# Generalization-Aware Structured Regression towards Balancing Bias and Variance

**Martin Pavlovski**[1,2], **Fang Zhou**[1], **Nino Arsov**[2], **Ljupco Kocarev**[2], **Zoran Obradovic**[1]

[1] Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA, USA
[2] Macedonian Academy of Sciences and Arts, Skopje, Republic of Macedonia

**Abstract**

Attaining the proper balance between underfitting and overfitting is one of the central challenges in machine learning. It has been approached mostly by deriving bounds on generalization risks of learning algorithms. Such bounds are, however, rarely controllable. In this study, a novel bias-variance balancing objective function is introduced in order to improve generalization performance. By utilizing distance correlation, this objective function is able to indirectly control a stability-based upper bound on a model's expected true risk. In addition, the Generalization-Aware Collaborative Ensemble Regressor (GLACER) is developed, a model that bags a crowd of structured regression models. Allowing its base components to collaborate in a fashion that minimizes the proposed objective function, GLACER has shown to outperform a broad range of both traditional and structured regression models, while sustaining stable predictions.

## The Notion of Generalization

- **Intuition:** Striking the proper balance between ***underfitting*** and ***overfitting***
  ⇒ *A fundamental challenge in supervised learning*
- ***Underfitting***
  - high bias
  - Avoided by **reducing** the *empirical risk* $R_{emp}$
- ***Overfitting***
  - high variance
  - Reduces as the *empirical risk* (training error) becomes a **valid estimate** of the *true unknown risk* (test error):

$$R_{gen} = |R_{emp} - R_{true}|$$

- **Objective:** Minimize $R_{emp}$, while maintaining low $R_{gen}$

## Main Theoretical Insight

- $R_{emp}$ can be easily minimized since it is **"measurable"** from the observed data
- $R_{gen}$ is often **impossible to determine** since $R_{true}$ is unknown
  - But, there are **stability-based upper bounds** derived on the *expected true risk* [1,2]:

$$\underbrace{\hat{R}_{true}(\mathcal{L})}_{\text{Expected true risk of a learning algorithm } \mathcal{L}} \leq \underbrace{\mathbb{E}_{\mathcal{D}}[\mathbb{E}_{h|\mathcal{D}}[R_{emp}(h,\mathcal{D})]]}_{\text{Expected empirical risk of a hypothesis } h \text{ w.r.t. a training set } \mathcal{D}} + \underbrace{1 - \mathcal{S}(\ell(\cdot,h), z_{trn})}_{\text{Mutual stability between the loss of } h \text{ and a random training example } z_{trn}} \qquad (*)$$

- Design of a **bias-variance balancing objective function**

$$R_{obj}(h,\mathcal{D}) = \sqrt{R_{emp}(h,\mathcal{D})^2 + dCorr(\ell(\cdot,h), z_{trn})^2}$$

- **Aims to tighten the upper bound** $(*)$ by:
  1) **minimizing** the **empirical risk** $R_{emp}(h,\mathcal{D})$
  2) utilizing **distance correlation** [3,4] to make the loss w.r.t. to given data as independent as possible of the data themselves and thus to indirectly control the **mutual stability term**

**Note:** $R_{obj}(h,\mathcal{D})$ is defined for a hypothesis $h$ selected by **any supervised learning algorithm** $\mathcal{L}$
⇒ In this study, this objective is utilized in a **structured regression** setting.

## Methodology

### Structured Regression by Gaussian CRFs

A Gaussian CRF (GCRF) models the conditional distribution:

$$P(\mathbf{y}|\mathbf{X}) = \frac{1}{Z}\exp\left\{-\alpha\sum_{i=1}^{N}(y_i - \phi(\mathbf{x}_i))^2 - \beta\sum_{i\sim j} S_{ij}\,(y_i - y_j)^2\right\}$$

### Proposed Model

**Genera**L**ization-**A**ware **C**ollaborative **E**nsemble **R**egressor **(GLACER)**

**Input:**
Training set $\mathcal{D}$
Similarity matrix $\mathbf{S}$
\# of components $M$
Sub-sampling fraction $\eta$

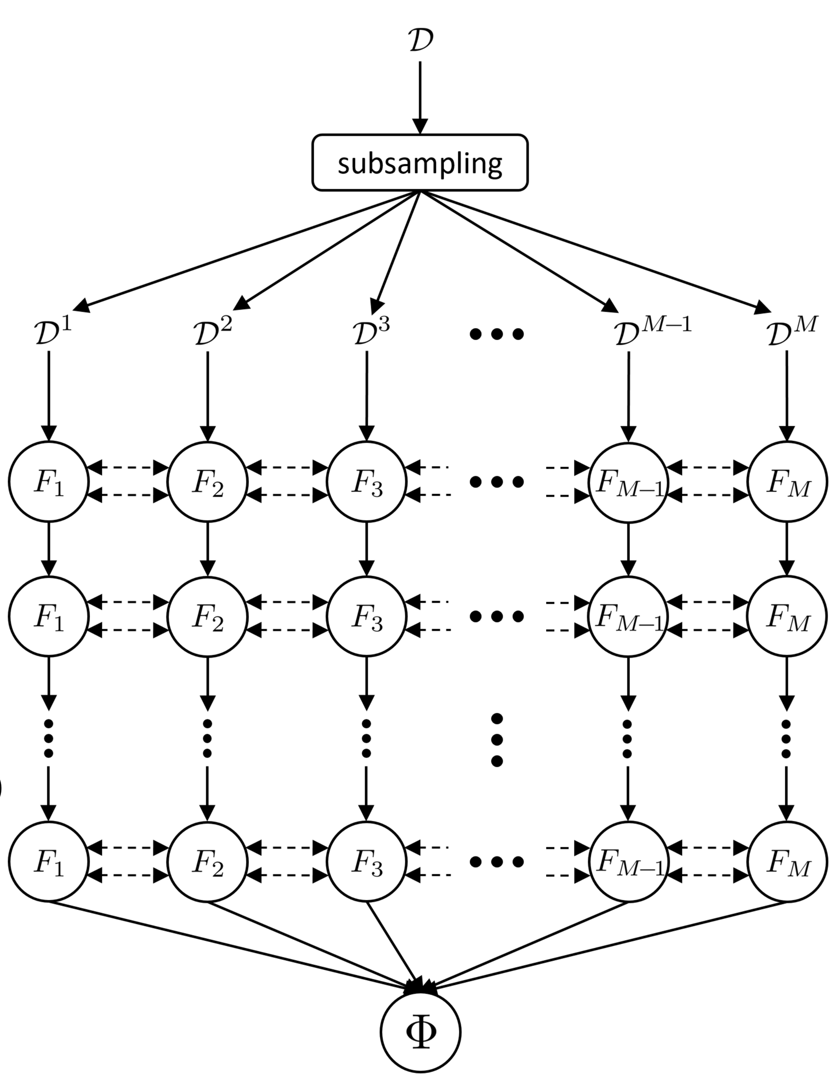
**Sample** $\mathcal{D}$ and $\mathbf{S}$ (omitted for brevity) $M$ times without replacement using the sub-sampling fraction $\eta$

**Train** a single GCRF component $F_m$ (on top of a least-squares booster) on each $\mathcal{D}^m$ and its corresponding $\mathbf{S}^m$

**Loop**
- **determine the worst-fit** example for each component
- **exchange** the worst-fit examples between the pair of GCRFs that fosters the highest decrease in $R_{obj}(\Phi,\mathcal{D})$

**Repeat until** no exchange can further decrease $R_{obj}(\Phi,\mathcal{D})$

**Prediction:** $\Phi(\mathbf{X},\mathbf{S}) = \frac{1}{M}\sum_{m=1}^{M} F_m(\mathbf{X},\mathbf{S})$




## Results

### Experiments on Synthetic Data

- **Examples:** 3000 input-output pairs
  - input features: normally distributed
  - outputs: parameterized polynomials with uniformly distributed parameters
- **Structure:** generated using an Erdős-Rényi random graph model

**Experiment #1: Generalization Capability**

- In general, structured variants perform better than unstructured
- While the baselines' MSEs decrease with the increased size of training data, GLACER is more accurate and sustains stable predictions when only 50% training data is available

  ⇒ This is consistent in case smaller/larger training fractions are used

| **Model \ Frac.** | **50%** |
|---|---|
| Linear Reg. | $2.3 \pm 0.05$ |
| Structured Linear Reg. | $1.5 \pm 0.05$ |
| Neural Network | $1.4 \pm 0.21$ |
| Structured Neural Network | $1.0 \pm 0.18$ |
| Support Vector Reg. | $2.4 \pm 0.12$ |
| Structured Support Vector Reg. | $1.8 \pm 0.12$ |
| Subbagging | $1.3 \pm 0.01$ |
| Structured Subbagging | $0.9 \pm 0.03$ |
| Random Forest | $1.6 \pm 0.05$ |
| Structured Random Forest | $1.4 \pm 0.03$ |
| LS Boosting | $2.8 \pm 0.05$ |
| Structured LS Boosting | $0.9 \pm 0.02$ |
| Convex Network Lasso | $1.2 \pm 0.04$ |
| Non-convex Network Lasso | $1.3 \pm 0.05$ |
| **GLACER** | $\mathbf{0.3 \pm 0.01}$ |

Average testing MSE when 50% of the training data is supplied.

**Experiment #2: Influence of** $dCorr$

| $R_{obj}$ / **Frac.** | **Without** $dCorr$ | **With** $dCorr$ |
|---|---|---|
| **10%** | 0.74 | 0.72 |
| **50%** | 0.44 | 0.25 |
| **100%** | 0.53 | 0.25 |

Average testing MSE, obtained before and after using $dCorr$ within $R_{obj}$.

- GLACER manifests lower average MSEs when $dCorr$ is used in $R_{obj}$
  ⇒ This is consistent as the training data increases
- Without $dCorr$, the avg. MSE deteriorates once the training fraction increases from 50% to 100%
  ⇒ Might be an indication of overfitting
- Incorporating $dCorr$ into $R_{obj}$ prevents large increases in MSE

### Sacramento Real-Estate

- **Nodes:** 985 real estate transactions were observed in the Greater Sacramento area
- **Features:** # of bedrooms and bathrooms, house area in square feet, location in terms of latitude and longitude
- **Structure:** based on geospatial similarity
- **Train/test** ratio used is the same as in [5]

**Task:** predict the housing prices

| **Model** | **MSE** |
|---|---|
| Linear Reg. | $0.507 \pm 0.025$ |
| Structured Linear Reg. | $0.465 \pm 0.024$ |
| Neural Network | $0.516 \pm 0.026$ |
| Structured Neural Network | $0.463 \pm 0.023$ |
| Support Vector Reg. | $0.515 \pm 0.031$ |
| Structured Support Vector Reg. | $0.479 \pm 0.034$ |
| Subbagging | $0.304 \pm 0.017$ |
| Structured Subbagging | $0.262 \pm 0.015$ |
| Random Forest | $0.283 \pm 0.020$ |
| Structured Random Forest | $0.249 \pm 0.015$ |
| LS Boosting | $0.288 \pm 0.015$ |
| Structured LS Boosting | $0.250 \pm 0.017$ |
| Convex Network Lasso | $0.368 \pm 0.013$ |
| Non-convex Network Lasso | $0.380 \pm 0.017$ |
| **GLACER** | $\mathbf{0.225 \pm 0.005}$ |

Testing MSE, averaged over 10 random splits.

### Medicare Readmissions

- **Nodes:** 1000 hospital records referring to hospitals with more than ∼150 readmissions
- **Features:** # of discharges, excess readmission ratio, estimated/expected readmission rates
- **Structure:** similarities between hospital readmissions
- **Period:** 36 months (July 2012 – June 2015)

**Task:** predict the number of hospital readmissions

| **Model** | **MSE** |
|---|---|
| Linear Reg. | $1755.708 \pm 616.119$ |
| Structured Linear Reg. | $525.551 \pm 196.065$ |
| Neural Network | $2037.421 \pm 1199.805$ |
| Structured Neural Network | $1618.547 \pm 1192.462$ |
| Support Vector Reg. | $1359.342 \pm 697.910$ |
| Structured Support Vector Reg. | $504.076 \pm 221.228$ |
| Subbagging | $441.524 \pm 101.065$ |
| Structured Subbagging | $234.505 \pm 74.378$ |
| Random Forest | $508.294 \pm 110.988$ |
| Structured Random Forest | $247.406 \pm 35.814$ |
| LS Boosting | $595.289 \pm 136.174$ |
| Structured LS Boosting | $182.006 \pm 24.919$ |
| (Non-)convex Network Lasso | $5012.614 \pm 768.945$ |
| **GLACER** | $\mathbf{73.183 \pm 9.032}$ |

Testing MSE, averaged over 10 random splits.

**GLACER - Discussion:**

- **Outperforms alternatives** by ∼**10-56%** and **more than 49%** when predicting housing prices and hospital readmissions, respectively.
- Achieves **statistically significant improvements** ⇒ $p$-values are smaller than 0.01 for Sacramento, and 0.021 for Medicare.
- Manifests **stable predictions** ⇒ tight confidence interval for its average MSE.

## Acknowledgments

This research was supported in part by the Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) under Contract no. DARPA-THOR-W911NF-16-C-0050, an IQVIA Big Data Analytics contract, NSF BIGDATA grant 14476570, and ONR/ONR Global grant N62909-16-1-2222.

## References

[1] Alabdulmohsin, I.: Algorithmic stability and uniform generalization. In: *Advances in Neural Information Processing Systems*, pp. 19-27 (2015)

[2] Alabdulmohsin, I.: An information-theoretic route from generalization in expectation to generalization in probability. In: *Artificial Intelligence and Statistics*, pp. 92-100 (2017)

[3] Székely, G. J., Rizzo, M. L., Bakirov, N. K.: Measuring and testing dependence by correlation of distances. *The annals of statistics* 35(6), 2769-2794 (2007)

[4] Székely, G. J., Rizzo, M. L.: Brownian distance covariance. *The annals of applied statistics*, 1236-1265 (2009)

[5] Hallac, D., Leskovec, J., Boyd, S.: Network lasso: Clustering and optimization in large graphs. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 387-396 (2015)

Corresponding author: Zoran Obradovic (zoran.obradovic@temple.edu)
**IJCAI-18**, THE 27TH INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, STOCKHOLM, SWEDEN, JULY 2018