Globally Convergent Newton Methods for Ill-conditioned Generalized Self-concordant Losses

U. Marteau-Ferey, A. Rudi and F. Bach — INRIA - Ecole Normale Supérieure - PSL Research University

Problem setting

Can we do fast logistic regression ?

Find $$\omega_\star^\lambda = \arg\min_{\omega \in \mathcal{H}} g^\lambda(\omega) := g(\omega) + \frac{\lambda}{2}\|\omega\|^2_{\mathcal{H}} \quad (1)$$

Supervised learning n data points $(x_i, y_i)_{1 \le i \le n} \in \mathcal{H} \times \mathcal{Y}$

$$g(\omega) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, \omega \cdot x_i)$$

- **logistic loss**: $\ell(y, y') = \log(1 + e^{-yy'})$
- **softmax loss**: $\ell(y, y') = -y \cdot y' - \log(\int_z e^{-y' \cdot z}\, dz)$

Ill conditioned $\lambda \le \frac{1}{n}$.

GSC function $\forall y,\ |\ell^{(3)}(y, \cdot)| \le R_\ell\, \ell^{(2)}(y, \cdot),\ R = R_\ell \sup_i \|x_i\|$

Ill-conditioned problems : second order method ?

Approximate Newton methods

Newton step at ω : minimizes the second-order approximation :

$$\Delta_\omega^\lambda = [\nabla^2 g^\lambda(\omega)]^{-1}\nabla g^\lambda(\omega) \qquad \textbf{hard to compute}$$

Approximate Newton step : $\widetilde{\Delta}_\omega^\lambda$ relative approximation of Δ_ω^λ

$$\|\widetilde{\Delta}_\omega^\lambda - \Delta_\omega^\lambda\|_{\nabla^2 g^\lambda(\omega)} \le \frac{1}{7}\|\Delta_\omega^\lambda\|_{\nabla^2 g^\lambda(\omega)} \qquad \textbf{easier using Hessian sketching}$$

algoritm: start from ω_0 set $\omega_{t+1} = \omega_t - \widetilde{\Delta}_{\omega_t}^\lambda$

For GSC functions : linear convergence in a small region

Dikin ellipsoid: $\mathsf{D}_\lambda(\mathsf{c}) = \left\{\omega \ :\ \|\nabla g^\lambda(\omega)\|_{\nabla^2 g^\lambda(\omega)^{-1}} \le \frac{\mathsf{c}\lambda^{1/2}}{7R}\right\}$,

$$\boxed{\omega_0 \in \mathsf{D}_\lambda(1) \implies g^\lambda(\omega_t) - g^\lambda(\omega_\star^\lambda) \le 2^{-t},\ \omega_t \in \mathsf{D}_\lambda(2^{-t})}$$

Problem : is it possible to get $\omega_0 \in \mathsf{D}_\lambda(1)$?

Globalization scheme

Main ingredient : inclusion of Dikin ellipsoids

$$\forall \mu \ge \lambda,\ \mathsf{D}_\mu(1/3) \subset \mathsf{D}_{q\mu}(1), \qquad q \ge 1 - 1/(1 + R\|\omega_\lambda^\star\|)$$

This guarantees that $\omega^k \in \mathsf{D}_{\mu_k}(1)$ in the following scheme

1. Start with $\omega^0 \in \mathsf{D}_{\mu_0}(1)$
2. Set $\omega^{k+1} = \mathsf{ANM}(g^{\mu_k}, \omega^k, t = 2)$ (2 iterations of approximate Newton method to g^{μ_k}); set $\mu_{k+1} = q\mu_k$

Global convergence guarantees

$$\boxed{\omega^K \in \mathsf{D}_\lambda(1), \qquad K = C(1 + R\|\omega_\lambda^\star\|)\log\frac{\mu_0}{\lambda}}$$

Note: one can make this algorithm adaptive.

Kernel methods

Data: n i.i.d. observations $(x_i, y_i)_{1 \le i \le n} \in \mathcal{X} \times \mathbb{R}$ with distribution ρ

Features: p.d. Kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$; defines a feature space $\mathcal{H}_K$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}_K$ s.t. $K(x, x') = \phi(x) \cdot \phi(x')$

Predictor: $f(x) = f \cdot \phi(x)$

Goal: minimize the *expected risk* $\mathcal{L}(f) := \mathbb{E}_\rho[\ell(y, f(x))]$

Classical solution : regularized ERM

$$\widehat{f}_\lambda = \arg\min_{f \in \mathcal{H}_K} \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i)) + \frac{\lambda}{2}\|f\|^2_{\mathcal{H}_K}$$

Kernel trick : $\exists(\alpha_i),\ \forall x,\ \widehat{f}_\lambda(x) = \sum_{i=1}^n \alpha_i K(x_i, x)$ *(n-dim problem)*

Issue : solving an n-dimensional problem is too costly

Statistical rates in the kernel setting

Assumptions ℓ GSC, K, Y bounded, $\exists f_\star \in \arg\min_{f \in \mathcal{H}_K} \mathcal{L}(f)$

Main quantities

$\mathbf{H} = \nabla^2\mathcal{L}(f_\star)$ Hessian at optimum, $\mathbf{H}_\lambda = \mathbf{H} + \lambda\mathbf{I}$

Bias term: $\mathsf{b}_\lambda = \lambda^2\|f_\star\|^2_{\mathbf{H}_\lambda^{-1}(f_\star)} \le \lambda\|f_\star\|^2$, **regularity of $f_\star$**

Effective dimension $\mathsf{d}_\lambda = \mathrm{Tr}(\mathbf{H}_\lambda^{-1/2}\mathbf{H}\mathbf{H}_\lambda^{-1/2}) \le C/\lambda$, **size of $\mathcal{H}_K$**

Statistical rates : with proba at least $1 - \delta$

$$\mathcal{L}(\widehat{f}_\lambda) - \mathcal{L}(f_\star) \le C\left(\mathsf{b}_\lambda + \frac{\mathsf{d}_\lambda}{n}\right)\log\frac{1}{\delta}, \qquad \text{if } \mathsf{b}_\lambda, \frac{\mathsf{d}_\lambda}{n} \le \frac{\lambda}{R^2}$$

Dimension reduction

Kernel trick: $\widehat{f}_\lambda \in \mathcal{H}_n = \{\sum_{i=1}^n \alpha_i K(x_i, \cdot)\}$

Principle of Nystrom sampling : dimension reduction

- Subsample points $(\tilde{x}_j)_{1 \le j \le M}$ from $(x_i)_{1 \le i \le n}$, $M << n$
- $\widehat{f}_{\lambda,M} = \arg\min_{f \in \mathcal{H}_M} \widehat{\mathcal{L}}_\lambda(f)$ where $\mathcal{H}_M = \left\{\sum_{j=1}^M \tilde{\alpha}_j K(\tilde{x}_j, \cdot)\right\}$.

Statistical guarantees with Nystrom

If (i) $M \ge C/\lambda \log(c/\lambda\delta)$ using *uniform sampling*
or (ii) $M \ge C\mathsf{d}_\lambda \log(c/\lambda\delta)$ using *Nystrom leverage scores*

$$\boxed{\mathcal{L}(\widehat{f}_{\lambda,M}) - \mathcal{L}(f_\star) \le C\left(\mathsf{b}_\lambda + \frac{\mathsf{d}_\lambda}{n}\right)\log\frac{1}{\delta}, \qquad \text{if } \mathsf{b}_\lambda, \frac{\mathsf{d}_\lambda}{n} \le \frac{\lambda}{R^2}}$$

M-dimensional optimization problem of type (1)

$\mathbf{T}$ s.t. $\mathbf{T}^\top\mathbf{T} = \mathbf{K}_{MM} = (K(\tilde{x}_i, \tilde{x}_j))_{1 \le i,j \le M}$, $\mathbf{K}_{nM} = (K(x_i, \tilde{x}_j))_{ij}$

Then $\widehat{f}_{\lambda,M} = \sum_{j=1}^M \tilde{\alpha}_j K(\tilde{x}_j, \cdot)$ where $\tilde{\alpha} = \mathbf{T}^{-1}\omega_\star^\lambda$ and

$$\omega_\star^\lambda = \arg\min_{\omega \in \mathbb{R}^M} g^\lambda(\omega),\ \ g(\omega) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, \omega \cdot \mathbf{T}^{-\top}\mathbf{K}_{Mn}e_i) \quad (2)$$

Main References

- Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. FALKON: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems* 30, pages 3888–3898. 2017.
- Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. Beyond least- squares: Fast rates for regularized empirical risk minimization through self-concordance. In *Proceedings of the Conference on Computational Learning Theory*, 2019.

Efficient approximate Newton steps

Gradient: $\nabla g^\mu = \mathbf{T}^{-\top}\mathbf{K}_{nM}\mathbf{P}_n + \mu\omega$, $O(nM)$ time, $O(n)$ memory

Hessian : $\nabla^2 g^\mu = \mathbf{T}^{-\top}\mathbf{K}_{Mn}\mathbf{W}_n\mathbf{K}_{nM}\mathbf{T}^{-1} + \mu\mathbf{I}$, $\mathbf{W}_n$ diagonal.

Solving the newton system $[\nabla^2 g^\mu]\Delta^\mu = \nabla g^\mu$

Problem : computing $\nabla^2 g^\mu \to O(nM^2)$ ops

Solution : computing $\nabla^2 g^\mu.x \to O(nM)$ ops. *Iterative methods* ?

Iterative methods to solve $\mathbf{A}x = \mathbf{b}$

\+ one matrix-vector product $\mathbf{A}.x$/epoch ($\nabla^2 g^\mu.x \to O(nM)$ time)

− #epochs needed : $O(\sqrt{\mathrm{Cond}(\mathbf{A})})$ (*conjugate gradient*)

Problem : $\mathrm{Cond}(\nabla^2 g^\mu) \approx \frac{C}{\mu}$: too big ($\mathrm{Cond}(\mathbf{A}) = (\lambda_{\max}/\lambda_{\min})(\mathbf{A})$)

Preconditioning the system $\mathbf{A}x = \mathbf{b}$

Find preconditioner $\mathbf{B} \in \mathbb{R}^{M \times M}$ s.t. $\mathrm{Cond}(\mathbf{B}^{-\top}\mathbf{A}\mathbf{B}^{-1}) \le 2$

Solve $\mathbf{B}^{-\top}\mathbf{A}\mathbf{B}^{-1}\ x = \mathbf{B}^{-\top}\mathbf{b}$ (using iterative method)

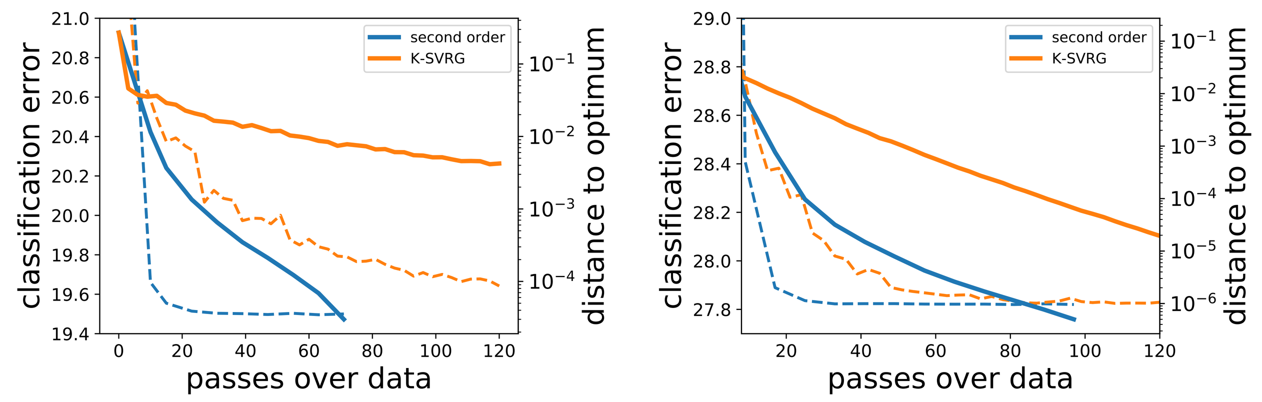
Return $\mathbf{B}^{-1}\ x \to O(nM + M^3)$ operations + cost of pre-conditioner

Pre-conditioner used

$$\mathbf{B} = \mathrm{chol}\left(\mathbf{T}^{-\top}\mathbf{K}_{MM}\mathbf{W}_M\mathbf{K}_{MM}\mathbf{T}^{-1} + \mu\mathbf{I}\right) \to O(M^3)\text{ time} \quad (3)$$

Guarantees of the algorithm

Logsitic regression on Susy and Higgs ($n \approx 10^7$)



Main assumption

(i) $M \ge C/\lambda \log(c/\lambda\delta)$ using *uniform sampling*
or (ii) $M \ge C\mathsf{d}_\lambda \log(c/\lambda\delta)$ using *Nystrom leverage scores*,

Method

- **Globalization scheme** to solve (2) with $K = C(1 + R\|f_\star\|)\log\frac{\mu_0}{\lambda}$
- **Approximate Newton steps** with (3), $C\log(1/\epsilon)$ iterations

Guarantees

$$\boxed{\mathcal{L}(f) - \mathcal{L}(f_\star) \le C\left(\mathsf{b}_\lambda + \frac{\mathsf{d}_\lambda}{n} + \epsilon\right)\log\frac{1}{\delta}, \qquad \text{if } \mathsf{b}_\lambda, \frac{\mathsf{d}_\lambda}{n} \le \frac{\lambda}{R^2}}$$

Time : $O(R\|f_\star\|(nM + M^3)\log(\mu_0/\lambda)\log(1/\epsilon))$

Memory : $O(n + M^2)$