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

Solving ill-conditioned logistic regression

Data : n points $(x_i, y_i)_{1\leq i\leq n} \in \mathcal{H}\times\mathbb{R}$, $\mathcal{H}$ Hilbert, $\sup_i \|x_i\| \leq R$

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

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

(ii) Potentially very small regularizer $\lambda \ll 1,\ e.g.\ \lambda \approx 10^{-12}$

Key property : Generalized Self Concordance (GSC)

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

Notations:

- $g^\lambda(\omega) = g(\omega) + \frac{\lambda}{2}\|\omega\|^2$ $\qquad$ $\mathbf{H}_\lambda(\omega) = \nabla^2 g^\lambda(\omega)$
- 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 ?

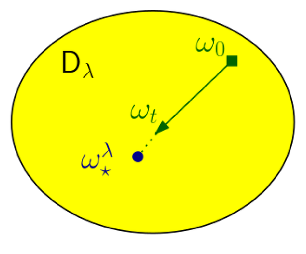
Approximate Newton Methods (ANM)

	Approximate Newton Step (ANS)
Start at ω_0	
Hessian sketch $\widetilde{\mathbf{H}}_\lambda(\omega_t)$	$\frac{1}{2}\widetilde{\mathbf{H}}_\lambda \preceq \mathbf{H}_\lambda \preceq 2\widetilde{\mathbf{H}}_\lambda$
Step $\omega_{t+1} = \omega_t - \mathrm{s}_t$	$\mathrm{s}_t = \widetilde{\mathbf{H}}_\lambda^{-1}(\omega_t)\nabla g^\lambda(\omega_t)$

Linear convergence near the optimum

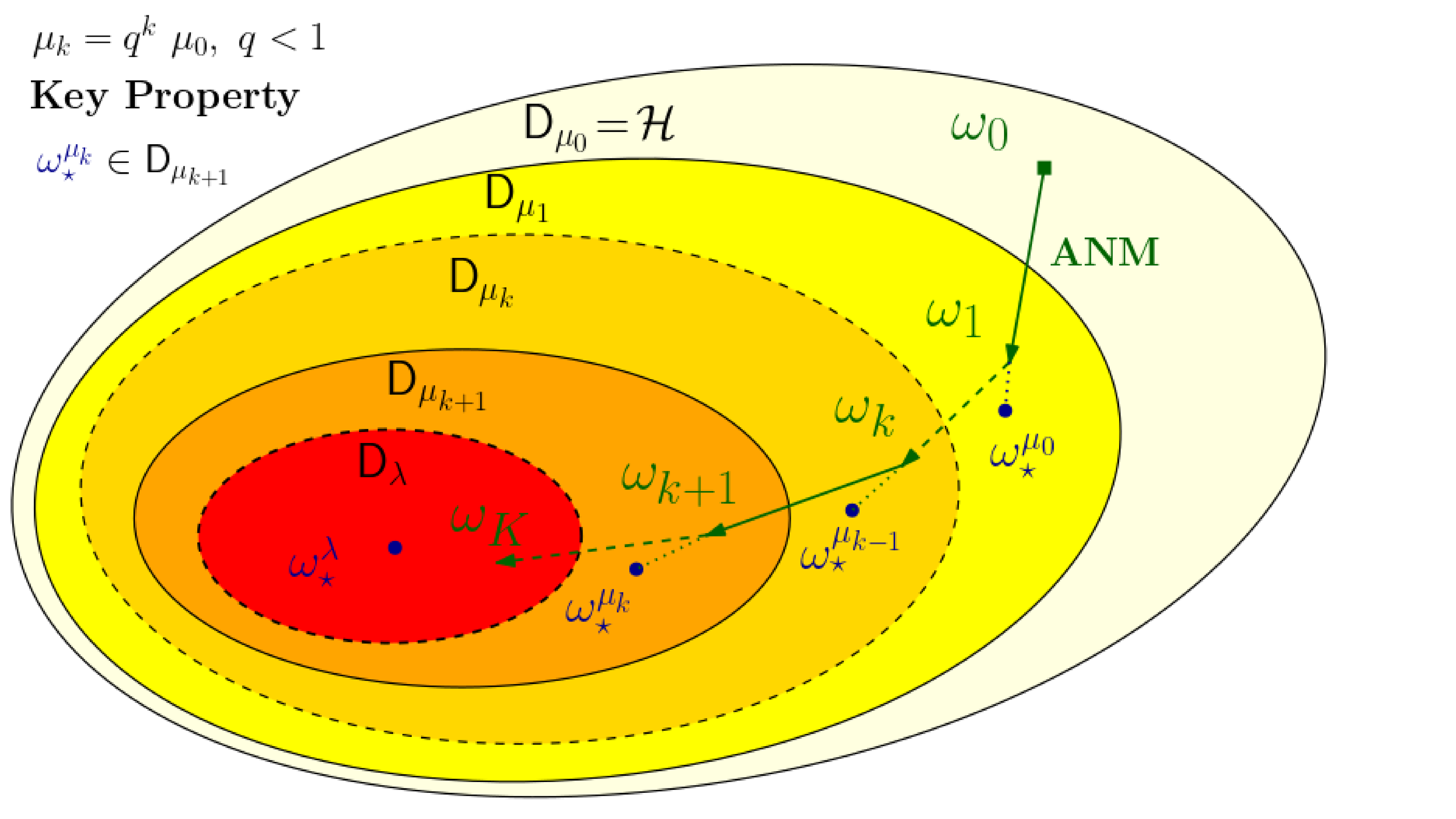
If $\omega_0 \in \mathrm{D}_\lambda$, $\mathrm{D}_\lambda := \left\{\omega : R\|\nabla g^\lambda(\omega)\|_{\mathbf{H}_\lambda^{-1}(\omega)} \leq \sqrt{\lambda}\right\}$,

$g^\lambda(\omega_t) - g^\lambda(\omega_\star^\lambda) \leq 2^{-t}$



Problem : is is possible to find $\omega\in\mathrm{D}_\lambda$?

Getting inside D_λ



Reaching D_λ in $O(\log(\mu_0/\lambda))$ approximate Newton steps

$\omega_K \in \mathrm{D}_\lambda,\ K = O(\log_q(\mu_0/\lambda)),\ 1/q \approx 1 - 1/(R\|\omega_\star^\lambda\|)$

Learning with GSC functions

- **Data**: $(x_i,y_i)_{1\leq i\leq n}\in\mathcal{X}\times\mathbb{R}$, i.i.d. with distribution ρ
- **Feature map**: $\Phi : \mathcal{X}\to\mathcal{H}$, $\mathcal{H}$ Hilbert
- **Linear predictor**: $f(x) \leftrightarrow \langle f, \Phi(x)\rangle_{\mathcal{H}},\ f\in\mathcal{H}$
- **Expected loss**: $\mathcal{L}(f) := \mathbb{E}_{(x,y)\sim\rho}[\ell(y,f(x))]$

Statistical goal

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

Classical estimator : Empirical Risk Minimization

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

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

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

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

Problem : finding $\widehat{f}_\lambda$ is a n-dimensional problem

Dimension reduction with Nyström

Finding a smaller set of candidate functions $\mathcal{H}_M$

- Subsample M points $(\tilde{x}_j)$ from the $(x_i)_{1\leq i\leq n}$, $\mathbf{M}\ll\mathbf{n}$
- Find the best possible f of the form $f = \sum_{j=1}^M \alpha_j\ \Phi(\tilde{x}_j),\ \alpha\in\mathbb{R}^M$:

$$\widehat{f}_{\lambda,M} := \arg\min_{f\in\mathcal{H}_M}\widehat{\mathcal{L}}_\lambda(f),\ \mathcal{H}_M = \left\{\sum_{j=1}^M \alpha_j\ \Phi(\tilde{x}_j)\ :\ \alpha\in\mathbb{R}^M\right\}$$

$\widehat{f}_{\lambda,M}$ has the same performance (2) as $\widehat{f}_\lambda$ if

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

(b) $M \geq \mathrm{d}_\lambda\log(c/\lambda\delta)$ (*Nystrom leverage scores*)

ANS for the Nyström problem

Optimization problem : $\widehat{f}_{\lambda,M} = \sum_{j=1}^M \widehat{\alpha}_j\ \Phi(\tilde{x}_j)$

$$\widehat{\alpha} = \arg\min_{\alpha\in\mathbb{R}^M} g^\lambda(\alpha) := \frac{1}{n}\sum_{i=1}^n \ell(y_i, \langle\alpha, \mathbf{K}_{nM}^\top e_i\rangle) + \frac{\lambda}{2}\alpha^\top\mathbf{K}_{MM}\alpha \qquad (3)$$

$(\mathbf{K}_{MM})_{ij} = \langle\Phi(\tilde{x}_i),\Phi(\tilde{x}_j)\rangle$, $(\mathbf{K}_{nM})_{ij} = \langle\Phi(x_i),\Phi(\tilde{x}_j)\rangle$.

Form of the Hessian

$\mathbf{H}_\mu := \mathbf{K}_{nM}^\top\mathbf{W}_n\mathbf{K}_{nM} + \mu\mathbf{K}_{MM}$, $\qquad$ $\mathbf{W}_n$ diagonal

Sketching the Hessian using Nyström

If **(a)** or **(b)** holds, then for all $\mu\geq\lambda$, defining

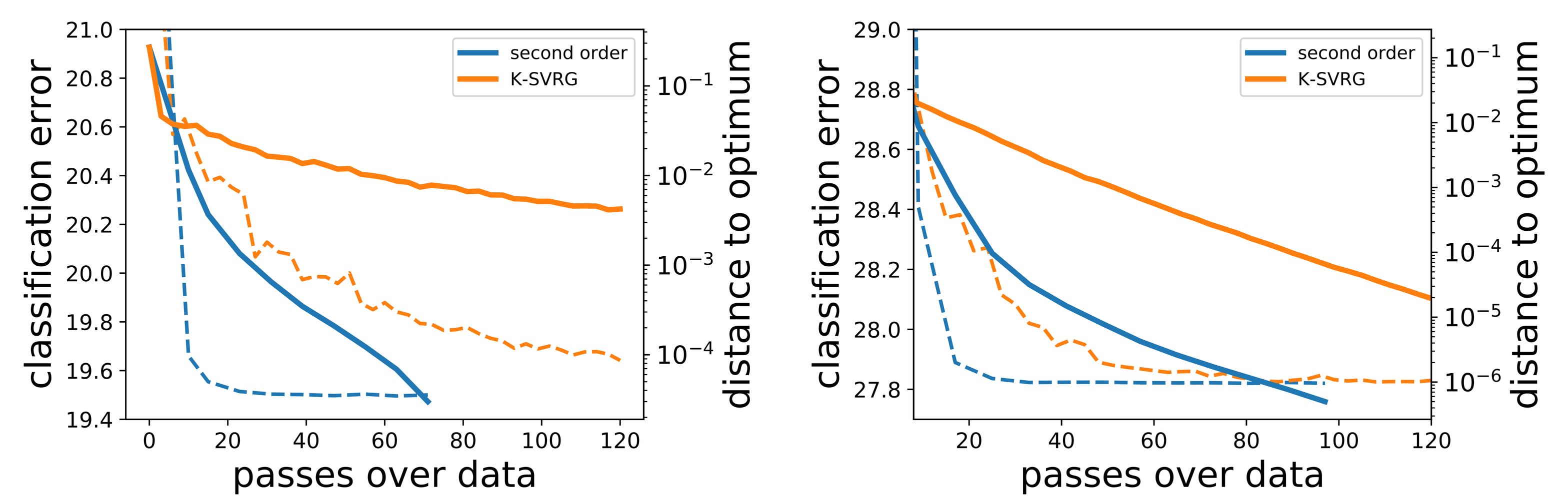
$\widetilde{\mathbf{H}}_\mu = \mathbf{K}_{MM}\mathbf{W}_M\mathbf{K}_{MM} + \mu\mathbf{K}_{MM}$, it holds

$$\frac{1}{2}\widetilde{\mathbf{H}}_\lambda \preceq \mathbf{H}_\lambda \preceq 2\widetilde{\mathbf{H}}_\lambda$$

complexity	Time	Memory
Computing $\widetilde{\mathbf{H}}_\mu$	$O(M^3)$	$O(M^2)$
Computing ∇g^μ	$O(nM+M^2)$	$O(n+M^2)$
Computing an ANS	$O(nM+M^3)$	$O(n+M^2)$

Second Order Strikes back

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



Algorithm (GS) to solve (3) with precision c/n

Returns α^{alg} s.t. $g^\lambda(\alpha^{\mathrm{alg}}) - g^\lambda(\widehat{\alpha}) \leq \mathrm{c}/n$

Predictor $f^{\mathrm{alg}} = \sum_{j=1}^M \alpha_j^{\mathrm{alg}}\ \Phi(\tilde{x}_j)$

Code and results: https://github.com/umarteau/Newton-Method-for-GSC-losses-

f^{alg} has the same guarantees (2) as $\widehat{f}_\lambda$

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

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

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.