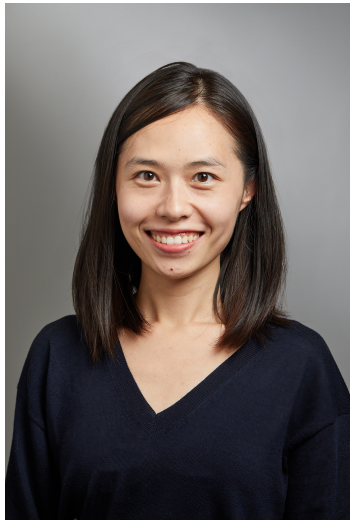
Gradient flows for empirical Bayes in high-dimensional linear models

Zhou Fan
Department of Statistics and Data Science, Yale University

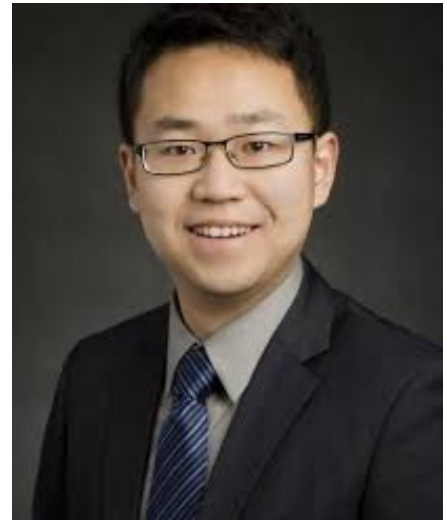
Storrs, CT
July 16, 2024

Joint work with



Leying Guan

Yandi Shen



Yihong Wu

Empirical Bayes in the sequence model

Gaussian sequence model

$$\mathbf{y} = \boldsymbol{\theta} + \boldsymbol{\varepsilon} \in \mathbb{R}^p$$

with parameters $\theta_j \overset{iid}{\sim} g_*$, noise $\varepsilon_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

Empirical Bayes in the sequence model

Gaussian sequence model

$$\mathbf{y} = \boldsymbol{\theta} + \boldsymbol{\varepsilon} \in \mathbb{R}^p$$

with parameters $\theta_j \overset{iid}{\sim} g_*$, noise $\varepsilon_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

Empirical Bayes: Estimate the prior g_* from $\mathbf{y}$, then use the estimated prior for Bayesian inference. [Robbins '50]

Empirical Bayes in the sequence model

Gaussian sequence model

$$\mathbf{y} = \boldsymbol{\theta} + \boldsymbol{\varepsilon} \in \mathbb{R}^p$$

with parameters $\theta_j \overset{iid}{\sim} g_*$, noise $\varepsilon_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

Empirical Bayes: Estimate the prior g_* from $\mathbf{y}$, then use the estimated prior for Bayesian inference. [Robbins '50]

70+ years of literature [Stein '55, Kiefer, Wolfowitz '56, Laird '78, Lindsay '83, Groeneboom, Wellner '92, Ghosal, van der Vaart '01, Zhang '09, ...]

Lindsay 1995, *Mixture models: Theory, geometry, and applications*
Efron 2010, *Large-scale inference*

Empirical Bayes in the linear model

Linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \in \mathbb{R}^n, \qquad \mathbf{X} \in \mathbb{R}^{n \times p}$$

with random effects $\theta_j \overset{iid}{\sim} g_*$, residual errors $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

Empirical Bayes in the linear model

Linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \in \mathbb{R}^n, \qquad \mathbf{X} \in \mathbb{R}^{n \times p}$$

with random effects $\theta_j \overset{iid}{\sim} g_*$, residual errors $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

Goal: Estimate the effect size prior g_* from $(\mathbf{X}, \mathbf{y})$ when it is unknown.

1.) When is accurate estimation of g_* statistically possible?
2.) How can we design fast algorithms to perform this estimation?

Motivation from statistical genetics

Bayesian linear models underlie most of the commonly-used methods for genetic analyses of complex (polygenic) traits, where

$\mathbf{y}$ = trait of interest (e.g. human height), $\mathbf{X}$ = genotypes at p variants

- Association testing [EMMAX (Kang et al. '10), BOLT-LMM (Loh et al. '15)]
- Genetic fine-mapping [FINEMAP (Benner et al. '16), SuSiE (Wang et al. '20)]
- Heritability analyses [GCTA (Yang et al. '11)]
- Phenotype prediction [SBayesR (Lloyd-Jones et al. '19), PRS-CS (Ge et al. '19)]

Motivation from statistical genetics

Bayesian linear models underlie most of the commonly-used methods for genetic analyses of complex (polygenic) traits, where

$\mathbf{y}$ = trait of interest (e.g. human height), $\mathbf{X}$ = genotypes at p variants

- Association testing [EMMAX (Kang et al. '10), BOLT-LMM (Loh et al. '15)]
- Genetic fine-mapping [FINEMAP (Benner et al. '16), SuSiE (Wang et al. '20)]
- Heritability analyses [GCTA (Yang et al. '11)]
- Phenotype prediction [SBayesR (Lloyd-Jones et al. '19), PRS-CS (Ge et al. '19)]

In these contexts, g_* represents the genetic architecture of the trait.

Many methods for estimating g_* have been developed by the genetics community [Zhang et al. '18, O'Connor '21, Spence et al. '22, Morgante et al. '23].

Correlated vs. "mean-field" designs

For mean-field designs (e.g. $x_{ij} \overset{iid}{\sim} \mathcal{N}(0, \frac{1}{n})$) many approaches can work:

- Direct Gaussian deconvolution of the z-scores $\mathbf{z} = \mathbf{X}^\top \mathbf{y}$

Correlated vs. "mean-field" designs

For mean-field designs (e.g. $x_{ij} \overset{iid}{\sim} \mathcal{N}(0, \frac{1}{n})$) many approaches can work:

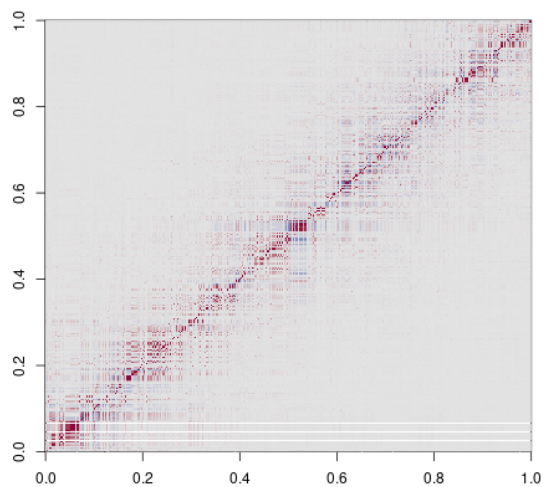
- Direct Gaussian deconvolution of the z-scores $\mathbf{z} = \mathbf{X}^\top \mathbf{y}$
- Variational inference using mean-field VB [Mukherjee et al. '23] or TAP approximation/AMP [Ahn, Lin, Mei '23], [Celentano, Fan, Lin, Mei '23]

Correlated vs. "mean-field" designs

For mean-field designs (e.g. $x_{ij} \overset{iid}{\sim} \mathcal{N}(0, \frac{1}{n})$) many approaches can work:

- Direct Gaussian deconvolution of the z-scores $\mathbf{z} = \mathbf{X}^\top \mathbf{y}$
- Variational inference using mean-field VB [Mukherjee et al. '23] or TAP approximation/AMP [Ahn, Lin, Mei '23], [Celentano, Fan, Lin, Mei '23]

Corresponds to estimating the (marginal) effect size distribution for sparsely sampled genetic variants.



This is different from estimating the conditional/causal effect size distribution in strongly correlated designs.

Outline

Gradient flow computation of the NPMLE

Theoretical results

- Statistical consistency of the NPMLE
- Mixing of Langevin dynamics
- Convergence of the joint gradient flow

Discussion

Outline

Nonparametric maximum likelihood

In the sequence model $\mathbf{y} = \boldsymbol{\theta} + \varepsilon$, the NPMLE $\hat{g}$ is the minimizer of

$$\bar{F}(g) = -\frac{1}{p}\sum_{j=1}^{p} \log[\mathcal{N}_\sigma * g](y_j)$$

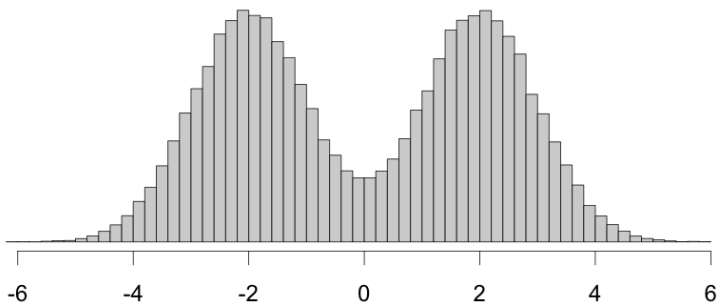
where $\mathcal{N}_\sigma \equiv \mathcal{N}(0, \sigma^2)$. [Robbins '50, Kiefer, Wolfowitz '56]

Nonparametric maximum likelihood

In the sequence model $\mathbf{y} = \boldsymbol{\theta} + \varepsilon$, the NPMLE $\hat{g}$ is the minimizer of

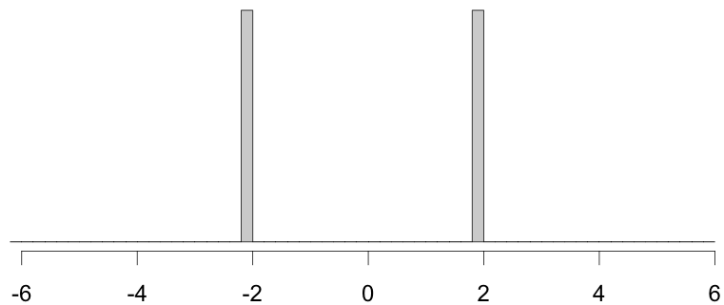
$$\bar{F}(g) = -\frac{1}{p}\sum_{j=1}^{p} \log[\mathcal{N}_\sigma * g](y_j)$$

where $\mathcal{N}_\sigma \equiv \mathcal{N}(0, \sigma^2)$. [Robbins '50, Kiefer, Wolfowitz '56]



Observed distribution of $\{y_j\}_{j=1}^p$

$\Rightarrow$



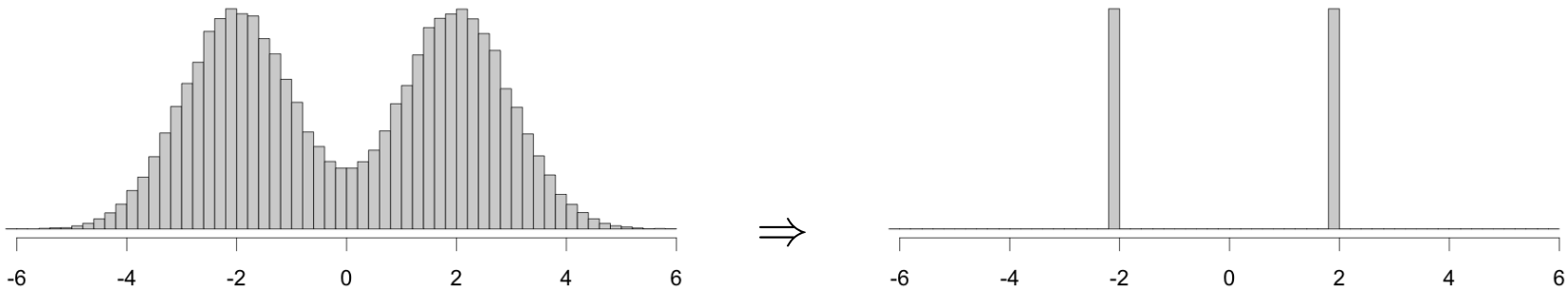
Inferred distribution of $\{\theta_j\}_{j=1}^p$

Nonparametric maximum likelihood

In the sequence model $\mathbf{y} = \boldsymbol{\theta} + \varepsilon$, the NPMLE $\hat{g}$ is the minimizer of

$$\bar{F}(g) = -\frac{1}{p}\sum_{j=1}^{p} \log[\mathcal{N}_\sigma * g](y_j)$$

where $\mathcal{N}_\sigma \equiv \mathcal{N}(0, \sigma^2)$. [Robbins '50, Kiefer, Wolfowitz '56]



Observed distribution of $\{y_j\}_{j=1}^p$ Inferred distribution of $\{\theta_j\}_{j=1}^p$

The optimization is convex (over probability measures) and efficiently solved by interior point methods [Koenker, Mizera '14]. Gradient flows for optimization were studied by [Yan, Wang, Rigollet '23].

Nonparametric maximum likelihood

In the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, the NPMLE $\hat{g}$ minimizes

$$\bar{F}(g) = -\frac{1}{p}\log \int \underbrace{\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2\right)}_{P(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})} \prod_{j=1}^{p} g(\theta_j)\mathrm{d}\theta_j$$

Nonparametric maximum likelihood

In the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \varepsilon$, the NPMLE $\hat{g}$ minimizes

$$\bar{F}(g) = -\frac{1}{p} \log \int \underbrace{\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2\right)}_{P(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})} \prod_{j=1}^{p} g(\theta_j) \mathrm{d}\theta_j$$

For general designs $\mathbf{X}$:

- The optimization is non-convex
- The marginal log-likelihood $\bar{F}(g)$ is not coordinate-separable
- Neither $\bar{F}(g)$ nor its gradient is explicitly computable

Existing methods in the linear model

Two common approaches used in practice:

- Monte Carlo Expectation-Maximization (MCEM)
  - **E-step:** Compute $Q(g) = \mathbb{E}_{\boldsymbol{\theta} \sim P(\cdot \mid g^{(t)})}[\log P(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}) + \sum_{j=1}^{p} \log g(\theta_j)]$.
  - **M-step:** Solve $g^{(t+1)} = \arg\max Q(g)$

  $\mathbb{E}_{\boldsymbol{\theta} \sim P(\cdot \mid g^{(t)})}[\cdot]$ is approximated by MCMC. [Wei, Tanner '90]

Existing methods in the linear model

Two common approaches used in practice:

- Monte Carlo Expectation-Maximization (MCEM)
  - **E-step:** Compute $Q(g) = \mathbb{E}_{\boldsymbol{\theta} \sim P(\cdot \mid g^{(t)})}[\log P(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}) + \sum_{j=1}^{p} \log g(\theta_j)]$.
  - **M-step:** Solve $g^{(t+1)} = \arg\max Q(g)$

  $\mathbb{E}_{\boldsymbol{\theta} \sim P(\cdot \mid g^{(t)})}[\cdot]$ is approximated by MCMC. [Wei, Tanner '90]

  Variations that put a hyperprior on g_* and perform Bayesian sampling of g_* are also common. [Lloyd-Jones et al. '19, Zhou, Zhao '21]

  Costly to run MCMC at every E-step, and hard to assess convergence.

Existing methods in the linear model

Two common approaches used in practice:

- Monte Carlo Expectation-Maximization (MCEM)
  - **E-step:** Compute $Q(g) = \mathbb{E}_{\boldsymbol{\theta} \sim P(\cdot \mid g^{(t)})}[\log P(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}) + \sum_{j=1}^{p} \log g(\theta_j)]$.
  - **M-step:** Solve $g^{(t+1)} = \arg\max Q(g)$

  $\mathbb{E}_{\boldsymbol{\theta} \sim P(\cdot \mid g^{(t)})}[\cdot]$ is approximated by MCMC. [Wei, Tanner '90]

  Variations that put a hyperprior on g_* and perform Bayesian sampling of g_* are also common. [Lloyd-Jones et al. '19, Zhou, Zhao '21]

  Costly to run MCMC at every E-step, and hard to assess convergence.
- Mean-field variational inference: Minimize a variational upper bound of $\bar{F}(g)$ using a product law approximation for the posterior $P(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y})$. [Bernardo et al. '03, Spence et al. '22, Morgante et al. '23]

Existing methods in the linear model

Two common approaches used in practice:

- Monte Carlo Expectation-Maximization (MCEM)
  - **E-step:** Compute $Q(g) = \mathbb{E}_{\boldsymbol{\theta} \sim P(\cdot \mid g^{(t)})}[\log P(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}) + \sum_{j=1}^{p} \log g(\theta_j)]$.
  - **M-step:** Solve $g^{(t+1)} = \arg\max Q(g)$

  $\mathbb{E}_{\boldsymbol{\theta} \sim P(\cdot \mid g^{(t)})}[\cdot]$ is approximated by MCMC. [Wei, Tanner '90]

  Variations that put a hyperprior on g_* and perform Bayesian sampling of g_* are also common. [Lloyd-Jones et al. '19, Zhou, Zhao '21]

  Costly to run MCMC at every E-step, and hard to assess convergence.

- Mean-field variational inference: Minimize a variational upper bound of $\bar{F}(g)$ using a product law approximation for the posterior $P(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y})$. [Bernardo et al. '03, Spence et al. '22, Morgante et al. '23]

  This approximation is inaccurate when $\mathbf{X}$ has correlated variables.

Reparametrization, variational form, and gradient flow

1. Reparametrize the linear model using smoothed regression coefficients

$$\mathbf{y} = \mathbf{X}\boldsymbol{\varphi} + \tilde{\varepsilon}, \qquad \varphi_1, \ldots, \varphi_p \overset{iid}{\sim} \mathcal{N}_\tau * g_*, \qquad \tilde{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I} - \tau^2 \mathbf{X}\mathbf{X}^\top)$$

   Inspired by a proof of LSI in [Bauerschmidt, Bodineau '18].

Reparametrization, variational form, and gradient flow

1. Reparametrize the linear model using smoothed regression coefficients

$$\mathbf{y} = \mathbf{X}\boldsymbol{\varphi} + \tilde{\varepsilon}, \qquad \varphi_1, \dots, \varphi_p \overset{iid}{\sim} \mathcal{N}_\tau * g_*, \qquad \tilde{\varepsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I} - \tau^2\mathbf{X}\mathbf{X}^\top)$$

   Inspired by a proof of LSI in [Bauerschmidt, Bodineau '18].

2. Write the Gibbs variational representation

$$\bar{F}(g) = \min_{q \in \mathcal{P}(\mathbb{R}^p)} F(q, g) \text{ where}$$

$$F(q,g) = -\frac{1}{p} \int \Bigg[\underbrace{\log P(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\varphi}) + \sum_{j=1}^{p} \log[\mathcal{N}_\tau * g](\varphi_j)}_{:= U_g(\boldsymbol{\varphi})} - \log q(\boldsymbol{\varphi}) \Bigg] \mathrm{d}q(\boldsymbol{\varphi})$$

   Here, g is the (1-dimensional) entrywise prior for $\boldsymbol{\theta}$, and q is the (p-dimensional) posterior for $\boldsymbol{\varphi} = \boldsymbol{\theta} + \mathcal{N}(0, \tau^2\mathbf{I})$.

Reparametrization, variational form, and gradient flow

1. Reparametrize the linear model using smoothed regression coefficients

$$\mathbf{y} = \mathbf{X}\boldsymbol{\varphi} + \tilde{\varepsilon}, \qquad \varphi_1, \dots, \varphi_p \overset{iid}{\sim} \mathcal{N}_\tau * g_*, \qquad \tilde{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I} - \tau^2 \mathbf{X}\mathbf{X}^\top)$$

   Inspired by a proof of LSI in [Bauerschmidt, Bodineau '18].

2. Write the Gibbs variational representation

$$\bar{F}(g) = \min_{q \in \mathcal{P}(\mathbb{R}^p)} F(q, g) \text{ where}$$

$$F(q,g) = -\frac{1}{p} \int \Bigg[\underbrace{\log P(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\varphi}) + \sum_{j=1}^{p} \log[\mathcal{N}_\tau * g](\varphi_j)}_{:= U_g(\boldsymbol{\varphi})} - \log q(\boldsymbol{\varphi}) \Bigg] \mathrm{d}q(\boldsymbol{\varphi})$$

   Here, g is the (1-dimensional) entrywise prior for $\boldsymbol{\theta}$, and q is the (p-dimensional) posterior for $\boldsymbol{\varphi} = \boldsymbol{\theta} + \mathcal{N}(0, \tau^2 \mathbf{I})$.

3. Optimize $F(q, g)$ jointly over (q, g) via the gradient flows

$$\partial_t q_t = -p \cdot \operatorname{grad}_q^{W_2} F(q_t, g_t), \qquad \partial_t g_t = -\operatorname{grad}_g^{\mathrm{FR}} F(q_t, g_t)$$

Wasserstein-2 gradient flow in q

The gradient flow equation in the posterior $q(\varphi)$ is

$$\begin{aligned}\partial_t q_t(\varphi) &= -p \cdot \operatorname{grad}_q^{W_2} F(q_t, g_t) \\ &= -\nabla \cdot \Big[q_t(\varphi) \nabla U_{g_t}(\varphi) \Big] + \Delta q_t(\varphi)\end{aligned}$$

This is the Fokker-Planck equation for the density evolution under a Langevin diffusion

$$\mathrm{d}\varphi_t = \nabla U_{g_t}(\varphi)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{B}_t$$

with time-evolving drift. [Jordan, Kinderlehrer, Otto '98]

Wasserstein-2 gradient flow in q

The gradient flow equation in the posterior $q(\varphi)$ is

$$\begin{aligned}\partial_t q_t(\varphi) &= -p \cdot \operatorname{grad}_q^{W_2} F(q_t, g_t) \\ &= -\nabla \cdot \Big[q_t(\varphi) \nabla U_{g_t}(\varphi) \Big] + \Delta q_t(\varphi)\end{aligned}$$

This is the Fokker-Planck equation for the density evolution under a Langevin diffusion

$$\mathrm{d}\varphi_t = \nabla U_{g_t}(\varphi)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{B}_t$$

with time-evolving drift. [Jordan, Kinderlehrer, Otto '98]

In our implementations we use a forward Euler discretization in time, leading to an Unadjusted Langevin Algorithm [Roberts, Tweedie '96].

Fisher-Rao gradient flow in g

The gradient flow equation in the prior $g(\theta)$ is

$$\partial_t g_t(\theta) = -\operatorname{grad}_g^{\mathrm{FR}} F(q_t, g_t) = g_t(\theta)\Big(\Big[\mathcal{N}_\tau * \frac{\bar{q}_t}{\mathcal{N}_\tau * g_t}\Big](\theta) - 1\Big)$$

where $\bar{q}_t$ is the average marginal distribution of $\varphi_1, \ldots, \varphi_p$ under q_t.

Fisher-Rao gradient flow in g

The gradient flow equation in the prior $g(\theta)$ is

$$\partial_t g_t(\theta) = -\operatorname{grad}_g^{\mathrm{FR}} F(q_t, g_t) = g_t(\theta)\Big(\Big[\mathcal{N}_\tau * \frac{\bar{q}_t}{\mathcal{N}_\tau * g_t}\Big](\theta) - 1\Big)$$

where $\bar{q}_t$ is the average marginal distribution of $\varphi_1, \ldots, \varphi_p$ under q_t.

We estimate $\bar{q}_t$ using the empirical distribution of coordinates of the current Langevin iterate, $\bar{q}_t \approx \frac{1}{p}\sum_{j=1}^p \delta_{\varphi_{t,j}}$,

$$\Big[\mathcal{N}_\tau * \frac{\bar{q}_t}{\mathcal{N}_\tau * g_t}\Big](\theta) \approx \frac{1}{p}\sum_{j=1}^p \frac{\mathcal{N}_\tau(\theta - \varphi_{t,j})}{\mathcal{N}_\tau * g_t(\varphi_{t,j})}$$

Fisher-Rao gradient flow in g

The gradient flow equation in the prior $g(\theta)$ is

$$\partial_t g_t(\theta) = -\operatorname{grad}_g^{\mathrm{FR}} F(q_t, g_t) = g_t(\theta)\Big(\Big[\mathcal{N}_\tau * \frac{\bar{q}_t}{\mathcal{N}_\tau * g_t}\Big](\theta) - 1\Big)$$

where $\bar{q}_t$ is the average marginal distribution of $\varphi_1, \ldots, \varphi_p$ under q_t.

We estimate $\bar{q}_t$ using the empirical distribution of coordinates of the current Langevin iterate, $\bar{q}_t \approx \frac{1}{p}\sum_{j=1}^p \delta_{\varphi_{t,j}}$,

$$\Big[\mathcal{N}_\tau * \frac{\bar{q}_t}{\mathcal{N}_\tau * g_t}\Big](\theta) \approx \frac{1}{p}\sum_{j=1}^p \frac{\mathcal{N}_\tau(\theta - \varphi_{t,j})}{\mathcal{N}_\tau * g_t(\varphi_{t,j})}$$

For fixed $\boldsymbol{\varphi} \equiv \boldsymbol{\varphi}_t$, the resulting gradient flow has the simple interpretation of a Fisher-Rao flow on the log-likelihood of the sequence model

$$\boldsymbol{\varphi} = \boldsymbol{\theta} + \mathcal{N}(0, \tau^2 \mathbf{I}), \qquad \theta_1, \ldots, \theta_p \overset{iid}{\sim} g_*$$

EBflow

To summarize, EBflow is a time-discretization of

$$\mathrm{d}\boldsymbol{\varphi}_t = \nabla U_{g_t}(\boldsymbol{\varphi}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{B}_t \tag{1}$$

$$\partial_t g_t(\theta) = g_t(\theta) \cdot \left(\frac{1}{p} \sum_{j=1}^{p} \frac{\mathcal{N}_\tau(\theta - \varphi_{t,j})}{\mathcal{N}_\tau * g_t(\varphi_{t,j})} - 1 \right) \tag{2}$$

where

- $\boldsymbol{\varphi}_t$ follows the Langevin diffusion (1) for the posterior law of $\boldsymbol{\varphi} = \boldsymbol{\theta} + \mathcal{N}(0, \tau^2)$ with time-evolving prior $\mathcal{N}_\tau * g_t$.
- g_t follows (2) which is a gradient flow for computing the NPMLE from the coordinates of $\boldsymbol{\varphi}_t$.

EBflow

To summarize, EBflow is a time-discretization of

$$\mathrm{d}\boldsymbol{\varphi}_t = \nabla U_{g_t}(\boldsymbol{\varphi}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{B}_t \tag{1}$$

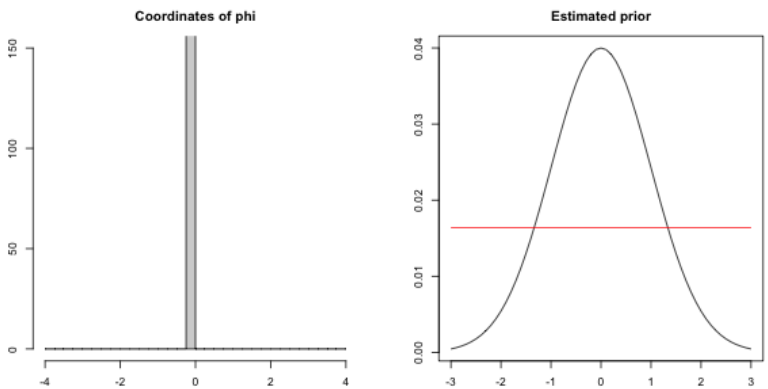
$$\partial_t g_t(\theta) = g_t(\theta) \cdot \left(\frac{1}{p} \sum_{j=1}^{p} \frac{\mathcal{N}_\tau(\theta - \varphi_{t,j})}{\mathcal{N}_\tau * g_t(\varphi_{t,j})} - 1 \right) \tag{2}$$

where

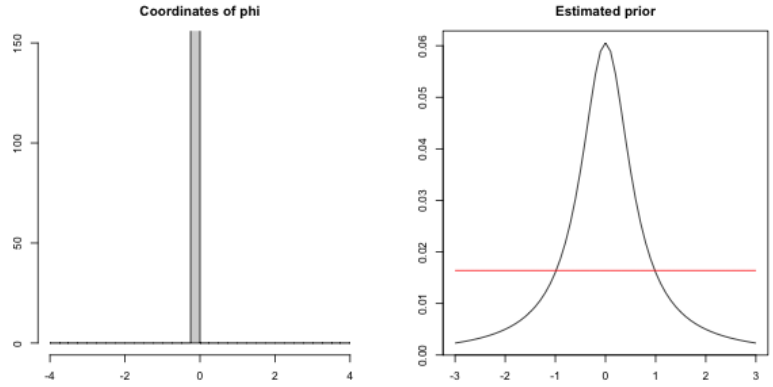
- $\boldsymbol{\varphi}_t$ follows the Langevin diffusion (1) for the posterior law of $\boldsymbol{\varphi} = \boldsymbol{\theta} + \mathcal{N}(0, \tau^2)$ with time-evolving prior $\mathcal{N}_\tau * g_t$.
- g_t follows (2) which is a gradient flow for computing the NPMLE from the coordinates of $\boldsymbol{\varphi}_t$.

A similar idea of jointly optimizing the Gibbs variational objective was proposed in [Kuntz, Lim, Johansen '23] for general latent variable models, focusing on "low-dimensional" problems with parametric priors.

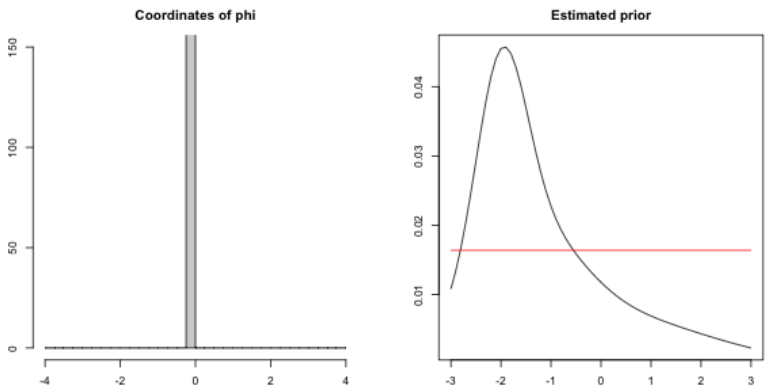
Illustrations for various priors



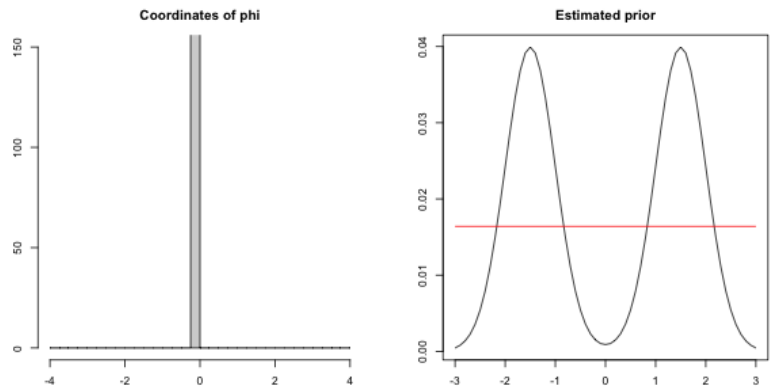
Gaussian



Cauchy



Skew



Bimodal

$n = 1000$, $p = 500$ variables in pairs w/ 90%-correlation.

EBflow was performed with a small spline smoothing regularizer, pre-conditioning of the Langevin diffusion, and log-linear step size decay.

Relation to mean-field VI and EM

Empirical Bayes within mean-field VI [Bernardo et al. '03] solves

$$(\hat{q}, \hat{g}) = \underset{q \in \mathcal{Q},\, g \in \mathcal{P}(\mathbb{R})}{\arg\min} \; F(q, g), \quad \mathcal{Q} = \{\text{product measures on } \mathbb{R}^p\}$$

Relation to mean-field VI and EM

Empirical Bayes within mean-field VI [Bernardo et al. '03] solves

$$(\hat{q}, \hat{g}) = \underset{q \in \mathcal{Q},\, g \in \mathcal{P}(\mathbb{R})}{\arg\min} \; F(q, g), \quad \mathcal{Q} = \{\text{product measures on } \mathbb{R}^p\}$$

EM performs coordinate descent [Neal, Hinton '98]

$$q^{(t)} = \underset{q \in \mathcal{P}(\mathbb{R}^p)}{\arg\min}\, F(q, g^{(t)}), \qquad g^{(t+1)} = \underset{g \in \mathcal{P}(\mathbb{R})}{\arg\min}\, F(q^{(t)}, g)$$

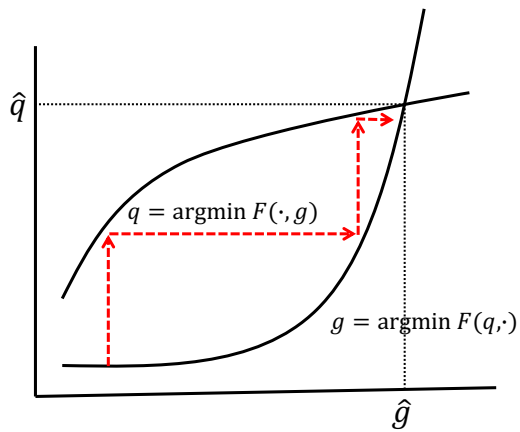
Relation to mean-field VI and EM

Empirical Bayes within mean-field VI [Bernardo et al. '03] solves

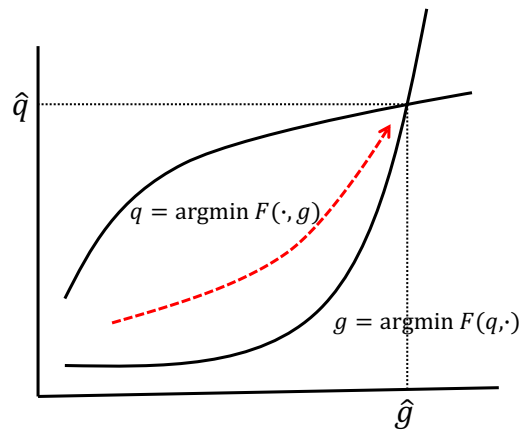
$$(\hat{q}, \hat{g}) = \underset{q \in \mathcal{Q},\, g \in \mathcal{P}(\mathbb{R})}{\arg\min} \; F(q, g), \quad \mathcal{Q} = \{\text{product measures on } \mathbb{R}^p\}$$

EM performs coordinate descent [Neal, Hinton '98]

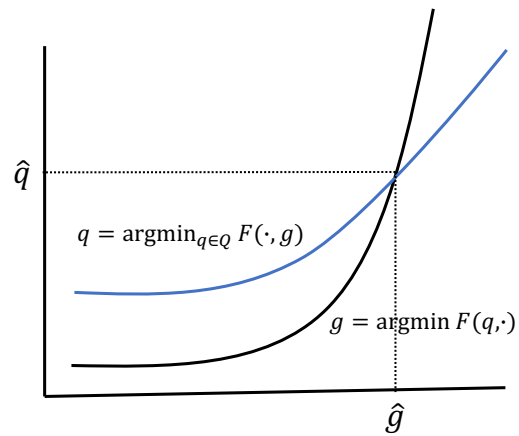
$$q^{(t)} = \underset{q \in \mathcal{P}(\mathbb{R}^p)}{\arg\min}\, F(q, g^{(t)}), \qquad g^{(t+1)} = \underset{g \in \mathcal{P}(\mathbb{R})}{\arg\min}\, F(q^{(t)}, g)$$



EM



EBflow



VI

Outline

Outline

NPMLE consistency when $n \geq p$

Linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \in \mathbb{R}^n, \qquad \mathbf{X} \in \mathbb{R}^{n \times p}, \qquad \theta_1, \ldots, \theta_p \overset{iid}{\sim} g_*$$

When is accurate estimation of g_* statistically possible?

NPMLE consistency when $n \geq p$

Linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \in \mathbb{R}^n, \qquad \mathbf{X} \in \mathbb{R}^{n \times p}, \qquad \theta_1, \ldots, \theta_p \overset{iid}{\sim} g_*$$

When is accurate estimation of g_* statistically possible?

If $n \geq p$ and $\mathbf{X}$ has full column rank, the model is equivalent to observing

$$\hat{\boldsymbol{\theta}}_{\mathrm{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \boldsymbol{\theta} + \underbrace{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon}}_{\text{colored Gaussian noise}}$$

Theorem (Mukherjee, Sen, Sen '23)

Suppose $g_* \in \mathcal{P}([-M, M])$, and $\bar{F}(\hat{g}) \leq \bar{F}(g_*) + o_{\mathbb{P}}(1)$. As $n, p \to \infty$, if $\mathbf{X}$ has full column rank and $\lambda_{\min}(\mathbf{X}^\top \mathbf{X}) \asymp \lambda_{\max}(\mathbf{X}^\top \mathbf{X}) \asymp 1$, then

$$\hat{g} \xrightarrow{\mathbb{P}} g_* \text{ weakly.}$$

NPMLE consistency when $n < p$

If $n < p$, then $\mathbf{y}$ depends on $\boldsymbol{\theta}$ only via its projection onto $\operatorname{row\,span}(\mathbf{X})$.

When is g_* (stably) recoverable from the projection of $g_*^{\otimes p}$ onto a subspace of $\mathbb{R}^p$ of dimension $n < p$?

NPMLE consistency when $n < p$

If $n < p$, then $\mathbf{y}$ depends on $\boldsymbol{\theta}$ only via its projection onto $\operatorname{row\,span}(\mathbf{X})$.

When is g_* (stably) recoverable from the projection of $g_*^{\otimes p}$ onto a subspace of $\mathbb{R}^p$ of dimension $n < p$?

Assumption

Suppose $n, p \to \infty$ with $\lambda_{\max}(\mathbf{X}^\top \mathbf{X}) \lesssim 1$, $\|\mathbf{X}\frac{\mathbf{1}}{\sqrt{p}}\|_2^2 \gtrsim 1$, and for a subset of variables $\mathcal{S}$ of cardinality $|\mathcal{S}| \asymp p$ and each $j \in \mathcal{S}$, there exists a test vector $\mathbf{z}_j \in \mathbb{R}^n$ such that

$$\langle \mathbf{z}_j, \mathbf{x}_j \rangle \to 1, \qquad \sum_{k \neq j} |\langle \mathbf{z}_j, \mathbf{x}_k \rangle|^{2+\varepsilon} \to 0.$$

[Intuition: Then by the CLT, $\mathbf{z}_j^\top \mathbf{y} \approx \theta_j +$ Gaussian noise.]

Condition is reminiscent of the debiased lasso in [Zhang, Zhang '14]. However, here $\mathbf{z}_j$ does *not* need to be estimable from the data $(\mathbf{X}, \mathbf{y})$.

NPMLE consistency when $n < p$

Theorem (F., Guan, Shen, Wu)

Suppose $g_* \in \mathcal{P}([-M, M])$, and $\bar{F}(\hat{g}) \le \bar{F}(g_*) + o_{\mathbb{P}}(1)$. Under the preceding assumption, as $n, p \to \infty$,

$$\hat{g} \xrightarrow{\mathbb{P}} g_* \text{ weakly.}$$

Example (random design)

Suppose $n, p \to \infty$ with

- $n/p \ge \gamma$ any positive constant
- $\mathbf{X}$ has i.i.d. subgaussian rows $\{\frac{1}{\sqrt{n}} x^{(i)}\}$ where $\mathbb{E} x^{(i)} = 0$, $\operatorname{Cov} x^{(i)} = \mathbf{\Sigma}$,

$$\lambda_{\min}(\mathbf{\Sigma}) \asymp \lambda_{\max}(\mathbf{\Sigma}) \asymp 1.$$

Then the assumptions hold, for z_j depending on Σ (typically unknown).

Outline

High-temperature log-Sobolev inequality

When does the gradient flow of q_t for an arbitrary *fixed* prior g converge in polynomial time to the posterior law

$$P_g(\boldsymbol{\varphi} \mid \mathbf{X}, \mathbf{y}) \propto P(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\varphi}) \prod_{j=1}^{p} [\mathcal{N}_\tau * g](\varphi_j)?$$

High-temperature log-Sobolev inequality

When does the gradient flow of q_t for an arbitrary *fixed* prior g converge in polynomial time to the posterior law

$$P_g(\varphi \mid \mathbf{X}, \mathbf{y}) \propto P(\mathbf{y} \mid \mathbf{X}, \varphi) \prod_{j=1}^{p} [\mathcal{N}_\tau * g](\varphi_j)?$$

Theorem (F., Guan, Shen, Wu)

Let $g \in \mathcal{P}([-M, M])$ and suppose $\sigma^2 > M^2 \|\mathbf{X}\|^2$. Then $P_g(\varphi \mid \mathbf{X}, \mathbf{y})$ satisfies the LSI, for a constant $C = C(M, \tau) > 0$ (uniform over g)

$$\operatorname{Ent} f(\varphi)^2 \leq C \cdot \mathbb{E} \|\nabla f(\varphi)\|^2$$

High-temperature log-Sobolev inequality

When does the gradient flow of q_t for an arbitrary *fixed* prior g converge in polynomial time to the posterior law

$$P_g(\varphi \mid \mathbf{X}, \mathbf{y}) \propto P(\mathbf{y} \mid \mathbf{X}, \varphi) \prod_{j=1}^{p} [\mathcal{N}_\tau * g](\varphi_j)?$$

Theorem (F., Guan, Shen, Wu)

Let $g \in \mathcal{P}([-M, M])$ and suppose $\sigma^2 > M^2 \|\mathbf{X}\|^2$. Then $P_g(\varphi \mid \mathbf{X}, \mathbf{y})$ satisfies the LSI, for a constant $C = C(M, \tau) > 0$ (uniform over g)

$$\operatorname{Ent} f(\varphi)^2 \leq C \cdot \mathbb{E} \|\nabla f(\varphi)\|^2$$

- Implies exponential contraction of $D_{\mathrm{KL}}(q_t(\varphi) \| P_g(\varphi \mid \mathbf{X}, \mathbf{y}))$ when the prior g is fixed (and possibly misspecified).
- LSI holds also for the posterior $P(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y})$ in the $\boldsymbol{\theta}$ parameter when $g(\cdot)$ has density bounded away from $\{0, \infty\}$.

High-temperature log-Sobolev inequality

Consider the generative process

$$\theta_j \overset{iid}{\sim} g \Rightarrow \varphi'_j = \theta_j + \mathcal{N}(0, \tau'^2) \Rightarrow \mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\varphi}', \sigma^2\mathbf{I} - \tau'^2\mathbf{X}\mathbf{X}^\top)$$

High-temperature log-Sobolev inequality

Consider the generative process

$$\theta_j \overset{iid}{\sim} g \Rightarrow \varphi'_j = \theta_j + \mathcal{N}(0, \tau'^2) \Rightarrow \mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\varphi}', \sigma^2\mathbf{I} - \tau'^2\mathbf{X}\mathbf{X}^\top)$$

This gives a mixture-of-products representation (analogous to [Bauerschmidt, Bodineau '18] for high-temperature spin systems)

$$P_g(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y}) = \int P_g(\boldsymbol{\theta} \mid \boldsymbol{\varphi}') P_g(\boldsymbol{\varphi}' \mid \mathbf{X}, \mathbf{y}) \, \mathrm{d}\boldsymbol{\varphi}'$$

High-temperature log-Sobolev inequality

Consider the generative process

$$\theta_j \overset{iid}{\sim} g \Rightarrow \varphi'_j = \theta_j + \mathcal{N}(0, \tau'^2) \Rightarrow \mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\varphi}', \sigma^2\mathbf{I} - \tau'^2\mathbf{X}\mathbf{X}^\top)$$

This gives a mixture-of-products representation (analogous to [Bauerschmidt, Bodineau '18] for high-temperature spin systems)

$$P_g(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y}) = \int P_g(\boldsymbol{\theta} \mid \boldsymbol{\varphi}') P_g(\boldsymbol{\varphi}' \mid \mathbf{X}, \mathbf{y}) \, \mathrm{d}\boldsymbol{\varphi}'$$

- $P_g(\boldsymbol{\theta} \mid \boldsymbol{\varphi}') = \prod_{j=1}^p P_g(\theta_j \mid \varphi_j)$. If $g(\cdot)$ is bounded away from $\{0, \infty\}$, then LSI holds by univariate characterization of [Bobkov, Götze '99].

High-temperature log-Sobolev inequality

Consider the generative process

$$\theta_j \overset{iid}{\sim} g \Rightarrow \varphi'_j = \theta_j + \mathcal{N}(0, \tau'^2) \Rightarrow \mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\varphi}', \sigma^2\mathbf{I} - \tau'^2\mathbf{X}\mathbf{X}^\top)$$

This gives a mixture-of-products representation (analogous to [Bauerschmidt, Bodineau '18] for high-temperature spin systems)

$$P_g(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y}) = \int P_g(\boldsymbol{\theta} \mid \boldsymbol{\varphi}') P_g(\boldsymbol{\varphi}' \mid \mathbf{X}, \mathbf{y}) \, \mathrm{d}\boldsymbol{\varphi}'$$

- $P_g(\boldsymbol{\theta} \mid \boldsymbol{\varphi}') = \prod_{j=1}^p P_g(\theta_j \mid \varphi_j)$. If $g(\cdot)$ is bounded away from $\{0, \infty\}$, then LSI holds by univariate characterization of [Bobkov, Götze '99].
- For $\tau'^2 > M^2$, the prior $\mathcal{N}_{\tau'} * g$ is strongly log-concave for any $g \in \mathcal{P}([-M, M])$, hence so is $P_g(\boldsymbol{\varphi}' \mid \mathbf{X}, \mathbf{y})$. LSI holds by Bakry-Emery.

This implies LSI for $P_g(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y})$.

High-temperature log-Sobolev inequality

Consider the generative process

$$\theta_j \overset{iid}{\sim} g \Rightarrow \varphi'_j = \theta_j + \mathcal{N}(0, \tau'^2) \Rightarrow \mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\varphi}', \sigma^2\mathbf{I} - \tau'^2\mathbf{X}\mathbf{X}^\top)$$

This gives a mixture-of-products representation (analogous to [Bauerschmidt, Bodineau '18] for high-temperature spin systems)

$$P_g(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y}) = \int P_g(\boldsymbol{\theta} \mid \boldsymbol{\varphi}') P_g(\boldsymbol{\varphi}' \mid \mathbf{X}, \mathbf{y}) \, \mathrm{d}\boldsymbol{\varphi}'$$

- $P_g(\boldsymbol{\theta} \mid \boldsymbol{\varphi}') = \prod_{j=1}^p P_g(\theta_j \mid \varphi_j)$. If $g(\cdot)$ is bounded away from $\{0, \infty\}$, then LSI holds by univariate characterization of [Bobkov, Götze '99].
- For $\tau'^2 > M^2$, the prior $\mathcal{N}_{\tau'} * g$ is strongly log-concave for any $g \in \mathcal{P}([-M, M])$, hence so is $P_g(\boldsymbol{\varphi}' \mid \mathbf{X}, \mathbf{y})$. LSI holds by Bakry-Emery.

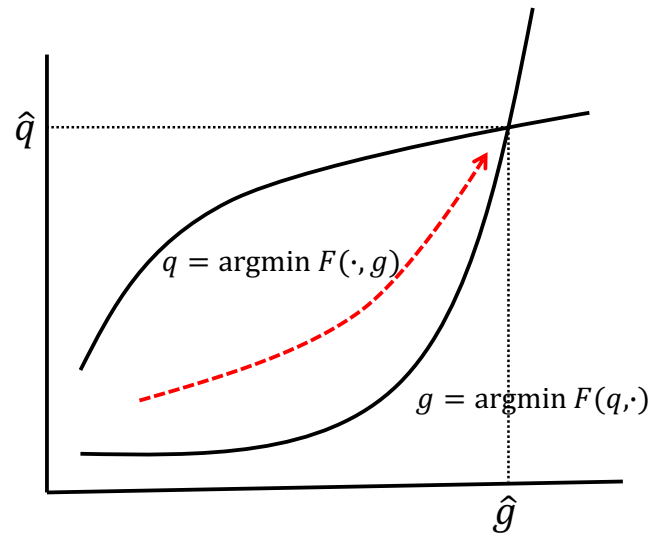
This implies LSI for $P_g(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y})$.

LSI for $P_g(\boldsymbol{\varphi} \mid \mathbf{X}, \mathbf{y})$ is similar, without needing bounds on $g(\cdot)$.

Outline

Convergence of the joint flow

When does g_t under the joint gradient flow of $\{q_t, g_t\}$ converge (in polynomial time) to an approximate NPMLE?



LSI implies contraction of $D_{\mathrm{KL}}(q_t(\varphi) \| P_{g_t}(\varphi \mid \mathbf{X}, \mathbf{y}))$, not of $D_{\mathrm{KL}}(q_t(\varphi) \| P_{g_*}(\varphi \mid \mathbf{X}, \mathbf{y}))$.

In general $F(q, g)$ may have multiple local minimizers, in 1-to-1 correspondence with local minimizers of $\bar{F}(g)$.

Convergence of the joint flow

Theorem (F., Guan, Shen, Wu)

Suppose $P_g(\varphi \mid \mathbf{X}, \mathbf{y})$ *satisfies a LSI uniformly over* $g \in \mathcal{P}([-M, M])$, *and* $\bar{F}$ *is convex over the sub-level set* $\{g : \bar{F}(g) \leq F(q_0, g_0)\}$.

Then under mild conditions for the initialization (q_0, g_0), *for any* $\varepsilon > 0$ *and some* $t \lesssim \varepsilon^{-2+o(1)}(n+p)$,

$$\bar{F}(g_t) \leq \bar{F}(g_*) + \varepsilon.$$

Convergence of the joint flow

Theorem (F., Guan, Shen, Wu)

Suppose $P_g(\varphi \mid \mathbf{X}, \mathbf{y})$ *satisfies a LSI uniformly over* $g \in \mathcal{P}([-M, M])$, *and* $\bar{F}$ *is convex over the sub-level set* $\{g : \bar{F}(g) \le F(q_0, g_0)\}$.

Then under mild conditions for the initialization (q_0, g_0), *for any* $\varepsilon > 0$ *and some* $t \lesssim \varepsilon^{-2+o(1)}(n+p)$,

$$\bar{F}(g_t) \le \bar{F}(g_*) + \varepsilon.$$

- Local convexity is required for $\bar{F}(g)$, not for $F(q, g)$.
- In asymptotic settings where $\bar{F}(g)$ converges to a deterministic limit, it is convex on a sub-level set that is dimension-free.

Roles of variable reparametrization

Parametrization of the model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\varphi} + \tilde{\varepsilon}, \qquad \boldsymbol{\varphi} = \boldsymbol{\theta} + \mathcal{N}(0, \tau^2 \mathbf{I})$$

is useful throughout the analyses, including:

- Ensuring a finite metric entropy for the space of priors, when proving consistency.

Roles of variable reparametrization

Parametrization of the model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\varphi} + \tilde{\varepsilon}, \qquad \boldsymbol{\varphi} = \boldsymbol{\theta} + \mathcal{N}(0, \tau^2 \mathbf{I})$$

is useful throughout the analyses, including:

- Ensuring a finite metric entropy for the space of priors, when proving consistency.
- Ensuring LSI uniformly over all $g \in \mathcal{P}([-M, M])$, irrespective of upper/lower bounds for the prior density.

Roles of variable reparametrization

Parametrization of the model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\varphi} + \tilde{\varepsilon}, \qquad \boldsymbol{\varphi} = \boldsymbol{\theta} + \mathcal{N}(0, \tau^2 \mathbf{I})$$

is useful throughout the analyses, including:

- Ensuring a finite metric entropy for the space of priors, when proving consistency.
- Ensuring LSI uniformly over all $g \in \mathcal{P}([-M, M])$, irrespective of upper/lower bounds for the prior density.
- Ensuring uniform smoothness of the Langevin diffusion drift, for existence, uniqueness, and smoothness of the solution $\{q_t, g_t\}$.

Outline

Summary

- Under mild assumptions for the design, i.i.d. priors in linear models may be estimated consistently and nonparametrically using the NPMLE.

Summary

- Under mild assumptions for the design, i.i.d. priors in linear models may be estimated consistently and nonparametrically using the NPMLE.
- EBflow introduces a reparametrization by a smoothed variable $\varphi = \theta + \mathcal{N}(0, \tau^2 \mathbf{I})$, and optimizes the Gibbs variational representation of the log-likelihood via the joint gradient flow

$$\partial_t g_t = -\operatorname{grad}_g^{\mathrm{FR}} F(q_t, g_t)$$
$$\partial_t q_t = -p \cdot \operatorname{grad}_q^{W_2} F(q_t, g_t)$$

Summary

- Under mild assumptions for the design, i.i.d. priors in linear models may be estimated consistently and nonparametrically using the NPMLE.
- EBflow introduces a reparametrization by a smoothed variable $\varphi = \theta + \mathcal{N}(0, \tau^2 \mathbf{I})$, and optimizes the Gibbs variational representation of the log-likelihood via the joint gradient flow

$$\partial_t g_t = -\operatorname{grad}_g^{\mathrm{FR}} F(q_t, g_t)$$
$$\partial_t q_t = -p \cdot \operatorname{grad}_q^{W_2} F(q_t, g_t)$$

- The prior g_t is 1-dimensional, and the form of $\partial_t g_t$ depends only on $\bar{q}_t$ which is easier to estimate than q_t.

Future directions / Open questions

- Can one analyze the optimization landscape of $\bar{F}(g)$, perhaps in restricted settings (e.g. high noise or i.i.d. Gaussian designs)?

Future directions / Open questions

- Can one analyze the optimization landscape of $\bar{F}(g)$, perhaps in restricted settings (e.g. high noise or i.i.d. Gaussian designs)?

- Does convergence of (q_t, g_t) occur over a dimension-independent time horizon, and are the estimates $\frac{1}{p}\sum_{j=1}^{p}\delta_{\varphi_{t,j}}$ consistent for $\bar{q}_t$ over such a time horizon with only a single Langevin chain, as $p \to \infty$?

Future directions / Open questions

- Can one analyze the optimization landscape of $\bar{F}(g)$, perhaps in restricted settings (e.g. high noise or i.i.d. Gaussian designs)?
- Does convergence of (q_t, g_t) occur over a dimension-independent time horizon, and are the estimates $\frac{1}{p}\sum_{j=1}^{p} \delta_{\varphi_{t,j}}$ consistent for $\bar{q}_t$ over such a time horizon with only a single Langevin chain, as $p \to \infty$?
- Extensions to other high-dimensional regression models (e.g. GLMs) and other Monte Carlo sampling procedures?

Reference

Z. Fan, L. Guan, Y. Shen, Y. Wu. "Gradient flows for empirical Bayes in high-dimensional linear models," https://arxiv.org/abs/2312.12708