Fast and Furious Convergence: Stochastic Second-Order Methods under Interpolation

Joint work with:

Sharan Vaswani*
(Mila, UdeM)

Issam Laradji
(UBC, Element AI)

Mark Schmidt
(UBC)

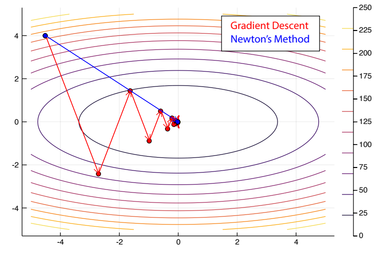
Simon Lacoste-Julien
(Mila, UdeM)

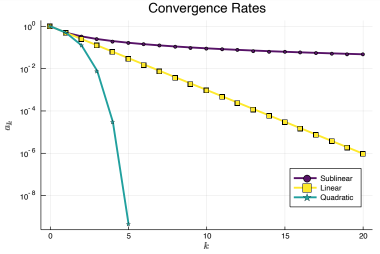
Si Yi (Cathy) Meng*

AISTATS 2020

Sub-sampled Newton's method under interpolation

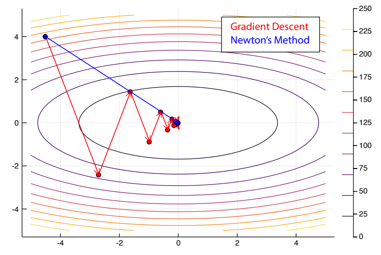
- First order methods:
  - Cheap iterations.
  - Slow convergence for ill-conditioned problems.
- Second order methods:
  - Faster convergence by explicitly adapting to the local curvature of the objective.
  - Forming the Hessian and computing the update direction is expensive.

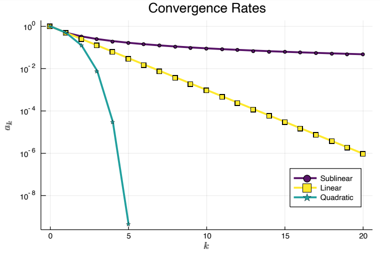




Sub-sampled Newton's method under interpolation

- First order methods:
  - Cheap iterations.
  - Slow convergence for ill-conditioned problems.
- Second order methods:
  - Faster convergence by explicitly adapting to the local curvature of the objective.
  - Forming the Hessian and computing the update direction is expensive.
- Sub-sampling the training set:
  - Reduces the iteration cost.
  - Slower convergence due to approximate update direction.





Sub-sampled Newton's method under interpolation

- In modern ML applications we often use overparameterized models that satisfies the **interpolation** condition.
  - Means that they can **complete fit the training data**.
- Examples:
  - Logistic regression on linearly-separable data.
  - Non-parametric regression.
  - Boosting.
  - Over-parameterized neural networks.

Sub-sampled Newton's method under interpolation

- In modern ML applications we often use overparameterized models that satisfies the **interpolation** condition.
  - Means that they can **complete fit the training data**.
- Examples:
  - Logistic regression on linearly-separable data.
  - Non-parametric regression.
  - Boosting.
  - Over-parameterized neural networks.
- It's been shown that sub-sampled first-order methods converge faster under interpolation [Vaswani et al., 2019a].

Sub-sampled Newton's method under interpolation

- In modern ML applications we often use overparameterized models that satisfies the **interpolation** condition.
  - Means that they can **complete fit the training data**.
- Examples:
  - Logistic regression on linearly-separable data.
  - Non-parametric regression.
  - Boosting.
  - Over-parameterized neural networks.
- It's been shown that sub-sampled first-order methods converge faster under interpolation [Vaswani et al., 2019a].

What's the behaviour of sub-sampled Newton's method in this setting?

Setup

Unconstrained minimization: finite-sum objective.

$$\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

where the f_i's are twice continuously differentiable, and n is the number of training examples.

- f is μ-strongly convex and L-smooth, $\implies \mu I \preceq \nabla^2 f(w) \preceq LI$.

Setup

Unconstrained minimization: finite-sum objective.

$$\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

where the f_i's are twice continuously differentiable, and n is the number of training examples.

- f is μ-strongly convex and L-smooth, $\implies \mu I \preceq \nabla^2 f(w) \preceq LI$.
- Define $\bar{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mu_i$ and $\bar{L} = \frac{1}{n} \sum_{i=1}^{n} L_i$.
  $\implies$ For any sub-sample S, the function $\frac{1}{|S|} \sum_{i \in S} f_i$ is L_S-smooth and μ_S-strongly convex.

Setup

Unconstrained minimization: finite-sum objective.

$$\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

where the f_i's are twice continuously differentiable, and n is the number of training examples.

- f is μ-strongly convex and L-smooth, $\implies \mu I \preceq \nabla^2 f(w) \preceq LI$.
- Define $\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mu_i$ and $\bar{L} = \frac{1}{n}\sum_{i=1}^{n} L_i$.
  $\implies$ For any sub-sample S, the function $\frac{1}{|S|}\sum_{i \in S} f_i$ is L_S-smooth and μ_S-strongly convex.
  Define $\tilde{\mu} = \min_S \mu_S \geq 0$ and $\tilde{L} = \max_S L_S$.

Setup

Unconstrained minimization: finite-sum objective.

$$\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

where the f_i's are twice continuously differentiable, and n is the number of training examples.

- f is μ-strongly convex and L-smooth, $\implies \mu I \preceq \nabla^2 f(w) \preceq LI$.
- Define $\bar{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mu_i$ and $\bar{L} = \frac{1}{n} \sum_{i=1}^{n} L_i$.
  $\implies$ For any sub-sample S, the function $\frac{1}{|S|} \sum_{i \in S} f_i$ is L_S-smooth and μ_S-strongly convex.
  Define $\tilde{\mu} = \min_S \mu_S \geq 0$ and $\tilde{L} = \max_S L_S$.
- **Interpolation:** $\nabla f(w^*) = 0 \implies \nabla f_i(w^*) = 0$ for all i. For smooth, strongly convex, finite-sum objectives, interpolation $\implies$ **strong growth condition**:

$$\rho\textbf{-SGC:} \quad \mathbb{E}_i \|\nabla f_i(w)\|^2 \leq \rho \|\nabla f(w)\|^2.$$

Regularized sub-sampled Newton method (R-SSN)

Algorithm

Stochastic gradient descent (SGD)

$$w_{k+1} = w_k - \eta_k \underbrace{\nabla f_{\mathcal{G}_k}(w_k)}_{\text{subsampled gradient}}$$

- η_k is the step size.

Algorithm

Stochastic gradient descent (SGD)

$$w_{k+1} = w_k - \eta_k \underbrace{\nabla f_{\mathcal{G}_k}(w_k)}_{\text{subsampled gradient}}$$

- η_k is the step size.
- $\mathcal{G}_k \subseteq [n]$ is chosen uniformly at random.

Algorithm

Stochastic gradient descent (SGD)

$$w_{k+1} = w_k - \eta_k \underbrace{\nabla f_{\mathcal{G}_k}(w_k)}_{\text{subsampled gradient}}$$

- η_k is the step size.
- $\mathcal{G}_k \subseteq [n]$ is chosen uniformly at random.
- Sub-sampled gradient:

$$\nabla f_{\mathcal{G}_k}(w_k) = \frac{1}{b_{g_k}} \sum_{i \in \mathcal{G}_k} \nabla f_i(w_k)$$

Algorithm

Regularized sub-sampled Newton method (R-SSN)

$$w_{k+1} = w_k - \eta_k \underbrace{\left[\mathbf{H}_{\mathcal{S}_k}(w_k)\right]^{-1} \nabla f_{\mathcal{G}_k}(w_k)}_{\text{regularized sub-sampled Newton direction}}$$

- η_k is the step size.
- $\mathcal{G}_k$, $\mathcal{S}_k \subseteq [n]$ are index sets chosen independently, uniformly at random.
- Sub-sampled gradient:

$$\nabla f_{\mathcal{G}_k}(w_k) = \frac{1}{b_{g_k}} \sum_{i \in \mathcal{G}_k} \nabla f_i(w_k)$$

- Levenberg-Marquardt (LM)-regularized sub-sampled Hessian:

$$\mathbf{H}_{\mathcal{S}_k}(w_k) = \frac{1}{b_{s_k}} \sum_{i \in \mathcal{S}_k} \nabla^2 f_i(w_k) + \tau I_d$$

Theorem I - Global convergence

- Similar to SGD [Vaswani et al., 2019a], we show that interpolation allows R-SSN with a constant batch size to obtain global Q-linear convergence rate.

Theorem I - Global convergence

- Similar to SGD [Vaswani et al., 2019a], we show that interpolation allows R-SSN with a constant batch size to obtain global Q-linear convergence rate.

Global linear convergence

Under (a) μ-strong convexity, (b) L-smoothness, (c) $[\tilde{\mu} + \tau, \tilde{L} + \tau]$-bounded eigenvalues of the regularized sub-sampled Hessian and (d) ρ-SGC, R-SSN with constant batch sizes converges at a **Q-linear** rate

$$\mathbb{E}[f(w_T)] - f(w^*) \leq (1 - \alpha)^T \left(f(w_0) - f(w^*)\right)$$

where $\alpha = \min\left\{\frac{(\bar{\mu}+\tau)^2}{2\kappa c_g\left(\tilde{L}+\tau\right)}, \frac{(\bar{\mu}+\tau)}{2\kappa\left(\tilde{L}+\tau\right)}\right\}$, $\kappa = \frac{L}{\mu}$ and $c_g = \frac{(\rho-1)\,(n-b_g)}{b_g(n-1)}$.

Theorem I - Global convergence

- Similar to SGD [Vaswani et al., 2019a], we show that interpolation allows R-SSN with a constant batch size to obtain global Q-linear convergence rate.

Global linear convergence

Under (a) μ-strong convexity, (b) L-smoothness, (c) $[\tilde{\mu} + \tau, \tilde{L} + \tau]$-bounded eigenvalues of the regularized sub-sampled Hessian and (d) ρ-SGC, R-SSN with constant batch sizes converges at a **Q-linear** rate

$$\mathbb{E}[f(w_T)] - f(w^*) \leq (1 - \alpha)^T \left(f(w_0) - f(w^*)\right)$$

where $\alpha = \min\left\{\frac{(\bar{\mu}+\tau)^2}{2\kappa c_g(\tilde{L}+\tau)}, \frac{(\bar{\mu}+\tau)}{2\kappa(\tilde{L}+\tau)}\right\}$, $\kappa = \frac{L}{\mu}$ and $c_g = \frac{(\rho-1)(n-b_g)}{b_g(n-1)}$.

Theorem I - Global convergence

- Similar to SGD [Vaswani et al., 2019a], we show that interpolation allows R-SSN with a constant batch size to obtain global Q-linear convergence rate.

Global linear convergence

Under (a) μ-strong convexity, (b) L-smoothness, (c) $[\tilde{\mu}+\tau, \tilde{L}+\tau]$-bounded eigenvalues of the regularized sub-sampled Hessian and (d) ρ-SGC, R-SSN with constant batch sizes converges at a **Q-linear** rate

$$\mathbb{E}[f(w_T)] - f(w^*) \leq (1-\alpha)^T \left(f(w_0) - f(w^*)\right)$$

where $\alpha = \min\left\{\frac{(\bar{\mu}+\tau)^2}{2\kappa c_g(\tilde{L}+\tau)}, \frac{(\bar{\mu}+\tau)}{2\kappa(\tilde{L}+\tau)}\right\}$, $\kappa = \frac{L}{\mu}$ and $c_g = \frac{(\rho-1)(n-b_g)}{b_g(n-1)}$.

Theorem I - Global convergence

- Similar to SGD [Vaswani et al., 2019a], we show that interpolation allows R-SSN with a constant batch size to obtain global Q-linear convergence rate.

Global linear convergence

Under (a) μ-strong convexity, (b) L-smoothness, (c) $[\tilde{\mu}+\tau, \tilde{L}+\tau]$-bounded eigenvalues of the regularized sub-sampled Hessian and (d) ρ-SGC, R-SSN with constant batch sizes converges at a **Q-linear** rate

$$\mathbb{E}[f(w_T)] - f(w^*) \leq (1-\alpha)^T \left(f(w_0) - f(w^*)\right)$$

where $\alpha = \min\left\{\frac{(\bar{\mu}+\tau)^2}{2\kappa c_g(\tilde{L}+\tau)}, \frac{(\bar{\mu}+\tau)}{2\kappa(\tilde{L}+\tau)}\right\}$, $\kappa = \frac{L}{\mu}$ and $c_g = \frac{(\rho-1)(n-b_g)}{b_g(n-1)}$.

Theorem I - Global convergence

- Similar to SGD [Vaswani et al., 2019a], we show that interpolation allows R-SSN with a constant batch size to obtain global Q-linear convergence rate.

Global linear convergence

Under (a) μ-strong convexity, (b) L-smoothness, (c) $[\tilde{\mu} + \tau, \tilde{L} + \tau]$-bounded eigenvalues of the regularized sub-sampled Hessian and (d) ρ-SGC, R-SSN with constant batch sizes converges at a **Q-linear** rate

$$\mathbb{E}[f(w_T)] - f(w^*) \leq (1 - \alpha)^T \left(f(w_0) - f(w^*)\right)$$

where $\alpha = \min\left\{\frac{(\bar{\mu}+\tau)^2}{2\kappa c_g(\tilde{L}+\tau)}, \frac{(\bar{\mu}+\tau)}{2\kappa(\tilde{L}+\tau)}\right\}$, $\kappa = \frac{L}{\mu}$ and $c_g = \frac{(\rho-1)(n-b_g)}{b_g(n-1)}$.

Theorem I - Global convergence

- Similar to SGD [Vaswani et al., 2019a], we show that interpolation allows R-SSN with a constant batch size to obtain global Q-linear convergence rate.

Global linear convergence

Under (a) μ-strong convexity, (b) L-smoothness, (c) $[\tilde{\mu} + \tau, \tilde{L} + \tau]$-bounded eigenvalues of the regularized sub-sampled Hessian and (d) ρ-SGC, R-SSN with constant batch sizes converges at a **Q-linear** rate

$$\mathbb{E}[f(w_T)] - f(w^*) \leq (1 - \alpha)^T \left(f(w_0) - f(w^*)\right)$$

where $\alpha = \min\left\{\frac{(\bar{\mu}+\tau)^2}{2\kappa c_g(\tilde{L}+\tau)}, \frac{(\bar{\mu}+\tau)}{2\kappa(\tilde{L}+\tau)}\right\}$, $\kappa = \frac{L}{\mu}$ and $c_g = \frac{(\rho-1)\,(n-b_g)}{b_g(n-1)}$.

Theorem I - Global convergence

- Similar to SGD [Vaswani et al., 2019a], we show that interpolation allows R-SSN with a constant batch size to obtain global Q-linear convergence rate.

Global linear convergence

Under (a) μ-strong convexity, (b) L-smoothness, (c) $[\tilde{\mu}+\tau, \tilde{L}+\tau]$-bounded eigenvalues of the regularized sub-sampled Hessian and (d) ρ-SGC, R-SSN with constant batch sizes converges at a **Q-linear** rate

$$\mathbb{E}[f(w_T)] - f(w^*) \leq (1-\alpha)^T \left(f(w_0) - f(w^*)\right)$$

where $\alpha = \min\left\{\frac{(\bar{\mu}+\tau)^2}{2\kappa c_g(\tilde{L}+\tau)}, \frac{(\bar{\mu}+\tau)}{2\kappa(\tilde{L}+\tau)}\right\}$, $\kappa = \frac{L}{\mu}$ and $c_g = \frac{(\rho-1)(n-b_g)}{b_g(n-1)}$.

- If $b_g = b_s = n$ (full-batch) and $\tau = 0$, we recover deterministic rate.
- In the absence of interpolation, SSN can only achieve an R-linear rate with geometrically increasing batch size for the sub-sampled gradient [Bollapragada et al., 2018a].

Theorem II - Local convergence

- With interpolation, we obtain linear-quadratic convergence in expectation with a geometric batch growth.

Theorem II - Local convergence

- With interpolation, we obtain linear-quadratic convergence in expectation with a geometric batch growth.

Local linear-quadratic convergence

Under the same assumptions (a) - (d) of Theorem I, along with (e) *M-Lipschitz continuity of the Hessian*, (f) bounded moments of iterates, and (g) σ-bounded variance of the regularized sub-sampled Hessian, R-SSN with (i) unit step size $\eta_k = 1$ and (ii) growing batch sizes satisfying

$$b_{g_k} \geq \frac{n}{\left(\frac{n-1}{\rho-1}\right) \|\nabla f(w_k)\|^2 + 1}, \quad \text{and} \quad b_{s_k} \geq \frac{n}{\frac{n}{\sigma^2} \|\nabla f(w_k)\| + 1}$$

converges to w^* in a local neighbourhood $\|w_0 - w^*\| \leq \delta$ at a **linear-quadratic** rate

$$\mathbb{E}\,\|w_{k+1} - w^*\| \leq c_1 \left(\mathbb{E}\,\|w_k - w^*\|\right)^2 + c_2\,\mathbb{E}\,\|w_k - w^*\| \quad \text{for some } c_1 > 0 \text{ and } c_2 \in (0,1).$$

Theorem II - Local convergence

- With interpolation, we obtain linear-quadratic convergence in expectation with a geometric batch growth.

Local linear-quadratic convergence

Under the same assumptions (a) - (d) of Theorem I, along with (e) M-Lipschitz continuity of the Hessian, (f) bounded moments of iterates, and (g) σ-bounded variance of the regularized sub-sampled Hessian, R-SSN with (i) unit step size $\eta_k = 1$ and (ii) growing batch sizes satisfying

$$b_{g_k} \geq \frac{n}{\left(\frac{n-1}{\rho-1}\right) \|\nabla f(w_k)\|^2 + 1}, \quad \text{and} \quad b_{s_k} \geq \frac{n}{\frac{n}{\sigma^2} \|\nabla f(w_k)\| + 1}$$

converges to w^* in a local neighbourhood $\|w_0 - w^*\| \leq \delta$ at a **linear-quadratic** rate

$$\mathbb{E}\, \|w_{k+1} - w^*\| \leq c_1 \left(\mathbb{E}\, \|w_k - w^*\|\right)^2 + c_2\, \mathbb{E}\, \|w_k - w^*\| \quad \text{for some } c_1 > 0 \text{ and } c_2 \in (0,1).$$

Theorem II - Local convergence

- With interpolation, we obtain linear-quadratic convergence in expectation with a geometric batch growth.

Local linear-quadratic convergence

Under the same assumptions (a) - (d) of Theorem I, along with (e) M-Lipschitz continuity of the Hessian, (f) bounded moments of iterates, and (g) σ-bounded variance of the regularized sub-sampled Hessian, R-SSN with (i) unit step size $\eta_k = 1$ and (ii) growing batch sizes satisfying

$$b_{g_k} \geq \frac{n}{\left(\frac{n-1}{\rho-1}\right) \|\nabla f(w_k)\|^2 + 1}, \quad \text{and} \quad b_{s_k} \geq \frac{n}{\frac{n}{\sigma^2} \|\nabla f(w_k)\| + 1}$$

converges to w^* in a local neighbourhood $\|w_0 - w^*\| \leq \delta$ at a **linear-quadratic** rate

$$\mathbb{E}\,\|w_{k+1} - w^*\| \leq c_1 \left(\mathbb{E}\,\|w_k - w^*\|\right)^2 + c_2\,\mathbb{E}\,\|w_k - w^*\| \quad \text{for some } c_1 > 0 \text{ and } c_2 \in (0,1).$$

Theorem II - Local convergence

- With interpolation, we obtain linear-quadratic convergence in expectation with a geometric batch growth.

Local linear-quadratic convergence

Under the same assumptions (a) - (d) of Theorem I, along with (e) M-Lipschitz continuity of the Hessian, (f) bounded moments of iterates, and (g) σ-bounded variance of the regularized sub-sampled Hessian, R-SSN with (i) unit step size $\eta_k = 1$ and (ii) growing batch sizes satisfying

$$b_{g_k} \geq \frac{n}{\left(\frac{n-1}{\rho-1}\right) \|\nabla f(w_k)\|^2 + 1}, \quad \text{and} \quad b_{s_k} \geq \frac{n}{\frac{n}{\sigma^2} \|\nabla f(w_k)\| + 1}$$

converges to w^* in a local neighbourhood $\|w_0 - w^*\| \leq \delta$ at a **linear-quadratic** rate

$$\mathbb{E}\,\|w_{k+1} - w^*\| \leq c_1 \left(\mathbb{E}\,\|w_k - w^*\|\right)^2 + c_2\, \mathbb{E}\,\|w_k - w^*\| \quad \text{for some } c_1 > 0 \text{ and } c_2 \in (0,1).$$

Theorem II - Local convergence

- With interpolation, we obtain linear-quadratic convergence in expectation with a geometric batch growth.

Local linear-quadratic convergence

Under the same assumptions (a) - (d) of Theorem I, along with (e) M-Lipschitz continuity of the Hessian, (f) bounded moments of iterates, and (g) σ-bounded variance of the regularized sub-sampled Hessian, R-SSN with (i) unit step size $\eta_k = 1$ and (ii) growing batch sizes satisfying

$$b_{g_k} \geq \frac{n}{\left(\frac{n-1}{\rho-1}\right) \|\nabla f(w_k)\|^2 + 1}, \quad \text{and} \quad b_{s_k} \geq \frac{n}{\frac{n}{\sigma^2} \|\nabla f(w_k)\| + 1}$$

converges to w^* in a local neighbourhood $\|w_0 - w^*\| \leq \delta$ at a **linear-quadratic** rate

$$\mathbb{E}\,\|w_{k+1} - w^*\| \leq c_1 \left(\mathbb{E}\,\|w_k - w^*\|\right)^2 + c_2\,\mathbb{E}\,\|w_k - w^*\| \quad \text{for some } c_1 > 0 \text{ and } c_2 \in (0, 1).$$

- Rate of growth for $\mathcal{G}_k$ is the same as that's required to obtain linear convergence by SGD without variance reduction or interpolation [Friedlander and Schmidt, 2012].

Theorem II - Local convergence

- With interpolation, we obtain linear-quadratic convergence in expectation with a geometric batch growth.

Local linear-quadratic convergence

Under the same assumptions (a) - (d) of Theorem I, along with (e) M-Lipschitz continuity of the Hessian, (f) bounded moments of iterates, and (g) σ-bounded variance of the regularized sub-sampled Hessian, R-SSN with (i) unit step size $\eta_k = 1$ and (ii) growing batch sizes satisfying

$$b_{g_k} \geq \frac{n}{\left(\frac{n-1}{\rho-1}\right) \|\nabla f(w_k)\|^2 + 1}, \quad \text{and} \quad b_{s_k} \geq \frac{n}{\frac{n}{\sigma^2} \|\nabla f(w_k)\| + 1}$$

converges to w^* in a local neighbourhood $\|w_0 - w^*\| \leq \delta$ at a **linear-quadratic** rate

$$\mathbb{E}\,\|w_{k+1} - w^*\| \leq c_1 \left(\mathbb{E}\,\|w_k - w^*\|\right)^2 + c_2\,\mathbb{E}\,\|w_k - w^*\| \quad \text{for some } c_1 > 0 \text{ and } c_2 \in (0,1).$$

- Rate of growth for $\mathcal{G}_k$ is the same as that's required to obtain linear convergence by SGD without variance reduction or interpolation [Friedlander and Schmidt, 2012].
- In the absence of interpolation, SSN can only achieve an asymptotic superlinear rate, with batch size $\mathcal{G}_k$ growing faster than a geometric rate [Bollapragada et al., 2018a].

Corollary

- If we decay the regularization sequence, we can obtain a stronger result, similar to the quadratic convergence of Newton's method in the deterministic setting.

Local quadratic convergence for decaying τ_k

Under the same assumptions as Theorem II, if we decrease the regularization term as $\tau_k \leq \|\nabla f(w_k)\|$, R-SSN can achieve local quadratic convergence

$$\mathbb{E}\|w_{k+1} - w^*\| \leq c_3 \left(\mathbb{E}\|w_k - w^*\|\right)^2 \quad \text{for some } c_3 \geq 0.$$

Corollary

- If we decay the regularization sequence, we can obtain a stronger result, similar to the quadratic convergence of Newton's method in the deterministic setting.

Local quadratic convergence for decaying τ_k

Under the same assumptions as Theorem II, if we decrease the regularization term as $\tau_k \leq \|\nabla f(w_k)\|$, R-SSN can achieve local quadratic convergence

$$\mathbb{E}\|w_{k+1} - w^*\| \leq c_3 \left(\mathbb{E}\|w_k - w^*\|\right)^2 \quad \text{for some } c_3 \geq 0.$$

- This decay rate is inversely proportional to the growth of the batch size for the sub-sampled Hessian, $b_{s_k} \geq \frac{n}{\frac{n}{\sigma^2}\|\nabla f(w_k)\|+1}$
  - Larger batch sizes require smaller regularization.

Self-concordance

Self-concordance

- Newton's method is invariant to affine transformations of the parameters [Boyd and Vandenberghe, 2004].
    - But this is not reflected in the classical analysis – the convergence rate obtained depends on the strong-convexity and Lipschitz constants that change with affine transformations.

Self-concordance

- Newton's method is invariant to affine transformations of the parameters [Boyd and Vandenberghe, 2004].
  - But this is not reflected in the classical analysis – the convergence rate obtained depends on the strong-convexity and Lipschitz constants that change with affine transformations.
- However, for self-concordant functions, the analysis yields an affine-invariant rate in the deterministic case.

Self-concordance

- Newton's method is invariant to affine transformations of the parameters [Boyd and Vandenberghe, 2004].
    - But this is not reflected in the classical analysis – the convergence rate obtained depends on the strong-convexity and Lipschitz constants that change with affine transformations.
- However, for self-concordant functions, the analysis yields an affine-invariant rate in the deterministic case.

Definition 1 (Self-concordance)

A convex function $f : \mathbb{R} \to \mathbb{R}$ is self-concordant if for all $w \in \mathbb{R}$,

$$|f'''(w)| \leq 2[f''(w)]^{3/2}.$$

Self-concordance

- Newton's method is invariant to affine transformations of the parameters [Boyd and Vandenberghe, 2004].
  - But this is not reflected in the classical analysis – the convergence rate obtained depends on the strong-convexity and Lipschitz constants that change with affine transformations.
- However, for self-concordant functions, the analysis yields an affine-invariant rate in the deterministic case.

Definition 1 (Self-concordance)

A convex function $f : \mathbb{R} \to \mathbb{R}$ is self-concordant if for all $w \in \mathbb{R}$,

$$|f'''(w)| \leq 2[f''(w)]^{3/2}.$$

Can we obtain an affine-invariant rate for R-SSN under self-concordance and interpolation?

R-SSN under self-concordance

Regularized Newton decrement:

$$\lambda := \|\nabla f(w)\|_{[\nabla^2 f(w)+\tau I]^{-1}} = \left\langle \nabla f(w)\,,\, [\nabla^2 f(w) + \tau I]^{-1} \nabla f(w) \right\rangle^{1/2}$$

Regularized stochastic Newton decrement:

$$\tilde{\lambda} := \|\nabla f_i(w)\|_{[\mathbf{H}_j(w)]^{-1}} = \left\langle \nabla f_i(w)\,,\, [\mathbf{H}_j(w)]^{-1} \nabla f_i(w) \right\rangle^{1/2}$$

R-SSN under self-concordance

Regularized Newton decrement:

$$\lambda := \|\nabla f(w)\|_{[\nabla^2 f(w)+\tau I]^{-1}} = \left\langle \nabla f(w) \, , \, [\nabla^2 f(w) + \tau I]^{-1} \nabla f(w) \right\rangle^{1/2}$$

Regularized stochastic Newton decrement:

$$\tilde{\lambda} := \|\nabla f_i(w)\|_{[\mathbf{H}_j(w)]^{-1}} = \left\langle \nabla f_i(w) \, , \, [\mathbf{H}_j(w)]^{-1} \nabla f_i(w) \right\rangle^{1/2}$$

Newton-decrement SGC:

$$\mathbb{E}_i[\tilde{\lambda}^2] \leq \rho_{nd} \lambda^2, \quad \text{for all } w, j.$$

R-SSN under self-concordance

Regularized Newton decrement:

$$\lambda := \|\nabla f(w)\|_{[\nabla^2 f(w)+\tau I]^{-1}} = \left\langle \nabla f(w)\,,\, [\nabla^2 f(w) + \tau I]^{-1} \nabla f(w) \right\rangle^{1/2}$$

Regularized stochastic Newton decrement:

$$\tilde{\lambda} := \|\nabla f_i(w)\|_{[\mathbf{H}_j(w)]^{-1}} = \left\langle \nabla f_i(w)\,,\, [\mathbf{H}_j(w)]^{-1} \nabla f_i(w) \right\rangle^{1/2}$$

Newton-decrement SGC:

$$\mathbb{E}_i[\tilde{\lambda}^2] \leq \rho_{nd} \lambda^2, \quad \text{for all } w, j.$$

Modified R-SSN update

$$w_{k+1} = w_k - \frac{c\,\eta}{1 + \eta \tilde{\lambda}_k} [\mathbf{H}_j(w_k)]^{-1} \nabla f_i(w_k)$$

where $\tilde{\lambda}_k$ is $\tilde{\lambda}$ evaluated at w_k.

Theorem III - R-SSN under self-concordance

Two-phased analysis

Under (a) self-concordance (b) L-smoothness, (c) $[\tilde{\mu}+\tau, \tilde{L}+\tau]$-bounded values of the regularized sub-sampled Hessian, (d) ρ_{nd}-Newton decrement SGC with $\rho_{nd} = \frac{\rho L}{\tilde{\mu}+\tau}$, and (e) bounded iterates $\|w - w^*\| \leq D$, then there exists $c \in (0, 1]$ and a constant step size η such that the first phase $\{w_k\}_{k \in [0,m]}$ converges at an R-linear rate

$$\mathbb{E}[f(w_{k+1})] \leq f(w_k) - \epsilon_k,$$

where ϵ_k is some positive sequence. Furthermore, in a local neighbourhood where $\lambda_m \leq 1/6$, the sequence $\{w_k\}_{k \geq m}$ converges to w^* at a Q-linear rate

$$\mathbb{E}\left[f(w_T)\right] - f(w^*) \leq (1-\beta)^{T-m}\left(\mathbb{E}\left[f(w_m)\right] - f(w^*)\right),$$

where $\beta \in (0, 1)$.

Theorem III - R-SSN under self-concordance

Two-phased analysis

Under (a) self-concordance (b) L-smoothness, (c) $[\tilde{\mu} + \tau, \tilde{L} + \tau]$-bounded values of the regularized sub-sampled Hessian, (d) ρ_{nd}-Newton decrement SGC with $\rho_{nd} = \frac{\rho L}{\tilde{\mu}+\tau}$, and (e) bounded iterates $\|w - w^*\| \leq D$, then there exists $c \in (0, 1]$ and a constant step size η such that the first phase $\{w_k\}_{k \in [0,m]}$ converges at an R-linear rate

$$\mathbb{E}[f(w_{k+1})] \leq f(w_k) - \epsilon_k,$$

where ϵ_k is some positive sequence. Furthermore, in a local neighbourhood where $\lambda_m \leq 1/6$, the sequence $\{w_k\}_{k \geq m}$ converges to w^* at a Q-linear rate

$$\mathbb{E}\left[f(w_T)\right] - f(w^*) \leq (1 - \beta)^{T-m} \left(\mathbb{E}\left[f(w_m)\right] - f(w^*)\right),$$

where $\beta \in (0, 1)$.

Theorem III - R-SSN under self-concordance

Two-phased analysis

Under (a) self-concordance (b) L-smoothness, (c) $[\tilde{\mu}+\tau, \tilde{L}+\tau]$-bounded values of the regularized sub-sampled Hessian, (d) ρ_{nd}-Newton decrement SGC with $\rho_{nd} = \frac{\rho L}{\tilde{\mu}+\tau}$, and (e) bounded iterates $\|w - w^*\| \leq D$, then there exists $c \in (0, 1]$ and a constant step size η such that the first phase $\{w_k\}_{k\in[0,m]}$ converges at an R-linear rate

$$\mathbb{E}[f(w_{k+1})] \leq f(w_k) - \epsilon_k,$$

where ϵ_k is some positive sequence. Furthermore, in a local neighbourhood where $\lambda_m \leq 1/6$, the sequence $\{w_k\}_{k\geq m}$ converges to w^* at a Q-linear rate

$$\mathbb{E}\left[f(w_T)\right] - f(w^*) \leq (1-\beta)^{T-m}\left(\mathbb{E}\left[f(w_m)\right] - f(w^*)\right),$$

where $\beta \in (0, 1)$.

Theorem III - R-SSN under self-concordance

Two-phased analysis

Under (a) self-concordance (b) L-smoothness, (c) $[\tilde{\mu} + \tau, \tilde{L} + \tau]$-bounded values of the regularized sub-sampled Hessian, (d) ρ_{nd}-Newton decrement SGC with $\rho_{nd} = \frac{\rho L}{\tilde{\mu}+\tau}$, and (e) bounded iterates $\|w - w^*\| \leq D$, then there exists $c \in (0, 1]$ and a constant step size η such that the first phase $\{w_k\}_{k\in[0,m]}$ converges at an R-linear rate

$$\mathbb{E}[f(w_{k+1})] \leq f(w_k) - \epsilon_k,$$

where ϵ_k is some positive sequence. Furthermore, in a local neighbourhood where $\lambda_m \leq 1/6$, the sequence $\{w_k\}_{k\geq m}$ converges to w^* at a Q-linear rate

$$\mathbb{E}\left[f(w_T)\right] - f(w^*) \leq (1-\beta)^{T-m} \left(\mathbb{E}\left[f(w_m)\right] - f(w^*)\right),$$

where $\beta \in (0, 1)$.

Theorem III - R-SSN under self-concordance

Two-phased analysis

Under (a) self-concordance (b) L-smoothness, (c) $[\tilde{\mu}+\tau, \tilde{L}+\tau]$-bounded values of the regularized sub-sampled Hessian, (d) ρ_{nd}-Newton decrement SGC with $\rho_{nd} = \frac{\rho L}{\tilde{\mu}+\tau}$, and (e) bounded iterates $\|w - w^*\| \leq D$, then there exists $c \in (0, 1]$ and a constant step size η such that the first phase $\{w_k\}_{k\in[0,m]}$ converges at an R-linear rate

$$\mathbb{E}[f(w_{k+1})] \leq f(w_k) - \epsilon_k,$$

where ϵ_k is some positive sequence. Furthermore, in a local neighbourhood where $\lambda_m \leq 1/6$, the sequence $\{w_k\}_{k\geq m}$ converges to w^* at a Q-linear rate

$$\mathbb{E}\left[f(w_T)\right] - f(w^*) \leq (1-\beta)^{T-m}\left(\mathbb{E}\left[f(w_m)\right] - f(w^*)\right),$$

where $\beta \in (0, 1)$.

Theorem III - R-SSN under self-concordance

Two-phased analysis

Under (a) self-concordance (b) L-smoothness, (c) $[\tilde{\mu}+\tau, \tilde{L}+\tau]$-bounded values of the regularized sub-sampled Hessian, (d) ρ_{nd}-Newton decrement SGC with $\rho_{nd} = \frac{\rho L}{\tilde{\mu}+\tau}$, and (e) bounded iterates $\|w - w^*\| \leq D$, then there exists $c \in (0, 1]$ and a constant step size η such that the first phase $\{w_k\}_{k \in [0,m]}$ converges at an R-linear rate

$$\mathbb{E}[f(w_{k+1})] \leq f(w_k) - \epsilon_k,$$

where ϵ_k is some positive sequence. Furthermore, in a local neighbourhood where $\lambda_m \leq 1/6$, the sequence $\{w_k\}_{k \geq m}$ converges to w^* at a Q-linear rate

$$\mathbb{E}\left[f(w_T)\right] - f(w^*) \leq (1-\beta)^{T-m}\left(\mathbb{E}\left[f(w_m)\right] - f(w^*)\right),$$

where $\beta \in (0, 1)$.

Theorem III - R-SSN under self-concordance

Two-phased analysis

Under (a) self-concordance (b) L-smoothness, (c) $[\tilde{\mu} + \tau, \tilde{L} + \tau]$-bounded values of the regularized sub-sampled Hessian, (d) ρ_{nd}-Newton decrement SGC with $\rho_{nd} = \frac{\rho L}{\tilde{\mu}+\tau}$, and (e) bounded iterates $\|w - w^*\| \leq D$, then there exists $c \in (0, 1]$ and a constant step size η such that the first phase $\{w_k\}_{k \in [0,m]}$ converges at an R-linear rate

$$\mathbb{E}[f(w_{k+1})] \leq f(w_k) - \epsilon_k,$$

where ϵ_k is some positive sequence. Furthermore, in a local neighbourhood where $\lambda_m \leq 1/6$, the sequence $\{w_k\}_{k \geq m}$ converges to w^* at a Q-linear rate

$$\mathbb{E}\left[f(w_T)\right] - f(w^*) \leq (1 - \beta)^{T-m} \left(\mathbb{E}\left[f(w_m)\right] - f(w^*)\right),$$

where $\beta \in (0, 1)$.

Theorem III - R-SSN under self-concordance

Two-phased analysis

Under (a) self-concordance (b) L-smoothness, (c) $[\tilde{\mu}+\tau, \tilde{L}+\tau]$-bounded values of the regularized sub-sampled Hessian, (d) ρ_{nd}-Newton decrement SGC with $\rho_{nd} = \frac{\rho L}{\tilde{\mu}+\tau}$, and (e) bounded iterates $\|w - w^*\| \leq D$, then there exists $c \in (0, 1]$ and a constant step size η such that the first phase $\{w_k\}_{k\in[0,m]}$ converges at an R-linear rate

$$\mathbb{E}[f(w_{k+1})] \leq f(w_k) - \epsilon_k,$$

where ϵ_k is some positive sequence. Furthermore, in a local neighbourhood where $\lambda_m \leq 1/6$, the sequence $\{w_k\}_{k\geq m}$ converges to w^* at a Q-linear rate

$$\mathbb{E}\left[f(w_T)\right] - f(w^*) \leq (1-\beta)^{T-m} \left(\mathbb{E}\left[f(w_m)\right] - f(w^*)\right),$$

where $\beta \in (0, 1)$.

Theorem III - R-SSN under self-concordance

Two-phased analysis

Under (a) self-concordance (b) L-smoothness, (c) $[\tilde{\mu}+\tau, \tilde{L}+\tau]$-bounded values of the regularized sub-sampled Hessian, (d) ρ_{nd}-Newton decrement SGC with $\rho_{nd} = \frac{\rho L}{\tilde{\mu}+\tau}$, and (e) bounded iterates $\|w - w^*\| \leq D$, then there exists $c \in (0, 1]$ and a constant step size η such that the first phase $\{w_k\}_{k \in [0,m]}$ converges at an R-linear rate

$$\mathbb{E}[f(w_{k+1})] \leq f(w_k) - \epsilon_k,$$

where ϵ_k is some positive sequence. Furthermore, in a local neighbourhood where $\lambda_m \leq 1/6$, the sequence $\{w_k\}_{k \geq m}$ converges to w^* at a Q-linear rate

$$\mathbb{E}\left[f(w_T)\right] - f(w^*) \leq (1-\beta)^{T-m} \left(\mathbb{E}\left[f(w_m)\right] - f(w^*)\right),$$

where $\beta \in (0, 1)$.

- Although strong-convexity is not required, the rate is still problem-dependent as β depends on $\tilde{\mu}$ and $\tilde{L}$ as in previous work [Zhang and Lin, 2015].

Stochastic BFGS as preconditioned SGD

Stochastic BFGS as preconditioned SGD

- Quasi-Newton methods allow us to incorporate approximate second-order information without computing the Hessian.

Stochastic BFGS update as preconditioned SGD

$$w_{k+1} = w_k - \eta_k \mathbf{B}_k \nabla f_{\mathcal{G}_k}(w_k)$$

where $\mathbf{B}_k$ is a positive-definite matrix constructed to approximate the inverse Hessian.

Stochastic BFGS as preconditioned SGD

- Quasi-Newton methods allow us to incorporate approximate second-order information without computing the Hessian.

Stochastic BFGS update as preconditioned SGD

$$w_{k+1} = w_k - \eta_k \mathbf{B}_k \nabla f_{\mathcal{G}_k}(w_k)$$

where $\mathbf{B}_k$ is a positive-definite matrix constructed to approximate the inverse Hessian.

What's the behaviour of stochastic quasi-Newton methods when interpolation is satisfied?

Theorem IV - Stochastic BFGS as preconditioned SGD

Global linear convergence

Under (a) μ-strongly convex, (b) L-smoothness, (c) ρ-SGC, and (d) $[\lambda_1, \lambda_d]$-bounded eigenvalues of the preconditioner $\mathbf{B}_k$, the sequence $\{w_k\}_{k\geq 0}$ generated by stochastic BFGS with constant step-size $\eta_k = \eta = \frac{\lambda_1}{c_g L \lambda_d^2}$ and constant batch size b_g converges globally to w^* at a Q-linear rate

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \left(1 - \frac{\mu \lambda_1^2}{c_g\, L \lambda_d^2}\right)^T (f(w_0) - f(w^*))$$

where $c_g = \frac{(n-b_g)\,(\rho-1)}{(n-1)\,b_g} + 1$.

Theorem IV - Stochastic BFGS as preconditioned SGD

Global linear convergence

Under (a) μ-strongly convex, (b) L-smoothness, (c) ρ-SGC, and (d) $[\lambda_1, \lambda_d]$-bounded eigenvalues of the preconditioner $\mathbf{B}_k$, the sequence $\{w_k\}_{k\geq 0}$ generated by stochastic BFGS with constant step-size $\eta_k = \eta = \frac{\lambda_1}{c_g L \lambda_d^2}$ and constant batch size b_g converges globally to w^* at a Q-linear rate

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \left(1 - \frac{\mu\lambda_1^2}{c_g\, L\lambda_d^2}\right)^T (f(w_0) - f(w^*))$$

where $c_g = \frac{(n-b_g)\,(\rho-1)}{(n-1)\,b_g} + 1$.

Theorem IV - Stochastic BFGS as preconditioned SGD

Global linear convergence

Under (a) μ-strongly convex, (b) L-smoothness, (c) ρ-SGC, and (d) $[\lambda_1, \lambda_d]$-bounded eigenvalues of the preconditioner $\mathbf{B}_k$, the sequence $\{w_k\}_{k\geq 0}$ generated by stochastic BFGS with constant step-size $\eta_k = \eta = \frac{\lambda_1}{c_g L \lambda_d^2}$ and constant batch size b_g converges globally to w^* at a Q-linear rate

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \left(1 - \frac{\mu \lambda_1^2}{c_g\, L \lambda_d^2}\right)^T (f(w_0) - f(w^*))$$

where $c_g = \frac{(n-b_g)\,(\rho-1)}{(n-1)\,b_g} + 1$.

Theorem IV - Stochastic BFGS as preconditioned SGD

Global linear convergence

Under (a) μ-strongly convex, (b) L-smoothness, (c) ρ-SGC, and (d) $[\lambda_1, \lambda_d]$-bounded eigenvalues of the preconditioner $\mathbf{B}_k$, the sequence $\{w_k\}_{k \geq 0}$ generated by stochastic BFGS with constant step-size $\eta_k = \eta = \frac{\lambda_1}{c_g L \lambda_d^2}$ and constant batch size b_g converges globally to w^* at a Q-linear rate

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \left(1 - \frac{\mu \lambda_1^2}{c_g \, L \lambda_d^2}\right)^T (f(w_0) - f(w^*))$$

where $c_g = \frac{(n - b_g)\,(\rho - 1)}{(n-1)\, b_g} + 1$.

Theorem IV - Stochastic BFGS as preconditioned SGD

Global linear convergence

Under (a) μ-strongly convex, (b) L-smoothness, (c) ρ-SGC, and (d) $[\lambda_1, \lambda_d]$-bounded eigenvalues of the preconditioner $\mathbf{B}_k$, the sequence $\{w_k\}_{k\geq 0}$ generated by stochastic BFGS with constant step-size $\eta_k = \eta = \frac{\lambda_1}{c_g L \lambda_d^2}$ and constant batch size b_g converges globally to w^* at a Q-linear rate

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \left(1 - \frac{\mu \lambda_1^2}{c_g\, L \lambda_d^2}\right)^T (f(w_0) - f(w^*))$$

where $c_g = \frac{(n-b_g)\,(\rho-1)}{(n-1)\,b_g} + 1$.

Theorem IV - Stochastic BFGS as preconditioned SGD

Global linear convergence

Under (a) μ-strongly convex, (b) L-smoothness, (c) ρ-SGC, and (d) $[\lambda_1, \lambda_d]$-bounded eigenvalues of the preconditioner $\mathbf{B}_k$, the sequence $\{w_k\}_{k\geq 0}$ generated by stochastic BFGS with constant step-size $\eta_k = \eta = \frac{\lambda_1}{c_g L \lambda_d^2}$ and constant batch size b_g converges globally to w^* at a Q-linear rate

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \left(1 - \frac{\mu \lambda_1^2}{c_g\, L \lambda_d^2}\right)^T (f(w_0) - f(w^*))$$

where $c_g = \frac{(n - b_g)\,(\rho - 1)}{(n-1)\, b_g} + 1$.

- This rate matches the global linear rate for R-SSN up to constants without having to compute the subsampled Hessian.
- Previous works required a growing batch size [Bollapragada et al., 2018b] or variance-reduction [Lucchi et al., 2015, Moritz et al., 2016].

Theorem IV - Stochastic BFGS as preconditioned SGD

Global linear convergence

Under (a) μ-strongly convex, (b) L-smoothness, (c) ρ-SGC, and (d) $[\lambda_1, \lambda_d]$-bounded eigenvalues of the preconditioner $\mathbf{B}_k$, the sequence $\{w_k\}_{k\geq 0}$ generated by stochastic BFGS with constant step-size $\eta_k = \eta = \frac{\lambda_1}{c_g L \lambda_d^2}$ and constant batch size b_g converges globally to w^* at a Q-linear rate

$$\mathbb{E}[f(w_T)] - f(w^*) \leq \left(1 - \frac{\mu\lambda_1^2}{c_g\, L\lambda_d^2}\right)^T (f(w_0) - f(w^*))$$

where $c_g = \frac{(n-b_g)\,(\rho-1)}{(n-1)\,b_g} + 1$.

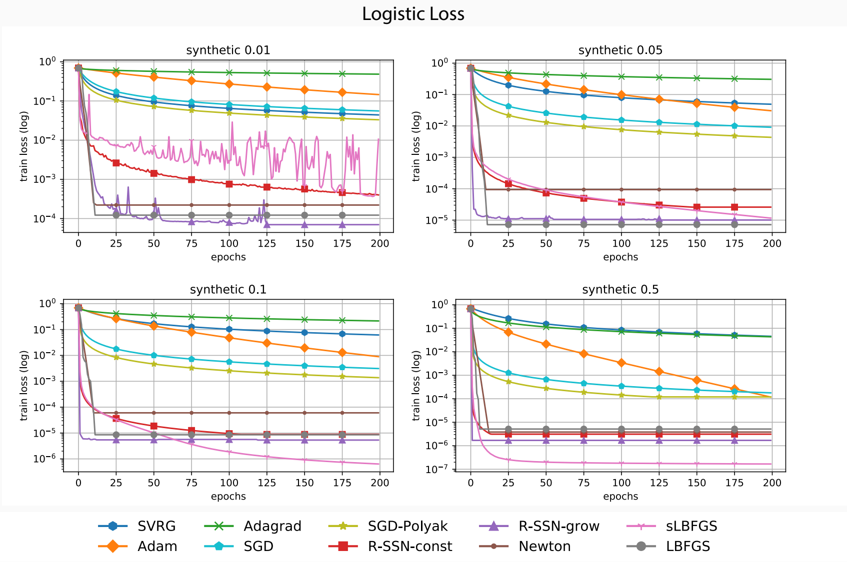
- This rate matches the global linear rate for R-SSN up to constants without having to compute the subsampled Hessian.
- Previous works required a growing batch size [Bollapragada et al., 2018b] or variance-reduction [Lucchi et al., 2015, Moritz et al., 2016].
- Our theoretical result holds for all preconditioners with bounded eigenvalues, but we only focus on stochastic L-BFGS in the experiments.

Experiments

Experimental setup

- Synthetic, linearly separable datasets, linear model $\implies$ Interpolation satisfied.
  - $n = 10k$ examples, $d = 20$ features, with varying margins : `[0.01, 0.05, 0.1, 0.5]`.
- `R-SSN-const`: constant batch size.
- `R-SSN-grow`: grow both batch sizes geometrically.
- **Hessian-free** implementation:
  - Inexact CG with tuned τ that decreases as the batch size grows.
- Compare against `SGD/Acceleration` (line search), `SVRG` (tuned step size), `Adam` and `AdaGrad` (default), and deterministic, unregularized `Newton`.
- All subsampled second-order methods use stochastic line search to select the step size [Vaswani et al., 2019b].

Experimental results



Conclusion

We showed that in the interpolation setting:

- Regularized Sub-sampled Newton (R-SSN) with a constant batch size can achieve global linear convergence.

Conclusion

We showed that in the interpolation setting:

- Regularized Sub-sampled Newton (R-SSN) with a constant batch size can achieve global linear convergence.
- Growing the batch size allows R-SSN to achieve local quadratic convergence.

Conclusion

We showed that in the interpolation setting:

- Regularized Sub-sampled Newton (R-SSN) with a constant batch size can achieve global linear convergence.
- Growing the batch size allows R-SSN to achieve local quadratic convergence.
- R-SSN for self-concordant functions achieves a linear rate.

Conclusion

We showed that in the interpolation setting:

- Regularized Sub-sampled Newton (R-SSN) with a constant batch size can achieve global linear convergence.
- Growing the batch size allows R-SSN to achieve local quadratic convergence.
- R-SSN for self-concordant functions achieves a linear rate.
- Stochastic BFGS converges globally at a Q-linear rate with only constant batch-size.

Conclusion

We showed that in the interpolation setting:

- Regularized Sub-sampled Newton (R-SSN) with a constant batch size can achieve global linear convergence.
- Growing the batch size allows R-SSN to achieve local quadratic convergence.
- R-SSN for self-concordant functions achieves a linear rate.
- Stochastic BFGS converges globally at a Q-linear rate with only constant batch-size.
- Stochastic second-order methods converge faster than first-order methods in practice.

Thank you!

Paper: https://arxiv.org/abs/1910.04920
Code: https://github.com/IssamLaradji/ssn

Raghu Bollapragada, Richard H. Byrd, and Jorge Nocedal. Exact and inexact subsampled Newton methods for optimization. *IMA Journal of Numerical Analysis*, 39(2):545–578, 2018a.

Raghu Bollapragada, Dheevatsa Mudigere, Jorge Nocedal, Hao-Jun Michael Shi, and Ping Tak Peter Tang. A progressive batching L-BFGS method for machine learning. In *ICML*, 2018b.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Michael P. Friedlander and Mark Schmidt. Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing*, 34(3):A1380–A1405, 2012.

Aurelien Lucchi, Brian McWilliams, and Thomas Hofmann. A variance reduced stochastic Newton method. *arXiv preprint arXiv:1503.08316*, 2015.

Philipp Moritz, Robert Nishihara, and Michael I. Jordan. A linearly-convergent stochastic L-BFGS algorithm. In *AISTATS*, 2016.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *AISTATS*, 2019a.

Sharan Vaswani, Aaron Mishkin, Issam H. Laradji, Mark W. Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless Stochastic Gradient: Interpolation, Line-Search, and Convergence Rates. In *NeurIPS*, 2019b.

Yuchen Zhang and Xiao Lin. Disco: Distributed optimization for self-concordant empirical loss. In *ICML*, 2015.