Average-case acceleration for bilinear games and normal matrices

Carles Domingo-Enrich[a], Fabian Pedregosa[b], Damien Scieur[c]

[a]Courant Institute (NYU), [b]Google Research, [c]Samsung SAIT AI Lab & Mila

ICLR 2021

May, 2021

Background

– Traditional theory in optimization is worst-case analysis: Not representative of typical behavior.

– Optimal average-case methods have recently been developed for quadratic minimization problems [Berthier et al., 2020, Pedregosa and Scieur, 2020, Lacotte and Pilanci, 2020].

– Optimal methods for smooth games only exist for the worst-case analysis [Azizian et al., 2020].

– **Gap**: Average-case optimal methods for smooth games.

Contributions of the paper

We combine average-case analysis with smooth games.

1. We develop novel average-case optimal algorithms for finding the root of a linear system determined by a (potentially non-symmetric) normal matrix.

2. We show that solving the Hamiltonian using an average-case optimal method is optimal to find equilibria in bilinear games.

Framework (1/2)

– For $\mathbf{A} \in \mathbb{R}^{d\times d}$ and $\mathbf{x}^\star \in \mathbb{R}^d$, we consider the non-symmetric operator problem (**NSO**):

$$\text{Find } \mathbf{x} \; : \; F(\mathbf{x}) \overset{\text{def}}{=} \mathbf{A}(\mathbf{x} - \mathbf{x}^\star) = \mathbf{0}\,.$$

– We define

$$\operatorname{dist}(\boldsymbol{x}, \mathcal{X}^\star) \overset{\text{def}}{=} \min_{\boldsymbol{v}\in\mathcal{X}^\star} \|\boldsymbol{x}-\boldsymbol{v}\|_2, \quad \text{with } \mathcal{X}^\star = \{\boldsymbol{x} \in \mathbb{R}^d \mid \boldsymbol{A}(\boldsymbol{x}-\boldsymbol{x}^\star) = \mathbf{0}\}\,.$$

Framework (2/2)

– **First-order methods**: For all $t \geq 0$,

$$\boldsymbol{x}_t = \boldsymbol{x}_0 + \sum_{i=0}^{t} \alpha_{t,i} F(\boldsymbol{x}_i),$$

for some coefficients $\alpha_{t,i}$.

– For some random $\boldsymbol{A}$, $\boldsymbol{x}^\star$ and initialization $\boldsymbol{x}_0$, average-case first-order optimal algorithms solve:

$$\min_{\boldsymbol{x}_t \text{ first-order method}} \mathbb{E}_{(\boldsymbol{A}, \boldsymbol{x}^\star, \boldsymbol{x}_0)} \text{dist}(\boldsymbol{x}_t, \mathcal{X}^\star)$$

Sketch of the theory

– *Residual polynomial*: polynomial P that satisfies $P(0) = 1$.

– Known: if $(\boldsymbol{x}_t)$ is the sequence generated by a first-order method, there exist residual polynomials P_t of degree at most t such that $\boldsymbol{x}_t - \boldsymbol{x}^\star = P_t(\boldsymbol{A})(\boldsymbol{x}_0 - \boldsymbol{x}^\star)$.

– Empirical spectral distribution of $\boldsymbol{A}$: $\hat{\mu}_{\boldsymbol{A}}(\lambda) = \frac{1}{d}\sum_{i=1}^{d}\delta_{\lambda_i}(\lambda)$, where $(\lambda_i)_{i=1}^{d}$ are the eigenvalues of $\boldsymbol{A}$. Expected spectral distribution: $\mu_{\boldsymbol{A}} = \mathbb{E}_{\boldsymbol{A}}\hat{\mu}_{\boldsymbol{A}}(\lambda)$.

– If $\boldsymbol{A}$ is a random normal matrix and $\boldsymbol{x}^\star$ are sampled appropriately, we show that **for any first order method with associated polynomials (P_t), we have** $\mathbb{E}[\mathrm{dist}(\boldsymbol{x}_t, \mathcal{X}^\star)] = R^2 \int_{\mathbb{C}\setminus\{0\}} |P_t|^2 \,\mathrm{d}\mu_{\boldsymbol{A}}$.

– For simple measures μ, we can compute the sequence of residual polynomials that optimize $\int_{\mathbb{C}\setminus\{0\}} |P_t|^2 \,\mathrm{d}\mu_{\boldsymbol{A}}$, and their corresponding first-order methods.

Average-case optimal methods for bilinear games

We want to find a Nash equilibrium of the zero-sum minimax game given by

$$\min_{\boldsymbol{\theta}_1} \max_{\boldsymbol{\theta}_2} \ell(\boldsymbol{\theta}_1, \, \boldsymbol{\theta}_2) \stackrel{\text{def}}{=} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_1^\star)^\top \boldsymbol{M}(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_2^\star) \, .$$

where $\boldsymbol{\theta}_1, \boldsymbol{\theta}_1^\star \in \mathbb{R}^{d_1}, \boldsymbol{\theta}_2, \boldsymbol{\theta}_2^\star \in \mathbb{R}^{d_2}, \boldsymbol{M} \in \mathbb{R}^{d_1 \times d_2}$. Defining

$$\boldsymbol{A} = \begin{bmatrix} 0 & \boldsymbol{M} \\ -\boldsymbol{M}^\top & 0 \end{bmatrix},$$

we recast the problem as solving the NSO $F(\boldsymbol{x}) \stackrel{\text{def}}{=} \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}^\star) = 0$.

Example: $\boldsymbol{M}$ with i.i.d components

Setting: Each entry of $\boldsymbol{M}$ is sampled from iid from distribution with mean 0 and variance σ^2, in the regime $d_1, d_2 \to \infty, \; d_1/d_2 = r$.

Optimal average-case algorithm.

Initialization. $\boldsymbol{x}_{-1} = \boldsymbol{x}_0 = \big(\boldsymbol{\theta}_{1,0}, \; \boldsymbol{\theta}_{2,0}\big)$.
Main loop. For $t \geq 0$,

$$\boldsymbol{g}_t = F(\boldsymbol{x}_t - F(\boldsymbol{x}_t)) - F(\boldsymbol{x}_t) \qquad \big(= \tfrac{1}{2}\nabla\|F(\boldsymbol{x}_t)\|^2\big)$$

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - h_{t+1}\boldsymbol{g}_t + m_{t+1}(\boldsymbol{x}_{t-1} - \boldsymbol{x}_t) \quad \text{where}$$

$$h_t = -\frac{\delta_t}{\sigma^2\sqrt{r}}, \quad m_t = 1 + \rho\delta_t, \quad \rho = \frac{1+r}{\sqrt{r}}, \quad \delta_t = (-\rho - \delta_{t-1})^{-1}, \quad \delta_0 = 0.$$

Example: M with i.i.d components

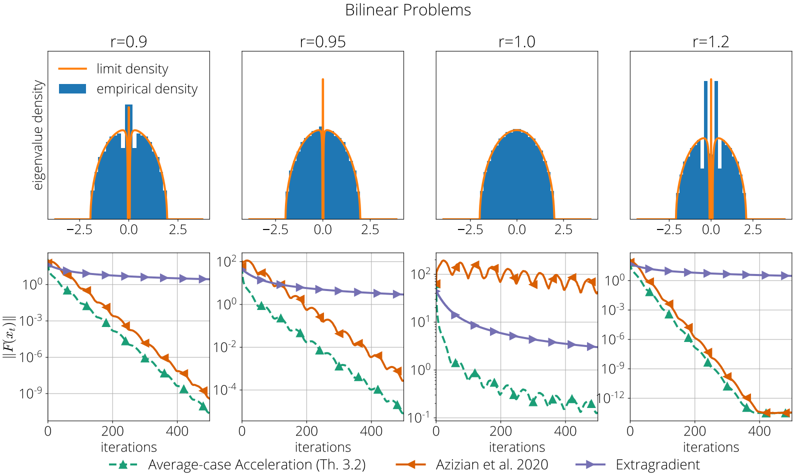

Figure: *First row*: spectral density associated with bilinear games for varying values of the ratio parameter $r = n/d$ (the x-axis represents the imaginary line). *Second row*: Comparison of gradient norm decay with benchmark. The largest gain is in the ill-conditioned regime ($r \approx 1$).

Normal matrices with circular spectral distribution

Setting: Assume that the expected spectral distribution $\mu_{\boldsymbol{A}}$ is the uniform probability measure on the complex disk of center $C \in \mathbb{R}$, $C > 0$ and radius $R < C$.

Optimal average-case algorithm.

Initialization. $y_{-1} = y_0 = \boldsymbol{x}_0$.
Main loop. For $t \geq 0$,

$$y_t = y_{t-1} - \tfrac{1}{C} F(y_{t-1}), \quad \beta_t = (\tfrac{C}{R})^{2t}(t+1), \quad B_t = B_{t-1} + \beta_{t-1},$$

$$\boldsymbol{x}_t = \frac{B_t}{B_t + \beta_t} \boldsymbol{x}_{t-1} + \frac{\beta_t}{B_t + \beta_t} y_t.$$

Moreover, $\mathbb{E}_{(\boldsymbol{A}, \boldsymbol{x}^\star, \boldsymbol{x}_0)} \mathrm{dist}(\boldsymbol{x}_t, \mathcal{X}^\star)$ converges to zero at rate $1/B_t$.

Uniform circular spectral distribution

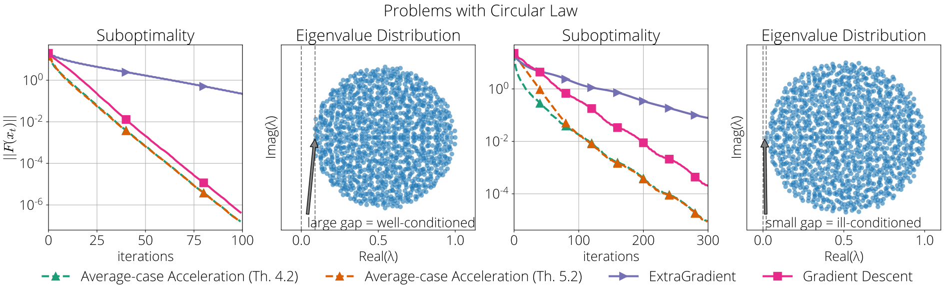

Figure: Benchmarks (columns 1 and 3) and eigenvalue distribution of a design matrix generated with iid entries for two different degrees of conditioning. Depite the normality assumption not being satisfied, we still observe an improvement of average-case optimal methods vs worst-case optimal ones.

References

Azizian, W., Scieur, D., Mitliagkas, I., Lacoste-Julien, S., and Gidel, G. (2020).
Accelerating smooth games by manipulating spectral shapes.
In *Proceedings of Machine Learning Research*.

Berthier, R., Bach, F., and Gaillard, P. (2020).
Accelerated gossip in networks of given dimension using Jacobi polynomial iterations.
SIAM Journal on Mathematics of Data Science, 2(1):24–47.

Lacotte, J. and Pilanci, M. (2020).
Optimal randomized first-order methods for least-squares problems.
Proceedings of the 37th International Conference on Machine Learning.

Pedregosa, F. and Scieur, D. (2020).
Average-case acceleration through spectral density estimation.
In *Proceedings of the 37th International Conference on Machine Learning*.