**OSMAC**

# General Subsampling Algorithm

**Algorithm**

1. Sampling: size = r (<<n), with distribution . Subsample: and .
2. Estimation on subsamples: max weighted () log-likelihood🡪

**Asymptotic Properties:**

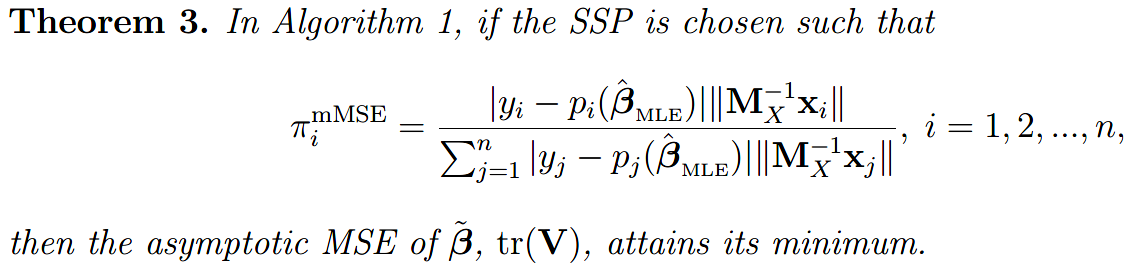
1. Consistency: likelihood log-concave (global max), is consistent to (conditional)
2. Asymptotic normal:

# Optimal Subsampling Strategies

**Min AMSE of <min tr(V), A-optimal >**

1. 🡪.





1. depends on covariates & responses
   1. Covariates: larger for larger 
   2. Response: prefer dpts more likely to be mis-classified (more information)
2. Logistic regression: separable of p 🡪 no MLE

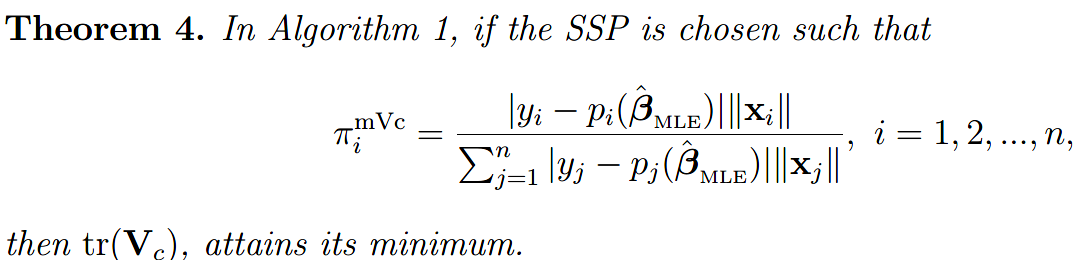
Subsampling SSP: increase overlap of p🡪 more likely to have MLE

**Min AMSE of <min tr(Vc), L-optimal, faster>**

1. Motivation:



Min tr(Vc)



1. Computational issue: for min; for min

# Two-Step Algorithm

**Algorithm (replace full data MLE by pilot estimation of MLE):**

1. Pilot subsample & estimation: size = . <empirically: 0.2\*(r0 + r)>

Distribution:

Balanced data: uniform

imbalanced: uniform proportion to number of y=1/ 0 <case-control sampling>

Estimate full MLE by r0 subsamples:

Replace   with , get SSP

1. Subsample: sample r by SSP from (1). Combine r + r0 subsamples to get estimation

**Asymptotic Properties** (assume 1) A-COV full rank & 2) COV distribution light tail):

1. Asymptotic consistent to
2. Asymptotic normal

**Standard error formula:**

use for CI and hypothesis test

Estimation is not good for rare event data (improve)