One-Step Estimation with Scaled Proximal Methods

Robert Bassett

Naval Postgraduate School

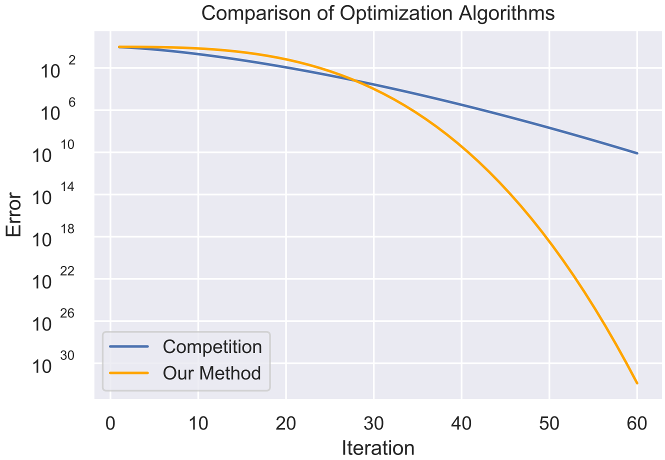
Stochastic Optimization and Statistical Learning Workshop

Erice, Italy 2022

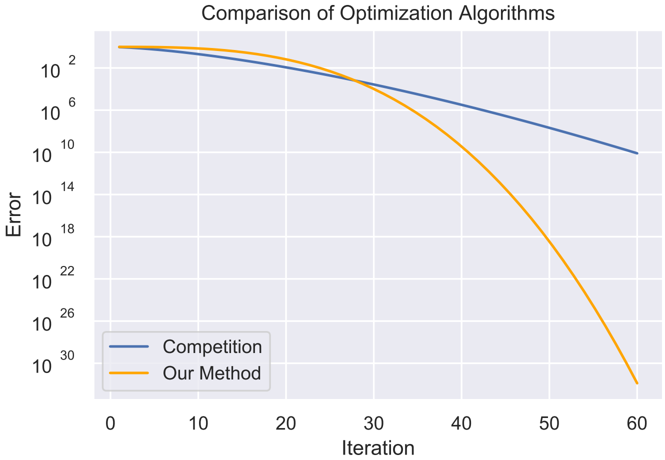
Acknowledgements

Joint with Julio Deride, Universidad Tecnica Federico Santa Maria

Motivation

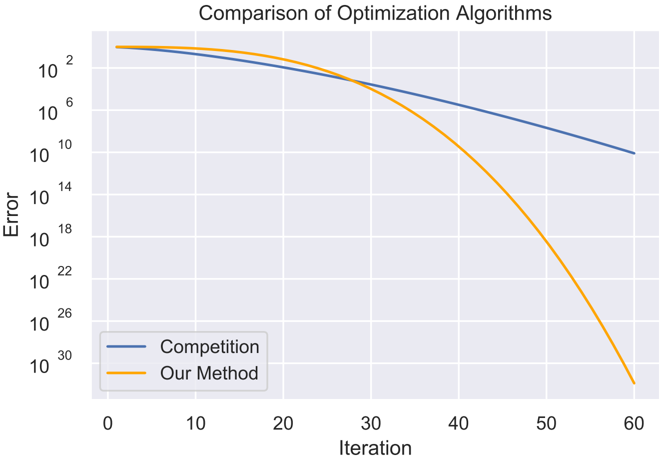

Comparison of Optimization Algorithms
Error
Iteration
10^2
10^6
10^10
10^14
10^18
10^22
10^26
10^30
0
10
20
30
40
50
60
Competition
Our Method

Motivation



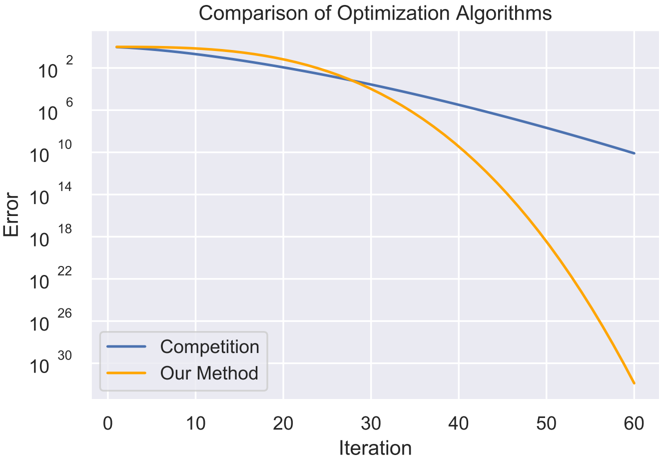
When does a graph like this make sense?

Motivation



Logistic Regression with a sample of size 100K?

Motivation



Logistic Regression with a sample of size 100?

Outline

Problem

- Should simultaneously focus on both numerical and statistical accuracy.
  - Statistical accuracy: How well do the data capture the problem we want to solve?
  - Numerical accuracy: How quickly can we can compute an estimator to (insert number) of digits?

Outline

Problem

- Should simultaneously focus on both numerical and statistical accuracy.
  - Statistical accuracy: How well do the data capture the problem we want to solve?
  - Numerical accuracy: How quickly can we can compute an estimator to (insert number) of digits?

Contributions

- We make a small contribution in this direction using *proximal methods*.
- We provide theoretical support for early stopping of *scaled* proximal methods.

Parametric Estimation

- We have a parametric family of densities $\{p(\cdot|\theta) : \theta \in \Theta \subseteq \mathbb{R}^d\}$.
- Observe n independent copies $X_1, ..., X_n$ of a random vector $X \sim p(\cdot|\theta_0)$.
- Do not know θ_0 and want to use $X_1, ..., X_n$ to estimate it.

Parametric Estimation

- We have a parametric family of densities $\{p(\cdot|\theta) : \theta \in \Theta \subseteq \mathbb{R}^d\}$.
- Observe n independent copies $X_1, ..., X_n$ of a random vector $X \sim p(\cdot|\theta_0)$.
- Do not know θ_0 and want to use $X_1, ..., X_n$ to estimate it.

Theorem (Cramer-Rao Bound)

Assume that the Fisher Information exists.

$$I_{\theta_0} := \mathrm{Var}\left[\frac{\partial}{\partial\theta}\log p(X|\theta)\bigg|_{\theta_0}\right].$$

Then any unbiased estimator $\hat{\theta}$ of θ_0 satisfies

$$\mathrm{Var}\left[\hat{\theta}\right] \succeq (nI_{\theta_0})^{-1}.$$

Parametric Estimation

We define the **Maximum Likelihood Estimator** as

$$\hat{\theta}_{MLE} \in \mathrm{argmax}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log p(X_i|\theta).$$

Parametric Estimation

We define the **Maximum Likelihood Estimator** as

$$\hat{\theta}_{MLE} \in \mathrm{argmin}_{\theta \in \Theta} -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i|\theta).$$

Parametric Estimation

We define the **Maximum Likelihood Estimator** as

$$\hat{\theta}_{MLE} \in \operatorname{argmin}_{\theta \in \Theta} F_n(\theta).$$

Parametric Estimation

We define the **Maximum Likelihood Estimator** as

$$\hat{\theta}_{MLE} \in \operatorname{argmin}_{\theta \in \Theta} F_n(\theta).$$

Theorem (Fisher 1920s, Cramer 1946)

As the sample size $n \to \infty$, the maximum likelihood estimator is unbiased. Its variance matches the Cramer-Rao bound. More precisely,

$$\hat{\theta}_{MLE} \to^{\mathcal{D}} N(\theta_0, (nI_{\theta_0})^{-1})$$

where $\to^{\mathcal{D}}$ denotes convergence in distribution.

Parametric Estimation

We define the **Maximum Likelihood Estimator** as

$$\hat{\theta}_{MLE} \in \operatorname{argmin}_{\theta \in \Theta} F_n(\theta).$$

Theorem (Fisher 1920s, Cramer 1946)

As the sample size $n \to \infty$, the maximum likelihood estimator is unbiased. Its variance matches the Cramer-Rao bound. More precisely,

$$\hat{\theta}_{MLE} \to^{\mathcal{D}} N(\theta_0, (nI_{\theta_0})^{-1})$$

where $\to^{\mathcal{D}}$ denotes convergence in distribution.

We can rewrite the conclusion of the theorem

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \to^{\mathcal{D}} N(0, I_{\theta_0}^{-1})$$

Parametric Estimation

"The justification through asymptotics appears to be the only general justification of the method of maximum likelihood"
- A. W. van der Vaart, *Asymptotic Statistics*.

- In "perfect data" regime, MLE has strong supporting theory.
- But these results were developed in the 1920s and 1940s!
- No computers $\Rightarrow$ limited ability to *compute* MLE.
- How was a respectable statistician supposed to use this insight?

Enter Le Cam

Lucien Le Cam (1924-2000)

One Step Estimators

Theorem (Le Cam, 1956)

- *Let $\tilde{\theta}_{init}$ be an initial estimator of θ_0, such that*

$$\sqrt{n}\|\tilde{\theta}_{init} - \theta_0\| = O_P(1).$$

- *Some mild regularity conditions hold.*

Then performing a single Newton step *on the objective function F_n, from* starting point $\tilde{\theta}_{init}$*, yields an estimator $\hat{\theta}_{ose}$ which is* asymptotically equivalent *to $\hat{\theta}_{MLE}$.*

This estimator

$$\hat{\theta}_{ose} := \tilde{\theta}_{init} - \nabla^2 F_n(\tilde{\theta}_{init})^{-1} \nabla F(\tilde{\theta}_{init})$$

is called the **one-step estimator**.

With Great Power...

- Starting within $M\, n^{-1/2}$ of $\hat{\theta}_{MLE}$, for some constant M satisfies the condition on $\tilde{\theta}_{init}$ in the theorem.
- This gives us "wiggle room" in the optimization of $n^{-1/2}$, where n is the sample size.
- One step of Newton's method is sufficient for an asymptotically optimal estimator (unbiased with variance equal to Cramer-Rao).

With Great Power...

- Starting within $M\,n^{-1/2}$ of $\hat{\theta}_{MLE}$, for some constant M satisfies the condition on $\tilde{\theta}_{init}$ in the theorem.
- This gives us "wiggle room" in the optimization of $n^{-1/2}$, where n is the sample size.
- One step of Newton's method is sufficient for an asymptotically optimal estimator (unbiased with variance equal to Cramer-Rao).

In practice this gave statisticians license to optimize poorly.

1. Choose starting point
2. Run a few iterations of Newton's method (by hand!?)
3. Cite Le Cam's theory suggesting this is good enough.

Modern Take

Why is one-step estimation relevant in the computer age?

1. Nonlinear optimization problems are still solved **approximately**, to a pre-specified numerical tolerance.
    - Where does numerical tolerance outpace statistical error?
    - One-step estimators link these two concepts by taking $\tilde{\theta}_{init}$ as penultimate value of Newton's method.
2. Maximum likelihood estimation with local maximizers.

   If one has access to **any** $\sqrt{n}$-consistent estimator $\tilde{\theta}_{init}$ (not necessarily MLE) of θ, the one-step estimator from this starting point is asymptotically efficient under some minimal moment conditions.

Modern Take

3. One-step estimation has been extended in a number of different directions since its inception. Primarily in statistics, as opposed to optimization, community.

- J. Fan and J. Chen. One-step local quasi-likelihood estimation. JRSSB. 1999.
- H. Zou and R. Li. One-step estimates in noncave penalized likelihood models. Annals of Statistics. 2008
- M. Taddy. One-step estimator paths for concave regularization. JCGS. 2017
- C. Huang and X. Huo. A distributed one-step estimator. Mathematical Programming. 2019.

Modern Take

4. Early stopping results have also been discussed is machine learning. The setup there is:

- Your model is overparametrized, so that your minimizer is **not actually a good estimator**.
- Early stopping can help avoid overfitting.

Here, we assume that the minimizer of your objective is a **good estimator**. Otherwise you might consider reformulating your objective, instead of **avoiding the minimizer** in your iterative method.

Only Newton's Method?

You may want to scale this beyond Newton's method.

Can we use gradient descent in Le Cam's theory?

Only Newton's Method?

You may want to scale this beyond Newton's method.

Can we use gradient descent in Le Cam's theory?

Answer: **No.**

Counterexample

We estimate the population mean from multivariate normal observations

$$X \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 100 & 0 \\ 0 & 1 \end{pmatrix}\right).$$

Take starting point $\tilde{\theta} \sim U\left([-n^{-1/2}, 0] \times [-n^{-1/2}, 0]\right)$

The one-step gradient descent estimator is <u>biased</u>.

Independent of n, this estimator underestimates the first coordinate of the mean

Counterexample

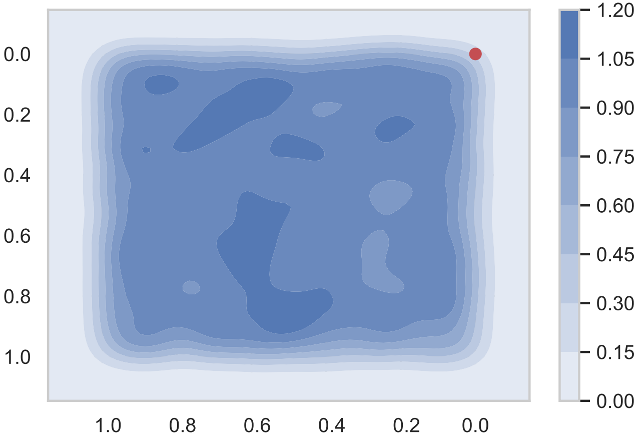


Figure: A kernel density estimate from a ($\sqrt{n}$ standardized) sample of the starting distribution

Counterexample

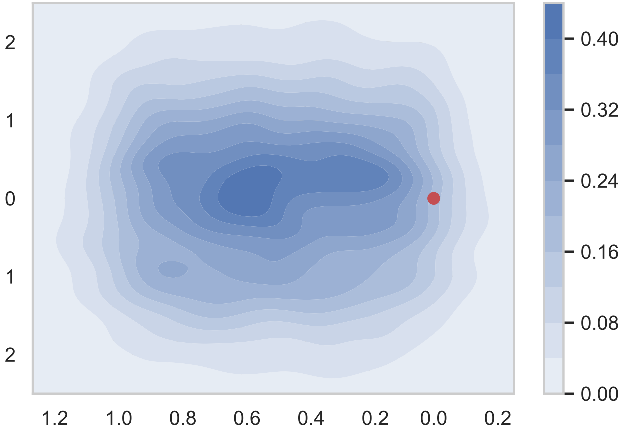


Figure: A kernel density estimate from a ($\sqrt{n}$ standardized) sample of the one-step estimator with gradient descent and optimal step length

Counterexample

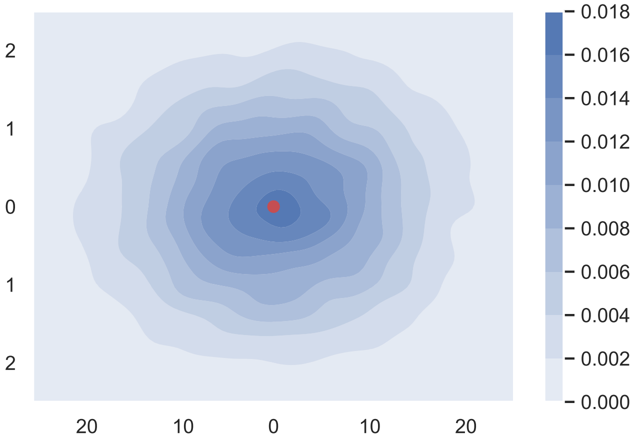


Figure: A kernel density estimate from a ($\sqrt{n}$ standardized) sample of the MLE

Only Unregularized Problems?

Regularized estimation problems are extremely important in statistical learning.

Can one-step estimation be extended to regularized problems?

Answer: **Yes**.

Composite Model & Proximal Methods

$$\min_{\theta \in \Theta} F(\theta) + G(\theta)$$

is often solved with the following, called proximal gradient descent

Composite Model & Proximal Methods

$$\min_{\theta \in \Theta} F(\theta) + G(\theta)$$

is often solved with the following, called proximal gradient descent

Initiate θ_0 and iterate the following for appropriate step lengths γ_k.

1. $\phi_k = \theta_k - \gamma_k \nabla F(\theta_k)$
2. $\theta_{k+1} \in \mathrm{argmin}_{\theta \in \Theta}\, G(\theta) + \frac{1}{2\gamma_k}\|\theta - \phi_k\|_2^2$.

Composite Model & Proximal Methods

$$\min_{\theta \in \Theta} F(\theta) + G(\theta)$$

is often solved with the following, called proximal gradient descent

Initiate θ_0 and iterate the following for appropriate step lengths γ_k.

1. $\phi_k = \theta_k - \gamma_k \nabla F(\theta_k)$
2. $\theta_{k+1} \in \operatorname{argmin}_{\theta \in \Theta} G(\theta) + \frac{1}{2\gamma_k} \|\theta - \phi_k\|_2^2$.

The **proximal operator** of G with parameter γ is

$$\operatorname{prox}_{G,\gamma}(y) = \operatorname{argmin}_{\theta \in \Theta} G(\theta) + \frac{1}{2\gamma} \|\theta - y\|_2^2.$$

Composite Model & Proximal Methods

$$\min_{\theta \in \Theta} F(\theta) + G(\theta)$$

is often solved with the following, called proximal gradient descent

Initiate θ_0 and iterate the following for appropriate step lengths γ_k.

1. $\phi_k = \theta_k - \gamma_k \nabla F(\theta_k)$
2. $\theta_{k+1} \in \operatorname{argmin}_{\theta \in \Theta} G(\theta) + \frac{1}{2\gamma_k} \|\theta - \phi_k\|_2^2$.

The **proximal operator** of G with parameter γ is

$$\operatorname{prox}_{G,\gamma}(y) = \operatorname{argmin}_{\theta \in \Theta} G(\theta) + \frac{1}{2\gamma} \|\theta - y\|_2^2.$$

So the proximal gradient method consists of applying a gradient step (in F) and proximal step (in G) for each iteration.

Scaled Proximal Gradient

Proximal gradient has an extension called *Scaled Proximal Gradient* for scaling matrices $C_k \succ 0$.

Prox Gradient

Iterate the following:

1. Gradient Step

$$\phi_k = \theta_k - \gamma_k \nabla F(\theta_k)$$

2. Proximal Step

$$\theta_{k+1} \in \mathrm{argmin}_{\theta \in \Theta} \; G(\theta) + \frac{1}{2\gamma_k}\|\theta - \phi_k\|_2^2$$

Scaled Prox Gradient

Iterate the following:

1. Newton Step

$$\phi_k = \theta_k - C_k^{-1} \nabla F(\theta_k)$$

2. Scaled Proximal Step

$$\theta_{k+1} \in \mathrm{argmin}_{\theta \in \Theta} \; G(\theta) + \frac{1}{2}\|\theta - \phi_k\|_{C_k}^2$$

Recall that $\|y\|_C^2 = y^T C y$ is the weighted euclidean norm

Prox Gradient vs Scaled Prox Gradient

Prox Gradient	Scaled Prox Gradient
(Often) Closed form prox	Rarely closed form prox
Linear convergence rate	Superlinear convergence rate

Scaled Prox Gradient is used by reputable packages such as `glmnet`, `newglmnet`, `QUIC` (QUadratic Inverse Covariance estimation).

see Lee, Sun & Saunders, 2014

Main Contribution

Theorem (Bassett & Deride, '21)

Assume we have the composite model, and form estimator

$$\hat{\theta}_M \in \text{argmin}_{\theta \in \Theta} F_n(\theta) + G_n(\theta)$$

where F_n is negative log likelihood and G_n is a regularizer. If

- $\tilde{\theta}_{init}$ *is an initial estimator such that* $\sqrt{n}\left\|\hat{\theta}_M - \tilde{\theta}_{init}\right\| = O_P(1)$.
- $G_n(\theta)$ *is convex.*
- *The scaling* C_n *is* $\succ 0$ *and* $C_n^{-1} I_{\theta_0} \to^P \mathrm{I}.^*$
- *Some mild regularity conditions hold.*

Then $\hat{\theta}_{ose}$, *the one-step estimator with* scaled proximal gradient, *is asymptotically equivalent to* $\hat{\theta}_M$.

That is, $\sqrt{n}(\hat{\theta} - \hat{\theta}_M) \to 0$ *in probability.*

Interpretation

When solving penalized log-likelihood with scaled proximal gradient,

Numerical error should scale like $n^{-1/2}$

in order to respect the statistical nature of the problem

Theorem

If F_n has Lipschitz continuous gradient, then

$$\sqrt{n}\|\hat{\theta}_{ose} - \tilde{\theta}_{init}\| = O_P(1) \Rightarrow \sqrt{n}\|\hat{\theta}_{init} - \hat{\theta}_M\| = O_P(1).$$

Thus, terminating scaled proximal gradient descent when the iterates change less than $1/\sqrt{n}$ gives the same asymptotic distribution of $\hat{\theta}_M$.

Application: Low Rank Logistic Regression

- Email-Eu-core data set: Emails sent between N members of an academic department.
- For each of T time steps, receive observations

$$X_{i,i,t} = \begin{cases} 1 & \text{Individual } i \text{ sent individual } j \text{ an email in time } t \\ 0 & \text{Otherwise.} \end{cases}$$

- Goal: Estimate $P_{i,j}$, probability of communication between individuals i and j. Assumed stationary.
- Assumptions: $\log\left(\frac{P}{1-P}\right)$ is low rank[1], i.e. individuals have similar communication patterns across all members of the department.

---

[1] Operations here are elementwise on $N \times N$ matrix of P communication probabilities

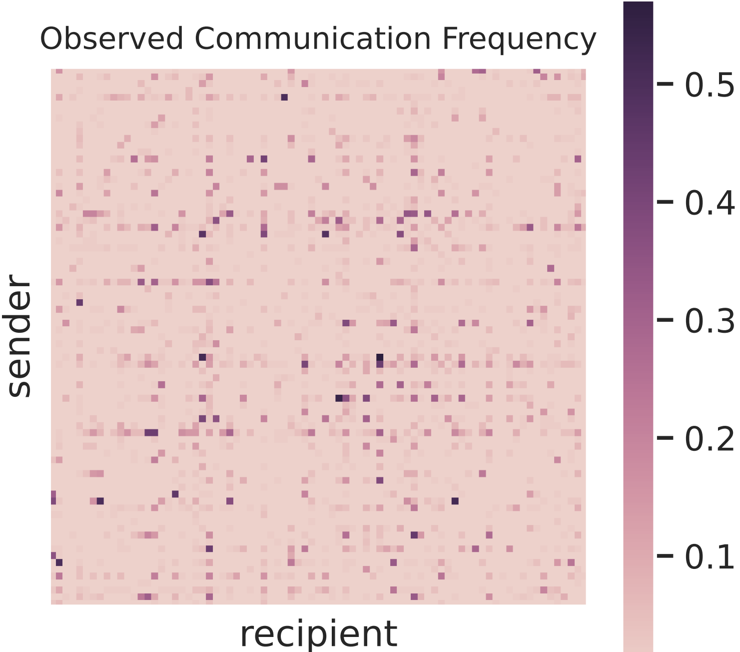
Application: Low Rank Logistic Regression

- Let $\theta = \log \frac{P}{1-P}$.
- Solve nuclear norm penalized logistic regression.
- Nuclear norm $\|\cdot\|_*$ is the ℓ_1 norm of a matrix's singular values. This penalty encourages low rank solutions.

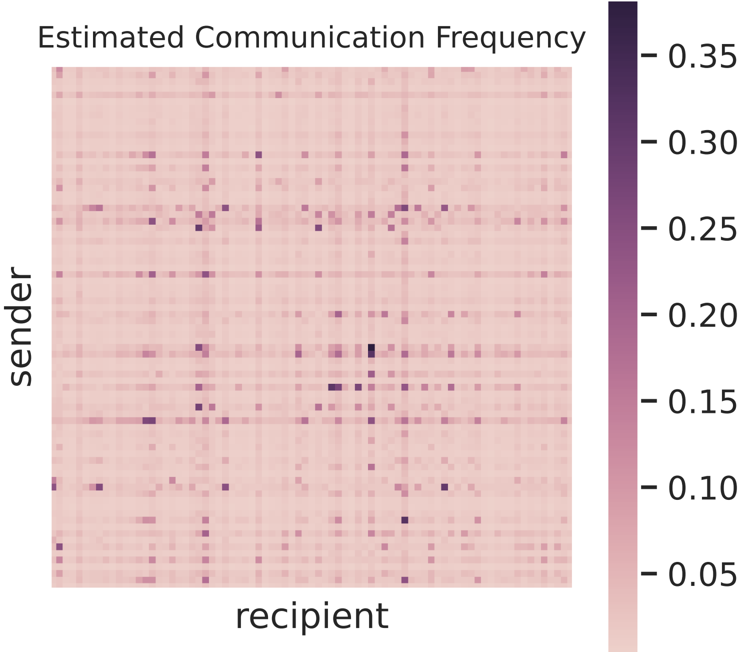
$$\min_{\theta \in \mathbb{R}^{N \times N}} \sum_{i,j \in [N] \times [N]} \left\{ \log(\exp(\theta_{i,j}) + 1) - \overline{X}_{i,j} \theta_{i,j} \right\} + \lambda \|\theta\|_*.$$

- Terminate scaled proximal gradient descent when step length between iterates is less than $T^{-1/2}$. This guarantees $\hat{\theta}_{ose}$ and $\hat{\theta}_M$ are asymptotically equivalent.

Application: Low Rank Logistic Regression



Application: Low Rank Logistic Regression



(Scaled) Proximal Descent

We've discussed one-step estimation in scaled proximal gradient descent for the regularized problem

$$\hat{\theta}_M \in \mathrm{argmin}_{\theta \in \Theta} F_n(\theta) + G_n(\theta).$$

We'll next discuss scaled proximal descent applied to the problem

$$\hat{\theta}_M \in \mathrm{argmin}_{\theta \in \Theta} F_n(\theta).$$

$$\hat{\theta} \in \underset{F_n, C_n}{prox}(\tilde{\theta}_{init}) = \mathrm{argmin}_{\theta \in \Theta} F_n(\theta) + \frac{1}{2} \left\| \theta - \tilde{\theta}_{init} \right\|_{C_n}^2 .$$

(Scaled) Proximal Descent

We have a similar result for scaled proximal descent, where we have the (unregularized) maximum likelihood estimator

$$\hat{\theta}_M \in \operatorname{argmin}_{\theta \in \Theta} F_n(\theta)$$

and we form the one-step estimator through the scaled proximal operator:

$$\hat{\theta}_{ose} \in \operatorname{argmin}_{\theta \in \Theta} F_n(\theta) + \frac{1}{2}\|\theta - \tilde{\theta}_{init}\|^2_{C_n}$$

Theorem (Bassett & Deride, '21)

If:
- $\sqrt{n}\|\tilde{\theta}_{init} - \hat{\theta}_M\| = O_P(1)$
- $\lambda_{max}\left(C_n\right) = o_P(1)$
- *Scaled prox is Lipschitz continuous.**

Then $\hat{\theta}_{ose}$ is asymptotically equivalent to $\hat{\theta}_M$.

Interpretation as a Smoother

Quasi-Newton methods are usually cheaper per iteration and have same convergence rate as scaled proximal descent.

So why would we use scaled proximal descent?

Answer: This result permits **smoothing** of a log-likelihood.

Let e_C give the scaled Moreau envelope with scaling C

$$e_C\, f(x) = \inf_{w \in \mathbb{R}^d} \left\{ f(w) + \frac{1}{2} \|x - w\|_C^2 \right\}.$$

The Moreau envelope smooths a function via infimal convolution.

Fact: Scaled proximal gradient descent is **Quasi-Newton Method** applied to the smoothed function.

$$x_{k+1} = x_k - C^{-1} \nabla e_C f(x)$$

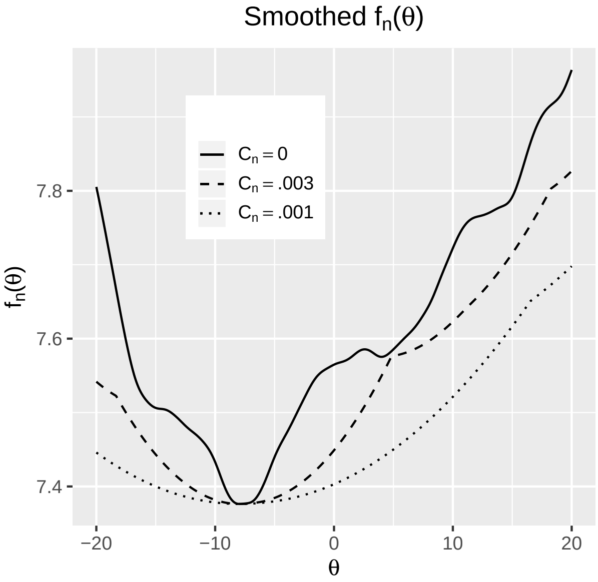
Example: Cauchy Likelihood

Goal: Estimate location parameter θ from a Cauchy distributed sample.

$$X_1, \ldots, X_n \sim^{\text{iid}} \pi^{-1}(1 + (x - \theta)^2)^{-1}.$$

- Sample mean has distribution as the X_i. **Very inefficient** (undefined mean and variance).
- Maximum likelihood estimator is **asymptotically efficient**.
- **But** there are **local maximizers** of likelihood.
- Global maximizer tends to be **well-separated**.

Example: Cauchy Likelihood



Cauchy negative log likelihood for a sample of 100 observations.

Example: Cauchy Likelihood with Laplacian Prior

Let's add an ℓ_1 regularizer to encourage sparse solutions.

We return to the setting of scaled proximal gradient descent.

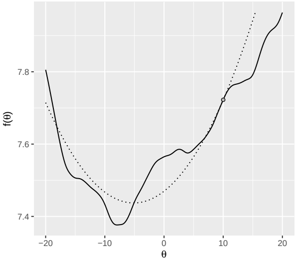
Iterations of **scaled proximal gradient descent** can be written.

$$x_{k+1} = \text{argmin}_x \left\{ f(x_k) + \nabla f(x_k)^T (x - x_k) + g(x) + \frac{1}{2} \|\theta - \theta_k\|_C^2 \right\}$$

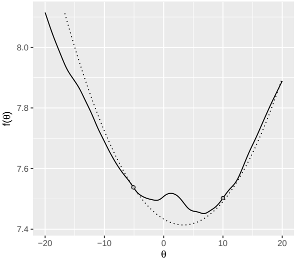
$$= \text{argmin}_x \underbrace{e_C}_{\text{Moreau Envelope}} \Big(f(x_k) + \nabla f(x_k)^T (x - x_k) + g(x) \Big)$$

Therefore our scaled proximal descent results also have a statistical smoothing interpretation, but here it is a local one.

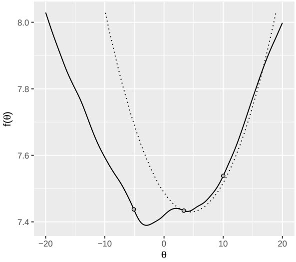
Example: Cauchy Likelihood with Laplacian Prior



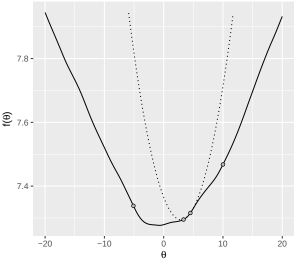
(a) n=100



(b) n=400



(c) n=700



(d) n=1000

Ongoing Work: Finite Sample Extensions

One-step estimation results depend critically on the theory of local asymptotic normality.

Local asymptotic normality (informally): The log-likelihood function derived from n iid samples can be locally approximated by a quadratic function. The approximation error converges to 0 in probability.

Finite sample results for one-step estimators require finite sample extensions of local asymptotic normality.

Such extensions exist, but they do not extend beyond the sub-gaussian setting. Example include:

- V. Spokoiny. Parametric Estimation. Finite Sample Theory. *Annals of Statistics*. 2012.
- S. Boucheron and P. Massart. A high-dimensional Wilks Phenomenon. *Probability Theory and Related Fields*. 2011.

Conclusion

- Le Cam worked on early stopping results for Newton's method applied to MLE.
- We extend this insight to penalized and constrained problems by considering **Scaled Proximal Gradient Descent** and **Scaled Proximal Descent**.
- Scaled Proximal Methods work similarly to Newton–a one-step estimator from a starting point within $n^{-1/2}$ of the minimum behaves like the minimum.
- When loss functions are well behaved these results inform stopping tolerance, by using the penulimate iteration as $\tilde{\theta}_{init}$.
- Applies to many problems where we want to build structured estimates from data.

References

- Bassett, Deride. One-Step Estimation with Scaled Proximal Methods. Mathematics of Operations Research. 2021.
- Slides Available at: https://faculty.nps.edu/rbassett
- Excellent summary on Le Cam's OSE work can be found in van der Vaart's *Asymptotic Statistics*.