Mitigating multiple descents: A general framework for risk monotonization [a]

Pratik Patil

Carnegie Mellon University

TOPML 2021

[a] Joint work with Arun Kuchibhotla, Yuting Wei, Alessandro Rinaldo

Table of contents

Outline

Double/multiple descent recap

- Risk behavior of several commonly used prediction procedures such as OLS linear regression, logistic regression, SVMs have been recently studied under the proportional asymptotics setting.
- Proportional asymptotics refers to the setting where the number of features p of the data scales proportionally to the number of observations n of the data (i.e., $p/n \to \gamma \in (0, \infty)$).
- This should be contrasted with the traditional "low-dimensional" setting where either p is fixed or p diverges but $p/n \to 0$.

Double/multiple descent recap

- Risk behavior of several commonly used prediction procedures such as OLS linear regression, logistic regression, SVMs have been recently studied under the proportional asymptotics setting.
- Proportional asymptotics refers to the setting where the number of features p of the data scales proportionally to the number of observations n of the data (i.e., $p/n \to \gamma \in (0, \infty)$).
- This should be contrasted with the traditional "low-dimensional" setting where either p is fixed or p diverges but $p/n \to 0$.

Double/multiple descent recap

- Risk behavior of several commonly used prediction procedures such as OLS linear regression, logistic regression, SVMs have been recently studied under the proportional asymptotics setting.
- Proportional asymptotics refers to the setting where the number of features p of the data scales proportionally to the number of observations n of the data (i.e., $p/n \to \gamma \in (0, \infty)$).
- This should be contrasted with the traditional "low-dimensional" setting where either p is fixed or p diverges but $p/n \to 0$.

Double/multiple descent recap

- A surprising phenomenon has been observed in the proportional asymptotics regime both empirically and theoretically (under some distributional assumptions).

- The risk of the common predictors first increases as p/n increases up to some threshold and then decreases.

- There are two ways to view this:
  - If p is thought of as fixed (large value), this implies that as sample size increases the risk first decreases and then increases.
    More data hurts.
  - If n is thought of as fixed (large value), this implies that as the number of features/covariates increase the risk first increases and then decreases.
    More features do not hurt.

- We will focus on the first interpretation: more data can hurt.

Double/multiple descent recap

- A surprising phenomenon has been observed in the proportional asymptotics regime both empirically and theoretically (under some distributional assumptions).
- The risk of the common predictors first increases as p/n increases up to some threshold and then decreases.
- There are two ways to view this:
  - If p is thought of as fixed (large value), this implies that as sample size increases the risk first decreases and then increases.
    More data hurts.
  - If n is thought of as fixed (large value), this implies that as the number of features/covariates increase the risk first increases and then decreases.
    More features do not hurt.
- We will focus on the first interpretation: more data can hurt.

Double/multiple descent recap

- A surprising phenomenon has been observed in the proportional asymptotics regime both empirically and theoretically (under some distributional assumptions).
- The risk of the common predictors first increases as p/n increases up to some threshold and then decreases.
- There are two ways to view this:
  - If p is thought of as fixed (large value), this implies that as sample size increases the risk first decreases and then increases.
    More data hurts.
  - If n is thought of as fixed (large value), this implies that as the number of features/covariates increase the risk first increases and then decreases.
    More features do not hurt.
- We will focus on the first interpretation: more data can hurt.

Double/multiple descent recap

- A surprising phenomenon has been observed in the proportional asymptotics regime both empirically and theoretically (under some distributional assumptions).
- The risk of the common predictors first increases as p/n increases up to some threshold and then decreases.
- There are two ways to view this:
  - If p is thought of as fixed (large value), this implies that as sample size increases the risk first decreases and then increases. More data hurts.
  - If n is thought of as fixed (large value), this implies that as the number of features/covariates increase the risk first increases and then decreases. More features do not hurt.
- We will focus on the first interpretation: more data can hurt.

Double/multiple descent recap

- A surprising phenomenon has been observed in the proportional asymptotics regime both empirically and theoretically (under some distributional assumptions).
- The risk of the common predictors first increases as p/n increases up to some threshold and then decreases.
- There are two ways to view this:
  - If p is thought of as fixed (large value), this implies that as sample size increases the risk first decreases and then increases.
    More data hurts.
  - If n is thought of as fixed (large value), this implies that as the number of features/covariates increase the risk first increases and then decreases.
    More features do not hurt.
- We will focus on the first interpretation: more data can hurt.

Double/multiple descent recap

- A surprising phenomenon has been observed in the proportional asymptotics regime both empirically and theoretically (under some distributional assumptions).
- The risk of the common predictors first increases as p/n increases up to some threshold and then decreases.
- There are two ways to view this:
  - If p is thought of as fixed (large value), this implies that as sample size increases the risk first decreases and then increases.
    More data hurts.
  - If n is thought of as fixed (large value), this implies that as the number of features/covariates increase the risk first increases and then decreases.
    More features do not hurt.
- We will focus on the first interpretation: more data can hurt.

Double/multiple descent in linear regression

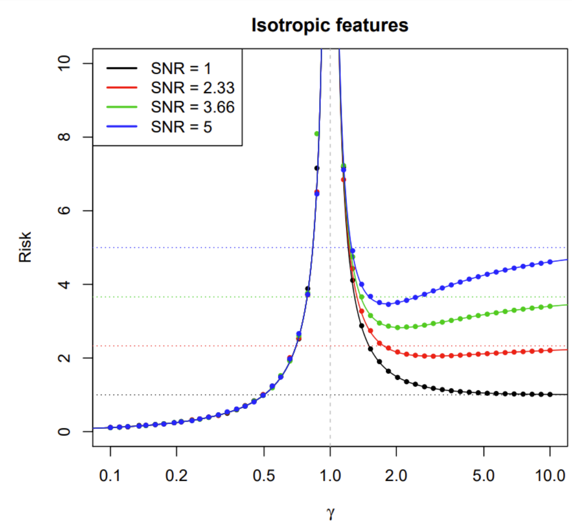


Figure 1: Risk of the min-norm least squares under $p/n \approx \gamma$ [HMRT19]

Outline

Double/multiple descent

Motivation and the problem

Risk monotonization (zero-step)

Risk monotonization (one-step)

Summary

Motivation

- When the data comprises of i.i.d. observations, we expect that more data will help in prediction or estimation.
- A procedure leading to worse risk as the number of observations increases is not using the data properly and can be labeled "sub-optimal."
- It is, thus, surprising to note that several procedures optimal in the "low-dimensional" settings are sub-optimal in the proportional asymptotics regime.
- Such procedures can be readily improved by simply using less number of observations than available for better risk behaviour.

Motivation

- When the data comprises of i.i.d. observations, we expect that more data will help in prediction or estimation.
- A procedure leading to worse risk as the number of observations increases is not using the data properly and can be labeled "sub-optimal."
- It is, thus, surprising to note that several procedures optimal in the "low-dimensional" settings are sub-optimal in the proportional asymptotics regime.
- Such procedures can be readily improved by simply using less number of observations than available for better risk behaviour.

Motivation

- When the data comprises of i.i.d. observations, we expect that more data will help in prediction or estimation.
- A procedure leading to worse risk as the number of observations increases is not using the data properly and can be labeled "sub-optimal."
- It is, thus, surprising to note that several procedures optimal in the "low-dimensional" settings are sub-optimal in the proportional asymptotics regime.
- Such procedures can be readily improved by simply using less number of observations than available for better risk behaviour.

Motivation

- When the data comprises of i.i.d. observations, we expect that more data will help in prediction or estimation.
- A procedure leading to worse risk as the number of observations increases is not using the data properly and can be labeled "sub-optimal."
- It is, thus, surprising to note that several procedures optimal in the "low-dimensional" settings are sub-optimal in the proportional asymptotics regime.
- Such procedures can be readily improved by simply using less number of observations than available for better risk behaviour.

Motivation

- When the data comprises of i.i.d. observations, we expect that more data will help in prediction or estimation.
- A procedure leading to worse risk as the number of observations increases is not using the data properly and can be labeled "sub-optimal."
- It is, thus, surprising to note that several procedures optimal in the "low-dimensional" settings are sub-optimal in the proportional asymptotics regime.
- Such procedures can be readily improved by simply using less number of observations than available for better risk behaviour.

Motivation and the problem

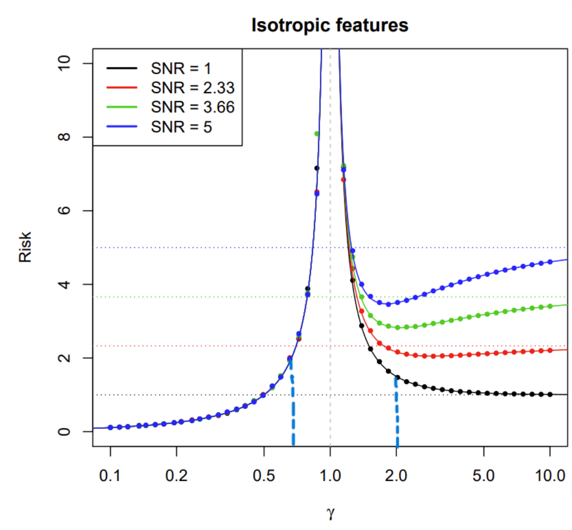


Figure 2: Risk of the min-norm least squares under $p/n \approx \gamma$.

The problem

- Given a number of observations (n) and a number of features (p), how do we know if a lesser number of observations would actually yield a better risk?
- What is the best sample size to reduce the dataset in order to attain the best possible risk?

Solution: cross-validation.

The problem

- Given a number of observations (n) and a number of features (p), how do we know if a lesser number of observations would actually yield a better risk?
- What is the best sample size to reduce the dataset in order to attain the best possible risk?

Solution: cross-validation.

The problem

- Given a number of observations (n) and a number of features (p), how do we know if a lesser number of observations would actually yield a better risk?
- What is the best sample size to reduce the dataset in order to attain the best possible risk?

Solution: cross-validation.

Outline

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works even with risk divergences at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works even with risk divergences at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works even with risk divergences at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works even with risk divergences at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works even with risk divergences at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works even with risk divergences at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works even with risk divergences at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works even with risk divergences at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

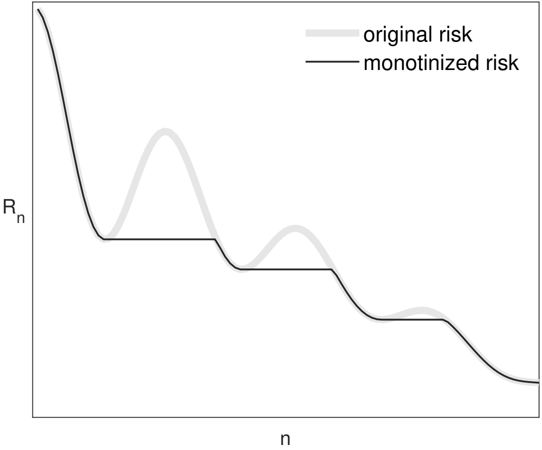
1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works even with risk divergences at some aspect ratios

Risk monotonization, illustration

If R_n represents the "risk" of a procedure at sample size n, then by risk monotonization we mean a procedure with risk $\min_{m \le n} R_m$.



Split sample cross-validation

- Given data $\mathcal{D}_n$ of n i.i.d. observations and a prediction procedure $\widetilde{f}$, split $\mathcal{D}_n$ into training data $\mathcal{D}_{\mathrm{tr}}$ with $n(1 - 1/\log n)$ observations and test data $\mathcal{D}_{\mathrm{te}}$ with $n/\log n$ observations.
- Note that
$$\lim_n \frac{p}{n} = \lim_n \frac{p}{n(1 - 1/\log n)}.$$
- For $n^{1/2} \leq k \leq |\mathcal{D}_{\mathrm{tr}}|$, obtain a predictor $\widetilde{f}_k$ by training $\widetilde{f}$ on a subset of $\mathcal{D}_{\mathrm{tr}}$ with k observations.
- If p/n converges to γ as $n \to \infty$, then
$$\left\{ \frac{p}{n^{1/2}}, \frac{p}{n^{1/2}+1}, \ldots, \frac{p}{|\mathcal{D}_{\mathrm{tr}}|} \right\} \text{ "} \to \text{" } [\gamma, \infty].$$
The set of aspect ratios for the predictors $\widetilde{f}_k$ covers $[\gamma, \infty]$.
- Choose one out of $\widetilde{f}_k, n^{1/2} \leq k \leq |\mathcal{D}_{\mathrm{tr}}|$ using an estimate of out-of-sample risk computed from $\mathcal{D}_{\mathrm{te}}$ This is split sample cross-validation.

Cross-validation risk estimate

- Traditionally, the risk of a predictor based on a test data is done via average loss. For example, with squared error loss, the traditional estimate of (prediction) risk of a predictor $\widetilde{f}_k$

$$\widehat{R}(\widetilde{f}_k) := \frac{1}{|\mathcal{D}_{\mathrm{te}}|} \sum_{j \in \mathcal{D}_{\mathrm{te}}} (Y_j - \widetilde{f}_k(X_j))^2.$$

- For a good performance simultaneously over $O(n)$ predictors and also to avoid strong tail assumptions on the loss, we also consider the median-of-means estimator.

- With either the average or median-of-means estimator of risk, we return the predictor $\widehat{f} := \widetilde{f}_{\widehat{k}}$ where

$$\widehat{k} \; := \operatorname*{argmin}_{n^{1/2} \le k \le |\mathcal{D}_{\mathrm{tr}}|} \; \widehat{R}(\widetilde{f}_k).$$

- $\widehat{k}$ represents the "best" sample size to use for the given number of features in the dataset and $\widetilde{f}_{\widehat{k}}$ is what we call a zero-step predictor that achieves risk monotonization.

Risk monotonization guarantee (informal statement)

Under the proportional asymptotics regime ($p/n \to \gamma$), and a mild assumption on the convergence of the prediction risk of $\widehat{f}$ trained on datasets with a limiting aspect ratio converges, we show that

$$R(\widehat{f}) = R(\widetilde{f}_{\widehat{k}}) \; = \; \inf_{\zeta \in [\gamma, \infty]} R^{\mathrm{det}}(\zeta; \widehat{f}) \; \times \; (1 + o_p(1)).$$

This shows that the zero-step predictor has a monotone risk in terms of the sample size and hence with respect to the limiting aspect ratio.

This is a model-free result in that no parametric model is assumed for the data. This is unlike most results in overparametrized learning which require stringent assumptions.

Risk monotonization (illustration)

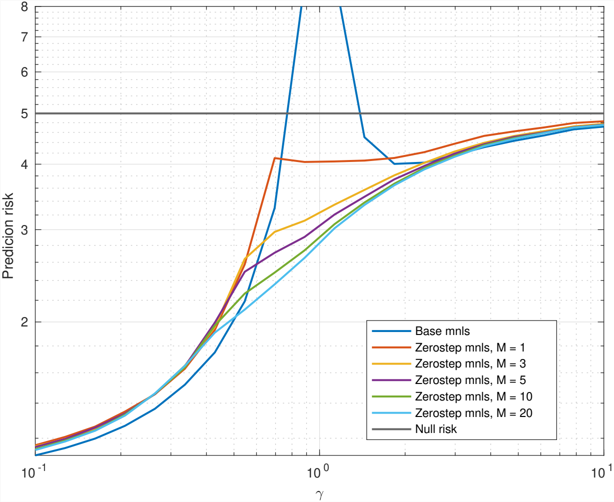


Figure 3: Risk monotonization of the minimum ℓ_2-norm interpolator

Outline

Double/multiple descent

Motivation and the problem

Risk monotonization (zero-step)

Risk monotonization (one-step)

Summary

Classical one-step estimation

- Idea: start with any arbitrary linear predictor, compute "residuals", fit least squares on residuals, and add to the original predictor.
- If the initial predictor is $\widetilde{f}(x) = x^\top \widehat{\beta}^{\text{init}}$, then the final predictor is:

$$\underbrace{X^\top \widetilde{\beta}^{\text{init}}}_{\text{initial predictor}} + \underbrace{X^\top \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i (Y_i - X_i^\top \widetilde{\beta}^{\text{init}}) \right)}_{\text{one-step component}}.$$

- It is well-known that in a low dimensional setting, starting with any consistent estimator, the final estimator is $n^{-1/2}$ consistent.

One-step estimation in high dimensions

- Question: can we perform one-step estimation in high dimensions?
- Issues:
  1. The inverse of sample covariance matrix $\sum_{i=1}^{n} XX_i^\top / n$ need not exist.
  2. In the overparameterized regime, the residuals $Y_i - X_i^\top \widehat{\beta}^{\mathrm{init}}$ are identically or approximately zero for many common estimators.
- Solutions:
  1. Use Moore-Penrose inverse in place of regular inverse
  2. Split the training data, use a part to compute initial estimator $\widehat{\beta}^{\mathrm{init}}$, and the other part to compute the residuals $Y_i - X_i^\top \widehat{\beta}^{\mathrm{init}}$.
- In summary:
  1. Start with a base predictor computed on subset of data.
  2. Evaluate residuals on a different subset of data.
  3. Fit min ℓ_2-norm estimator on the residuals.
  4. Add to the original predictor.
  5. Cross-validate the split proportions.

One-step monotonization guarantee (informal)

Under the proportional asymptotics regime ($p/n \to \gamma$), and a mild assumption on the convergence of the prediction risk of the base procedure trained on datasets with a limiting aspect ratio converges, we show that the one-step achieves the risk of

$$\inf_{1/\zeta_1 + 1/\zeta_2 \le 1/\gamma} R^{\mathrm{det}}(\zeta_1, \zeta_2; \widetilde{f}) \ \times \ (1 + o_p(1)).$$

The above function is monotone with respect to the limiting aspect ratio.

Furthermore, the risk of the one-step procedure is no smaller than that the zero-step procedure:

$$\min_{1/\zeta_1 + 1/\zeta_2 \le 1/\gamma} R^{\mathrm{det}}(\zeta_1, \zeta_2; \widetilde{f}) \ \le \ \min_{1/\zeta_1 \le 1/\gamma} R^{\mathrm{det}}(\zeta_1; \widetilde{f}),$$

One-step risk monotonization (illustration)

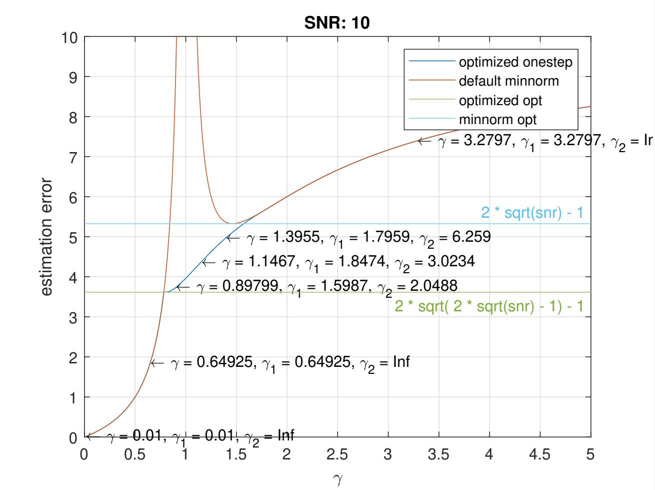


Figure 4: Risk monotonization of the min ℓ_2-norm interpolator

Outline

Summary

- We have introduced a general-purpose method to potentially improve any given predictor by monotonizing its risk in terms of n.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.
- In the paper, we study both average as well as median-of-means estimator of the prediction risk.
- Further, we provide additive and multiplicative oracle inequalities for the cross-validated risk and can handle diverging risks.
- We introduced the zero-step prediction procedure with a tuning parameter M that monotonizes the risk of a given predictor.
- For several commonly used predictors (min-ℓ_1, ℓ_2-norm LS), zero step predictor with $M > 1$ is strictly better than that with $M = 1$.
- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure.

Summary

- We have introduced a general-purpose method to potentially improve any given predictor by monotonizing its risk in terms of n.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.
- In the paper, we study both average as well as median-of-means estimator of the prediction risk.
- Further, we provide additive and multiplicative oracle inequalities for the cross-validated risk and can handle diverging risks.
- We introduced the zero-step prediction procedure with a tuning parameter M that monotonizes the risk of a given predictor.
- For several commonly used predictors (min-ℓ_1, ℓ_2-norm LS), zero step predictor with $M > 1$ is strictly better than that with $M = 1$.
- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure.

Summary

- We have introduced a general-purpose method to potentially improve any given predictor by monotonizing its risk in terms of n.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.
- In the paper, we study both average as well as median-of-means estimator of the prediction risk.
- Further, we provide additive and multiplicative oracle inequalities for the cross-validated risk and can handle diverging risks.
- We introduced the zero-step prediction procedure with a tuning parameter M that monotonizes the risk of a given predictor.
- For several commonly used predictors (min-ℓ_1, ℓ_2-norm LS), zero step predictor with $M > 1$ is strictly better than that with $M = 1$.
- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure.

Summary

- We have introduced a general-purpose method to potentially improve any given predictor by monotonizing its risk in terms of n.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.
- In the paper, we study both average as well as median-of-means estimator of the prediction risk.
- Further, we provide additive and multiplicative oracle inequalities for the cross-validated risk and can handle diverging risks.
- We introduced the zero-step prediction procedure with a tuning parameter M that monotonizes the risk of a given predictor.
- For several commonly used predictors (min-ℓ_1, ℓ_2-norm LS), zero step predictor with $M > 1$ is strictly better than that with $M = 1$.
- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure.

Summary

- We have introduced a general-purpose method to potentially improve any given predictor by monotonizing its risk in terms of n.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.
- In the paper, we study both average as well as median-of-means estimator of the prediction risk.
- Further, we provide additive and multiplicative oracle inequalities for the cross-validated risk and can handle diverging risks.
- We introduced the zero-step prediction procedure with a tuning parameter M that monotonizes the risk of a given predictor.
- For several commonly used predictors (min-ℓ_1, ℓ_2-norm LS), zero step predictor with $M > 1$ is strictly better than that with $M = 1$.
- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure.

Summary

- We have introduced a general-purpose method to potentially improve any given predictor by monotonizing its risk in terms of n.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.
- In the paper, we study both average as well as median-of-means estimator of the prediction risk.
- Further, we provide additive and multiplicative oracle inequalities for the cross-validated risk and can handle diverging risks.
- We introduced the zero-step prediction procedure with a tuning parameter M that monotonizes the risk of a given predictor.
- For several commonly used predictors (min-ℓ_1, ℓ_2-norm LS), zero step predictor with $M > 1$ is strictly better than that with $M = 1$.
- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure.

Summary

- We have introduced a general-purpose method to potentially improve any given predictor by monotonizing its risk in terms of n.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.
- In the paper, we study both average as well as median-of-means estimator of the prediction risk.
- Further, we provide additive and multiplicative oracle inequalities for the cross-validated risk and can handle diverging risks.
- We introduced the zero-step prediction procedure with a tuning parameter M that monotonizes the risk of a given predictor.
- For several commonly used predictors (min-ℓ_1, ℓ_2-norm LS), zero step predictor with $M > 1$ is strictly better than that with $M = 1$.
- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure.

Thanks for listening!

Questions/comments/thoughts?

Supplement

Recall: simple cross-validation

- Given data $\mathcal{D}_n$ of n i.i.d. observations and a prediction procedure $\widetilde{f}$, split $\mathcal{D}_n$ into training data $\mathcal{D}_{\mathrm{tr}}$ with $n(1 - 1/\log n)$ observations and test data $\mathcal{D}_{\mathrm{te}}$ with $n/\log n$ observations.
- Note that
$$\lim_n \frac{p}{n} = \lim_n \frac{p}{n(1 - 1/\log n)}.$$
- For $n^{1/2} \le k \le |\mathcal{D}_{\mathrm{tr}}|$, obtain a predictor $\widetilde{f}_k$ by training $\widetilde{f}$ on a subset of $\mathcal{D}_{\mathrm{tr}}$ with k observations.
- If p/n converges to γ as $n \to \infty$, then
$$\left\{ \frac{p}{n^{1/2}}, \frac{p}{n^{1/2}+1}, \ldots, \frac{p}{|\mathcal{D}_{\mathrm{tr}}|} \right\} \ \text{"} \to \text{"} \ [\gamma, \infty].$$
The set of aspect ratios for the predictors $\widetilde{f}_k$ covers $[\gamma, \infty]$.
- Now choose one out of $\widetilde{f}_k, n^{1/2} \le k \le |\mathcal{D}_{\mathrm{tr}}|$ using an estimate of out-of-sample risk computed from $\mathcal{D}_{\mathrm{te}}$.

Recall: simple cross-validation

- Given data $\mathcal{D}_n$ of n i.i.d. observations and a prediction procedure $\widetilde{f}$, split $\mathcal{D}_n$ into training data $\mathcal{D}_{\mathrm{tr}}$ with $n(1-1/\log n)$ observations and test data $\mathcal{D}_{\mathrm{te}}$ with $n/\log n$ observations.
- Note that
$$\lim_n \frac{p}{n} = \lim_n \frac{p}{n(1-1/\log n)}.$$
- For $n^{1/2} \leq k \leq |\mathcal{D}_{\mathrm{tr}}|$, obtain a predictor $\widetilde{f}_k$ by training $\widetilde{f}$ on a subset of $\mathcal{D}_{\mathrm{tr}}$ with k observations.
- Because there are $\binom{|\mathcal{D}_{\mathrm{tr}}|}{k}$ subsets of $\mathcal{D}_{\mathrm{tr}}$, one can alternatively consider
$$\widetilde{f}_k(x) := \frac{1}{M}\sum_{j=1}^{M} \widetilde{f}(x; \mathcal{D}_{\mathrm{tr}}^{k,j}).$$
- This reduces variance of the predictor $\widetilde{f}_k$, while keeping its expectation the same. Larger the M, better the predictor.

Risk monotonization (illustration)

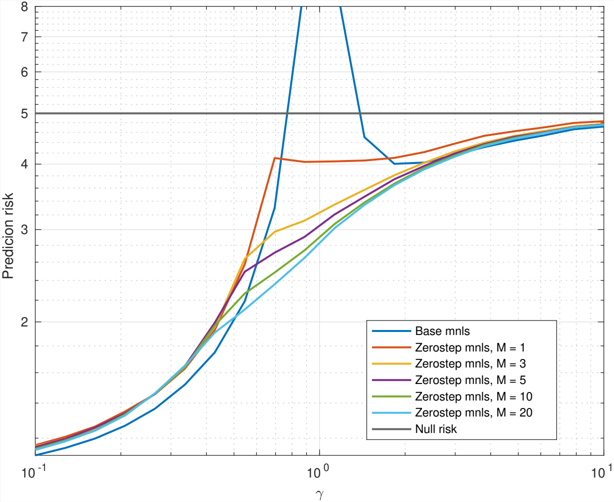


Figure 5: Risk monotonization of the min ℓ_2-norm interpolator