Loss factorization, weakly supervised learning and label noise robustness

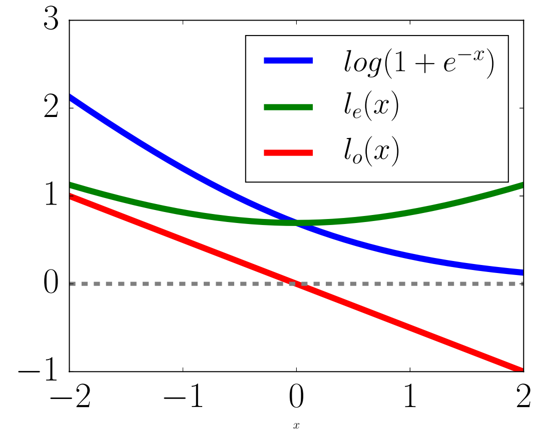
Giorgio Patrini, **Frank Nielsen, Richard Nock, Marcello Carioni**
Australian National University, Data61 (ex NICTA),
Ecole Polytechnique, Sony CS Labs,
Max Planck Institute of Mathematics in the Sciences

In 1 slide

Loss functions **factor**

$$\ell(x) = \ell_e(x) + \ell_o(x)$$

and so their risks, isolating a sufficient statistic for the labels, $\boldsymbol{\mu}$.

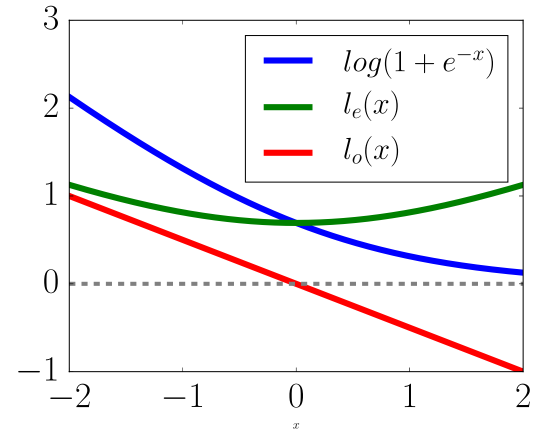


In 1 slide

Loss functions **factor**

$$\ell(x) = \ell_e(x) + \ell_o(x)$$

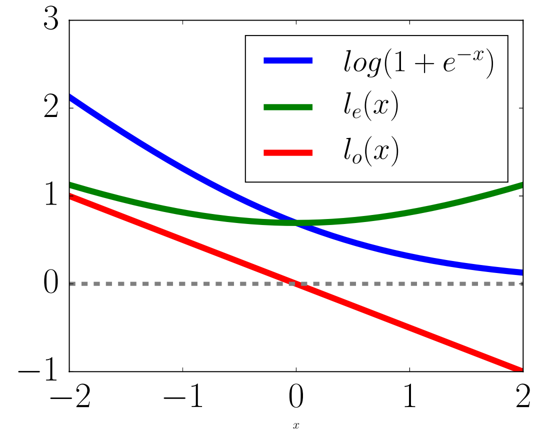
and so their risks, isolating a sufficient statistic for the labels, $\boldsymbol{\mu}$.



Weakly supervised learning: (1) estimate $\boldsymbol{\mu}$ and (2) plug it into $\ell(x)$ and call standard algorithms. E.g., SGD:

$$\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^{t} - \eta \nabla \ell(\pm\langle \boldsymbol{\theta}^{t}, \boldsymbol{x}_i \rangle) - \frac{1}{2}\eta a \boldsymbol{\mu}$$

In 1 slide



Loss functions **factor**

$$\ell(x) = \ell_e(x) + \ell_o(x)$$

and so their risks, isolating a sufficient statistic for the labels, $\boldsymbol{\mu}$.

Weakly supervised learning: (1) estimate $\boldsymbol{\mu}$ and (2) plug it into $\ell(x)$ and call standard algorithms. E.g., SGD:

$$\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^{t} - \eta \nabla \ell(\pm\langle \boldsymbol{\theta}^{t}, \boldsymbol{x}_i \rangle) - \frac{1}{2}\eta a \boldsymbol{\mu}$$

For asymmetric **label noise** with rates p_+, p_-, an unbiased estimator is

$$\hat{\boldsymbol{\mu}} \doteq \mathbb{E}_{(\boldsymbol{x},y)} \left[\frac{y - (p_- - p_+)}{1 - p_- - p_+} \boldsymbol{x} \right]$$

Preliminary

{1, …, m}

- Binary classification
  $\mathcal{S} = \{(\boldsymbol{x}_i, y_i), i \in [m]\}$ sampled from $\mathcal{D}$ over $\mathbb{R}^d \times \{-1, 1\}$
- Learn a linear (or kernel) model $h \in \mathcal{H}$
- Minimize the empirical risk associated with a surrogate loss $\ell(x)$

$$\operatorname*{argmin}_{h \in \mathcal{H}} \mathbb{E}_{\mathcal{S}} \left[\ell(yh(\boldsymbol{x})) \right] = \operatorname*{argmin}_{h \in \mathcal{H}} R_{\mathcal{S},\ell}(h)$$

Mean operator & linear-odd losses

- Mean operator

$$\boldsymbol{\mu}_{\mathcal{S}} \doteq \mathbb{E}_{\mathcal{S}}[y\boldsymbol{x}] = \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{x}_i$$

Mean operator & linear-odd losses

- Mean operator

$$\boldsymbol{\mu}_{\mathcal{S}} \doteq \mathbb{E}_{\mathcal{S}}[y\boldsymbol{x}] = \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{x}_i$$

- Linear-odd loss, a-LOL

$$\exists a \in \mathbb{R},\ \frac{1}{2}\ (\ell(x) - \ell(-x)) = \ell_o(x) = ax$$

generic x
argument

Loss factorization

- Linear model h
- Linear-odd loss $\frac{1}{2}(\ell(x) - \ell(-x)) = \ell_o(x) = ax$
- Given a sample $\mathcal{S}$, define a "double sample"

$$\mathcal{S}_{2x} \doteq \{(\boldsymbol{x}_i, \sigma), i \in [m], \sigma \in \{\pm 1\}\}$$

Loss factorization

- Linear model h
- Linear-odd loss $\frac{1}{2}(\ell(x) - \ell(-x)) = \ell_o(x) = ax$
- Given a sample $\mathcal{S}$, define a "double sample"

$$\mathcal{S}_{2x} \doteq \{(\boldsymbol{x}_i, \sigma), i \in [m], \sigma \in \{\pm 1\}\}$$

$$R_{\mathcal{S},\ell}(h) = \frac{1}{2} R_{\mathcal{S}_{2x},\ell}(h) + a \cdot h(\boldsymbol{\mu}_{\mathcal{S}})$$

smoothness
nor convexity
of ℓ required

Loss factorization: proof

$$R_{\mathcal{S},\ell}(h) =$$

$$= \mathbb{E}_{\mathcal{S}}\left[\ell(yh(\boldsymbol{x}))\right]$$

Loss factorization: proof

$$R_{\mathcal{S},\ell}(h) =$$

$$= \mathbb{E}_{\mathcal{S}}\Big[\ell(yh(\boldsymbol{x}))\Big]$$

even + odd

$$= \frac{1}{2}\mathbb{E}_{\mathcal{S}}\Big[\ell(yh(\boldsymbol{x})) + \ell(-yh(\boldsymbol{x})) + \ell(yh(\boldsymbol{x})) - \ell(-yh(\boldsymbol{x}))\Big]$$

Loss factorization: proof

$$R_{\mathcal{S},\ell}(h) =$$

$$= \mathbb{E}_{\mathcal{S}}\Big[\ell(yh(\boldsymbol{x}))\Big]$$

even + odd

$$= \frac{1}{2}\mathbb{E}_{\mathcal{S}}\Big[\ell(yh(\boldsymbol{x})) + \ell(-yh(\boldsymbol{x})) + \ell(yh(\boldsymbol{x})) - \ell(-yh(\boldsymbol{x}))\Big]$$

$$= \frac{1}{2}\mathbb{E}_{\mathcal{S}_{2x}}\Big[\ell(\sigma h(\boldsymbol{x}))\Big] + \mathbb{E}_{\mathcal{S}}\Big[\ell_o(h(y\boldsymbol{x}))\Big]$$

Loss factorization: proof

$$
\begin{aligned}
R_{\mathcal{S},\ell}(h) &= \\
&= \mathbb{E}_{\mathcal{S}}\Big[\ell(yh(\boldsymbol{x}))\Big] \\
&= \frac{1}{2}\mathbb{E}_{\mathcal{S}}\Big[\ell(yh(\boldsymbol{x})) + \ell(-yh(\boldsymbol{x})) + \ell(yh(\boldsymbol{x})) - \ell(-yh(\boldsymbol{x}))\Big] \\
&= \frac{1}{2}\mathbb{E}_{\mathcal{S}_{2x}}\Big[\ell(\sigma h(\boldsymbol{x}))\Big] + \mathbb{E}_{\mathcal{S}}\Big[\ell_o(h(y\boldsymbol{x}))\Big] \\
&= \frac{1}{2}R_{\mathcal{S}_{2x},\ell}(h) + a \cdot h(\boldsymbol{\mu}_{\mathcal{S}})
\end{aligned}
$$

even + odd

linear ℓ_o and h

sufficiency of $\boldsymbol{\mu}$ for y

Linear-odd losses: examples

- Logistic loss & exponential family

$$\sum_{i=1}^{m} \log \sum_{y \in \mathcal{Y}} e^{y\langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle} - \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle = \sum_{i=1}^{m} \log \left(1 + e^{-2y_i \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle} \right)$$

Linear-odd losses: examples

- Logistic loss & exponential family

$$\sum_{i=1}^{m} \log \sum_{y \in \mathcal{Y}} e^{y \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle} - \langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle = \sum_{i=1}^{m} \log \left(1 + e^{-2 y_i \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle} \right)$$

	loss ℓ	odd term ℓ_o
LOL	$\ell(x)$	$-ax$
ρ-loss	$\rho\|x\| - \rho x + 1$	$-\rho x \ (\rho \geq 0)$
unhinged	$1 - x$	$-x$
perceptron	$\max(0, -x)$	$-x$
double-hinge	$\max(-x, 1/2 \max(0, 1 - x))$	$-x$
SPL	$a_\ell + \ell^\star(-x)/b_\ell$	$-x/(2b_\ell)$
logistic	$\log(1 + e^{-x})$	$-x/2$
square	$(1 - x)^2$	$-2x$
Matsushita	$\sqrt{1 + x^2} - x$	$-x$

Generalization bound

- Loss ℓ is a-LOL and L-Lipschitz
- Bounds $\mathbb{R}^d \supseteq \mathcal{X} = \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq X < \infty\}$ and $\mathcal{H} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_2 \leq B < \infty\}$
- Bounded loss $c(X, B) \doteq \max_{y \in \{\pm 1\}} \ell(yXB)$
- Let $\hat{\boldsymbol{\theta}} \doteq \mathrm{argmin}_{\boldsymbol{\theta} \in \mathcal{H}} R_{\mathcal{S},\ell}(\boldsymbol{\theta})$

Generalization bound

- Loss ℓ is a-LOL and L-Lipschitz
- Bounds $\mathbb{R}^d \supseteq \mathcal{X} = \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq X < \infty\}$ and $\mathcal{H} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_2 \leq B < \infty\}$
- Bounded loss $c(X, B) \doteq \max_{y \in \{\pm 1\}} \ell(yXB)$
- Let $\hat{\boldsymbol{\theta}} \doteq \operatorname{argmin}_{\boldsymbol{\theta} \in \mathcal{H}} R_{\mathcal{S},\ell}(\boldsymbol{\theta})$

Then for any $\delta > 0$, with probability at least $1 - \delta$:

$$R_{\mathcal{D},\ell}(\hat{\boldsymbol{\theta}}) - \inf_{\boldsymbol{\theta} \in \mathcal{H}} R_{\mathcal{D},\ell}(\boldsymbol{\theta}) \leq \left(\frac{\sqrt{2}+1}{4}\right) \cdot \frac{XBL}{\sqrt{m}} +$$

$$\frac{c(X,B)L}{2} \cdot \sqrt{\frac{1}{m} \log\left(\frac{1}{\delta}\right)} + 2|a|B \cdot \|\boldsymbol{\mu}_{\mathcal{D}} - \boldsymbol{\mu}_{\mathcal{S}}\|_2$$

Generalization bound

- Loss ℓ is a-LOL and L-Lipschitz
- Bounds $\mathbb{R}^d \supseteq \mathcal{X} = \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq X < \infty\}$ and $\mathcal{H} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_2 \leq B < \infty\}$
- Bounded loss $c(X, B) \doteq \max_{y \in \{\pm 1\}} \ell(yXB)$
- Let $\hat{\boldsymbol{\theta}} \doteq \operatorname{argmin}_{\boldsymbol{\theta} \in \mathcal{H}} R_{\mathcal{S},\ell}(\boldsymbol{\theta})$

Then for any $\delta > 0$, with probability at least $1 - \delta$:

$$R_{\mathcal{D},\ell}(\hat{\boldsymbol{\theta}}) - \inf_{\boldsymbol{\theta} \in \mathcal{H}} R_{\mathcal{D},\ell}(\boldsymbol{\theta}) \leq \left(\frac{\sqrt{2}+1}{4}\right) \cdot \frac{XBL}{\sqrt{m}} +$$

complexity of space $\mathcal{H}$

$$\frac{c(X,B)L}{2} \cdot \sqrt{\frac{1}{m} \log\left(\frac{1}{\delta}\right)} + 2|a|B \cdot \|\boldsymbol{\mu}_{\mathcal{D}} - \boldsymbol{\mu}_{\mathcal{S}}\|_2$$

non-linearity

$$2|a|XB\sqrt{\frac{d}{m}\log\left(\frac{2d}{\delta}\right)}$$

Weakly supervised learning

$$\mathcal{D} \xrightarrow{corrupt} \tilde{\mathcal{D}} \xrightarrow{sample} \tilde{\mathcal{S}}$$

- **Weak** labels may be wrong (noisy), missing, multi-instance, etc.

Weakly supervised learning

$$\mathcal{D} \xrightarrow{corrupt} \tilde{\mathcal{D}} \xrightarrow{sample} \tilde{\mathcal{S}}$$

- **Weak** labels may be wrong (noisy), missing, multi-instance, etc.

- 2-step approach:
  (1) Estimate $\boldsymbol{\mu}_{\mathcal{S}}$ from weak labels
  (2) Plug it into ℓ and call any algorithm for risk minimization on $\mathcal{S}_{2x}$

Example: SGD (step 2)

Algorithm $\boldsymbol{\mu}\textsc{SGD}$

Input: $\mathcal{S}_{2x}, \boldsymbol{\mu}$, ℓ is a-LOL;

$\boldsymbol{\theta}^0 \leftarrow \mathbf{0}$

For any $t = 1, 2, \ldots$ until convergence

  Pick $i \in \{1, \ldots, |\mathcal{S}_{2x}|\}$ at random

  $\eta \leftarrow 1/t$

  Pick any $\boldsymbol{v} \in \partial\ell(y_i\langle\boldsymbol{\theta}^t, \boldsymbol{x}_i\rangle)$

  $\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t - \eta(\boldsymbol{v} + a\boldsymbol{\mu}/2)$

Output: $\boldsymbol{\theta}^{t+1}$

Example: SGD (step 2)

Algorithm $\boldsymbol{\mu}$SGD

Input: $\mathcal{S}_{2x}, \boldsymbol{\mu}$, ℓ is a-LOL;
$\boldsymbol{\theta}^0 \leftarrow \mathbf{0}$
For any $t = 1, 2, \ldots$ until convergence
  Pick $i \in \{1, \ldots, |\mathcal{S}_{2x}|\}$ at random
  $\eta \leftarrow 1/t$
  Pick any $\boldsymbol{v} \in \partial\ell(y_i\langle\boldsymbol{\theta}^t, \boldsymbol{x}_i\rangle)$
  $\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t - \eta(\boldsymbol{v} + a\boldsymbol{\mu}/2)$
Output: $\boldsymbol{\theta}^{t+1}$

only changes wrt SGD

- In the paper: proximal algorithms

A unifying approach

Learning from label proportions with

- logistic loss [N.Quadrianto et al. ’09]
- symmetric proper loss [G. Patrini et al. ’14]

Learning with noisy labels with

- logistic loss [Gao et al. ’16]

Asymmetric label noise

Sample $\tilde{\mathcal{S}} = \{(\boldsymbol{x}_i, \tilde{y}_i)\}_{i=1}^{m}$ corrupted by **asymmetric** noise rates p_+, p_-

Asymmetric label noise

Sample $\tilde{\mathcal{S}} = \{(\boldsymbol{x}_i, \tilde{y}_i)\}_{i=1}^{m}$ corrupted by **asymmetric** noise rates p_+, p_-

By the method of [Natarajan et al. '13] an unbiased estimator of $\boldsymbol{\mu}_{\mathcal{S}}$ is

$$\hat{\boldsymbol{\mu}}_{\mathcal{S}} \doteq \mathbb{E}_{\tilde{\mathcal{S}}}\left[\frac{y - (p_- - p_+)}{1 - p_- - p_+}\boldsymbol{x}\right]$$

This is step (1), then run $\boldsymbol{\mu}$-SGD for (2).

Generalization bound under noise

- Same as before, except that now $\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathcal{H}} \hat{R}_{\tilde{\mathcal{S}},\ell}(\boldsymbol{\theta})$

Then for any $\delta > 0$, with probability at least $1 - \delta$:

$$R_{\mathcal{D},\ell}(\hat{\boldsymbol{\theta}}) - \inf_{\boldsymbol{\theta} \in \mathcal{H}} R_{\mathcal{D},\ell}(\boldsymbol{\theta}) \leq \left(\frac{\sqrt{2}+1}{4} \right) \cdot \frac{XBL}{\sqrt{m}} + \frac{c(X,B)L}{2} \sqrt{\frac{1}{m} \log\left(\frac{2}{\delta}\right)} + \frac{2|a|XB}{1 - p_- - p_+} \sqrt{\frac{d}{m} \log\left(\frac{2d}{\delta}\right)}$$

Generalization bound under noise

- Same as before, except that now $\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathcal{H}} \hat{R}_{\tilde{\mathcal{S}},\ell}(\boldsymbol{\theta})$

Then for any $\delta > 0$, with probability at least $1 - \delta$:

$$R_{\mathcal{D},\ell}(\hat{\boldsymbol{\theta}}) - \inf_{\boldsymbol{\theta} \in \mathcal{H}} R_{\mathcal{D},\ell}(\boldsymbol{\theta}) \leq \left(\frac{\sqrt{2}+1}{4} \right) \cdot \frac{XBL}{\sqrt{m}} + \frac{c(X,B)L}{2} \sqrt{\frac{1}{m} \log\left(\frac{2}{\delta}\right)} + \frac{2|a|XB}{1 - p_- - p_+} \sqrt{\frac{d}{m} \log\left(\frac{2d}{\delta}\right)}$$

complexity untouched

noise affects the linear term only

Empirics

- Artificially corrupted data. Noise rates up to ~50%
- SGD vs μ-SGD with the same parameters
- Test error average difference over 25 runs

$(p_-, p_+) \rightarrow$	(.00, .00)		(.20, .00)		(.20, .10)		(.20, .20)		(.20, .30)		(.20, .40)		(.20, .49)	
dataset	SGD	$\boldsymbol{\mu}$SGD	SGD	$\boldsymbol{\mu}$SGD	SGD	$\boldsymbol{\mu}$SGD	SGD	$\boldsymbol{\mu}$SGD	SGD	$\boldsymbol{\mu}$SGD	SGD	$\boldsymbol{\mu}$SGD	SGD	$\boldsymbol{\mu}$SGD
australian	0.13	+.01	0.15	**−.01**	0.14	±.00	0.14	+.01	0.16	**−.01**	0.26	**−.09**	0.45	**−.25**
breast-can.	0.02	+.01	0.03	±.00	0.03	±.00	0.03	±.00	0.05	**−.01**	0.11	**−.06**	0.17	**−.08**
diabetes	0.28	**−.03**	0.29	**−.03**	0.29	**−.03**	0.27	**−.02**	0.28	**−.02**	0.39	**−.13**	0.59	**−.22**
german	0.27	**−.02**	0.26	±.00	0.27	**−.02**	0.29	**−.02**	0.31	**−.01**	0.31	±.00	0.31	±.00
heart	0.15	+.01	0.17	**−.01**	0.16	±.00	0.17	±.00	0.18	**−.01**	0.26	**−.08**	0.35	**−.15**
housing	0.17	**−.03**	0.23	**−.05**	0.22	**−.04**	0.20	**−.02**	0.22	**−.03**	0.34	**−.12**	0.41	**−.13**
ionosphere	0.14	+05	0.19	**−.05**	0.20	**−.05**	0.20	**−.03**	0.21	**−.03**	0.35	**−.13**	0.54	**−.29**
sonar	0.27	±.00	0.29	+.02	0.29	+.01	0.34	**−.04**	0.36	**−.03**	0.43	**−.10**	0.45	**−.05**

Empirics

- Artificially corrupted data. Noise rates up to ~50%
- SGD vs μ-SGD with the same parameters
- Test error average difference over 25 runs

$(p_-, p_+) \rightarrow$	(.00, .00)		(.20, .00)		(.20, .10)		(.20, .20)		(.20, .30)		(.20, .40)		(.20, .49)	
dataset	SGD	μSGD	SGD	μSGD	SGD	μSGD	SGD	μSGD	SGD	μSGD	SGD	μSGD	SGD	μSGD
australian	0.13	+.01	0.15	**−.01**	0.14	±.00	0.14	+.01	0.16	**−.01**	0.26	**−.09**	0.45	**−.25**
breast-can.	0.02	+.01	0.03	±.00	0.03	±.00	0.03	±.00	0.05	**−.01**	0.11	**−.06**	0.17	**−.08**
diabetes	0.28	**−.03**	0.29	**−.03**	0.29	**−.03**	0.27	**−.02**	0.28	**−.02**	0.39	**−.13**	0.59	**−.22**
german	0.27	**−.02**	0.26	±.00	0.27	**−.02**	0.29	**−.02**	0.31	**−.01**	0.31	±.00	0.31	±.00
heart	0.15	+.01	0.17	**−.01**	0.16	±.00	0.17	±.00	0.18	**−.01**	0.26	**−.08**	0.35	**−.15**
housing	0.17	**−.03**	0.23	**−.05**	0.22	**−.04**	0.20	**−.02**	0.22	**−.03**	0.34	**−.12**	0.41	**−.13**
ionosphere	0.14	+05	0.19	**−.05**	0.20	**−.05**	0.20	**−.03**	0.21	**−.03**	0.35	**−.13**	0.54	**−.29**
sonar	0.27	±.00	0.29	+.02	0.29	+.01	0.34	**−.04**	0.36	**−.03**	0.43	**−.10**	0.45	**−.05**

=> Still able to learn with one label ~ random

Bonus: data-dependent robustness

- The mean operator bounds the effect of noise

Let $\epsilon = 4|a|B \max\{p_+, p_-\}\|\boldsymbol{\mu}_{\mathcal{D}}\|_2$.
Any a-LOL ℓ is such that

$$R_{\tilde{\mathcal{D}},\ell}(\boldsymbol{\theta}^\star) - R_{\tilde{\mathcal{D}},\ell}(\tilde{\boldsymbol{\theta}}^\star) \leq \epsilon$$

$\boldsymbol{\theta}^\star$ and $\tilde{\boldsymbol{\theta}}^\star$ minimizers under $\mathcal{D}$ and $\tilde{\mathcal{D}}$

Moreover, if ℓ is differentiable and γ-strongly convex, then

$$\|\boldsymbol{\theta}^\star - \tilde{\boldsymbol{\theta}}^\star\|_2^2 \leq 2/\gamma \cdot \epsilon$$

a data-dependent statistic

More in the paper

- Mean and covariance operators
- Non linear models & kernel mean map
- Learning reductions
- Data-dependent bounds
- …

- Ongoing work: factorization and noise-correction for deep neural networks