Robust empirical risk minimization via Newton's method

Po-Ling Loh

University of Cambridge, Department of Pure Mathematics and Mathematical Statistics

CRiSM Seminar
University of Warwick

11 January 2023

Joint work with Eirini Ioannou (Edinburgh) and Muni Sreenivas Pydi (Université Paris Dauphine)

Introduction

- **Goal:** Parametric estimation in contaminated data
- Huber's ϵ-contamination model: Observations $z_i \sim (1 - \epsilon)P_{\theta^*} + \epsilon Q$, where Q is arbitrary

Introduction

- **Goal:** Parametric estimation in contaminated data
- Huber's ϵ-contamination model: Observations $z_i \sim (1-\epsilon)P_{\theta^*} + \epsilon Q$, where Q is arbitrary
- Our method is also applied to heavy-tailed parametric estimation (no contamination)

Introduction

- Traditional approach via M-estimators: Suppose

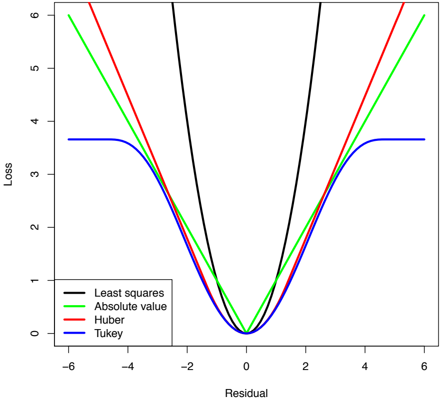
$$\theta^* = \arg\min_\theta \underbrace{\mathbb{E}_{x_i \sim P_{\theta^*}}[\mathcal{L}(\theta, x_i)]}_{\mathcal{R}(\theta)}$$

- Use empirical risk minimizer

$$\widehat{\theta} \in \arg\min_\theta \frac{1}{n} \sum_{i=1}^n \mathcal{L}(\theta, z_i),$$

for appropriately defined $\mathcal{L}$

Introduction



- Alternative approach: Use "non-robust" $\mathcal{L}$ (e.g., based on log-likelihood of P_θ) and robustify *optimization* procedure

Previous work

- Robust gradient descent algorithm (Prasad, Suggala, Balakrishnan, and Ravikumar (2020)):

$$\theta_{t+1} = \theta_t - \eta g(\theta_t),$$

where $g(\theta_t)$ is an estimate of $\nabla \mathcal{R}(\theta_t)$

Previous work

- SEVER algorithm (Diakonikolas, Kamath, Kane, Li, Steinhardt, and Stewart (2019)) uses an "approximate learner" algorithm which finds approximately critical points
- Iteratively filters out data points with outlying gradients computed at θ_t, chosen by the approximate learner

Previous work

- Median-of-means minimization approach (Lecué, Lerasle, and Mathieu (2020)) performs gradient descent by computing gradients w.r.t. a median block (computed w.r.t. empirical mean of $\mathcal{L}$) on each iterate
- Derives excess risk bounds on final iterate

Robust Newton's method

- We analyze a *second-order* version of Prasad et al. (2020), based on Newton's method:

$$\theta_{t+1} = \theta_t - \alpha_t H(\theta_t)^{-1} g(\theta_t),$$

where $(g(\theta_t), H(\theta_t))$ are estimates of $\left(\nabla \mathcal{R}(\theta_t), \nabla^2 \mathcal{R}(\theta_t)\right)$ and α_t is a step size

Robust Newton's method

- We analyze a *second-order* version of Prasad et al. (2020), based on Newton's method:

$$\theta_{t+1} = \theta_t - \alpha_t H(\theta_t)^{-1} g(\theta_t),$$

  where $(g(\theta_t), H(\theta_t))$ are estimates of $\left(\nabla \mathcal{R}(\theta_t), \nabla^2 \mathcal{R}(\theta_t)\right)$ and α_t is a step size
- Benefit of second-order algorithm: faster convergence to optimum (quadratic rather than linear convergence)

Robust estimators: Huber contamination

- Algorithm of Lai, Rao, and Vempala (2016) for multivariate mean estimation

Algorithm 3: AgnosticMean(S)

Input: $S \subset \mathbb{R}^n$, and a routine OutlierRemoval($\cdot$).
Output: $\widehat{\boldsymbol{\mu}} \in \mathbb{R}^n$.

1. Let $(\widetilde{S}, \boldsymbol{w}) = \text{OutlierRemoval}(S)$.
2. **if** $n = 1$:
   (a) **if** $\boldsymbol{w} = -1$, **Return** median($\widetilde{S}$). //Gaussian case
   (b) **else Return** mean($\widetilde{S}$). //General case
3. Let $\boldsymbol{\Sigma}_{\widetilde{S},\boldsymbol{w}}$ be the weighted covariance matrix of $\widetilde{S}$ with weights $\boldsymbol{w}$, and V be the span of the top $n/2$ principal components of $\boldsymbol{\Sigma}_{\widetilde{S},\boldsymbol{w}}$, and W be its complement.
4. Set $S_1 := \boldsymbol{P}_V(S)$ where $\boldsymbol{P}_V$ is the projection operation on to V.
5. Let $\widehat{\boldsymbol{\mu}}_V := \text{AgnosticMean}(S_1)$ and $\widehat{\boldsymbol{\mu}}_W := \text{mean}(\boldsymbol{P}_W \widetilde{S})$.
6. Let $\widehat{\boldsymbol{\mu}} \in \mathbb{R}^n$ be such that $\boldsymbol{P}_V \widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_V$ and $\boldsymbol{P}_W \widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_W$.
7. **Return** $\widehat{\boldsymbol{\mu}}$.

Robust estimators: Huber contamination

- For gradients, we treat vectors $\{\nabla \mathcal{L}(\theta, z_i)\}_{i=1}^n$ as contaminated samples from a distribution with mean $\nabla \mathcal{R}(\theta) \in \mathbb{R}^p$

Robust estimators: Huber contamination

- For gradients, we treat vectors $\{\nabla \mathcal{L}(\theta, z_i)\}_{i=1}^n$ as contaminated samples from a distribution with mean $\nabla \mathcal{R}(\theta) \in \mathbb{R}^p$
- For Hessians, we vectorize matrices $\{\nabla^2 \mathcal{L}(\theta, z_i)\}_{i=1}^n$ and treat them as contaminated samples from a distribution with mean $\nabla^2 \mathcal{R}(\theta) \in \mathbb{R}^{p \times p}$

Backtracking linesearch

- Traditional Newton's method analysis (e.g., Boyd and Vandenberghe (2004)) involves picking step size α_t using a linesearch algorithm

Backtracking linesearch

- Traditional Newton's method analysis (e.g., Boyd and Vandenberghe (2004)) involves picking step size α_t using a linesearch algorithm
- Robust version involves a slightly modified version of loss function evaluation and introduction of error parameter

Set $\alpha = 1$
while $\textsc{RobustEstimate}(\{\mathcal{L}(\theta + \alpha \Delta\theta_{nt}, z_i)\}_{i=1}^n) >$
$\textsc{RobustEstimate}(\{\mathcal{L}(\theta, z_i)\}_{i=1}^n) + \kappa_1 \alpha g(\theta) \Delta\theta_{nt} + \zeta$ **do**
  Update $\alpha = \kappa_2 \alpha$
end while

Backtracking linesearch

- Traditional Newton's method analysis (e.g., Boyd and Vandenberghe (2004)) involves picking step size α_t using a linesearch algorithm
- Robust version involves a slightly modified version of loss function evaluation and introduction of error parameter

  Set $\alpha = 1$
  while $\textsc{RobustEstimate}(\{\mathcal{L}(\theta + \alpha\Delta\theta_{nt}, z_i)\}_{i=1}^n) > \textsc{RobustEstimate}(\{\mathcal{L}(\theta, z_i)\}_{i=1}^n) + \kappa_1 \alpha g(\theta)\Delta\theta_{nt} + \zeta$ **do**
      Update $\alpha = \kappa_2 \alpha$
  end while

- Newton direction is $\Delta\theta_{nt} := -H(\theta_t)^{-1} g(\theta_t)$, contraction parameter is $\kappa_2 \in (0, 1)$, and step size is output of backtracking algorithm

Convergence guarantees

- Assume population-level objective satisfies strong convexity/smoothness:

$$mI \preceq \nabla^2 \mathcal{R}(\theta) \preceq MI$$

  (in a local region around θ^*)
- Also assume $\nabla^2 \mathcal{R}$ is L-Lipschitz

Convergence guarantees

- Assume population-level objective satisfies strong convexity/smoothness:

$$mI \preceq \nabla^2 \mathcal{R}(\theta) \preceq MI$$

  (in a local region around θ^*)
- Also assume $\nabla^2 \mathcal{R}$ is L-Lipschitz
- In traditional Newton's method analysis, iterates decrease objective by constant increments during damped Newton phase, then exhibit fast convergence with step size $\alpha_t = 1$ (pure Newton phase)

Convergence guarantees

- Assume gradient/Hessian errors are small:

$$\begin{aligned} \|g(\theta_t) - \nabla \mathcal{R}(\theta_t)\|_2 &\le \alpha_g \|\theta_t - \theta^*\|_2 + \beta_g, \\ \|H(\theta_t) - \nabla^2 \mathcal{R}(\theta_t)\|_2 &\le \alpha_h \|\theta_t - \theta^*\|_2 + \beta_h, \end{aligned}$$

for all $1 \le t \le T$

Convergence guarantees

- Assume gradient/Hessian errors are small:

$$\|g(\theta_t) - \nabla \mathcal{R}(\theta_t)\|_2 \leq \alpha_g \|\theta_t - \theta^*\|_2 + \beta_g,$$
$$\|H(\theta_t) - \nabla^2 \mathcal{R}(\theta_t)\|_2 \leq \alpha_h \|\theta_t - \theta^*\|_2 + \beta_h,$$

  for all $1 \leq t \leq T$
- Also assume robust loss estimates are smaller than $\frac{\zeta}{4}$ for all evaluations of backtracking linesearch

Convergence guarantees

Theorem (Pure Newton phase)

Suppose $\|\nabla \mathcal{R}(\theta_0)\|_2 < \eta(m, L)$*. Then backtracking linesearch chooses* $\alpha_t = 1$ *on all successive iterates, and* $\|\nabla \mathcal{R}(\theta_t)\|_2 < \eta$ *and*

$$\|\theta_t - \theta^*\|_2 \leq \frac{m}{L}\left(\frac{1}{2}\right)^{2^t} + \underbrace{c(m, L)\Big(O(\alpha_g + \beta_g + \alpha_h + \beta_h)\Big)}_{\omega},$$

for all $1 \leq t \leq T$.

Convergence guarantees

Theorem (Pure Newton phase)

Suppose $\|\nabla \mathcal{R}(\theta_0)\|_2 < \eta(m, L)$. *Then backtracking linesearch chooses* $\alpha_t = 1$ *on all successive iterates, and* $\|\nabla \mathcal{R}(\theta_t)\|_2 < \eta$ *and*

$$\|\theta_t - \theta^*\|_2 \leq \frac{m}{L}\left(\frac{1}{2}\right)^{2^t} + \underbrace{c(m, L)\Big(O(\alpha_g + \beta_g + \alpha_h + \beta_h)\Big)}_{\omega},$$

for all $1 \leq t \leq T$.

- For Huber contamination, parameters $(\alpha_g, \beta_g, \alpha_h, \beta_h)$ will be functions of ϵ (e.g., all are $O(\sqrt{\epsilon})$ in GLMs)

Convergence guarantees

Theorem (Pure Newton phase)

Suppose $\|\nabla \mathcal{R}(\theta_0)\|_2 < \eta(m, L)$. *Then backtracking linesearch chooses* $\alpha_t = 1$ *on all successive iterates, and* $\|\nabla \mathcal{R}(\theta_t)\|_2 < \eta$ *and*

$$\|\theta_t - \theta^*\|_2 \le \frac{m}{L}\left(\frac{1}{2}\right)^{2^t} + \underbrace{c(m, L)\Big(O(\alpha_g + \beta_g + \alpha_h + \beta_h)\Big)}_{\omega},$$

for all $1 \le t \le T$.

- For Huber contamination, parameters $(\alpha_g, \beta_g, \alpha_h, \beta_h)$ will be functions of ϵ (e.g., all are $O(\sqrt{\epsilon})$ in GLMs)
- Proper choice of ζ is also $O(\alpha_g + \beta_g + \alpha_h + \beta_h)$

Convergence guarantees

Theorem (Pure Newton phase)

Suppose $\|\nabla \mathcal{R}(\theta_0)\|_2 < \eta(m, L)$. *Then backtracking linesearch chooses* $\alpha_t = 1$ *on all successive iterates, and* $\|\nabla \mathcal{R}(\theta_t)\|_2 < \eta$ *and*

$$\|\theta_t - \theta^*\|_2 \leq \frac{m}{L}\left(\frac{1}{2}\right)^{2^t} + \underbrace{c(m, L)\Big(O(\alpha_g + \beta_g + \alpha_h + \beta_h)\Big)}_{\omega},$$

for all $1 \leq t \leq T$.

- For Huber contamination, parameters $(\alpha_g, \beta_g, \alpha_h, \beta_h)$ will be functions of ϵ (e.g., all are $O(\sqrt{\epsilon})$ in GLMs)
- Proper choice of ζ is also $O(\alpha_g + \beta_g + \alpha_h + \beta_h)$
- After $\log\log\left(\frac{1}{\omega}\right)$ iterations (as opposed to $\log\left(\frac{1}{\omega}\right)$, for robust gradient descent), error becomes $O(\omega)$

Convergence guarantees

Theorem (Damped Newton phase)

Suppose $\|\nabla \mathcal{R}(\theta_t)\|_2 \geq \eta(m, L)$. *There exists some* $\gamma(m, M, L) > 0$ *such that after a constant number of function evaluations, backtracking linesearch chooses a step size such that*

$$\mathcal{R}(\theta_{t+1}) - \mathcal{R}(\theta_t) < -\gamma(m, M, L).$$

Convergence guarantees

Theorem (Damped Newton phase)

Suppose $\|\nabla \mathcal{R}(\theta_t)\|_2 \geq \eta(m, L)$*. There exists some* $\gamma(m, M, L) > 0$ *such that after a constant number of function evaluations, backtracking linesearch chooses a step size such that*

$$\mathcal{R}(\theta_{t+1}) - \mathcal{R}(\theta_t) < -\gamma(m, M, L).$$

- Thus, number of iterates in damped Newton phase is upper-bounded

Proof elements

- At some level, "just add some error terms to usual Newton analysis"

Proof elements

- At some level, “just add some error terms to usual Newton analysis”
- In pure Newton phase, need to show backtracking linesearch still only chooses $\alpha_t = 1$

Proof elements

- At some level, "just add some error terms to usual Newton analysis"
- In pure Newton phase, need to show backtracking linesearch still only chooses $\alpha_t = 1$
- In damped Newton phase, show backtracking linesearch still chooses descent directions (in fact, $\mathcal{R}(\theta_{t+1}) - \mathcal{R}(\theta_t) < -\gamma$)

Proof elements

- At some level, "just add some error terms to usual Newton analysis"
- In pure Newton phase, need to show backtracking linesearch still only chooses $\alpha_t = 1$
- In damped Newton phase, show backtracking linesearch still chooses descent directions (in fact, $\mathcal{R}(\theta_{t+1}) - \mathcal{R}(\theta_t) < -\gamma$)
- Although linesearch exit condition is

$$\mathcal{R}(\theta_t + \alpha \Delta\theta_t) \leq \mathcal{R}(\theta_t) - \kappa\alpha\lambda^2(\theta_t) + \zeta,$$

can show lower bound on step size, leading to sufficient decrease

Proof elements

- At some level, "just add some error terms to usual Newton analysis"
- In pure Newton phase, need to show backtracking linesearch still only chooses $\alpha_t = 1$
- In damped Newton phase, show backtracking linesearch still chooses descent directions (in fact, $\mathcal{R}(\theta_{t+1}) - \mathcal{R}(\theta_t) < -\gamma$)
- Although linesearch exit condition is

$$\mathcal{R}(\theta_t + \alpha\Delta\theta_t) \leq \mathcal{R}(\theta_t) - \kappa\alpha\lambda^2(\theta_t) + \zeta,$$

  can show lower bound on step size, leading to sufficient decrease
- Also need to show that iterates lie in a ball around θ^*, in order to obtain uniform upper bound on gradient/Hessian errors:

$$\|\theta_t - \theta^*\|_2 \leq \gamma_0$$

Heavy-tailed distributions

- Can use same Newton's method framework to obtain parameter estimates for heavy-tailed data

Heavy-tailed distributions

- Can use same Newton's method framework to obtain parameter estimates for heavy-tailed data
- Median-of-means algorithm of Minsker (2015)

Require: Samples $S = \{s_i\}_{i=1}^n$, Failure probability δ

function $\textsc{HeavyTailedEstimator}(S = \{s_i\}_{i=1}^n, \delta)$

  Set $b = 1 + \lfloor 3.5 \log 1/\delta \rfloor$, the number of buckets.

  Partition S into b blocks $B_1, \dots, B_b$, each of size $\lfloor n/b \rfloor$.

  for $i = 1 \dots n$ **do**

    $\widehat{\mu}_i = \frac{1}{|B_i|} \sum_{s \in B_i} s.$

  end for

  Set $\widehat{\mu} = \arg\min_{\mu} \sum_{i=1}^{b} \|\mu - \widehat{\mu}_i\|_2.$

  return $\widehat{\mu}$.

end function

Application to GLMs

- Assume

$$P_{\theta^*}(y|x) \propto \exp\left(\frac{yx^T\theta^* - \Phi(x^T\theta^*)}{c(\sigma)}\right),$$

where Φ is the link function and

$$\mathcal{L}(\theta, (x_i, y_i)) = -yx^T\theta + \Phi(x^T\theta)$$

is the negative log-likelihood

Application to GLMs

- Assume
$$P_{\theta^*}(y|x) \propto \exp\left(\frac{yx^T\theta^* - \Phi(x^T\theta^*)}{c(\sigma)}\right),$$
where Φ is the link function and
$$\mathcal{L}(\theta, (x_i, y_i)) = -yx^T\theta + \Phi(x^T\theta)$$
is the negative log-likelihood
- Assume regularity conditions on Φ (bounded derivatives and moments of derivatives)

Application to GLMs

- Assume
$$P_{\theta^*}(y|x) \propto \exp\left(\frac{yx^T\theta^* - \Phi(x^T\theta^*)}{c(\sigma)}\right),$$
where Φ is the link function and
$$\mathcal{L}(\theta, (x_i, y_i)) = -yx^T\theta + \Phi(x^T\theta)$$
is the negative log-likelihood
- Assume regularity conditions on Φ (bounded derivatives and moments of derivatives)
- Assume bounded eighth moments of x_i's

Application to GLMs

Theorem (Huber contamination)

Suppose $\{z_i\}_{i=1}^n$ *are i.i.d. draws from a Huber* ϵ*-contaminated GLM. Suppose* $n = \Omega\left(p + \epsilon p^2 + \frac{1}{\sqrt{\delta}}\right)$*. Then the robust Newton method with* $T \asymp \log\log\left(\frac{1}{\epsilon}\right)$ *returns an output satisfying*

$$\|\theta_T - \theta^*\|_2 = O\left(p^2\sqrt{\epsilon \log p}\right),$$

with probability at least $1 - T'\delta$

Application to GLMs

Theorem (Huber contamination)

Suppose $\{z_i\}_{i=1}^n$ *are i.i.d. draws from a Huber* ϵ*-contaminated GLM. Suppose* $n = \Omega\left(p + \epsilon p^2 + \frac{1}{\sqrt{\delta}}\right)$*. Then the robust Newton method with* $T \asymp \log\log\left(\frac{1}{\epsilon}\right)$ *returns an output satisfying*

$$\|\theta_T - \theta^*\|_2 = O\left(p^2\sqrt{\epsilon \log p}\right),$$

with probability at least $1 - T'\delta$

- Under additional assumptions on the covariates (e.g., 4-wise independence of coordinates), estimation error can be reduced to $O(\sqrt{\epsilon \log p})$

Application to GLMs

- In order to apply earlier theorem, need to determine $(\alpha_g, \beta_g, \alpha_h, \beta_h)$
- Analysis of Lai et al. (2016) shows

$$\|g(\theta) - \nabla\mathcal{R}(\theta)\|_2 = O\left(\sqrt{\|\operatorname{Cov}(\nabla\mathcal{R}(\theta))\|_2 \epsilon \log p}\right)$$

- Thus, we need bounds on $\|\operatorname{Cov}(\nabla\mathcal{R}(\theta))\|_2$ (and similarly, on $\|\operatorname{Cov}(\text{flatten}(\nabla^2\mathcal{R}(\theta)))\|_2$)

Application to GLMs

Theorem (Heavy-tailed distributions)

Suppse $\{z_i\}_{i=1}^n$ are i.i.d. draws from a heavy-tailed distribution. Suppose $n = \Omega\left(p^2 \log\left(\frac{1}{\delta}\right)\right)$. Then the robust Newton method with $T \asymp \log\log\left(\frac{n}{p^2}\right)$ returns an output satisfying

$$\|\theta_T - \theta^*\|_2 = O\left(\sqrt{\frac{p^2}{n}}\right),$$

with probability at least $1 - T'\delta$

Application to GLMs

Theorem (Heavy-tailed distributions)

Suppse $\{z_i\}_{i=1}^n$ *are i.i.d. draws from a heavy-tailed distribution. Suppose* $n = \Omega\left(p^2 \log\left(\frac{1}{\delta}\right)\right)$. *Then the robust Newton method with* $T \asymp \log\log\left(\frac{n}{p^2}\right)$ *returns an output satisfying*

$$\|\theta_T - \theta^*\|_2 = O\left(\sqrt{\frac{p^2}{n}}\right),$$

with probability at least $1 - T'\delta$

- Again, assuming 4-wise independence of coordinates of the covariates, we can tighten the error bound to $O\left(\sqrt{\frac{p}{n}}\right)$

Application to GLMs

Theorem (Heavy-tailed distributions)

Suppse $\{z_i\}_{i=1}^n$ are i.i.d. draws from a heavy-tailed distribution. Suppose $n = \Omega\left(p^2 \log\left(\frac{1}{\delta}\right)\right)$. Then the robust Newton method with $T \asymp \log\log\left(\frac{n}{p^2}\right)$ returns an output satisfying

$$\|\theta_T - \theta^*\|_2 = O\left(\sqrt{\frac{p^2}{n}}\right),$$

with probability at least $1 - T'\delta$

- Again, assuming 4-wise independence of coordinates of the covariates, we can tighten the error bound to $O\left(\sqrt{\frac{p}{n}}\right)$
- Here, we can show that $\alpha_g, \beta_g, \alpha_h, \beta_h = O\left(\frac{p^2 \log(1/\delta)}{n}\right)$

Conjugate gradient method

- Alternative version of robust Newton's method, inspired by Martens (2010), approximates Hessian-vector products via finite differences:

$$\nabla^2 f(\theta) v \approx \frac{\nabla f(\theta + \delta v) - \nabla f(\theta)}{\delta}$$

Conjugate gradient method

- Alternative version of robust Newton's method, inspired by Martens (2010), approximates Hessian-vector products via finite differences:

$$\nabla^2 f(\theta)v \approx \frac{\nabla f(\theta + \delta v) - \nabla f(\theta)}{\delta}$$

- Newton direction $\Delta\theta_t$ (for population-level objective) satisfies $\nabla^2 \mathcal{R}(\theta_t)\Delta\theta_t = -\nabla \mathcal{R}(\theta_t)$

Conjugate gradient method

- Conjugate gradient algorithm (Wright & Nocedal (1999)) provides iterative method for solving linear system $Ax = b$, where only products of the form Av are required for updates

  Set $r_0 = h_{\Delta\theta^{(0)}}(\theta) + g(\theta)$
  Set $p_0 = -r_0$
  for $k = 1$ to $p - 1$ **do**
  Compute Hessian-vector product estimate,
  $h_{p_k}(\theta) = \mathrm{HVP}\textsc{roduct}(\theta, p_k)$
  Set $\alpha_k = \frac{r_k^T r_k}{p_k^T h_{p_k}(\theta)}$
  Set $\Delta\theta^{(k+1)} = \Delta\theta^{(k)} + \alpha_k p_k$
  Set $r_{k+1} = r_k + \alpha_k h_{p_k}(\theta)$
  Set $\beta_{k+1} = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$
  Set $p_{k+1} = -r_{k+1} + \beta_{k+1} p_k$
  end for

Conjugate gradient method

- Our idea: Run conjugate gradient algorithm to obtain approximate Newton steps, so only robust *gradient vector* evaluations are required

Conjugate gradient method

- Our idea: Run conjugate gradient algorithm to obtain approximate Newton steps, so only robust *gradient vector* evaluations are required
- (In practice, also need to choose parameter $\delta > 0$ for finite difference calculations)

Conjugate gradient method: Theory?

- Preceding analysis of robust Newton's method only requires that Newton directions satisfy

$$\nabla \mathcal{R}(\theta_t) = -\nabla^2 \mathcal{R}(\theta_t) \Delta\theta_t + \chi_t,$$

where χ_t is a small, bounded error

Conjugate gradient method: Theory?

- Preceding analysis of robust Newton's method only requires that Newton directions satisfy

$$\nabla \mathcal{R}(\theta_t) = -\nabla^2 \mathcal{R}(\theta_t) \Delta\theta_t + \chi_t,$$

  where χ_t is a small, bounded error
- We can also think of conjugate gradient method as providing a direction that satisfies this approximate equation

Conjugate gradient method: Theory?

- Preceding analysis of robust Newton's method only requires that Newton directions satisfy

$$\nabla \mathcal{R}(\theta_t) = -\nabla^2 \mathcal{R}(\theta_t) \Delta\theta_t + \chi_t,$$

  where χ_t is a small, bounded error
- We can also think of conjugate gradient method as providing a direction that satisfies this approximate equation
- However, we need to quantify propagation of errors through conjugate gradient iterates, due to robust estimators/finite difference approximation

Conjugate gradient method: Theory?

- Preceding analysis of robust Newton's method only requires that Newton directions satisfy

$$\nabla \mathcal{R}(\theta_t) = -\nabla^2 \mathcal{R}(\theta_t) \Delta\theta_t + \chi_t,$$

  where χ_t is a small, bounded error
- We can also think of conjugate gradient method as providing a direction that satisfies this approximate equation
- However, we need to quantify propagation of errors through conjugate gradient iterates, due to robust estimators/finite difference approximation
- This seems to be an open question in optimization ...

Conjugate gradient method: Theory?

- Conjecture: Approximate conjugate gradient method may converge geometrically to a small ball around true solution to linear system: If $Ax^* = b$, then

$$\|x_s - x^*\|_A \leq 2\kappa^s \|x_0 - x^*\|_A + \text{err}$$

Conjugate gradient method: Theory?

- Conjecture: Approximate conjugate gradient method may converge geometrically to a small ball around true solution to linear system: If $Ax^* = b$, then

$$\|x_s - x^*\|_A \leq 2\kappa^s \|x_0 - x^*\|_A + \text{err}$$

- Taylor expansion implies optimal choice of δ would be $C\epsilon^{1/4}$, leading to overall error rate of $O(\epsilon^{1/4})$ (possibly more widely applicable than the robust Newton method, which gives $O(\epsilon^{1/2})$ error rate in analyzable settings, e.g., GLMs)

Contributions

- Established framework of analysis for robust second-order optimization algorithm for parameter estimation
- Noisy analysis of backtracking linesearch succeeds in finding approximate Newton directions
- Proposed alternative robust Newton method based on conjugate gradient method

Open questions

- Better robust matrix estimators
- High-dimensional extensions
- Theory for conjugate gradient version
- Inexact Newton methods

Reference

- Ioannou, Pydi & Loh (2023). Robust empirical risk minimization via Newton's method. *arXiv version coming soon*.

Thank you!!