Inheriting regularization through **Knowledge Distillation**

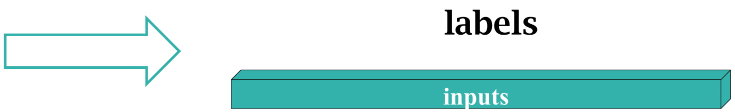
Statistical Physics of Computation Laboratory

Luca Saglietti, Lenka Zdeborová
SPOC lab – EPFL

MSLS 2021

EPFL

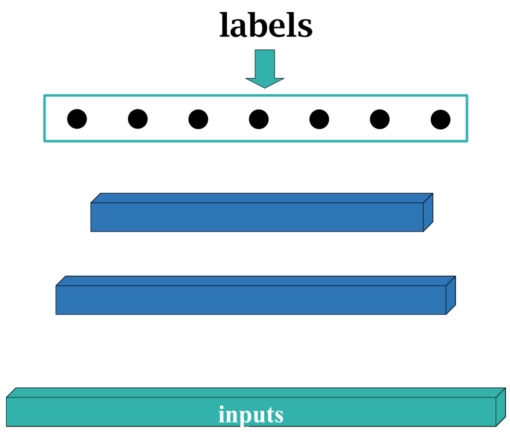
Knowledge distillation for network compression



G. Hinton, O. Vinyals, J. Dean, NIPS 2015

R. Anil, G. Hinton et al., ICLR 2018

Knowledge distillation for network compression



Small NN → optimization harder → POOR GENERALIZATION

Knowledge distillation for network compression

Larger NN → optimization bias (**implicit regularization**) → GOOD GENERALIZATION

Knowledge distillation for network compression

Larger NN → optimization bias (**implicit regularization**) → GOOD GENERALIZATION

The pre-activations retain:

- **uncertainty** estimation
- **relational information** between categories
- **reweight** the training samples

KD loss function

$$\mathcal{L}_{KD} = \sum_{\mu} \Big((1-\chi)\mathcal{H}(y^{\mu}, f(\boldsymbol{x};\boldsymbol{W})) + \chi\,\mathcal{H}(\tilde{f}(\boldsymbol{x};\tilde{\boldsymbol{W}}, T), f(\boldsymbol{x};\boldsymbol{W}, T)) \Big) + \lambda \|\boldsymbol{W}\|$$

KD **loss function**

usual logistic regression with **true labels**

logistic regression with **teacher pre-activations**

$$\mathcal{L}_{KD} = \sum_{\mu} \Big((1-\chi)\mathcal{H}(y^{\mu}, f(\boldsymbol{x}; \boldsymbol{W})) + \chi\, \mathcal{H}(\tilde{f}(\boldsymbol{x}; \tilde{\boldsymbol{W}}, T), f(\boldsymbol{x}; \boldsymbol{W}, T)) \Big) + \lambda \|\boldsymbol{W}\|$$

Cross-entropy: $\mathcal{H}(y,p) = -\left(y \log(p) + (1-y)\log(1-p)\right)$

KD loss function

usual logistic regression with **true labels**

logistic regression with **teacher pre-activations**

$$\mathcal{L}_{KD} = \sum_{\mu} \Big((1 - \chi) \mathcal{H}(y^{\mu}, f(\boldsymbol{x}; \boldsymbol{W})) + \chi \mathcal{H}(\tilde{f}(\boldsymbol{x}; \tilde{\boldsymbol{W}}, T), f(\boldsymbol{x}; \boldsymbol{W}, T)) \Big) + \lambda \|\boldsymbol{W}\|$$

Cross-entropy: $\mathcal{H}(y, p) = -\left(y \log(p) + (1 - y) \log(1 - p)\right)$

KD mixing parameter

KD loss function

usual logistic regression with **true labels**

logistic regression with **teacher pre-activations**

$$\mathcal{L}_{KD} = \sum_{\mu} \Big((1-\chi)\mathcal{H}(y^{\mu}, f(\boldsymbol{x}; \boldsymbol{W})) + \chi\,\mathcal{H}(\tilde{f}(\boldsymbol{x}; \tilde{\boldsymbol{W}}, T), f(\boldsymbol{x}; \boldsymbol{W}, T)) \Big) + \lambda \|\boldsymbol{W}\|$$

- **Cross-entropy**: $\mathcal{H}(y,p) = -\left(y\,\log(p) + (1-y)\log(1-p)\right)$
- **KD mixing parameter**
- **Direct (explicit) regularization**

Theory?

Studying **distillation: stat phys** approach

- **2-level problem**:
  - **a) pre-train** teacher $\tilde{W}$
  - **b) train** student W
- Step **a)** is **unaffected** by step **b)**
- Both levels share the **training set**

Studying **distillation: stat phys** approach

- **2-level problem**:
  - **a) pre-train** teacher $\tilde{W}$
  - **b) train** student W
- Step **a)** is **unaffected** by step **b)**
- Both levels share the **training set**

- **Franz-Parisi potential** formalism:

S. Franz and G. Parisi, PRL 1997

$$\mathcal{S}_{FP} = \int \boxed{d\tilde{\boldsymbol{W}} \frac{e^{-\tilde{\beta}\tilde{E}(\tilde{\boldsymbol{W}})}}{\tilde{Z}(\tilde{\beta})}} \log \int d\boldsymbol{W}\, e^{-\beta E(\boldsymbol{W}, \tilde{\boldsymbol{W}})}$$

$$E(\boldsymbol{W}, \xi = \{\boldsymbol{x}^\mu, y^\mu\}) = \sum_\mu \ell(\hat{y}(\boldsymbol{W}, \boldsymbol{x}^\mu), y^\mu) + \lambda \|\boldsymbol{W}\|$$

Studying **distillation: stat phys** approach

- **2-level problem**:
  - **a) pre-train** teacher $\tilde{W}$
  - **b) train** student W
- Step **a)** is **unaffected** by step **b)**
- Both levels share the **training set**

- **Franz-Parisi potential** formalism:

S. Franz and G. Parisi, PRL 1997

$$\mathcal{S}_{FP} = \int d\tilde{\boldsymbol{W}} \frac{e^{-\tilde{\beta}\tilde{E}(\tilde{\boldsymbol{W}})}}{\tilde{Z}(\tilde{\beta})} \log \int \boxed{d\boldsymbol{W}\, e^{-\beta E(\boldsymbol{W},\tilde{\boldsymbol{W}})}}$$

Studying **distillation: stat phys** approach

- **2-level problem**:
  - **a) pre-train** teacher $\tilde{W}$
  - **b) train** student W
- Step **a)** is **unaffected** by step **b)**
- Both levels share the **training set**

Franz-Parisi potential formalism:

S. Franz and G. Parisi, PRL 1997

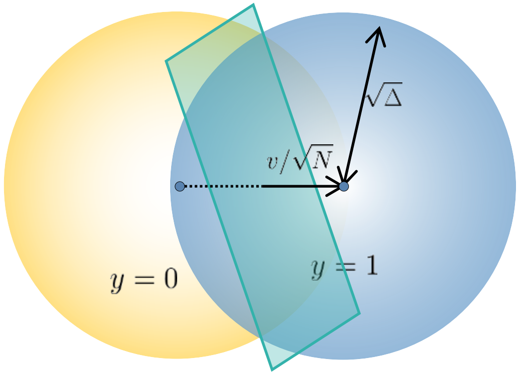
$$\mathcal{S}_{FP} = \int \boxed{d\tilde{\boldsymbol{W}} \frac{e^{-\tilde{\beta}\tilde{E}(\tilde{\boldsymbol{W}})}}{\tilde{Z}(\tilde{\beta})}} \log \int d\boldsymbol{W} \, e^{-\beta E(\boldsymbol{W},\tilde{\boldsymbol{W}})}$$

Quenched and **annealed** disorder -> **REPLICA METHOD**

Model assumptions

Data model: **Isotropic Gaussian mixture**
(2 clusters, **M** points in dimension **N**)

Learning model: **L2**–regularized **logistic regression**

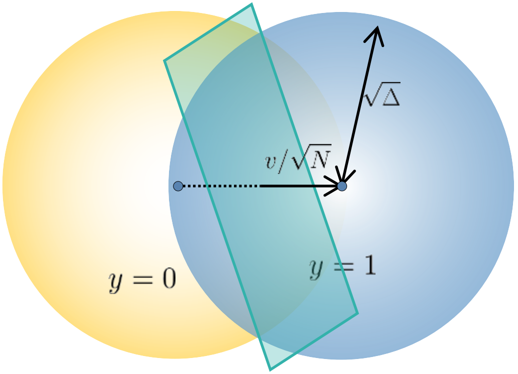


Model assumptions

Data model: Isotropic Gaussian mixture
(2 clusters, **M** points in dimension **N**)

Learning model: L2–regularized **logistic regression**

Asymptotic limit: $N, M \to \infty \qquad M/N = \alpha$



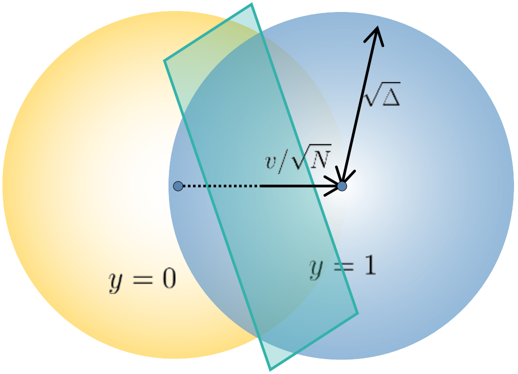
Tuning regularization intensity λ is key!!!

Model assumptions

Data model: **Isotropic Gaussian mixture**
(2 clusters, **M** points in dimension **N**)

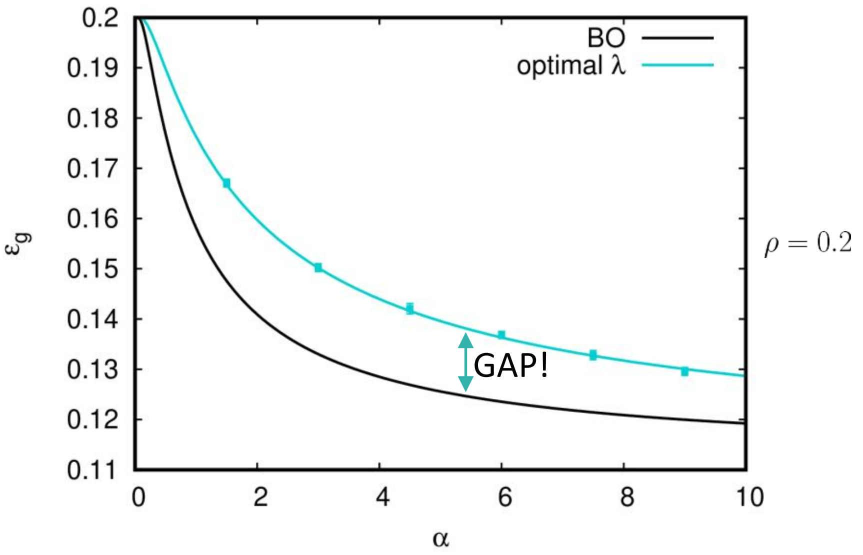
Learning model: **L2**–regularized **logistic regression**

Asymptotic limit: $N, M \to \infty \qquad M/N = \alpha$



Tuning regularization intensity λ is key!!!

Unbalanced clusters: $y^{\mu} = \mathrm{Bern}(\rho)$ **hard!**



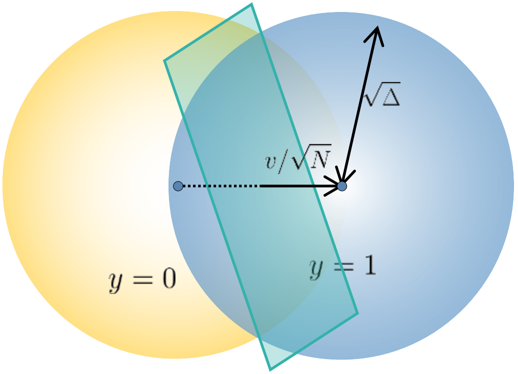
F. Mignacco, F. Krzakala, Y. Lu, P. Urbani, L. Zdeborova. ICML, 2020

Model assumptions

Data model: **Isotropic Gaussian mixture**
(2 clusters, **M** points in dimension **N**)

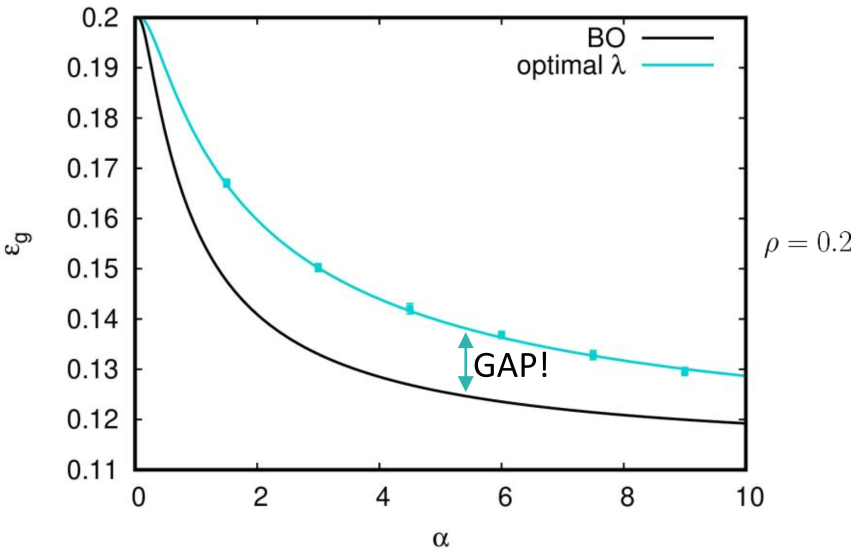
Learning model: **L2**–regularized **logistic regression**

Asymptotic limit: $N, M \to \infty \quad M/N = \alpha$



Tuning regularization intensity λ is key!!!

Unbalanced clusters: $y^\mu = \mathrm{Bern}(\rho)$ **hard!**

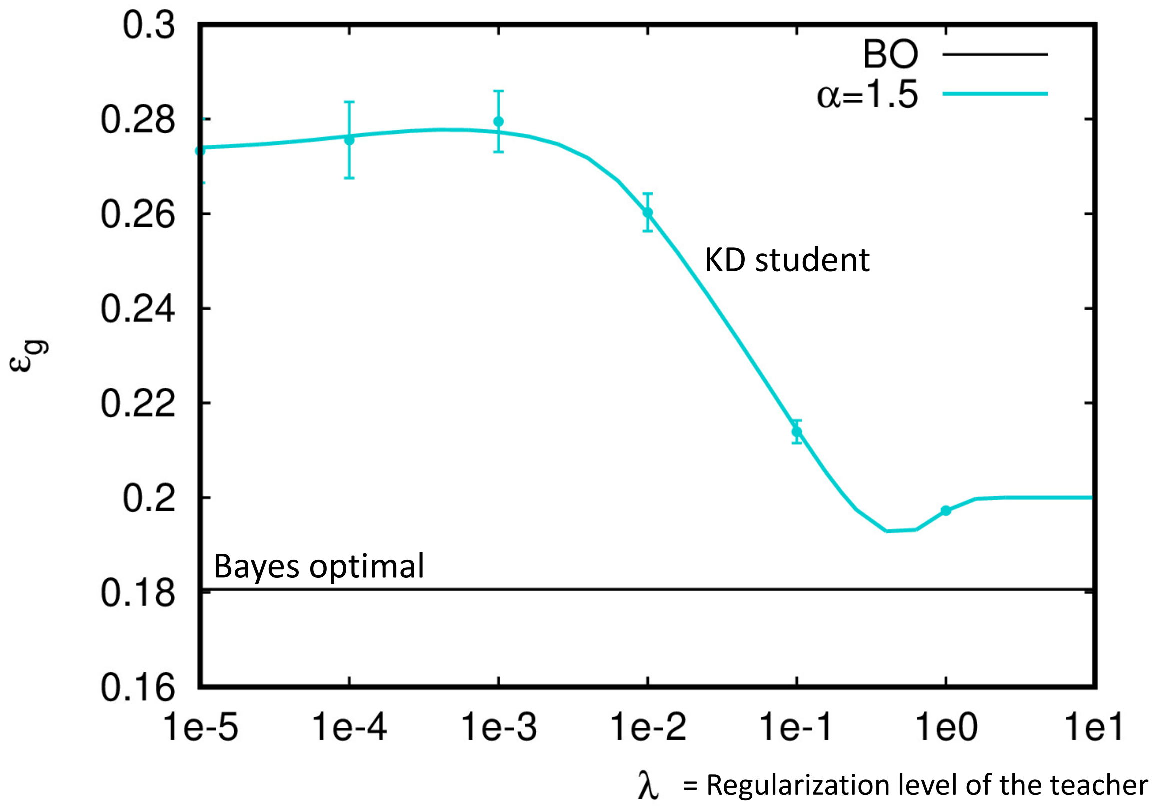


Teacher-student mismatch: **weaker student model**

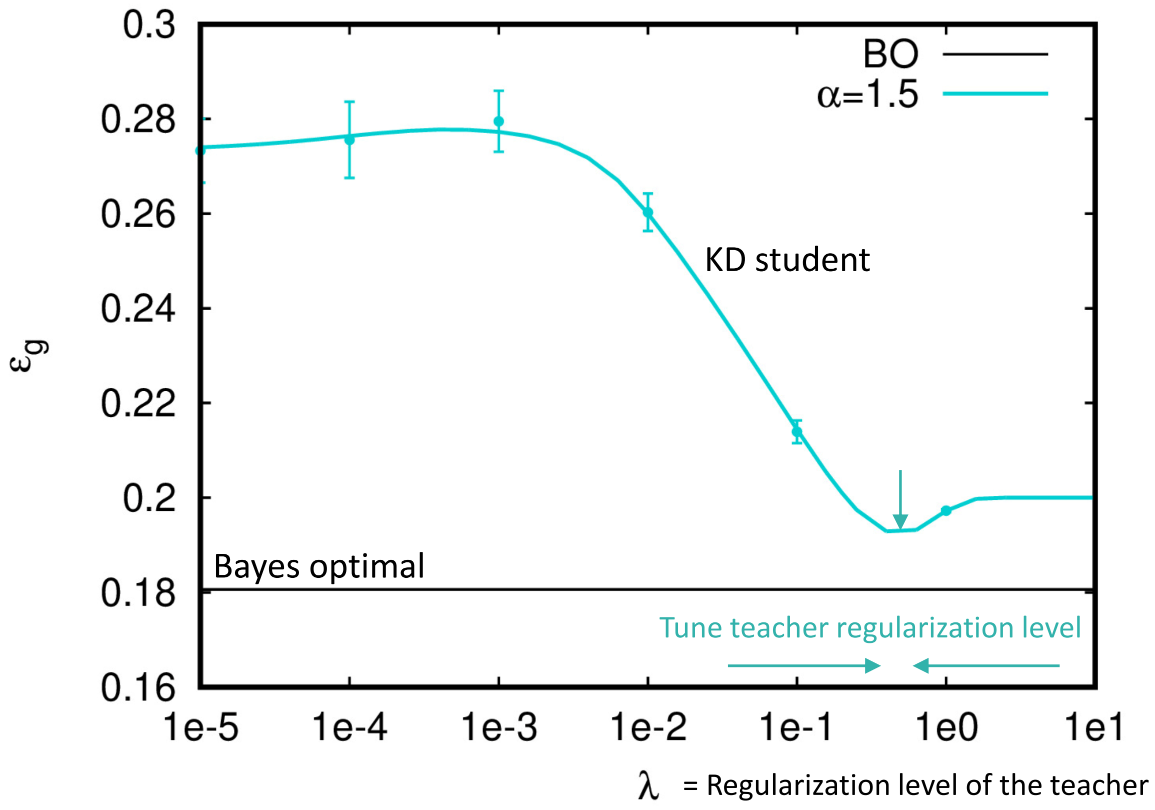
Fixed **student sparsity**: **fraction η=0.5** of the weights are **trained**, the **rest set to 0** a priori

F. Mignacco, F. Krzakala, Y. Lu, P. Urbani, L. Zdeborova. ICML, 2020

L_2-regularized logistic regression teacher: **effect of KD loss** on the student



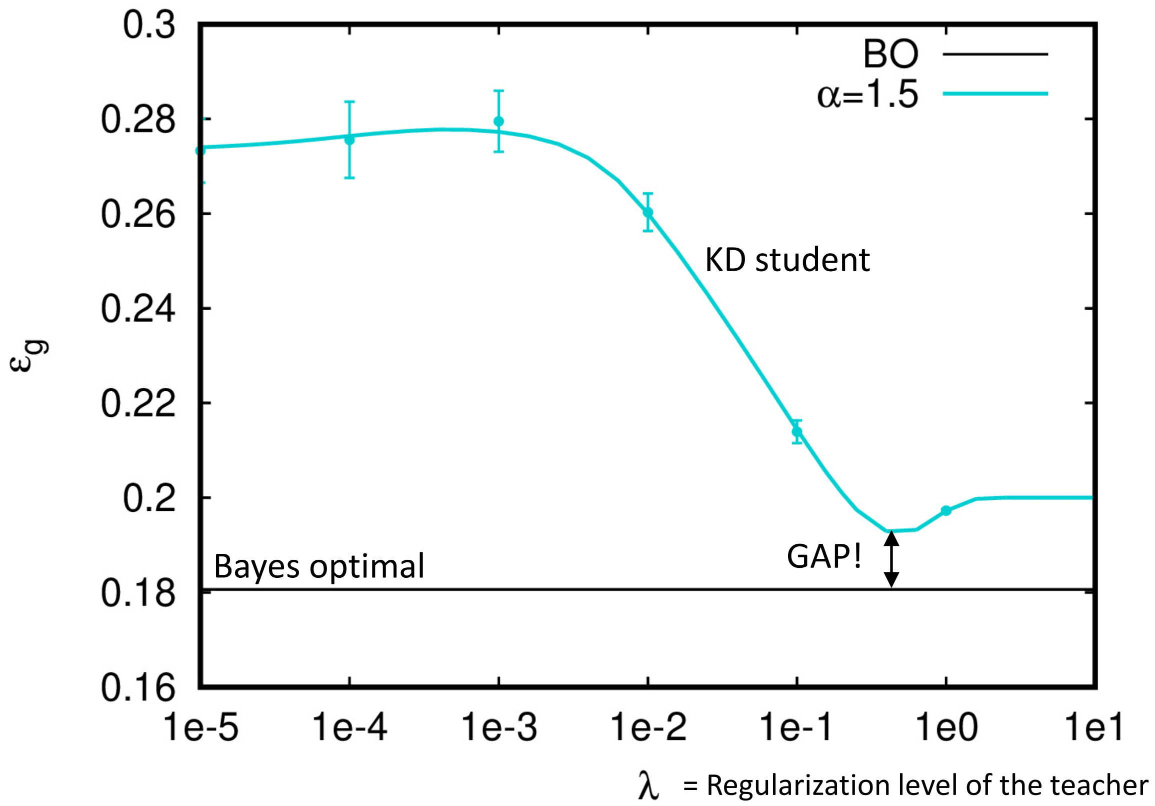
L_2-regularized logistic regression teacher: effect of KD loss on the student



Better teacher=better student

The student inherits the teacher regularization through KD!

L_2-regularized logistic regression teacher: **effect of KD loss** on the student



Better teacher=better student

The student inherits the teacher regularization through KD!

The obtained generalization performance is still sub-optimal!

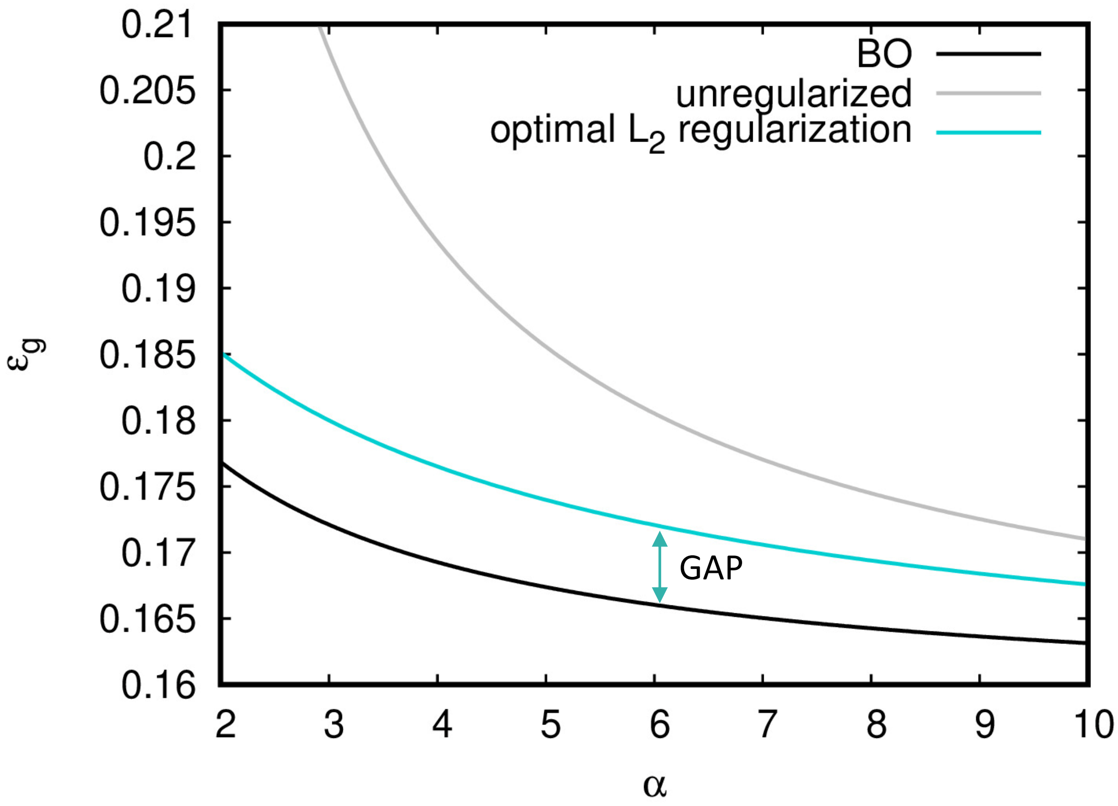
The KD student improves
together with the teacher

With a **sub-optimal teacher** the
student remains sub-optimal
(as the logistic regression estimator)

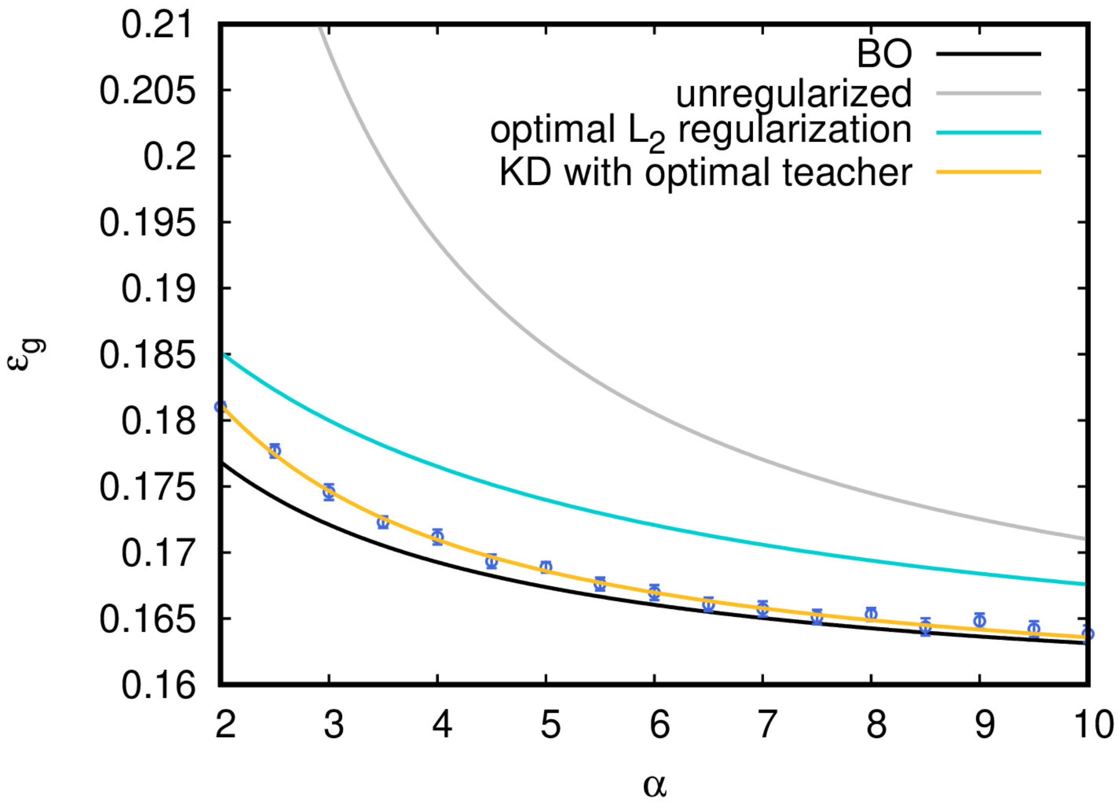
What if the teacher is **not just regularized "explicitly"**?

Is **KD still effective** in transferring
the generalization properties?

Bayes-Optimal teacher:
KD better than logistic regression?



Bayes-Optimal teacher:
KD better than logistic regression?



The KD student **closes** the **generalization gap!!!**

even though the **form of regularization of the teacher** is **not known explicitly!**
DEEP LEARNING READY

TAKE-HOME MESSAGE

With Knowledge Distillation the **student can inherit the teacher regularization** properties:

CONS

Cannot beat an explicit regularization of the same type!

TAKE-HOME MESSAGE

With Knowledge Distillation the **student can inherit the teacher regularization** properties:

CONS

Cannot beat an explicit regularization of the same type!

PROS

Achieve better generalization **even when**
the **form of regularization of the teacher** is **not known explicitly**
DEEP LEARNING READY :)

TAKE-HOME MESSAGE

With **Knowledge Distillation** the **student can inherit the teacher regularization** properties:

CONS

Cannot beat an explicit regularization of the same type!

PROS

Achieve better generalization **even when**
the **form of regularization of the teacher** is **not known explicitly**
DEEP LEARNING READY :)

Thank you for your attention!