One-Step Estimation with Scaled Proximal Methods

Robert Bassett

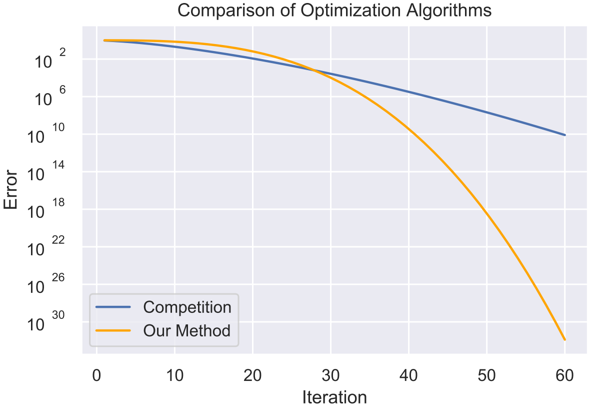
Naval Postgraduate School

SIAM Optimization Conference, 2021

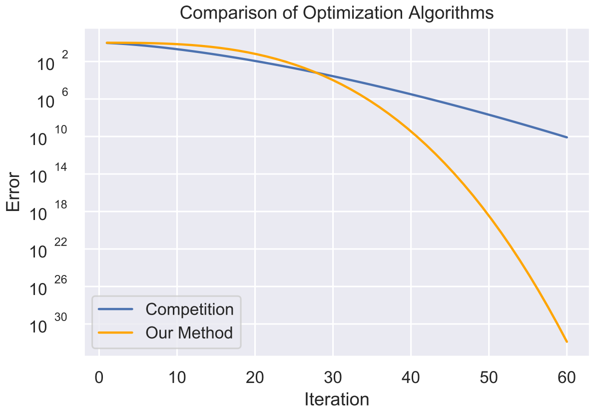
Acknowledgements

Joint with Julio Deride, Universidad Tecnica Federico Santa Maria

The Problem

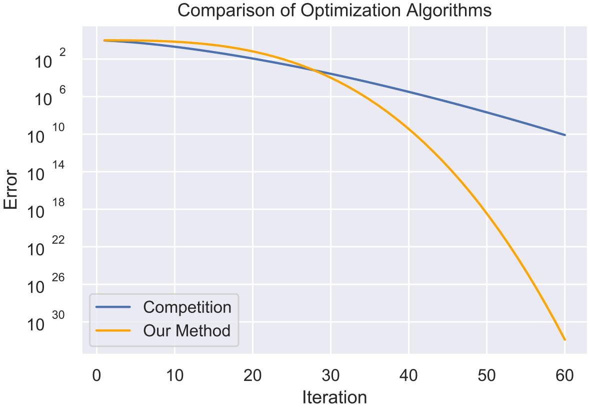

Comparison of Optimization Algorithms
Error
10^{2}
10^{6}
10^{10}
10^{14}
10^{18}
10^{22}
10^{26}
10^{30}
Competition
Our Method
0
10
20
30
40
50
60
Iteration

The Problem



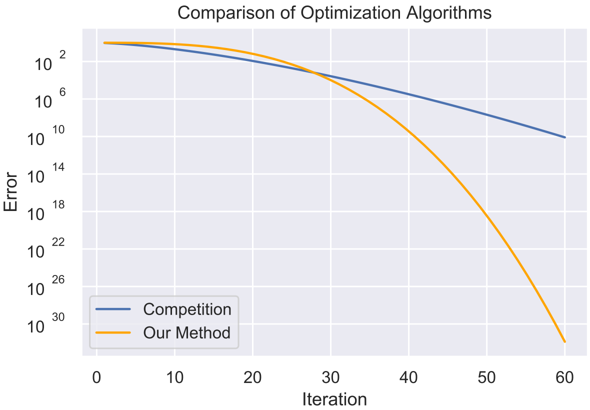
When does a graph like this make sense?

The Problem



Logistic Regression with a sample of size 100K?

The Problem



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

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

Parametric Estimation

We define the **Maximum Likelihood Estimator** as

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

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

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

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**

$$\sqrt{n}\|\tilde{\theta}_{init} - \theta_0\| < M$$

for some M and n large enough.

- *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 $M \cdot n^{-1/2}$ 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 $M \cdot n^{-1/2}$ 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.

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

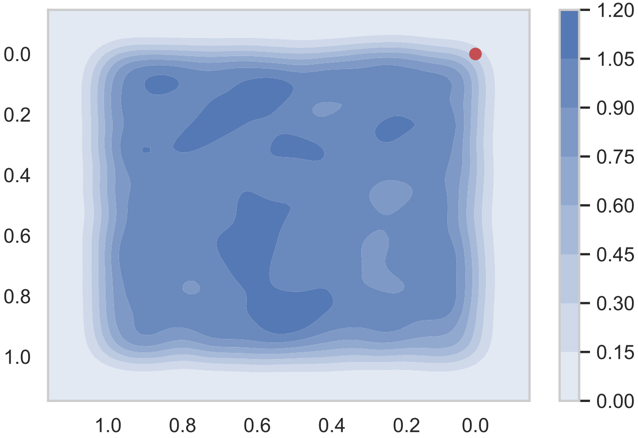


Figure: A kernel density estimate from a ($\sqrt{n}$ standardized) sample of the starting distribution

Counterexample

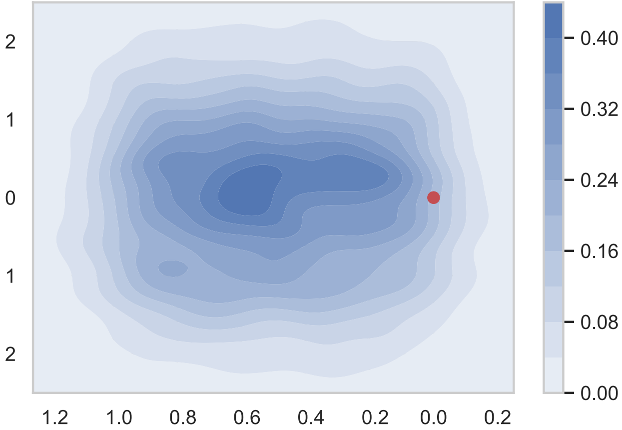


Figure: A kernel density estimate from a ($\sqrt{n}$ standardized) sample of the one step estimator with gradient descent and optimal step length

Counterexample

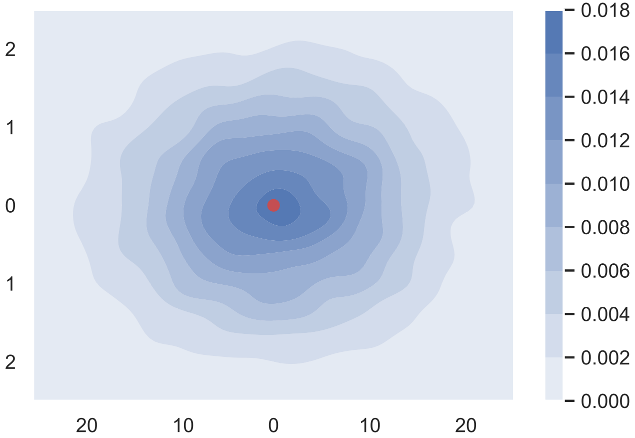


Figure: A kernel density estimate from a ($\sqrt{n}$ standardized) sample of the MLE

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

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

The proximal operator of G with parameter γ is

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

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

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

Prox Newton

Iterate the following:

1. Newton Step

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

2. Scaled Proximal Step

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

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

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

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

- $\tilde{\theta}_{init}$ *is an initial estimator within* * $M \cdot n^{-1/2}$ *of* $\hat{\theta}_M$.
- $G(\theta)$ *is convex.*
- *The scaling* C_n *is* $\succ 0$ *and* $C_n^{-1} I_{\theta_0} \to^{n \to \infty} \mathrm{I}$.
- *Some mild regularity conditions hold.*

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

Main Contribution

Theorem (Bassett & Deride, '21)

Assume we have the composite model, and form estimator

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

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

- $\tilde{\theta}_{init}$ *is an initial estimator within* * $M \cdot n^{-1/2}$ *of* $\hat{\theta}_M$.
- $G(\theta)$ *is convex.*
- *The scaling* C_n *is* $\succ 0$ *and* $C_n^{-1} I_{\theta_0} \to^{n \to \infty} \mathrm{I}$.
- *Some mild regularity conditions hold.*

Then $\hat{\theta}$, 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,

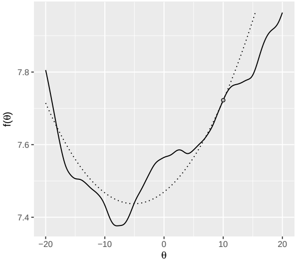
Numerical error should scale like $n^{-1/2}$

in order to respect the statistical nature of the problem

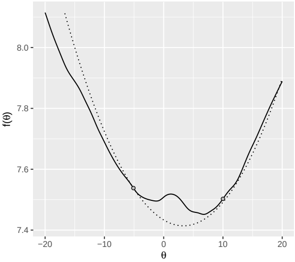
Interpretation as a Smoother

- The (scaled) proximal operator has a well known interpretation as a smoother, via the infimal convolution of epigraphs.
- Therefore our results provide theoretical justification for smoothing of a statistical objective using infimal convolution.

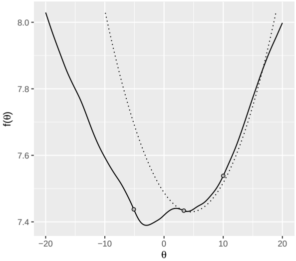
Example: Cauchy Likelihood with Laplacian Prior



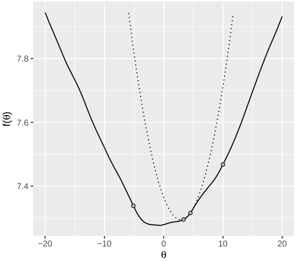
(a) n=100



(b) n=400



(c) n=700



(d) n=1000

Proximal Descent and Scaled Proximal Descent

We have a similar result for scaled proximal descent, where we have the estimator

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

and we iterate the scaled proximal operator:

$$\theta_{n+1} \in \text{argmin}_{\theta \in \Theta} F_n(\theta) + \frac{1}{2}\|\theta - \theta_k\|^2_{C_n}$$

Theorem (Bassett & Deride, '21)

If $C_n \to 0$, $\|\tilde{\theta}_{init} - \hat{\theta}_M\| \leq M/\sqrt{n}$, and the scaled prox is Lipschitz continuous, then $\text{argmin}_{\theta \in \Theta} F_n(\theta) + \frac{1}{2}\|\theta - \tilde{\theta}_{init}\|^2_{C_n}$ is asymptotically equivalent to $\hat{\theta}_M$.

Summary

- 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 Methods**.
- 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.
- Applies to many problems where we want to build structured estimates from data.