Beyond Least-Squares: Fast Rates for Regularized Empirical Risk Minimization through Self-Concordance

U. Marteau-Ferey, D. Ostrovskii, F. Bach, A. Rudi

SIERRA Project-Team, INRIA, Département d'Informatique de l'ÉNS, PSL Research University

Regularized Empirical Risk Minimization

Problem Setting:

Unknown distribution : rv $Z \in \mathcal{Z}$ with distribution ρ

Parameter $\theta \in \mathcal{H}$, $\mathcal{H}$ a Hilbert space

Problem: Minimize an expected loss:

$$\min_{\theta \in \mathcal{H}} L(\theta) := \mathbb{E}\left[\ell_Z(\theta)\right], \qquad \ell_z(\theta) \text{ loss function} \tag{1}$$

Well-specified assumption $\theta^\star \in \operatorname{argmin}_{\theta\in\mathcal{H}} L(\theta)$ exists

Statistical performance: $L(\theta) - L(\theta^\star)$

Data: access to ρ through n i.i.d observations $(z_i)_{1 \leq i \leq n} \in \mathcal{Z}^n$ from Z

Regularized ERM:

$$\widehat{\theta}_\lambda = \operatorname{argmin}_{\theta\in\mathcal{H}} \widehat{L}(\theta) + \frac{\lambda}{2}\|\theta\|^2_{\mathcal{H}}, \qquad \widehat{L}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell_{z_i}(\theta)$$

Basic result : slow rates

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leq \frac{\|\nabla L\|^2_\infty + \|\theta^\star\|^2}{\sqrt{n}}, \qquad \lambda = \frac{1}{\sqrt{n}}$$

Bias-variance trade-off for least-squares

Loss: $\ell_{x,y}(\theta) = \|y - \theta\cdot x\|^2$, $L(\theta) = \mathbb{E}\left[\|Y - \theta \cdot X\|^2\right]$

Covariance operator: $\Sigma = \mathbb{E}\left[X \otimes X\right]$

$$\forall \theta \in \mathcal{H},\ L(\theta) - L(\theta^\star) = \|\Sigma^{1/2}(\theta - \theta^\star)\|^2 = \|\theta - \theta^\star\|^2_{\Sigma}.$$

Two main terms :

Effective dimension: $\mathsf{df}_\lambda = \operatorname{Tr}(\Sigma_\lambda^{-1}\Sigma)$, $\quad \Sigma_\lambda = \Sigma + \lambda I$

Bias term: $\mathsf{b}_\lambda = \lambda\|\Sigma_\lambda^{-1}\theta^\star\|$

Bias-Variance trade-off :

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leq \mathsf{b}_\lambda^2 + \frac{\mathsf{df}_\lambda}{n}.$$

Parametrization and optimal rates

Effective dimension $\leftrightarrow$ spectrum of covariance matrix

$(\lambda_i)_i$ eigenvalues of Σ in decreasing order.

assumption: $\mathsf{df}_\lambda \leq \mathsf{Q}^2\lambda^{-1/\alpha} \ \leftrightarrow\ \lambda_i = O\left(i^{-\alpha}\right)$

Bias term $\leftrightarrow$ difficulty of the learning problem

assumption : $\mathsf{b}_\lambda \leq \mathsf{L}\lambda^{1/2+r} \ \leftrightarrow\ \|\Sigma^{-r}\theta^\star\| < \infty$

Optimal fast rates for $\lambda = (\mathsf{Q}/\mathsf{L})^2\ n^{-\alpha/((1+2r)\alpha+1)}$

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leq \mathsf{Q}^{2\gamma}\mathsf{L}^{2(1-\gamma)} n^{-\gamma}, \qquad \gamma = (1+2r)\alpha/(1+(1+2r)\alpha)$$

Acknowledgments and References

We acknowledge support from the ERCIM Alain Bensoussan Fellowship and the European Reasearch Council (SEQUOIA 724063)

- A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm.Found.Comput. Math., 7(3):331–368, July 2007.
- Dmitrii Ostrovskii and Francis Bach.Finite-sample analysis of M-estimators using self-concordance. Technical Report 1810.06838, arXiv, 2018.

Generalized bias-variance trade-off

- **Hessian at optimum:** $\mathbf{H}(\theta^\star) = \nabla^2 L(\theta^\star) = \mathbb{E}\left[\nabla^2 \ell_Z(\theta^\star)\right]$

Main terms :

- **Effective dimension:** $\mathsf{df}_\lambda = \mathbb{E}\left[\|\mathbf{H}_\lambda^{-1/2}(\theta^\star)\nabla\ell_Z(\theta^\star)\|^2\right]$
- **Bias term:** $\mathsf{b}_\lambda = \lambda\|\mathbf{H}_\lambda^{-1}(\theta^\star)\theta^\star\| = \|\nabla L_\lambda(\theta^\star)\|_{\mathbf{H}_\lambda^{-1}(\theta^\star)}$
- **Dikin radius:** $\mathsf{r}_\lambda = \sqrt{\lambda}/\mathsf{R}$

Bias-Variance trade-off :

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leq \mathsf{b}_\lambda^2 + \frac{\mathsf{df}_\lambda}{n}, \qquad \mathsf{b}_\lambda, \sqrt{\mathsf{df}_\lambda/n} \leq \mathsf{r}_\lambda$$

Generalized optimal rates

Effective dimension $\leftrightarrow$ renormalized Fischer information

assumption: $\mathsf{df}_\lambda \leq \mathsf{Q}^2\lambda^{-1/\alpha}$.

For GLMs in the well specified case, $\mathsf{df}_\lambda = \operatorname{Tr}(\mathbf{H}_\lambda^{-1}(\theta^\star)\mathbf{H}(\theta^\star))$.

Bias term $\leftrightarrow$ difficulty of the learning problem

assumption : $\mathsf{b}_\lambda \leq \mathsf{L}\lambda^{1/2+r} \ \leftrightarrow\ \|\mathbf{H}(\theta^\star)^{-r}\theta^\star\| < \infty$

Optimal fast rates for $\lambda = (\mathsf{Q}/\mathsf{L})^2\ n^{-\alpha/((1+2r)\alpha+1)}$

$$L(\widehat{\theta}_\lambda) - L(\theta^\star) \leq \mathsf{Q}^{2\gamma}\mathsf{L}^{2(1-\gamma)} n^{-\gamma}$$

Sketch of proof

Define $\theta^\star_\lambda = \operatorname{argmin}_{\theta\in\mathcal{H}} L(\theta) + \frac{\lambda}{2}\|\theta\|^2$.

Idea: Decompose the statistical performance $\theta^\star \to \theta^\star_\lambda \to \widehat{\theta}_\lambda$.

Bias term: $\theta^\star \leftrightarrow \theta^\star_\lambda$ Using localisation on L_λ:

$$\mathsf{b}_\lambda \leq \mathsf{r}_\lambda \implies \begin{cases} \mathsf{R}\ \|\theta^\star_\lambda - \theta^\star\| \leq \log 2 \\ \|\theta^\star_\lambda - \theta^\star\|_{\mathbf{H}_\lambda} \leq 2\mathsf{b}_\lambda \end{cases}$$

Equivalence of norms: $\mathbf{H}_\lambda \sim \widehat{\mathbf{H}}_\lambda$ Concentration inequality.

Variance term: $\widehat{\theta}_\lambda \leftrightarrow \theta^\star_\lambda$ Localization + Concentration inequality :

1. Localization on $\widehat{L}_\lambda$ + Equivalence of norms

$$\|\nabla\widehat{L}_\lambda(\theta^\star_\lambda)\|_{\mathbf{H}_\lambda^{-1}} \leq \mathsf{r}_\lambda \implies \begin{cases} \mathsf{R}\ \|\theta^\star_\lambda - \widehat{\theta}_\lambda\| \leq \log 2 \\ \|\theta^\star_\lambda - \widehat{\theta}_\lambda\|_{\mathbf{H}_\lambda} \leq 2\|\nabla\widehat{L}_\lambda(\theta^\star_\lambda)\|_{\mathbf{H}_\lambda^{-1}} \end{cases}$$

2. Concentration inequality

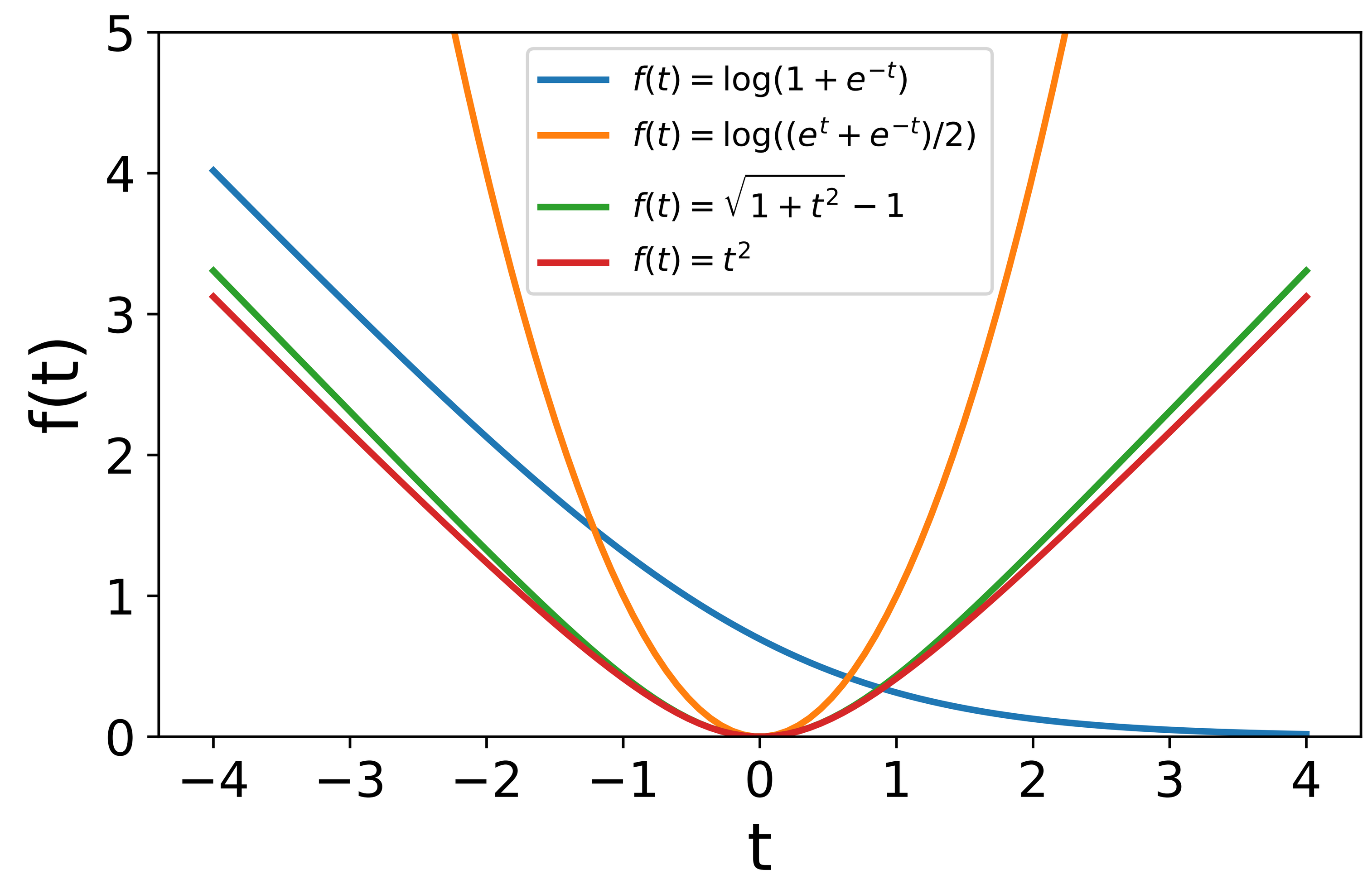
$$\|\nabla\widehat{L}_\lambda(\theta^\star_\lambda)\|_{\mathbf{H}_\lambda^{-1}} \leq \sqrt{\mathsf{df}_\lambda/n}$$

Putting things together

$$\sqrt{\mathsf{df}_\lambda/n}, \mathsf{b}_\lambda \leq \mathsf{r}_\lambda \implies \begin{cases} \mathsf{R}\ \|\theta^\star - \widehat{\theta}_\lambda\| \leq 2\log 2 \\ \|\theta^\star - \widehat{\theta}_\lambda\|_{\mathbf{H}_\lambda} \leq \sqrt{\mathsf{df}_\lambda/n} + \mathsf{b}_\lambda \end{cases}$$

Quadratic approximation: $L(\widehat{\theta}_\lambda) - L(\theta^\star) \leq 4(\sqrt{\mathsf{df}_\lambda/n} + \mathsf{b}_\lambda)^2$

Generalized Self-Concordance



Definition (GSC function)

$$\nabla^{(3)}F(\theta)[h,k,k] \leq \ \mathsf{R}\ \|h\|\ \nabla^2 F(\theta)[k,k]$$

Assumption: the $(\ell_z)_{z\in\operatorname{supp}(\rho)}$ are all GSC functions for R.

Consequence: L, $L_\lambda = L + \frac{\lambda}{2}\|\cdot\|^2$, $\widehat{L}_\lambda$ are GSC for R.

Examples from Supervised Learning

Distribution: $Z = (X, Y)$, where we want to predict Y from X.

Predictor: $\theta\cdot\Phi(x)$ ($\theta\cdot\Phi(x,y)$ for multiclass)

Assumption: Y is bounded, $\Phi(X)$ is bounded

Regression: $\ell_z(\theta) = \psi(y - \theta\cdot\Phi(x))$

- Square loss: $\psi(t) = \frac{1}{2}t^2$
- Huber loss 1: $\psi(t) = \sqrt{1+t^2} - 1$
- Huber loss 2: $\psi(t) = \log\frac{e^t + e^{-t}}{2}$

Classification:

- Logistic loss: $\ell_z(\theta) = \log(1 + e^{-y\ \theta\cdot\Phi(x)})$
- GLMs: $\ell_z(\theta) = -\theta\cdot\Phi(x,y) + \log\int_{\mathcal{Y}}\exp\left(\theta\cdot\Phi(x,y')\right)d\mu(y')$

Properties of GSC functions

Assume F is GSC, with minimizer $\theta^\star$. Let $\theta \in \mathcal{H}$, $\mathsf{t} = \mathsf{R}\ \|\theta - \theta^\star\|$.

Quadratic approximation

$$F(\theta) - F(\theta^\star) \leq e^{\mathsf{t}}\ \|\theta - \theta^\star\|^2_{\nabla^2 F(\theta)}$$

Localization using gradients Define $F_\lambda = F + \frac{\lambda}{2}\|\cdot\|^2$.

$$\|\nabla F_\lambda(\theta)\|_{\nabla^2 F_\lambda(\theta)^{-1}} \leq \mathsf{r}_\lambda \implies \begin{cases} \mathsf{t} \leq \log 2 \\ \|\theta - \theta^\star\|_{\nabla^2 F_\lambda(\theta)} \leq 2\|\nabla F_\lambda(\theta)\|_{\nabla^2 F_\lambda(\theta)^{-1}} \end{cases}$$