Towards a Statistical Theory of Data Selection Under Weak Supervision

Germain Kolossov, Andrea Montanari, Pulkit Tandon

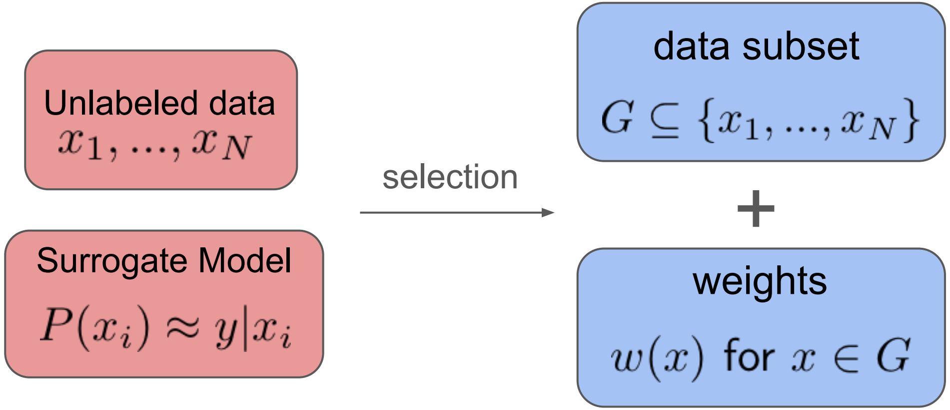
REFORM reading group, 11/20/24

Data Selection Setup

Unlabeled data
$x_1, \ldots, x_N$

Surrogate Model
$P(x_i) \approx y|x_i$

Data Selection Setup



Unlabeled data
$x_1, ..., x_N$

Surrogate Model
$P(x_i) \approx y|x_i$

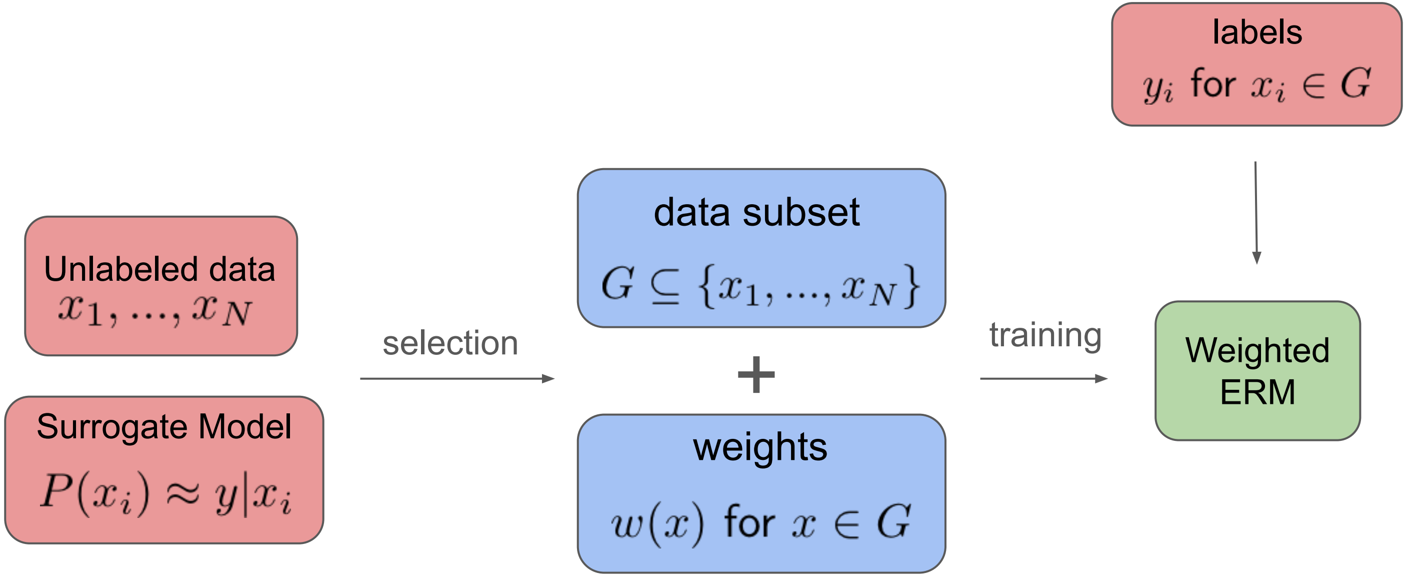
selection

data subset
$G \subseteq \{x_1, ..., x_N\}$

+

weights
$w(x)$ for $x \in G$

Data Selection Setup



labels

y_i for $x_i \in G$

Unlabeled data

$x_1, ..., x_N$

Surrogate Model

$P(x_i) \approx y|x_i$

selection

data subset

$G \subseteq \{x_1, ..., x_N\}$

+

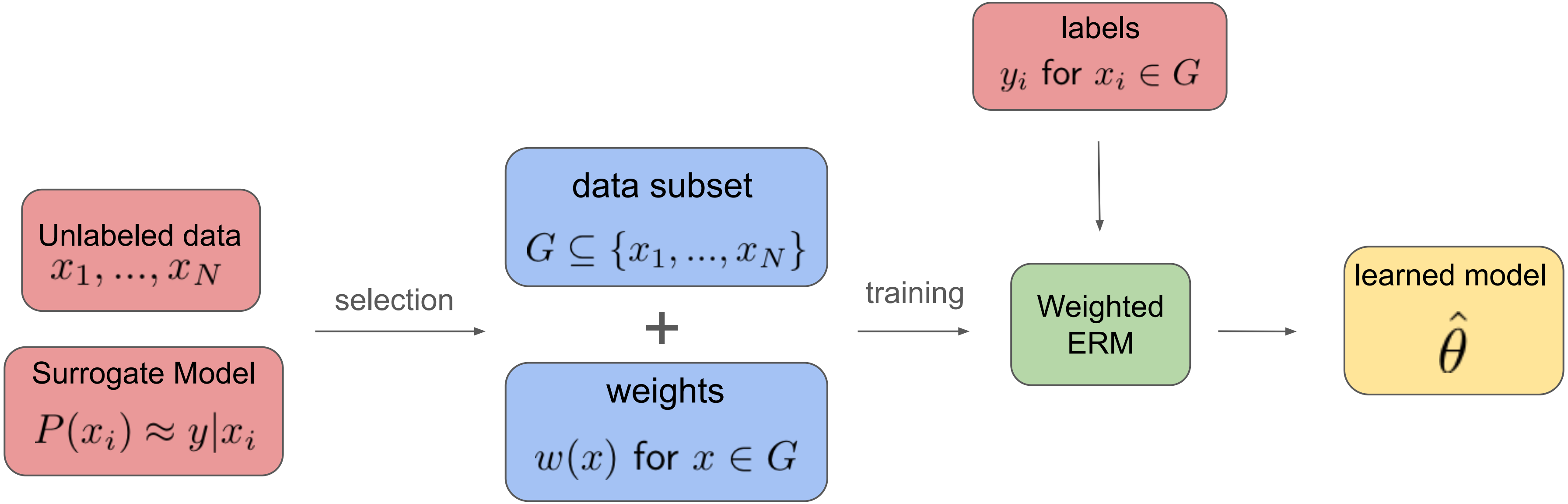
weights

$w(x)$ for $x \in G$

training

Weighted ERM

Data Selection Setup



Unlabeled data $x_1, ..., x_N$

Surrogate Model $P(x_i) \approx y|x_i$

selection

data subset $G \subseteq \{x_1, ..., x_N\}$

+

weights $w(x)$ for $x \in G$

training

labels y_i for $x_i \in G$

Weighted ERM

learned model $\hat{\theta}$

Selection Algorithm Details

Selection scheme = (π, w)

Selection Algorithm Details

Selection scheme = (π, w)

Each x included in dataset independently with probability

$$\pi(x, P(x))$$

If included, appears with weight

$$w(x, P(x)) > 0$$

Selection Algorithm Details

Target size achieved in expectation:

$$\sum_{i=1}^{N} \mathbb{E}[\pi(x_i)] = n$$

Selection scheme = (π, w)

Each x included in dataset independently with probability

$$\pi(x, P(x))$$

If included, appears with weight

$$w(x, P(x)) > 0$$

Selection Algorithm Details

Target size achieved in expectation:

$$\sum_{i=1}^{N} \mathbb{E}[\pi(x_i)] = n$$

Selection scheme = (π, w)

Each x included in dataset independently with probability

$$\pi(x, P(x))$$

If included, appears with weight

$$w(x, P(x)) > 0$$

Unbiased selection scheme: $w(x) = \dfrac{1}{\pi(x)}$

Selection Algorithm Details

Target size achieved in expectation:

$$\sum_{i=1}^{N} \mathbb{E}[\pi(x_i)] = n$$

Selection scheme = (π, w)

Each x included in dataset independently with probability

$$\pi(x, P(x))$$

If included, appears with weight

$$w(x, P(x)) > 0$$

Unbiased selection scheme: $w(x) = \dfrac{1}{\pi(x)}$

Non-reweighting selection scheme: $w(x) = 1$

Method of Analysis: Asymptotics

- Compare selections schemes based on performance as the original dataset size (N) grows to infinity.
  - Number of datapoints selected (n) also grows with N
  - Approaches some fixed fraction $\gamma \in (0, 1)$

Method of Analysis: Asymptotics

- Compare selections schemes based on performance as the original dataset size (N) grows to infinity.
  - Number of datapoints selected (n) also grows with N
  - Approaches some fixed fraction $\gamma \in (0, 1)$

“Low-Dimensional Regime”: Keep dimension fixed as N grows

“High-Dimensional Regime”: Grow dimension with N, converging to fixed ratio

Setting 1: Low-Dimension, Perfect Surrogate

Low-Dimension
Perfect Surrogate

Quantity of Interest: Asymptotic Error Coefficient

$$\rho(\overset{(\pi, w)}{S}, Q) = \lim_{M\to\infty} \lim_{N\to\infty} \mathbb{E}[\min\{N \underbrace{\|\hat{\theta} - \overset{\text{Population minimizer}}{\theta_*}\|_Q^2}_{\langle\hat{\theta} - \theta_*, Q(\hat{\theta} - \theta_*)\rangle}, M\}]$$

Low-Dimension
Perfect Surrogate

Quantity of Interest: Asymptotic Error Coefficient

$$\rho(\overset{(\pi, w)}{\overset{\downarrow}{S}}, Q) = \lim_{M\to\infty}\lim_{N\to\infty}\mathbb{E}[\min\{N\underbrace{\|\hat{\theta} - \overset{\text{Population minimizer}}{\overset{\downarrow}{\theta_*}}\|_Q^2}_{\overset{\uparrow}{\langle\hat{\theta} - \theta_*, Q(\hat{\theta} - \theta_*)\rangle}}, M\}]$$

Under some assumptions:

- Limit exists
- Has a closed form in terms of
    - S
    - conditional gradient covariance of loss at θ_*
    - conditional hessian of the loss at θ_*

Quantity of Interest: Asymptotic Error Coefficient

$$\rho(\overset{(\pi, w)}{S}, Q) = \lim_{M\to\infty} \lim_{N\to\infty} \mathbb{E}[\min\{N \underbrace{\|\hat{\theta} - \overset{\text{Population minimizer}}{\theta_*}\|_Q^2}_{\langle \hat{\theta} - \theta_*, Q(\hat{\theta} - \theta_*)\rangle}, M\}]$$

Under some assumptions:

- Limit exists
- Has a closed form in terms of
  - S
  - conditional gradient covariance of loss at θ_*
  - conditional hessian of the loss at θ_*

Best selection scheme in a class C:

$$\arg\min_{S\in C} \rho(S, Q)$$

Up next: used closed-form to solve for optimal strategy in certain classes.

Low-Dimension
Perfect Surrogate

Optimal Unbiased Strategy

$$w(x) = \frac{1}{\pi(x)}$$

- Can simplify and optimize the asymptotic error coefficient to solve for the optimal unbiased strategy:

Low-Dimension
Perfect Surrogate

Optimal Unbiased Strategy

$$\boxed{w(x) = \frac{1}{\pi(x)}}$$

- Can simplify and optimize the asymptotic error coefficient to solve for the optimal unbiased strategy:
    - Recovers selection based on influence function

Influence function

$$\psi(\boldsymbol{x}, y) = -\boldsymbol{H}^{-1}\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_*; y, \boldsymbol{x}) \longrightarrow \pi(\boldsymbol{x}_i) \propto \mathbb{E}\{\|\psi(\boldsymbol{x}_i, y_i)\|_{\boldsymbol{Q}}^2 | \boldsymbol{x}_i\}^{1/2}$$

Low-Dimension
Perfect Surrogate

Optimal Unbiased Strategy

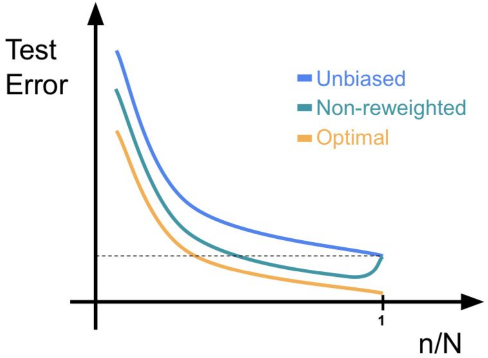
$$\boxed{w(x) = \frac{1}{\pi(x)}}$$

- Can simplify and optimize the asymptotic error coefficient to solve for the optimal unbiased strategy:
  - Recovers selection based on influence function

Influence function

$$\psi(\boldsymbol{x}, y) = -\boldsymbol{H}^{-1}\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_*; y, \boldsymbol{x}) \longrightarrow \pi(\boldsymbol{x}_i) \propto \mathbb{E}\left\{\left\|\psi(\boldsymbol{x}_i, y_i)\right\|_{\boldsymbol{Q}}^2 \middle| \boldsymbol{x}_i\right\}^{1/2}$$

  - Monotone non-increasing in selected proportion
  - Better than just random unbiased sampling



Low-Dimension
Perfect Surrogate

Optimal Non-reweighting Strategy $w(x) = 1$

- Again, simplify and optimize the asymptotic error coefficient to characterize the optimal non-reweighting strategy.
- Show that optimal strategy must be a fixed point of the following process:

Low-Dimension
Perfect Surrogate

Optimal Non-reweighting Strategy $w(x) = 1$

- Again, simplify and optimize the asymptotic error coefficient to characterize the optimal non-reweighting strategy.
- Show that optimal strategy must be a fixed point of the following process:
  - Strategy induces a score function for each point $Z(x; \pi)$
  - Optimal strategy must be a threshold decision based on score
  - Threshold parameters chosen to meet selection proportion

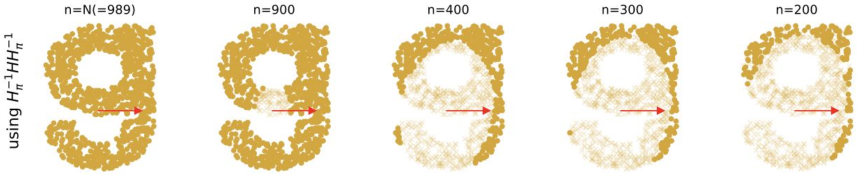
$$\pi_{\mathrm{nr}}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } Z(\boldsymbol{x}; \pi_{\mathrm{nr}}) > \lambda\,, \\ 0 & \text{if } Z(\boldsymbol{x}; \pi_{\mathrm{nr}}) < \lambda\,, \\ b(\boldsymbol{x}) \in [0,1] & \text{if } Z(\boldsymbol{x}; \pi_{\mathrm{nr}}) = \lambda\,. \end{cases}$$

Low-Dimension
Perfect Surrogate

Optimal Non-reweighting Strategy $\boxed{w(x) = 1}$

- Again, simplify and optimize the asymptotic error coefficient to characterize the optimal non-reweighting strategy.
- Show that optimal strategy must be a fixed point of the following process:
  - Strategy induces a score function for each point $Z(x;\pi)$
  - Optimal strategy must be a threshold decision based on score
  - Threshold parameters chosen to meet selection proportion

$$\pi_{\text{nr}}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } Z(\boldsymbol{x};\pi_{\text{nr}}) > \lambda\,, \\ 0 & \text{if } Z(\boldsymbol{x};\pi_{\text{nr}}) < \lambda\,, \\ b(\boldsymbol{x}) \in [0,1] & \text{if } Z(\boldsymbol{x};\pi_{\text{nr}}) = \lambda\,. \end{cases}$$



Linear regression setting: Z measures how different x is from selected data (related to *leverage scores*)

Optimal Non-reweighting Strategy: Proof Idea

- Suppose we have a minimizer π_{nr}
- Imagine perturbing π_{nr} to some $\pi_t := (1 - t)\pi_{\mathrm{nr}} + t\pi.$

Low-Dimension
Perfect Surrogate

Optimal Non-reweighting Strategy: Proof Idea

- Suppose we have a minimizer π_{nr}
- Imagine perturbing π_{nr} to some $\pi_t := (1-t)\pi_{\mathrm{nr}} + t\pi.$
- Can express perturbed error value as

$$\rho(\pi_t; \boldsymbol{Q}) = \rho(\pi_{\mathrm{nr}}; \boldsymbol{Q}) + t \int \big(\pi(\boldsymbol{x}) - \pi_{\mathrm{nr}}(\boldsymbol{x})\big) Z(\boldsymbol{x}; \pi_{\mathrm{nr}}) \, \mathbb{P}(\mathrm{d}\boldsymbol{x}) + o(t)$$

↑

Comes from closed-form
for asymptotic error

Low-Dimension
Perfect Surrogate

Optimal Non-reweighting Strategy: Proof Idea

- Suppose we have a minimizer π_{nr}
- Imagine perturbing π_{nr} to some $\pi_t := (1-t)\pi_{\rm nr} + t\pi.$
- Can express perturbed error value as

$$\rho(\pi_t;\boldsymbol{Q}) = \rho(\pi_{\rm nr};\boldsymbol{Q}) + t\int \big(\pi(\boldsymbol{x}) - \pi_{\rm nr}(\boldsymbol{x})\big) Z(\boldsymbol{x};\pi_{\rm nr})\,\mathbb{P}({\rm d}\boldsymbol{x}) + o(t)$$

$$\rho(\pi;\boldsymbol{Q}) = \mathsf{Tr}\Big(\mathbb{E}\{\pi(\boldsymbol{x})\boldsymbol{G}(\boldsymbol{x})\}\mathbb{E}\{\pi(\boldsymbol{x})\boldsymbol{H}(\boldsymbol{x})\}^{-1}\boldsymbol{Q}\mathbb{E}\{\pi(\boldsymbol{x})\boldsymbol{H}(\boldsymbol{x})\}^{-1}\Big)$$

Comes from closed-form
for asymptotic error

Conditional hessian

$$\boldsymbol{H}(\boldsymbol{x}) := \mathbb{E}\{\nabla^2_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_*;y,\boldsymbol{x})|\boldsymbol{x}\}$$

Conditional gradient covariance

$$\boldsymbol{G}(\boldsymbol{x}) := \mathbb{E}\{\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_*;y,\boldsymbol{x})\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_*;y,\boldsymbol{x})^{\mathsf{T}}|\boldsymbol{x}\}$$

Low-Dimension
Perfect Surrogate

Optimal Non-reweighting Strategy: Proof Idea

- Suppose we have a minimizer π_{nr}
- Imagine perturbing π_{nr} to some $\pi_t := (1-t)\pi_{\mathrm{nr}} + t\pi.$
- Can express perturbed error value as

$$\rho(\pi_t; \boldsymbol{Q}) = \rho(\pi_{\mathrm{nr}}; \boldsymbol{Q}) - t\int \big(\pi(\boldsymbol{x}) - \pi_{\mathrm{nr}}(\boldsymbol{x})\big) Z(\boldsymbol{x}; \pi_{\mathrm{nr}})\, \mathbb{P}(\mathrm{d}\boldsymbol{x}) + o(t)$$

Comes from closed-form
for asymptotic error

$$\rho(\pi; \boldsymbol{Q}) = \mathsf{Tr}\Big(\mathbb{E}\{\pi(\boldsymbol{x})\boldsymbol{G}(\boldsymbol{x})\}\mathbb{E}\{\pi(\boldsymbol{x})\boldsymbol{H}(\boldsymbol{x})\}^{-1}\boldsymbol{Q}\mathbb{E}\{\pi(\boldsymbol{x})\boldsymbol{H}(\boldsymbol{x})\}^{-1}\Big)$$

$$Z(\boldsymbol{x}; \pi) := -\mathsf{Tr}\{\boldsymbol{G}(\boldsymbol{x})\boldsymbol{H}_\pi^{-1}\boldsymbol{Q}\boldsymbol{H}_\pi^{-1}\} + 2\mathsf{Tr}\{\boldsymbol{H}(\boldsymbol{x})\boldsymbol{H}_\pi^{-1}\boldsymbol{Q}\boldsymbol{H}_\pi^{-1}\boldsymbol{G}_\pi\boldsymbol{H}_\pi^{-1}\}$$

$$\boldsymbol{G}_\pi := \mathbb{E}_\pi \boldsymbol{G}(\boldsymbol{x}), \qquad \boldsymbol{H}_\pi := \mathbb{E}_\pi \boldsymbol{H}(\boldsymbol{x}), \quad \textit{where} \quad \mathbb{E}_\pi f(\boldsymbol{x}) := \frac{\mathbb{E}\{f(\boldsymbol{x})\pi(\boldsymbol{x})\}}{\mathbb{E}\{\pi(\boldsymbol{x})\}}$$

Conditional hessian

$$\boldsymbol{H}(\boldsymbol{x}) := \mathbb{E}\{\nabla^2_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_*; y, \boldsymbol{x})|\boldsymbol{x}\}$$

Conditional gradient covariance

$$\boldsymbol{G}(\boldsymbol{x}) := \mathbb{E}\{\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_*; y, \boldsymbol{x})\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_*; y, \boldsymbol{x})^{\mathsf{T}}|\boldsymbol{x}\}$$

Low-Dimension
Perfect Surrogate

Optimal Non-reweighting Strategy: Proof Idea

- Suppose we have a minimizer π_{nr}
- Imagine perturbing π_{nr} to some $\pi_t := (1 - t)\pi_{\mathrm{nr}} + t\pi.$
- Can express perturbed error value as

Must be non-negative

$$\rho(\pi_t; \boldsymbol{Q}) = \rho(\pi_{\mathrm{nr}}; \boldsymbol{Q}) - t \int \big(\pi(\boldsymbol{x}) - \pi_{\mathrm{nr}}(\boldsymbol{x})\big) Z(\boldsymbol{x}; \pi_{\mathrm{nr}})\, \mathbb{P}(\mathrm{d}\boldsymbol{x}) + o(t)$$

Low-Dimension
Perfect Surrogate

Optimal Non-reweighting Strategy: Proof Idea

- Suppose we have a minimizer π_{nr}
- Imagine perturbing π_{nr} to some $\pi_t := (1-t)\pi_{\text{nr}} + t\pi.$
- Can express perturbed error value as

Must be non-negative

$$\rho(\pi_t; \boldsymbol{Q}) = \rho(\pi_{\text{nr}}; \boldsymbol{Q}) - t \int \big(\pi(\boldsymbol{x}) - \pi_{\text{nr}}(\boldsymbol{x})\big) Z(\boldsymbol{x}; \pi_{\text{nr}})\, \mathbb{P}(\mathrm{d}\boldsymbol{x}) + o(t)$$

- Claim: $Z(\boldsymbol{x}; \pi_{\text{nr}})$ constant on all x with $\pi_{nr}(x) \in (0,1)$
  - If not, can construct feasible strategy that breaks non-negativity

Low-Dimension
Perfect Surrogate

Optimal Non-reweighting Strategy: Proof Idea

- Suppose we have a minimizer π_{nr}
- Imagine perturbing π_{nr} to some $\pi_t := (1-t)\pi_{\mathrm{nr}} + t\pi.$
- Can express perturbed error value as

Must be non-positive

$$\rho(\pi_t;\boldsymbol{Q}) = \rho(\pi_{\mathrm{nr}};\boldsymbol{Q}) - t\int \big(\pi(\boldsymbol{x}) - \pi_{\mathrm{nr}}(\boldsymbol{x})\big) Z(\boldsymbol{x};\pi_{\mathrm{nr}})\,\mathbb{P}(\mathrm{d}\boldsymbol{x}) + o(t)$$

- Claim: $Z(\boldsymbol{x};\pi_{\mathrm{nr}})$ constant on all x with $\pi_{nr}(x) \in (0,1)$
  - If not, can construct feasible strategy that breaks non-positivity
- Implies the threshold structure

$$\pi_{\mathrm{nr}}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } Z(\boldsymbol{x};\pi_{\mathrm{nr}}) > \lambda\,, \\ 0 & \text{if } Z(\boldsymbol{x};\pi_{\mathrm{nr}}) < \lambda\,, \\ b(\boldsymbol{x}) \in [0,1] & \text{if } Z(\boldsymbol{x};\pi_{\mathrm{nr}}) = \lambda\,. \end{cases}$$

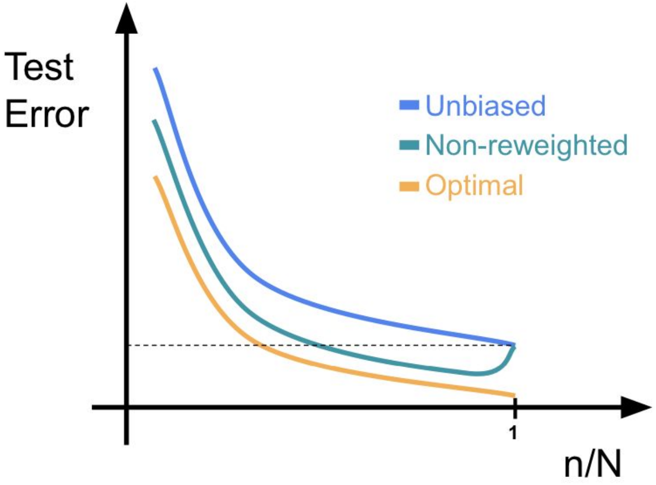
the constant value

Low-Dimension
Perfect Surrogate

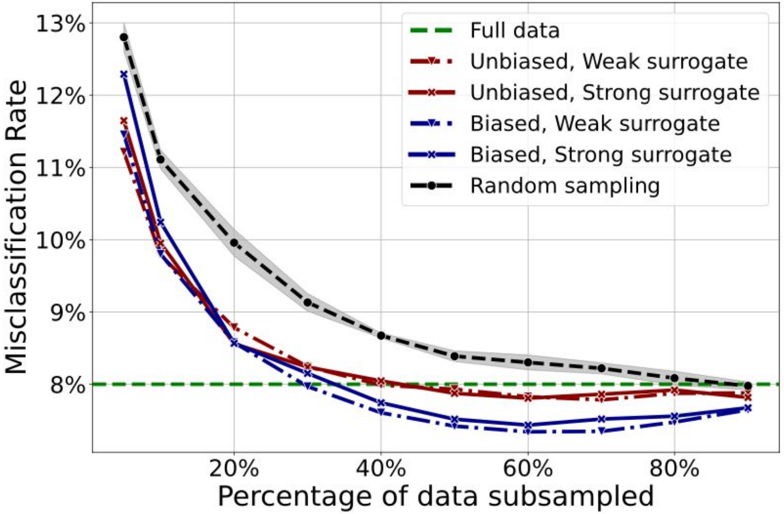
Non-reweighting vs Unbiased Strategies

Thm: Unbiased can be arbitrarily worse than non-reweighting strategies.

- In terms of ratio of asymptotic error coefficients



theorized



empirically

Part 2: Imperfect surrogates, High-dimensional asymptotics

Imperfect surrogates

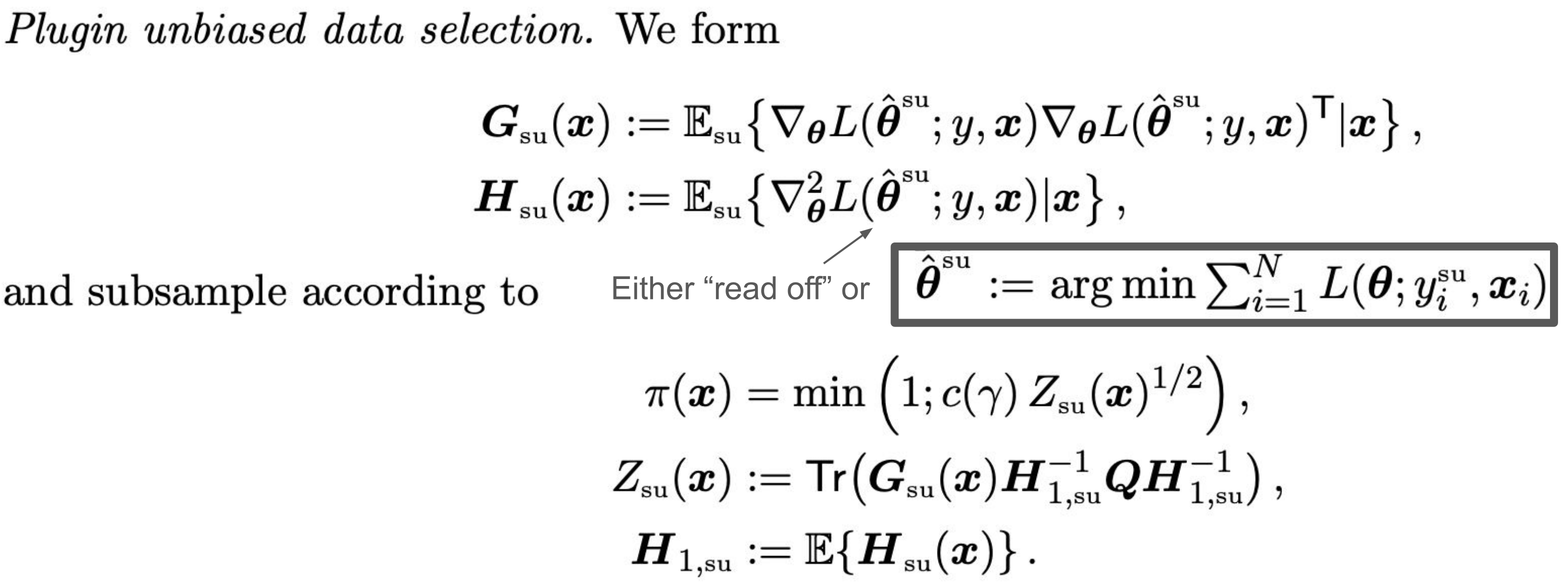
So far, we assume access to (essentially) $P(y|x)$ via a perfect surrogate

What happens if the surrogate is imperfect, so $P_{su}(y|x) \neq P(y|x)$

First idea: “Plug in” the surrogate (treat it as if it were truly $P(y|x)$)

Main result from this paper: This is suboptimal, but something close to it is roughly (minimax) optimal

First approach: Plug-in estimation



Plugin unbiased data selection. We form

$$\boldsymbol{G}_{\rm su}(\boldsymbol{x}) := \mathbb{E}_{\rm su}\big\{\nabla_{\boldsymbol{\theta}} L(\hat{\boldsymbol{\theta}}^{\rm su}; y, \boldsymbol{x}) \nabla_{\boldsymbol{\theta}} L(\hat{\boldsymbol{\theta}}^{\rm su}; y, \boldsymbol{x})^{\sf T} | \boldsymbol{x}\big\}\,,$$

$$\boldsymbol{H}_{\rm su}(\boldsymbol{x}) := \mathbb{E}_{\rm su}\big\{\nabla^2_{\boldsymbol{\theta}} L(\hat{\boldsymbol{\theta}}^{\rm su}; y, \boldsymbol{x}) | \boldsymbol{x}\big\}\,,$$

and subsample according to

Either “read off” or $\boxed{\hat{\boldsymbol{\theta}}^{\rm su} := \arg\min \sum_{i=1}^{N} L(\boldsymbol{\theta}; y_i^{\rm su}, \boldsymbol{x}_i)}$

$$\pi(\boldsymbol{x}) = \min\Big(1; c(\gamma)\, Z_{\rm su}(\boldsymbol{x})^{1/2}\Big)\,,$$

$$Z_{\rm su}(\boldsymbol{x}) := {\sf Tr}\big(\boldsymbol{G}_{\rm su}(\boldsymbol{x}) \boldsymbol{H}_{1,{\rm su}}^{-1} \boldsymbol{Q} \boldsymbol{H}_{1,{\rm su}}^{-1}\big)\,,$$

$$\boldsymbol{H}_{1,{\rm su}} := \mathbb{E}\{\boldsymbol{H}_{\rm su}(\boldsymbol{x})\}\,.$$

Approach for studying optimality: minimax framework

Assume that the surrogate predictor is close to (but not equal to) true model

$$\mathscr{K}_d(\mathsf{P}_{\rm su}; r) := \left\{\mathsf{P}: \ \ \mathbb{E}_{\boldsymbol{x}} \|\mathsf{P}(\,\cdot\,|\boldsymbol{x}) - \mathsf{P}_{\rm su}(\,\cdot\,|\boldsymbol{x})\|_{\rm TV} \le r\right\}.$$

(Doesn’t have to be TV, but we need the set to be convex)

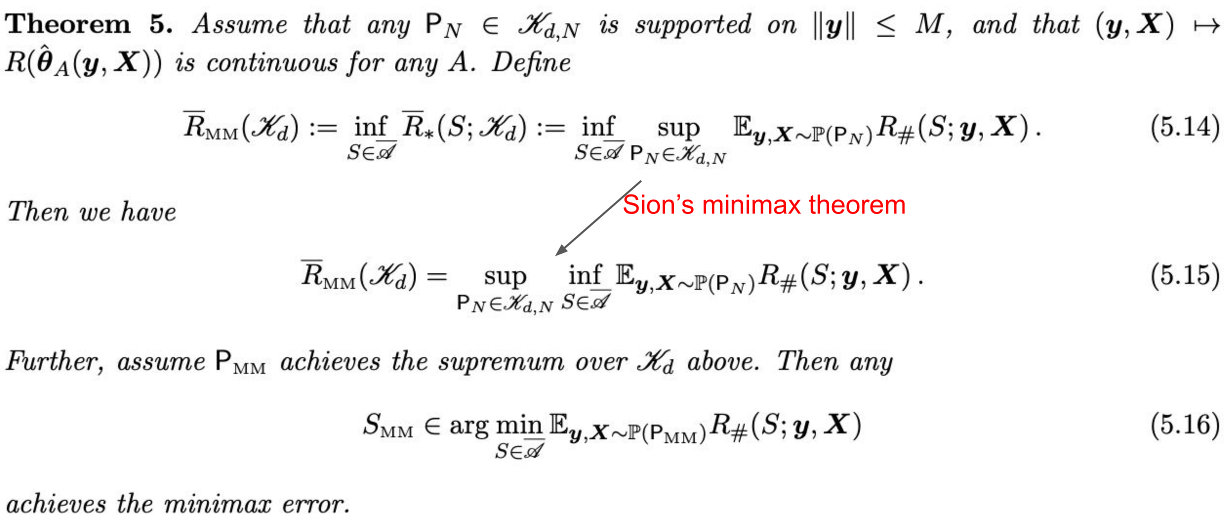
This allows us to define minimax risk:

Test risk of estimator under selection scheme S

$$R_*(S; \mathscr{K}_d) := \sup_{\mathsf{P} \in \mathscr{K}_d} \mathbb{E}_{\boldsymbol{y}, \boldsymbol{X} \sim \mathbb{P}(\mathsf{P})} R_\#(S; \boldsymbol{y}, \boldsymbol{X}),$$

$$R_{\rm MM}(\mathscr{K}_d) := \inf_{S \in \mathscr{A}} R_*(S; \mathscr{K}_d).$$

What is the optimal way to use the surrogate?



Theorem 5. *Assume that any* $\mathsf{P}_N \in \mathscr{K}_{d,N}$ *is supported on* $\|\boldsymbol{y}\| \le M$, *and that* $(\boldsymbol{y}, \boldsymbol{X}) \mapsto R(\hat{\boldsymbol{\theta}}_A(\boldsymbol{y}, \boldsymbol{X}))$ *is continuous for any* A. *Define*

$$\overline{R}_{\mathrm{MM}}(\mathscr{K}_d) := \inf_{S \in \overline{\mathscr{A}}} \overline{R}_*(S; \mathscr{K}_d) := \inf_{S \in \overline{\mathscr{A}}} \sup_{\mathsf{P}_N \in \mathscr{K}_{d,N}} \mathbb{E}_{\boldsymbol{y}, \boldsymbol{X} \sim \mathbb{P}(\mathsf{P}_N)} R_{\#}(S; \boldsymbol{y}, \boldsymbol{X}) \,. \tag{5.14}$$

Then we have

Sion’s minimax theorem

$$\overline{R}_{\mathrm{MM}}(\mathscr{K}_d) = \sup_{\mathsf{P}_N \in \mathscr{K}_{d,N}} \inf_{S \in \overline{\mathscr{A}}} \mathbb{E}_{\boldsymbol{y}, \boldsymbol{X} \sim \mathbb{P}(\mathsf{P}_N)} R_{\#}(S; \boldsymbol{y}, \boldsymbol{X}) \,. \tag{5.15}$$

Further, assume P_{MM} *achieves the supremum over* $\mathscr{K}_d$ *above. Then any*

$$S_{\mathrm{MM}} \in \arg\min_{S \in \overline{\mathscr{A}}} \mathbb{E}_{\boldsymbol{y}, \boldsymbol{X} \sim \mathbb{P}(\mathsf{P}_{\mathrm{MM}})} R_{\#}(S; \boldsymbol{y}, \boldsymbol{X}) \tag{5.16}$$

achieves the minimax error.

Intuition: we should use the “worst P(y|x)” that is near the given surrogate

High-dimensional asymptotics

$$\frac{n}{N} \to \gamma\,, \qquad \frac{N}{p} \to \delta_0\,,$$

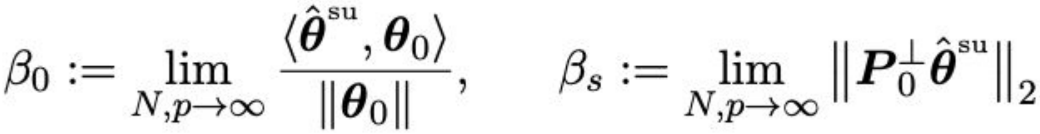
Specify to:

- Gaussian covariates (so x is drawn from isotropic Gaussian with dim p)
- Response dependent only on linear function of x: $\mathbb{P}(y_i \in A|\boldsymbol{x}_i) = \mathsf{P}(A|\langle \boldsymbol{\theta}_0, \boldsymbol{x}_i \rangle)$
- Generalized linear models + Ridge

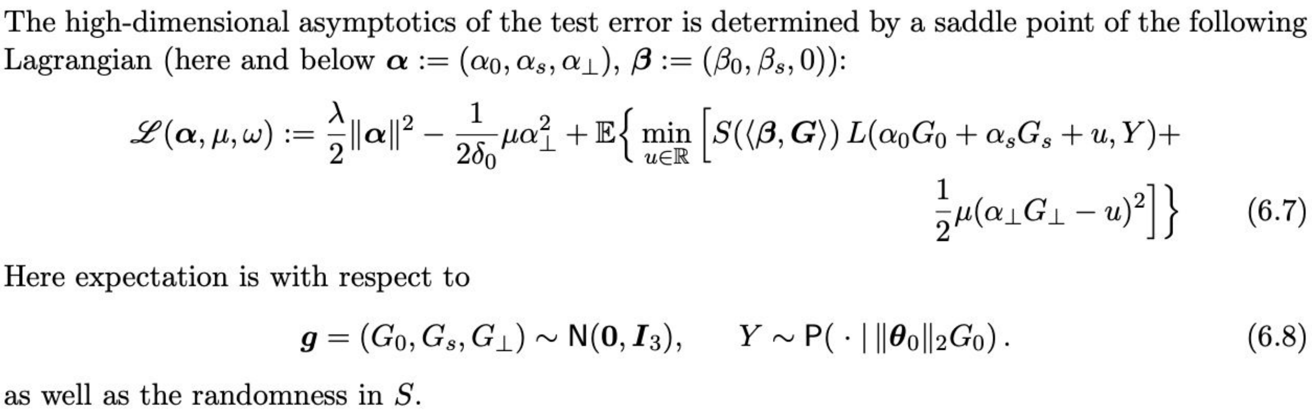
$$\hat{R}_N(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i=1}^{N} S_i(\langle \hat{\boldsymbol{\theta}}^{\mathrm{su}}, \boldsymbol{x}_i \rangle)\, L(\langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle, y_i) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2\,.$$

In the next slide, we will just ~~try to understand~~ show the theorem statement :)

Theorem statement: Setup



$$\beta_0 := \lim_{N,p\to\infty} \frac{\langle \hat{\boldsymbol{\theta}}^{\mathrm{su}}, \boldsymbol{\theta}_0 \rangle}{\|\boldsymbol{\theta}_0\|}, \qquad \beta_s := \lim_{N,p\to\infty} \left\| \boldsymbol{P}_0^{\perp} \hat{\boldsymbol{\theta}}^{\mathrm{su}} \right\|_2$$



The high-dimensional asymptotics of the test error is determined by a saddle point of the following Lagrangian (here and below $\boldsymbol{\alpha} := (\alpha_0, \alpha_s, \alpha_\perp)$, $\boldsymbol{\beta} := (\beta_0, \beta_s, 0)$):

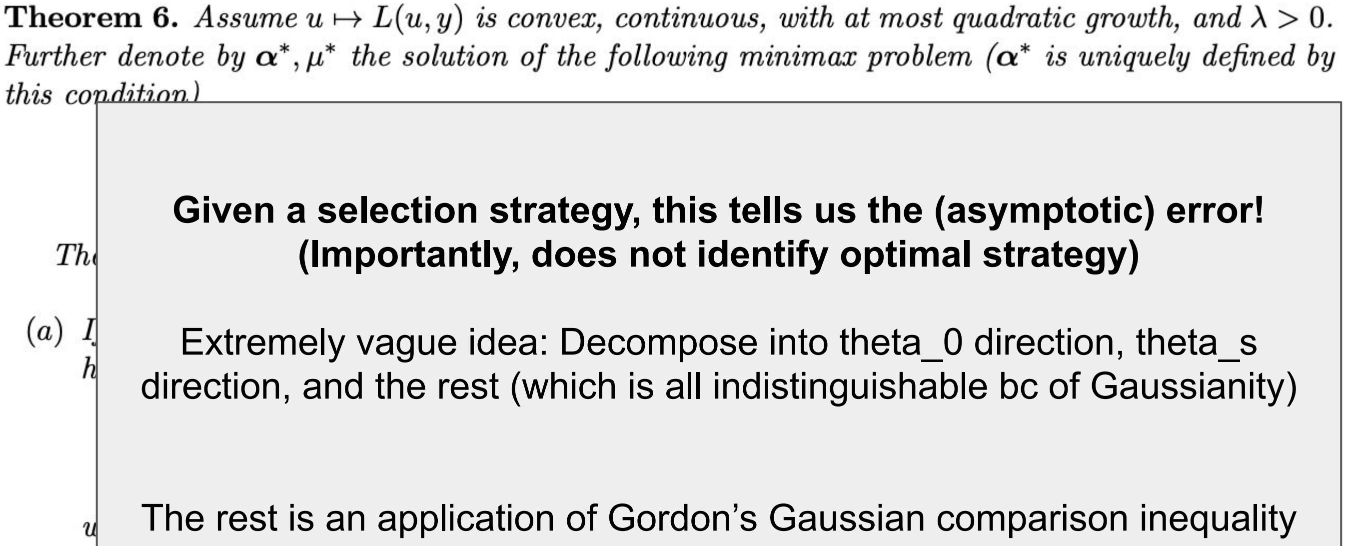
$$\mathscr{L}(\boldsymbol{\alpha}, \mu, \omega) := \frac{\lambda}{2}\|\boldsymbol{\alpha}\|^2 - \frac{1}{2\delta_0}\mu\alpha_\perp^2 + \mathbb{E}\Big\{ \min_{u\in\mathbb{R}} \Big[S(\langle \boldsymbol{\beta}, \boldsymbol{G}\rangle)\, L(\alpha_0 G_0 + \alpha_s G_s + u, Y) + \frac{1}{2}\mu(\alpha_\perp G_\perp - u)^2 \Big] \Big\} \tag{6.7}$$

Here expectation is with respect to

$$\boldsymbol{g} = (G_0, G_s, G_\perp) \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{I}_3), \qquad Y \sim \mathsf{P}(\,\cdot\,|\, \|\boldsymbol{\theta}_0\|_2 G_0)\,. \tag{6.8}$$

as well as the randomness in S.

Theorem statement: result



Theorem 6. *Assume* $u \mapsto L(u, y)$ *is convex, continuous, with at most quadratic growth, and* $\lambda > 0$. *Further denote by* $\boldsymbol{\alpha}^*, \mu^*$ *the solution of the following minimax problem (*$\boldsymbol{\alpha}^*$ *is uniquely defined by this condition)*

Th

(*a*) *I*

h

u

Given a selection strategy, this tells us the (asymptotic) error!
(Importantly, does not identify optimal strategy)

Extremely vague idea: Decompose into theta_0 direction, theta_s direction, and the rest (which is all indistinguishable bc of Gaussianity)

The rest is an application of Gordon’s Gaussian comparison inequality (generalization of Slepian’s inequality)

Special case: Gaussian x, binary y

Posit data selection mechanism:

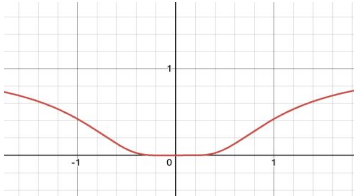
Normalizing constant

Second derivative of log-MGF (for logistic regression, 1 - tanh(t)^2)

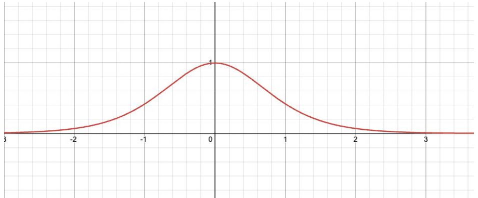
$$\pi(\boldsymbol{x}_i) = \min\left(c(\gamma)\,\phi''(\langle\hat{\boldsymbol{\theta}}^{\mathrm{su}}, \boldsymbol{x}_i\rangle)^{\alpha};\ 1\right)$$

For alpha = 1/2 , roughly equivalent to influence function-based sampling (only because the data is Gaussian and so the Hessian has a simple closed form)

Alpha controls whether we prefer hard examples or easy:



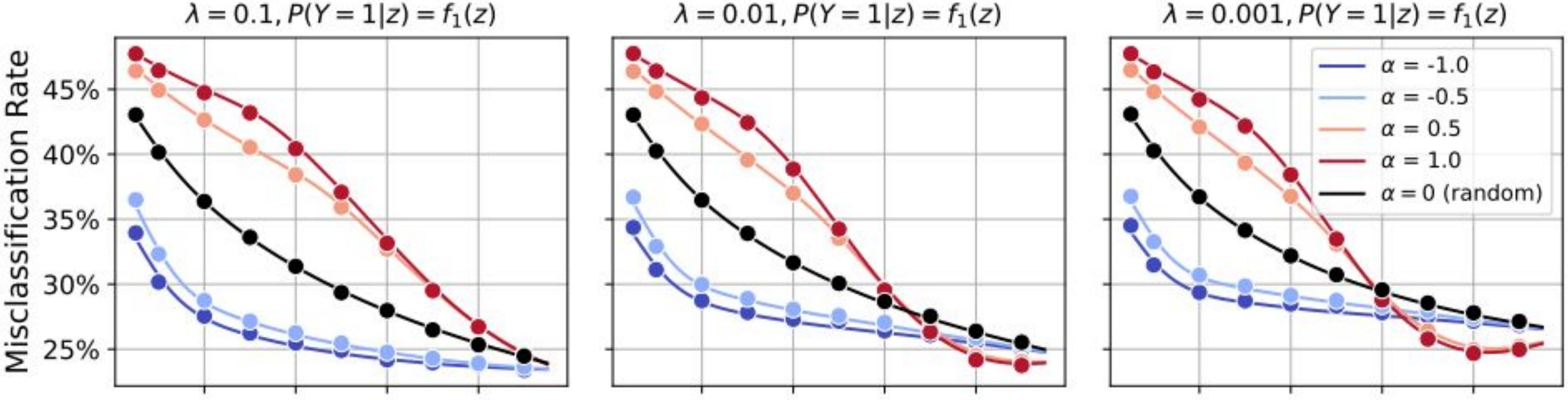
Alpha = -1 (prefer easy)



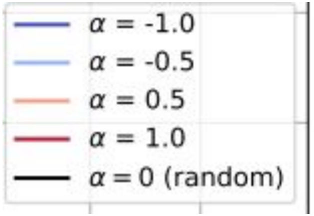
Alpha = 1 (prefer hard)

Results: perfect surrogate

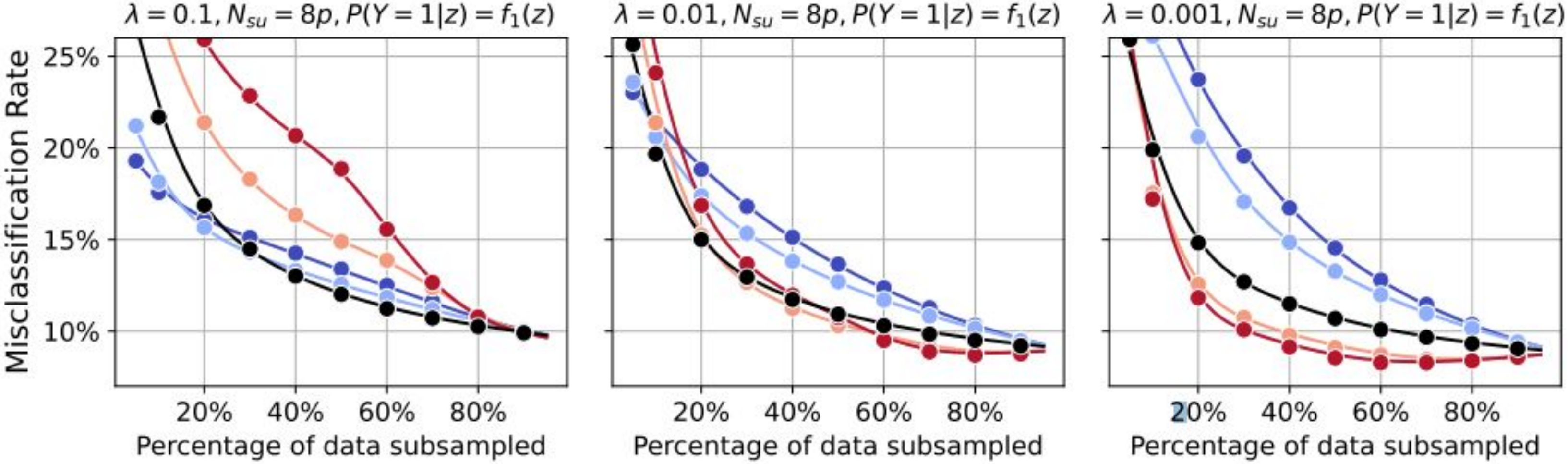
Setting: “misspecified” linear model, so P(y=1|x) = f_1(x^T \theta)



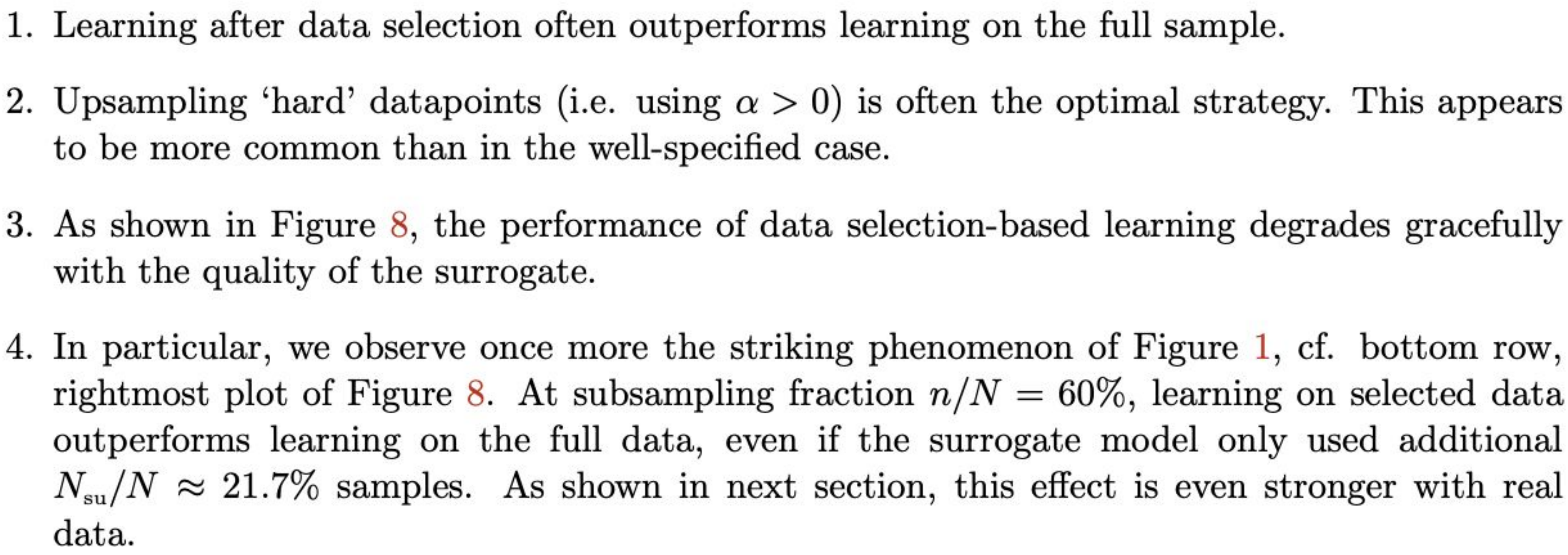
Results: imperfect surrogate



Surrogate is trained with additional N_{su} samples



Takeaways (imperfect surrogate)



1. Learning after data selection often outperforms learning on the full sample.
2. Upsampling 'hard' datapoints (i.e. using $\alpha > 0$) is often the optimal strategy. This appears to be more common than in the well-specified case.
3. As shown in Figure 8, the performance of data selection-based learning degrades gracefully with the quality of the surrogate.
4. In particular, we observe once more the striking phenomenon of Figure 1, cf. bottom row, rightmost plot of Figure 8. At subsampling fraction $n/N = 60\%$, learning on selected data outperforms learning on the full data, even if the surrogate model only used additional $N_{\text{su}}/N \approx 21.7\%$ samples. As shown in next section, this effect is even stronger with real data.

Experimental Verification