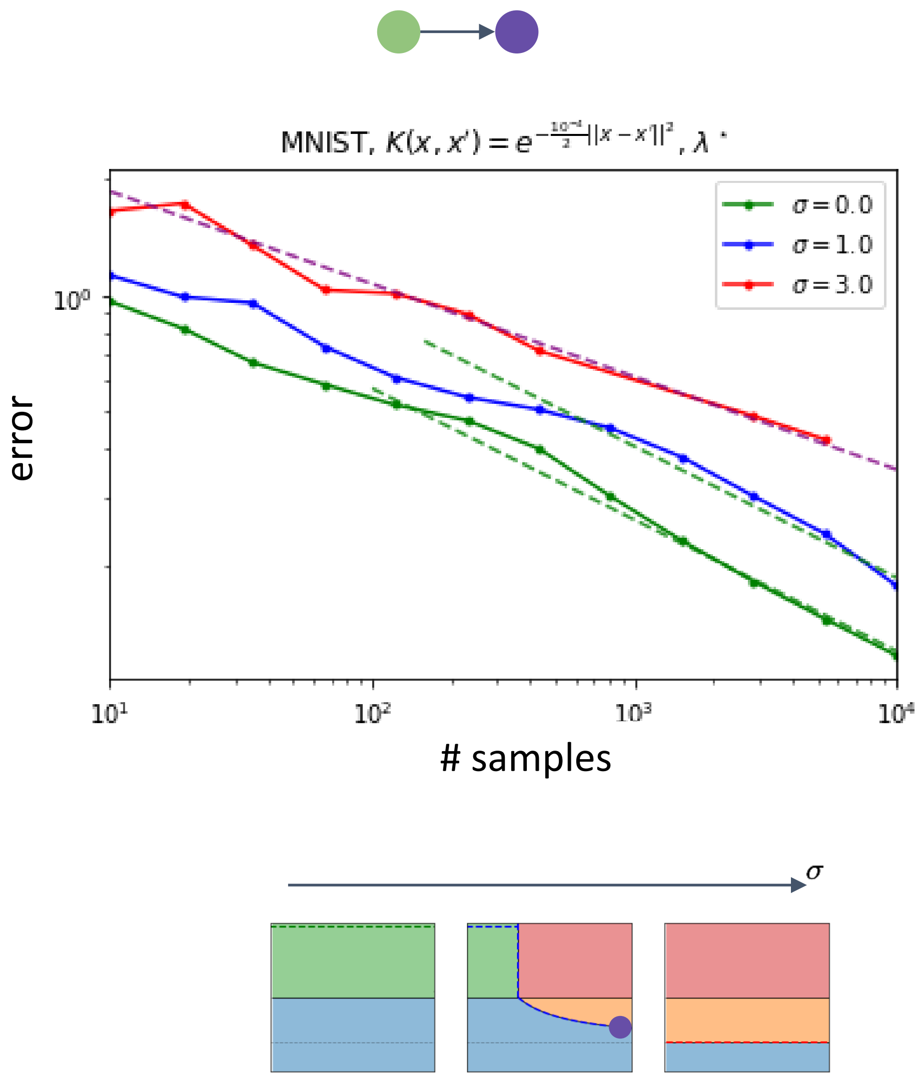
α = 2.0, r = 0.375, ℓ = 1.5
error
σ = 0.0
σ = 0.032
σ = 0.1
σ = 1.0
n
samples
σ



α = 2.5, r = 0.6, ℓ*
error
σ = 0.0
σ = 0.0
σ = 0.001
σ = 1.0
λ*
samples
σ

Crossovers


Fashion MNIST, $K(x,x') = (1+10^{-3}\langle x,x'\rangle)^5$, $\lambda = 0$
10^0
10^{-1}
error
10^1
10^2
10^3
samples
$\sigma = 0.0$
$\sigma = 0.5$
$\sigma = 1.0$
σ


error
10^0
10^1
10^2
10^3
10^4
samples
$\sigma = 0.0$
$\sigma = 1.0$
$\sigma = 3.0$
σ

Thank you for your attention!

For questions, see you at the virtual poster session