Second Order Strikes Back: 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

- **Supervised data :** n input-output pairs $(x_i, y_i)_{1\leq i \leq n} \in \mathcal{H} \times \mathcal{Y}$
- **Assumption :** $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ Hilbert space, $\mathcal{Y} = \mathbb{R}$, $\sup_i \|x_i\| \leq R$
- **Linear predictor :** Find ω such that $y \approx \langle \omega, x\rangle$

Goal : ill-conditioned logistic regression

$$\text{Compute } \omega_\star^\lambda := \arg\min_{\omega\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \ell(y_i, \langle \omega, x_i\rangle) + \frac{\lambda}{2}\|\omega\|^2 \qquad (1)$$

(i) Logistic loss : $\ell(y, y') = \log(1+e^{-yy'})$

(ii) Small regularizer : $\lambda \leq \frac{1}{n}$

Key property : Generalized Self Concordance (GSC)

$$\forall y \in \mathcal{Y},\ |\ell^{(3)}(y,\cdot)| \leq \ell^{(2)}(y,\cdot)$$

Notations:

- Functions : $g(\omega) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, \langle\omega, x_i\rangle)$, $\quad g^\lambda(\omega) = g(\omega) + \frac{\lambda}{2}\|\omega\|^2$
- Hessians : $\mathbf{H}(\omega) = \nabla^2 g(\omega)$, $\quad \mathbf{H}_\lambda(\omega) = \nabla^2 g^\lambda(\omega)$
- For any p.s.d. operator $\mathbf{A}$ on $\mathcal{H}$, $\|\cdot\|_{\mathbf{A}} = \|\mathbf{A}^{1/2}\cdot\|_{\mathcal{H}}$.

Ill-conditioned problems : second order method ?

Approximate Newton Methods (ANM)

Start ω_0	Newton step s_t^0	Approx Newton step s_t
Step $\omega_{t+1} = \omega_t - \mathsf{s}_t$	$\mathbf{H}_\lambda^{-1}(\omega_t)\nabla g^\lambda(\omega_t)$	$\|\mathsf{s}_t - \mathsf{s}_t^0\|_{\mathbf{H}_\lambda} \leq \rho\|\mathsf{s}_t^0\|_{\mathbf{H}_\lambda}$

Key region for GSC functions : the Dikin ellipsoid

$$\forall \mathsf{c} > 0,\ \mathsf{D}_\lambda(\mathsf{c}) = \left\{\omega \in \mathcal{H} \ :\ 7R\,\|\nabla g^\lambda(\omega)\|_{\mathbf{H}_\lambda^{-1}(\omega)} \leq \mathsf{c}\,\sqrt{\lambda}\right\}$$

Linear convergence of ANMs in the Dikin ellipsoid

$$\rho \leq \frac{1}{7},\ \omega_0 \in \mathsf{D}_\lambda(1) \implies g^\lambda(\omega_t) - g^\lambda(\omega_\star^\lambda) \leq 2^{-t},\ \omega_t \in \mathsf{D}_\lambda(2^{-t})$$

Globalization scheme (GS)

Main ingredient : inclusion of Dikin ellipsoids

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

Start	$\omega_0 \in \mathsf{D}_{\mu_0}(1)$ for a certain μ_0 (explicit)
For $0 \leq k < \mathsf{K}$	$\omega_{2k+2} \leftarrow 2$ iterations of ANM to g^{μ_k} from ω_{2k}
$\mathsf{K} = \lceil \log_{1/\mathsf{q}}(\mu_0/\lambda)\rceil$	$\mu_{k+1} \leftarrow \min(\mathsf{q}\mu_k, \lambda)$ **invariant:** $\omega_{2k} \in \mathsf{D}_{\mu_k}(1)$
Final ANM	$\omega_{2\mathsf{K}+t} \leftarrow t$ iterations of ANM to g^λ from $\omega_{2\mathsf{K}}$

Global convergence reaching precision ϵ

$$\forall k \geq 2(1+R\|\omega_\lambda^\star\|)\left\lceil \log\frac{\mu_0}{\lambda}\right\rceil + \left\lceil \log \epsilon^{-1}\right\rceil,\ g^\lambda(\omega_k) - g^\lambda(\omega_\star^\lambda) \leq \epsilon$$

Kernel methods

- **Data**: $(x_i, y_i)_{1\leq i\leq n} \in \mathcal{X}\times\mathbb{R}$, i.i.d. with distribution ρ
- **Feature space** $\mathcal{H}_K$: defined from p.d. Kernel $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$

Elementary functions of $\mathcal{H}_K$	$K_x = [x' \in \mathcal{X} \mapsto K(x, x')] \in \mathcal{H}_K$
Hilbert structure	$\langle K_x, K_{x'}\rangle = K(x, x')$

- **Predictor**: $f \in \mathcal{H}_K$; satisfies $f(x) = \langle f, K_x\rangle$: **linear**
- **Expected loss**: $\mathcal{L}(f) := \mathbb{E}_{(x,y)\sim\rho}[\ell(y, f(x))]$

Statistical goal

Construct $\widehat{f}$ s.t. $\mathcal{L}(\widehat{f}) - \inf_{f\in\mathcal{H}_K}\mathcal{L}(f)$ is small with high probability

Classical estimator : Empirical Risk Minimization

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

Statistical performance of $\widehat{f}_\lambda$

- **Assumptions**: ℓ GSC, $K(\cdot,\cdot) \leq R$, $\exists f_\star \in \arg\min_{f\in\mathcal{H}_K}\mathcal{L}(f)$
- **Notations**: $\mathbf{H}^\star = \nabla^2\mathcal{L}(f_\star)$; $\mathbf{H}_\lambda^\star = \mathbf{H}^\star + \lambda\mathbf{I}$

	Key quantity	Measures
bias	$\mathsf{b}_\lambda := \lambda^2\|f_\star\|^2_{\mathbf{H}_\lambda^{\star-1}}$	regularity of $f_\star$
variance	$\mathsf{d}_\lambda := \mathrm{Tr}(\mathbf{H}_\lambda^{\star-1/2}\mathbf{H}^\star\mathbf{H}_\lambda^{\star-1/2})$	dimension of $\mathcal{H}_K$

Performance of $\widehat{f}_\lambda$ with proba $1-\delta$

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

Reducing dimension with same performance

Kernel trick: $\widehat{f}_\lambda = \sum_{i=1}^n \alpha_i K_{x_i} \implies$ n**-dimensional problem**

Dimension reduction with Nyström sampling

- Subsample M points $(\tilde{x}_j)$ from the $(x_i)_{1\leq i\leq n}$, $\mathbf{M} \ll \mathbf{n}$
- $\widehat{f}_{\lambda,M} = \sum_{j=1}^M \tilde{\alpha}_j K_{\tilde{x}_j} = \arg\min_{f\in\mathcal{H}_M}\widehat{\mathcal{L}}_\lambda(f), \mathcal{H}_M = \left\{\sum_{j=1}^M \alpha_j K_{\tilde{x}_j}\right\}$.

$\widehat{f}_{\lambda,M}$ has the same performance as $\widehat{f}_\lambda$ with proba $1-\delta$

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

(a) $M \geq C_M(1/\lambda)\log(c/\lambda\delta), C_M = \Omega(1)$ (*uniform sampling*), or

(b) $M \geq C_M \mathsf{d}_\lambda \log(c/\lambda\delta), C_M = \Omega(1)$ (*Nystrom leverage scores*)

Remark: $\mathbf{K}_{MM} = (K(\tilde{x}_i, \tilde{x}_j))$, $\mathbf{K}_{nM} = (K(x_i, \tilde{x}_j))$.

$$\tilde{\alpha} = \arg\min_{\alpha\in\mathbb{R}^M} g^\lambda(\alpha) := \frac{1}{n}\sum_{i=1}^n \ell(y_i, \langle\alpha, \mathbf{K}_{nM}^\top e_i\rangle) + \frac{\lambda}{2}\alpha^\top\mathbf{K}_{MM}\alpha \qquad (2)$$

Computing approximate Newton steps

Form of a Newton step : $\mathsf{s}^0 = \mathbf{H}_\mu^{-1}\mathsf{g}^\mu$

gradient	$\mathsf{g}^\mu := \mathbf{K}_{nM}\mathbf{P}_n + \mu\mathbf{K}_{MM}\alpha$	$\mathbf{P}_n \in \mathbb{R}^n$
hessian	$\mathbf{H}_\mu := \mathbf{K}_{nM}^\top\mathbf{W}_n\mathbf{K}_{nM} + \mu\mathbf{K}_{MM}$	$\mathbf{W}_n$ diagonal

Sketching the Hessian using Nyström

If **(a)** or **(b)** holds with $C_M = \Omega(\log 1/\rho)$,

$$(1-\rho)\widetilde{\mathbf{H}}_\lambda \preceq \mathbf{H}_\lambda \preceq (1+\rho)\widetilde{\mathbf{H}}_\lambda, \qquad \widetilde{\mathbf{H}}_\lambda = \mathbf{K}_{MM}\mathbf{W}_M\mathbf{K}_{MM} + \mu\mathbf{K}_{MM}$$

Option I	$\mathsf{s} \leftarrow \widetilde{\mathbf{H}}_\mu^{-1}\mathbf{g}^\mu$, $C_M = \Omega(\log 1/\rho)$
Option II	$\mathsf{s} \leftarrow \Omega(\log 1/\rho)$ steps of iterative method solving
(more stable)	$\mathbf{H}_\mu\mathsf{s} = \mathsf{g}^\mu$ pre-conditioned with $\widetilde{H}_\mu$, $C_M = \Omega(1)$

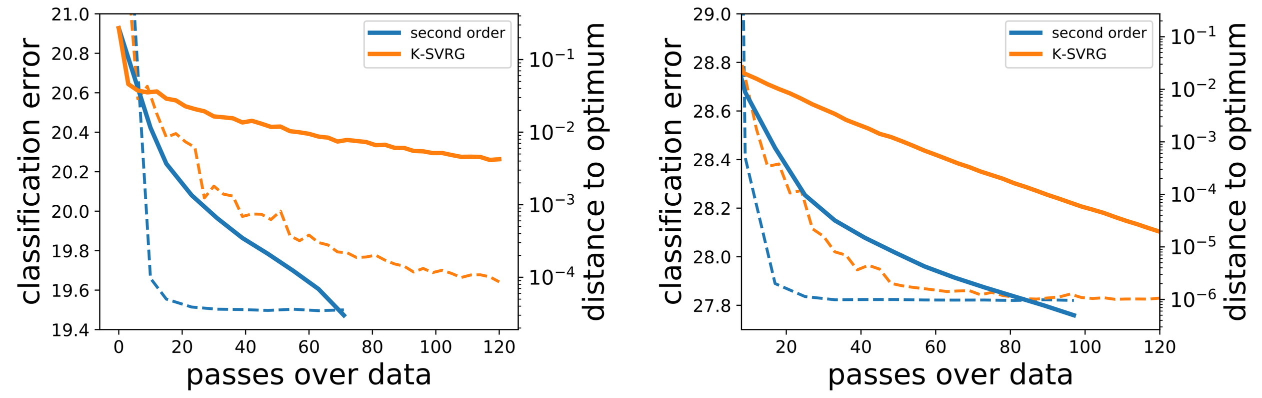
complexity	Time	Memory	Best M
Computing $\widetilde{\mathbf{H}}_\lambda$	$O(M^3)$	$O(M^2)$	
Computing g^μ	$O(nM + M^2)$	$O(n + M^2)$	
Computing $\mathbf{H}_\lambda.x$	$O(nM + M^2)$	$O(n + M^2)$	
Option I	$O(nM + M^3)$	$O(M^2 + n)$	$\Omega(\log(1/\rho)\mathsf{d}_\lambda)$
Option II	$O((nM + M^3)\log(1/\rho))$	$O(M^2 + n)$	$\Omega(\mathsf{d}_\lambda)$

In both cases, s is an approximate Newton step (ANS)

$$\|\mathsf{s} - \mathsf{s}^0\|_{\mathbf{H}_\mu} \leq \rho\|\mathsf{s}^0\|_{\mathbf{H}_\mu}$$

A fast optimal second order algorithm

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



Algorithm GS applied to g^λ, precision ϵ, **returns** $\alpha^{\mathtt{alg}}$

ANS used see option I or II with $\rho = 1/7$

Optimal guarantees for $f^{\mathtt{alg}} = \sum_{j=1}^M \alpha_j^{\mathtt{alg}} K_{\tilde{x}_j}$, proba $1-\delta$

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

Time complexity : $O\left(\mathsf{N}\ [n\mathsf{d}_\lambda + \mathsf{d}_\lambda^3]\right)$, $\mathsf{N} = R\|f_\star\|\log\frac{\mu_0}{\lambda} + \log\frac{1}{\epsilon}$

Memory complexity : $O(n + \mathsf{d}_\lambda^2)$

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.