A Nonnegative Gauss-Newton Method

Antonio Orvieto (ETH Zürich)
and **Lin Xiao** (Meta AI, FAIR)

The 14th International Conference on Numerical Optimization
and Numerical Linear Algebra (ICNONLA)

Taiyuan, August 15-18, 2023

Motivation

empirical risk minimization (ERM) (in modern deep learning)

$$\text{minimize } f(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(x)$$

many previous and concurrent works . . .

- stochastic optimization: minimize $f(x) := \mathbf{E}_z[f_z(x)]$
- minimizing finite-sums: variance reduction, acceleration, . . .

Motivation

empirical risk minimization (ERM) (in modern deep learning)

$$\text{minimize } f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x)$$

many previous and concurrent works . . .

- stochastic optimization: minimize $f(x) := \mathbf{E}_z[f_z(x)]$
- minimizing finite-sums: variance reduction, acceleration, . . .

focus of this talk

- f_i smooth but can be nonconvex
- $f_i \geq 0$: loss functions in machine learning mostly nonnegative

Outline

- minimizing single nonnegative function
  - nonnegative Gauss-Newton (NGN) step size rule
  - connection with Polyak step size
- generalized Gauss-Newton (prox-linear)
- stochastic NGN method
- extensions and summary

Minimizing single nonnegative function

$$\underset{x \in \mathbf{R}^d}{\text{minimize}} \;\; f(x)$$

- f smooth but can be nonconvex
- f non-negative: $f(x) \geq 0$ for all $x \in \mathbf{R}^n$

can apply to any f that has nontrivial lower bound $f^{\mathrm{lb}} \leq f^\star$:

$$f(x) \leftarrow f(x) - f^{\mathrm{lb}}$$

Minimizing single nonnegative function

$$\underset{x \in \mathbf{R}^d}{\text{minimize}} \;\; f(x)$$

- f smooth but can be nonconvex
- f non-negative: $f(x) \geq 0$ for all $x \in \mathbf{R}^n$

can apply to any f that has nontrivial lower bound $f^{\text{lb}} \leq f^\star$:

$$f(x) \leftarrow f(x) - f^{\text{lb}}$$

the trick: let $r(x) = \sqrt{f(x)}$ and consider

$$\underset{x \in \mathbf{R}^n}{\text{minimize}} \;\; r^2(x)$$

$r(x)$ may not arise naturally, e.g., $f(x) = -log(s(x))$ with $s(x) < 1$

Gauss-Newton method

$$x^{k+1} = x^k + p^k$$

where

$$p^k = \arg\min_p \left(r(x^k) + \nabla r(x^k)^T p \right)^2$$

Gauss-Newton method

$$x^{k+1} = x^k + p^k$$

where

$$p^k = \arg\min_p \left(r(x^k) + \nabla r(x^k)^T p\right)^2$$

- solving $\left(r(x^k) + \nabla r(x^k)^T p\right) \nabla r(x^k) = 0$ (under-determined)

$$p^k = -\left(\nabla r(x^k) \nabla r(x^k)^T\right)^\dagger r(x^k) \nabla r(x^k) = -\frac{r(x^k)}{\|\nabla r(x^k)\|^2} \nabla r(x^k)$$

Gauss-Newton method

$$x^{k+1} = x^k + p^k$$

where

$$p^k = \arg\min_p \left(r(x^k) + \nabla r(x^k)^T p\right)^2$$

- solving $\left(r(x^k) + \nabla r(x^k)^T p\right) \nabla r(x^k) = 0$ (under-determined)

$$p^k = -\left(\nabla r(x^k) \nabla r(x^k)^T\right)^\dagger r(x^k) \nabla r(x^k) = -\frac{r(x^k)}{\|\nabla r(x^k)\|^2} \nabla r(x^k)$$

- using $\nabla r(x) = \frac{1}{2\sqrt{f(x)}} \nabla f(x)$

$$x^{k+1} = x^k - \frac{2f(x^k)}{\|\nabla f(x^k)\|^2} \nabla f(x^k)$$

Gauss-Newton method

$$x^{k+1} = x^k + p^k$$

where

$$p^k = \arg\min_p \left(r(x^k) + \nabla r(x^k)^T p\right)^2$$

- solving $\left(r(x^k) + \nabla r(x^k)^T p\right) \nabla r(x^k) = 0$ (under-determined)

$$p^k = -\left(\nabla r(x^k) \nabla r(x^k)^T\right)^\dagger r(x^k) \nabla r(x^k) = -\frac{r(x^k)}{\|\nabla r(x^k)\|^2} \nabla r(x^k)$$

- using $\nabla r(x) = \frac{1}{2\sqrt{f(x)}} \nabla f(x)$

$$x^{k+1} = x^k - \frac{2f(x^k)}{\|\nabla f(x^k)\|^2} \nabla f(x^k)$$

gradient descent with variable step size (c.f. Polyak if $f^\star = 0$)

Polyak step size

(sub)gradient method for convex optimization

$$x^{k+1} = x^k - \gamma_k \nabla f(x^k)$$

Polyak step size rule:

$$\gamma_k = \frac{f(x^k) - f^\star}{\|\nabla f(x^k)\|^2}$$

1935-2023

Polyak step size

(sub)gradient method for convex optimization

$$x^{k+1} = x^k - \gamma_k \nabla f(x^k)$$

Polyak step size rule:

$$\gamma_k = \frac{f(x^k) - f^\star}{\|\nabla f(x^k)\|^2}$$

1935-2023

derivation:

$$\begin{aligned}
\|x^{k+1} - x^\star\|^2 &= \|x^k - \gamma_k \nabla f(x^k) - x^\star\|^2 \\
&= \|x^k - x^\star\|^2 - 2\gamma_k \nabla f(x^k)^T (x^k - x^\star) + \gamma_k^2 \|\nabla f(x^k)\|^2 \\
&\leq \|x^k - x^\star\|^2 - 2\gamma_k \big(f(x^k) - f^\star\big) + \gamma_k^2 \|\nabla f(x^k)\|^2
\end{aligned}$$

minimizing the upper bound to obtain optimal γ_k

Universal optimality of Polyak step size

theorem (Hazan and Kakade 2019)

$$f(\underline{x}^k) - f^\star \leq \min\left\{\frac{GD_0}{\sqrt{k}},\ \frac{2LD_0^2}{k},\ \frac{G^2}{\mu k},\ LD_0^2\left(1-\frac{\mu}{2L}\right)^k\right\}$$

where $\underline{x}^k = \arg\min_{x\in\{x^1,\ldots,x^k\}} f(x)$ and $D_0 = \|x^0 - x^\star\|$

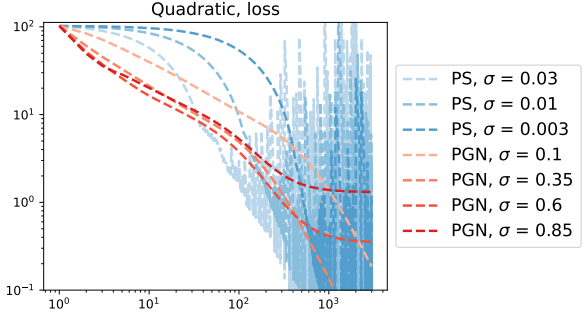
assumptions	convex & $\|\nabla f\| \leq G$	convex & L-smooth	μ-convex & $\|\nabla f\| \leq G$	μ-convex & L-smooth
rate of $f(\underline{x}^k) - f^\star$	$\frac{1}{\sqrt{k}}$	$\frac{L}{k}$	$\frac{1}{\mu k}$	$e^{-\frac{L}{\mu}k}$
step size γ_k	$\frac{1}{\sqrt{k}}$	$\frac{1}{L}$	$\frac{1}{\mu k}$	$\frac{1}{L}$

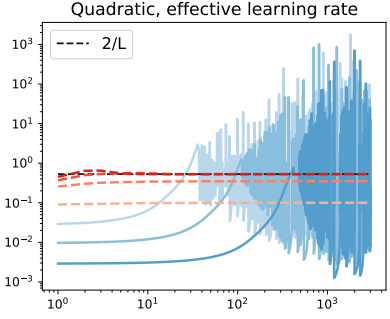
Instability when $f^\star$ unknown

Gauss-Newton step size rule:

$$x^{k+1} = x^k - \sigma \frac{f(x^k) - 0}{\|\nabla f(x^k)\|^2} \nabla f(x^k)$$

example: linear regression with $f^\star > 0$





Modifying Polyak step size

- gradually tighten lower bound (Hazan and Kakade 2019)

Modifying Polyak step size

- gradually tighten lower bound (Hazan and Kakade 2019)
- **clipping the step size:** $\gamma_k = \min\left\{\frac{f(x^k)}{\|\nabla f(x^k)\|^2}, \sigma\right\}$

$$p^k = \arg\min_p \left\{\max\left\{0, f(x^k) + \nabla f(x^k)^T p\right\} + \frac{1}{2\sigma}\|p\|^2\right\}$$

Modifying Polyak step size

- gradually tighten lower bound (Hazan and Kakade 2019)
- **clipping the step size:** $\gamma_k = \min\left\{\frac{f(x^k)}{\|\nabla f(x^k)\|^2}, \sigma\right\}$

$$p^k = \arg\min_p \left\{\max\left\{0, f(x^k) + \nabla f(x^k)^T p\right\} + \frac{1}{2\sigma}\|p\|^2\right\}$$

- **Gauss-Newton + trust region:** $\gamma_k = \min\left\{\frac{2f(x^k)}{\|\nabla f(x^k)\|^2}, \frac{\sigma}{\|\nabla f(x^k)\|}\right\}$

$$p^k = \arg\min_{\|p\| \le \sigma} \left(r(x^k) + \nabla r(x^k)^T p\right)^2$$

Modifying Polyak step size

- gradually tighten lower bound (Hazan and Kakade 2019)
- **clipping the step size:** $\gamma_k = \min\left\{\frac{f(x^k)}{\|\nabla f(x^k)\|^2}, \sigma\right\}$

$$p^k = \arg\min_p \left\{\max\left\{0, f(x^k) + \nabla f(x^k)^T p\right\} + \frac{1}{2\sigma}\|p\|^2\right\}$$

- **Gauss-Newton + trust region:** $\gamma_k = \min\left\{\frac{2f(x^k)}{\|\nabla f(x^k)\|^2}, \frac{\sigma}{\|\nabla f(x^k)\|}\right\}$

$$p^k = \arg\min_{\|p\| \le \sigma} \left(r(x^k) + \nabla r(x^k)^T p\right)^2$$

- **regularized Gauss-Newton** (Levenberg-Marquardt)

$$p^k = \arg\min_p \left\{\left(r(x^k) + \nabla r(x^k)^T p\right)^2 + \frac{1}{2\sigma}\|p\|^2\right\}$$

Regularized Gauss-Newton step size

$$x^{k+1} = x^k + p^k$$

where

$$p^k = \arg\min_p \left\{ \left(r(x^k) + \nabla r(x^k)^T p \right)^2 + \frac{1}{2\sigma} \|p\|^2 \right\}$$

Regularized Gauss-Newton step size

$$x^{k+1} = x^k + p^k$$

where

$$p^k = \arg\min_p \left\{ \left(r(x^k) + \nabla r(x^k)^T p \right)^2 + \frac{1}{2\sigma} \|p\|^2 \right\}$$

- optimality condition

$$\left(r(x^k) + \nabla r(x^k)^T p \right) \nabla r(x^k) + \frac{1}{\sigma} p = 0$$

- unique solution

$$\begin{aligned} p^k &= -\left(\frac{1}{\sigma} I + \nabla r(x^k) \nabla r(x^k)^T \right)^{-1} r(x^k) \nabla r(x^k) \\ &= -\frac{\sigma}{1 + \frac{\sigma}{2f(x^k)} \|\nabla f(x^k)\|^2} \nabla f(x^k) \end{aligned}$$

Properties of NGN step size

$$x^{k+1} = x^k - \gamma_k \nabla f(x^k), \qquad \gamma_k = \frac{\sigma}{1 + \frac{\sigma}{2f(x^k)}\|\nabla f(x^k)\|^2}$$

- range of step size (similar to clipping)

$$\gamma_k \in \left[\frac{1}{L + \sigma^{-1}},\ \sigma\right]$$

Properties of NGN step size

$$x^{k+1} = x^k - \gamma_k \nabla f(x^k), \qquad \gamma_k = \frac{\sigma}{1 + \frac{\sigma}{2f(x^k)}\|\nabla f(x^k)\|^2}$$

- range of step size (similar to clipping)

$$\gamma_k \in \left[\frac{1}{L + \sigma^{-1}},\ \sigma\right]$$

- harmonic mean of (naively modified) Polyak and constant

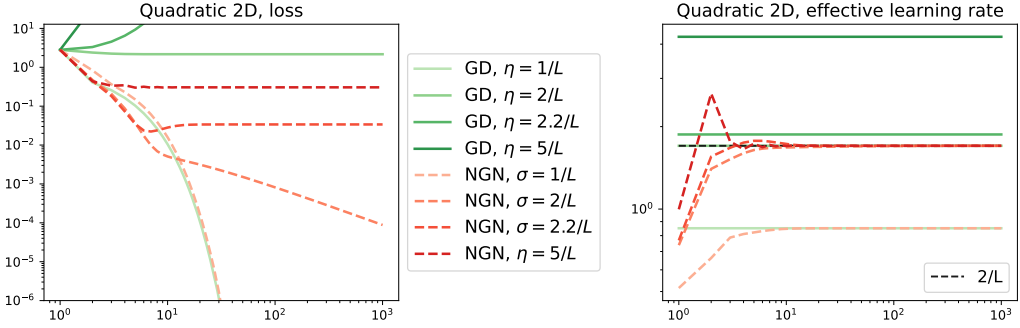
$$\gamma_k = \frac{1}{\frac{1}{\sigma} + \frac{1}{\gamma_k^{\mathrm{GN}}}}, \qquad \gamma_k^{\mathrm{GN}} = \frac{2f(x^k) - 0}{\|\nabla f(x^k)\|^2}$$

  - $\sigma \to 0$: almost constant step size σ
  - $\sigma \to \infty$: naive modification of Polyak step size

Numerical experiments

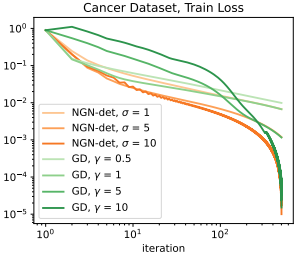
minimizing 2-D convex quadratic with $f^\star = 0$

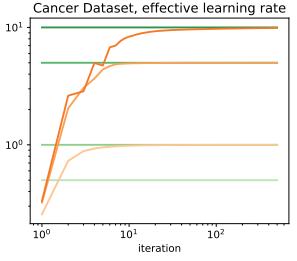
- gradient descent (GD) with constant step size η
- NGN with hyperparmaeter σ (variable step size γ_k)

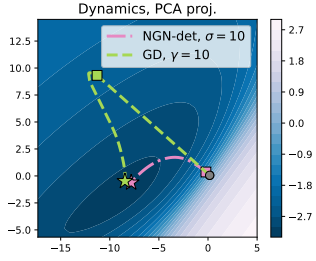


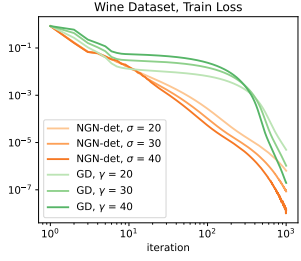
Numerical experiments

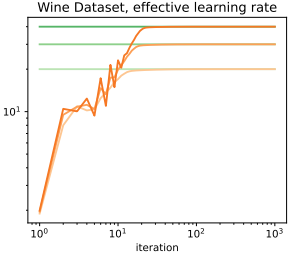
logistic regression on LIBSVM datasets (Breast Cancer, Wine)

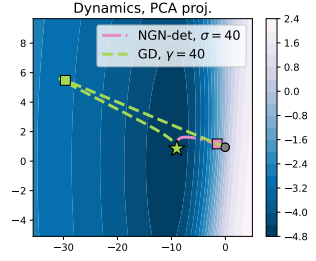












Convergence results: strongly convex

$$x^{k+1} = x^k - \gamma_k \nabla f(x^k), \qquad \gamma_k = \frac{\sigma}{1 + \frac{\sigma}{2f(x^k)} \|\nabla f(x^k)\|^2}$$

thm: if f nonnegative, μ-convex, L-smooth, and $\sigma < 1/\mu$, then

$$\|x^k - x^\star\|^2 \leq \left(1 - \frac{\mu}{L + \sigma^{-1}}\right)^k \|x^0 - x^\star\|^2 + \max\left\{0, \frac{2\sigma}{\mu}\left(\frac{\sigma L - 1}{\sigma L + 1}\right)\right\} f^\star$$

- $\sigma \leq 1/L$ (regardless of $f^\star$): linear convergence to solution
- $\sigma < 1/\mu$ and $f^\star = 0$: linear convergence to solution
- $\sigma \in (1/L, 1/\mu)$ and $f^\star > 0$: linear convergence to a ball

Convergence results: convex

$$x^{k+1} = x^k - \gamma_k \nabla f(x^k), \qquad \gamma_k = \frac{\sigma}{1 + \frac{\sigma}{2f(x^k)} \|\nabla f(x^k)\|^2}$$

thm: if f nonnegative, convex and L-smooth, then for any $\sigma > 0$,

$$f(\bar{x}^k) - f^\star \leq \frac{\|x^0 - x^\star\|^2}{\bar{\sigma} k} + \max\left\{0, \frac{2\sigma L - 1}{\bar{\sigma}}\right\} f^\star$$

where $\bar{x}^k = \frac{1}{k} \sum_{i=0}^{k-1} x^i$ and $\bar{\sigma} = \frac{\sigma}{\sigma L + 1}$

- $\sigma \leq 1/(2L)$: sublinear convergence to optimal value
- $f^\star = 0$ (for all $\sigma > 0$): sublinear convergence to optimal value
- $\sigma > 1/(2L)$ and $f^\star > 0$: sublinear convergence to a ball

Convergence results: nonconvex

$$x^{k+1} = x^k - \gamma_k \nabla f(x^k), \qquad \gamma_k = \frac{\sigma}{1 + \frac{\sigma}{2f(x^k)} \|\nabla f(x^k)\|^2}$$

thm: if f nonnegative and L-smooth and $\sigma < 2/L$, then

$$\min_{k \in [K-1]} \|\nabla f(x^k)\|^2 \;\leq\; \frac{1}{K} \sum_{k=0}^{K-1} \|\nabla f(x^k)\|^2 \;\leq\; \left[\frac{(L + \sigma^{-1})}{2 - L\sigma} \right] \frac{2f(x^0)}{K}$$

(cannot have σ arbitrarily large)

Outline

- minimizing single nonnegative function
  - nonnegative Gauss-Newton step size rule
  - connection with Polyak step size
- generalized Gauss-Newton (prox-linear)
- stochastic NGN method
- extensions and summary

Generalized Gauss-Newton (prox-linear)

$$\underset{x \in \mathbf{R}^d}{\text{minimize}} \ \{f(x) := h(c(x))\}$$

Generalized Gauss-Newton (prox-linear)

$$\underset{x \in \mathbf{R}^d}{\text{minimize}} \ \{f(x) := h(c(x))\}$$

- **update rule:**

$$x^{k+1} = x^k + p^k$$

where

$$p^k = \arg\min_p \ \left\{ h\Big(c(x^k) + \nabla c(x^k)^T p\Big) + \frac{1}{2\sigma}\|p\|^2 \right\}$$

Generalized Gauss-Newton (prox-linear)

$$\underset{x \in \mathbf{R}^d}{\text{minimize}} \ \{f(x) := h(c(x))\}$$

- **update rule:**

$$x^{k+1} = x^k + p^k$$

where

$$p^k = \arg\min_p \ \left\{ h\Big(c(x^k) + \nabla c(x^k)^T p\Big) + \frac{1}{2\sigma}\|p\|^2 \right\}$$

- **what if h does not have simple proximal mapping?**

Generalized Gauss-Newton (prox-linear)

$$\underset{x \in \mathbf{R}^d}{\text{minimize}} \ \{f(x) := h(c(x))\}$$

- **update rule:**

$$x^{k+1} = x^k + p^k$$

where

$$p^k = \arg\min_p \ \left\{ h\Big(c(x^k) + \nabla c(x^k)^T p\Big) + \frac{1}{2\sigma}\|p\|^2 \right\}$$

- **what if h does not have simple proximal mapping?**

$$p^k = \arg\min_p \left\{ f(x^k) + \nabla f(x^k)^T p + \frac{1}{2} p^T \nabla^2 f(x^k) p \right\}$$

idea: use composite structure to approximate $\nabla^2 f(x^k)$

Generalized Gauss-Newton (prox-linear)

approximate Hessian of $f(x) = h(c(x))$

$$\begin{aligned}\nabla^2 f(x) &= h'(c(x))\nabla^2 c(x) + h''(c(x))\nabla c(x)\nabla c(x)^T \\ &= h'(c(x))\nabla^2 c(x) + \tfrac{h''(c(x))}{h'(c(x))^2}\nabla f(x)\nabla f(x)^T \\ &\approx \tfrac{1}{\sigma} I + \underbrace{\tfrac{h''(c(x))}{h'(c(x))^2}}_{q(x)} \nabla f(x)\nabla f(x)^T\end{aligned}$$

Generalized Gauss-Newton (prox-linear)

approximate Hessian of $f(x) = h(c(x))$

$$\begin{aligned}\nabla^2 f(x) &= h'(c(x))\nabla^2 c(x) + h''(c(x))\nabla c(x)\nabla c(x)^T \\ &= h'(c(x))\nabla^2 c(x) + \tfrac{h''(c(x))}{h'(c(x))^2}\nabla f(x)\nabla f(x)^T \\ &\approx \tfrac{1}{\sigma} I + \underbrace{\tfrac{h''(c(x))}{h'(c(x))^2}}_{q(x)} \nabla f(x)\nabla f(x)^T\end{aligned}$$

determining step size (no need to use prox mapping of h)

$$\begin{aligned}p^k &= \arg\min_p \left\{ \nabla f(x^k)^T p + \frac{1}{2} p^T \left(\frac{1}{\sigma} I + q(x)\nabla f(x)\nabla f(x)^T \right) p \right\} \\ &= -\frac{\sigma}{1 + \sigma q(x^k)\|\nabla f(x^k)\|^2} \nabla f(x^k)\end{aligned}$$

Generalized Gauss-Newton (prox-linear)

$$\underset{x \in \mathbf{R}^d}{\text{minimize}} \ \{f(x) := h(c(x))\}$$

update rule:

$$x^{k+1} = x^k - \gamma_k \nabla f(x^k), \qquad \gamma_k = \frac{\sigma}{1 + \sigma q(x^k) \|\nabla f(x^k)\|^2}$$

where $q(x) = \frac{h''(c(x))}{h'(c(x))^2}$

Generalized Gauss-Newton (prox-linear)

$$\underset{x \in \mathbf{R}^d}{\text{minimize}} \ \{f(x) := h(c(x))\}$$

update rule:

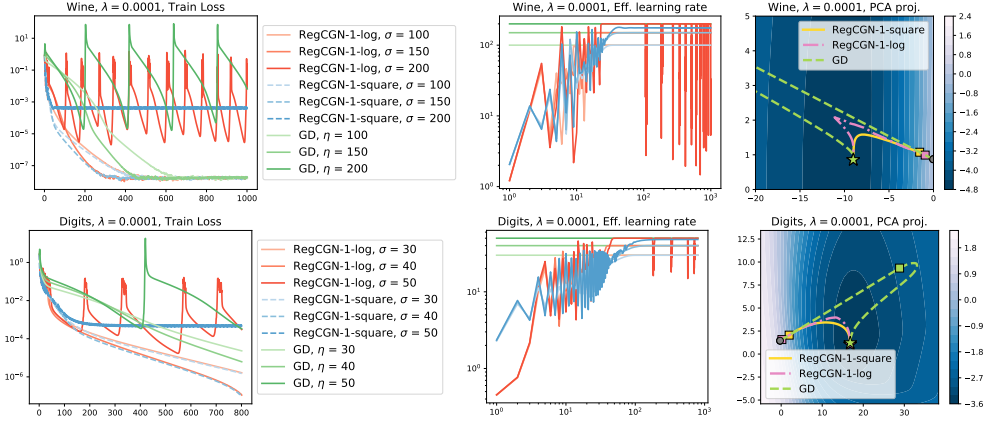
$$x^{k+1} = x^k - \gamma_k \nabla f(x^k), \qquad \gamma_k = \frac{\sigma}{1 + \sigma q(x^k) \|\nabla f(x^k)\|^2}$$

where $q(x) = \frac{h''(c(x))}{h'(c(x))^2}$

- **quadratic**: $h(y) = y^2$, $h'(y) = 2y$, $h''(y) = 2$, $q(x) = \frac{1}{2f(x)}$
- **monomial**: $h(y) = \alpha y^p$, $\ldots$, $q(x) = \frac{1}{\frac{p}{p-1} f(x)}$
- **negative log** (with $0 < y < 1$ in cross-entropy): $h(y) = -\log(y)$, $h'(y) = -1/y$, $h''(y) = 1/y^2$, and thus $q(x) = 1$

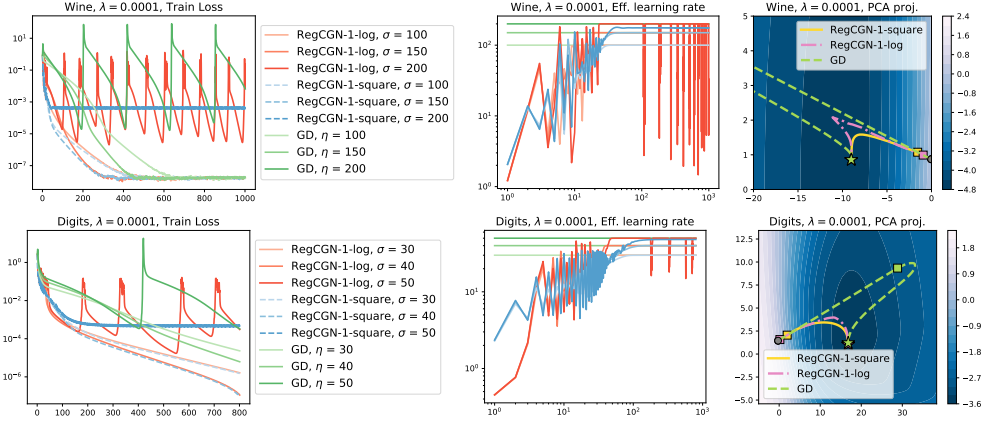
Numerical experiments

comparing $h(y) = -\log(y)$ vs $h(y) = y^2$



Numerical experiments

comparing $h(y) = -\log(y)$ vs $h(y) = y^2$



$h(y) = y^2$ best in practice even for cross-entropy loss

Outline

- minimizing single nonnegative function
  - nonnegative Gauss-Newton step size rule
  - connection with Polyak step size
- generalized Gauss-Newton (prox-linear)
- stochastic NGN method
- extensions and summary

Stochastic NGN method

empirical risk minimization (ERM)

$$\text{minimize } f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x)$$

Stochastic NGN method

empirical risk minimization (ERM)

$$\text{minimize } f(x) := \frac{1}{N}\sum_{i=1}^{N} f_i(x)$$

SGD with NGN step size rule:

$$x^{k+1} = x^k - \gamma_k \nabla f_{i_k}(x^k)$$

where i_k picked randomly and

$$\gamma_k = \frac{\sigma}{1 + \frac{\sigma}{2f_{i_k}(x^k)}\|\nabla f_{i_k}(x^k)\|^2} = \frac{1}{\frac{1}{\sigma} + \frac{\|\nabla f_{i_k}(x^k)\|^2}{2f_{i_k}(x^k)}}$$

Convergence analysis: Some definitions

define

$$x^\star := \arg\min f(x), \qquad f_i^\star := \min_x f_i(x)$$

and

$$\Delta_{\text{int}} := \mathbf{E}[f_i(x^\star) - f_i^\star], \qquad \Delta_{\text{pos}} := \mathbf{E}[f_i^\star]$$

- Δ_{pos} measures average positivity: zero if $f_i^\star = 0$ for all i
- Δ_{int} measures interpolation: zero if $x^\star = \arg\min_x f_i(x)$ for all i

Convergence analysis: Some definitions

define

$$x^\star := \arg\min f(x), \qquad f_i^\star := \min_x f_i(x)$$

and

$$\Delta_{\text{int}} := \mathbf{E}[f_i(x^\star) - f_i^\star], \qquad \Delta_{\text{pos}} := \mathbf{E}[f_i^\star]$$

- Δ_{pos} measures average positivity: zero if $f_i^\star = 0$ for all i
- Δ_{int} measures interpolation: zero if $x^\star = \arg\min_x f_i(x)$ for all i

possible scenarios

- in general $\Delta_{\text{pos}} > 0$ and $\Delta_{\text{int}} > 0$
- overparametrized models may have $\Delta_{\text{pos}} = \Delta_{\text{int}} = 0$
- both $\Delta_{\text{pos}} = 0, \Delta_{\text{int}} > 0$ and $\Delta_{\text{pos}} > 0, \Delta_{\text{int}} = 0$ feasible in theory

Convergence results: convex

$$x^{k+1} = x^k - \gamma_k \nabla f_{i_k}(x^k), \qquad \gamma_k = \frac{\sigma}{1 + \frac{\sigma}{2f_{i_k}(x^k)} \|\nabla f_{i_k}(x^k)\|^2}$$

thm: assume each f_i nonnegative, convex and L_i-smooth, and let $L = \max_i L_i$, then for any $\sigma > 0$

$$\mathbf{E}\left[f(\bar{x}^K) - f(x^*)\right] \leq \frac{\mathbf{E}\|x^0 - x^*\|^2}{\eta_\sigma K} + 3\sigma L \cdot (1 + \sigma L)\Delta_{\text{int}} + \sigma L \cdot \max\{0, 2\sigma L - 1\}\,\Delta_{\text{pos}}$$

where $\bar{x}^K = \frac{1}{K}\sum_{k=0}^{K-1} x^k$ and $\eta_\sigma := \frac{2\sigma}{(1+2\sigma L)^2}$

- if $\Delta_{\text{int}} = \Delta_{\text{pos}} = 0$, then convergence for any $\sigma > 0$
- reducing σ: $\sigma_k = \sigma/\sqrt{k+1}$ leads to $O(\log(k)/\sqrt{k})$ rate

Convergence results: nonconvex

$$x^{k+1} = x^k - \gamma_k \nabla f_{i_k}(x^k), \qquad \gamma_k = \frac{\sigma}{1 + \frac{\sigma}{2f_{i_k}(x^k)} \|\nabla f_{i_k}(x^k)\|^2}$$

thm: assume each f_i nonnegative, L_i-smooth, and let $L = \max_i L_i$, then for $\sigma \leq 1/(2L)$

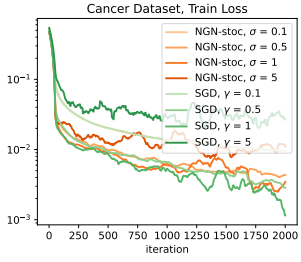
$$\mathbf{E}\left[\frac{1}{K}\sum_{k=0}^{K-1} \|\nabla f(x^k)\|^2\right] \leq \frac{12 \cdot \mathbf{E}[f(x^0) - f^*]}{\sigma K} + 18\sigma L \Delta_{\text{noise}}^2$$

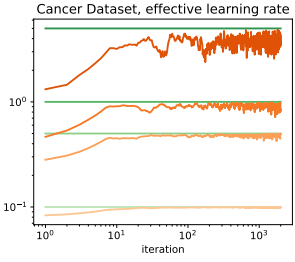
where $\Delta_{\text{noise}}^2 = \sup_x \mathbf{E}[\|\nabla f(x) - \nabla f_i(x)\|^2]$.

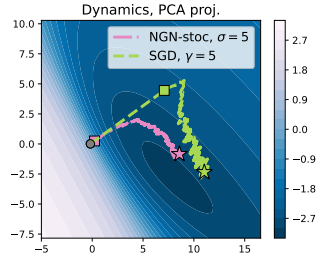
- decreasing $\sigma_k = \sigma/\sqrt{k+1}$ leads to $O(log(k)/\sqrt{k})$ rate

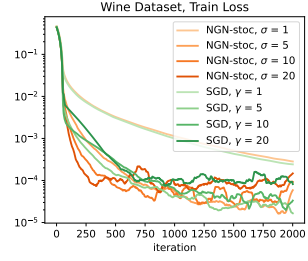
Numerical experiments: I

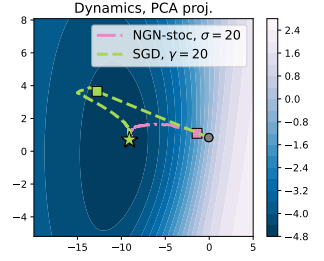
logistic regression on LIBSVM datasets (Breast Cancer, Wine)



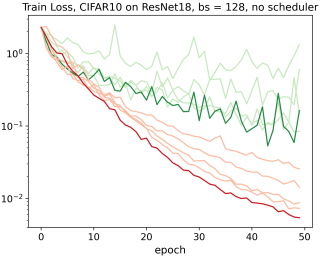


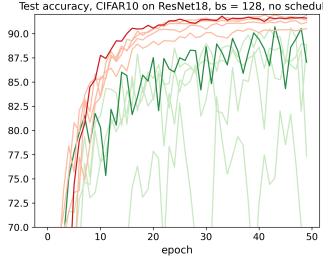


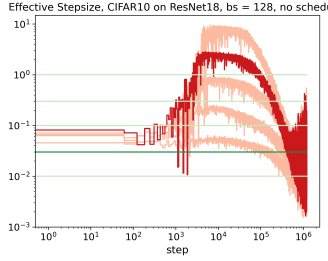




Numerical experiments: II



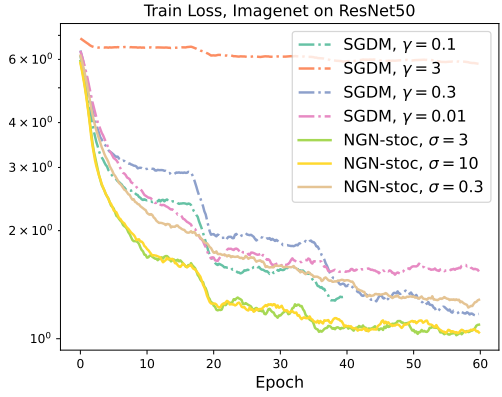


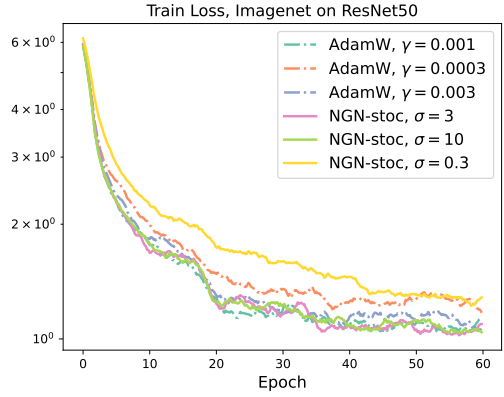


training ResNet18 (11 million parameters) on CIFAR10 dataset

- green: SGD with step sizes $\gamma \in \{0.01,\ 0.03,\ 0.1,\ 0.3,\ 1\}$
- red: stochastic NGN with $\sigma \in \{0.1,\ 0.3,\ 1,\ 3,\ 10\}$

Numerical experiments: III





training ResNet50 (23 million parameters) on ImageNet dataset

- **training loss**: better than SGDM, similar to AdamW
- **test performance**: slightly worse than AdamW

Outline

- minimizing single nonnegative function
  - nonnegative Gauss-Newton step size rule
  - connection with Polyak step size
- generalized Gauss-Newton (prox-linear)
- stochastic NGN method
- extensions and summary

Matrix update version

- regularizd Gauss-Newton step (Levenberg-Marquardt)

$$p^k = \arg\min_p \frac{1}{N}\sum_{i=1}^{N}\left(r_i(x) + \nabla r_i(x)^\top p\right)^2 + \frac{1}{2\sigma}\|p\|^2$$

- general form: $x^{k+1} = x^k - G_\sigma(x^k)^{-1}\nabla f(x^k)$, where

$$G_\sigma(x) = \frac{1}{\sigma}I + \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2f_i(x)}\nabla f_i(x)\nabla f_i(x)^\top$$

- reducing storage and computation cost:
  - sampling subsets, low-rank update
  - diagonal or block-diagonal approximations

Summary

- exploiting nonnegativity to derive adaptive step size rule
  - create virtual structure to use Gauss-Newton trick
  - close connection with Polyak step size
- extension to generalized Gauss-Newton setting
- convergence theory + promising empirical results

current and future work

- experiments and analysis of NGN + momentum
- reduction schedule of σ using ideas from trust-region methods