Achieving the Optimal Convergence Rate in Stochastic Optimisation

Derek Driggs
CCIMI

Advisor: Dr Carola Schönlieb
Collaborators: Dr Jingwei Liang, Dr Clarice Poon

Problem Setting

Many machine learning problems can be formulated as

$$\min_x \quad \boxed{\frac{1}{n}\sum_{i=1}^{n} f_i(x)} + \boxed{g(x)}$$

Convex, smooth functions with L-Lipschitz continuous gradients

Convex, non-smooth function

Problem Setting

$$\min_x \quad \boxed{\frac{1}{n}\sum_{i=1}^{n} f_i(x)} + \boxed{g(x)}$$

$\|\mathcal{A}(\cdot)\|_2$ – ℓ_2-norm with linear operator $\mathcal{A}$

$\|\cdot\|_1$ – ℓ_1-norm
$\|\cdot\|_*$ – nuclear norm

Applications: LASSO, Robust PCA, Logistic Regression

Proximal Gradient Descent

$$\min_x \quad \boxed{\frac{1}{n}\sum_{i=1}^{n} f_i(x)} + \boxed{g(x)}$$

This problem can be solved using *proximal gradient descent*:

$$x_{k+1} = \mathrm{prox}_{\gamma g}\left(x_k - \gamma \nabla f(x_k)\right)$$

where the *proximal operator* is defined as

$$\mathrm{prox}_h(y) := \mathrm{argmin}_x \quad \frac{1}{2}\|x - y\|^2 + h(x)$$

This algorithm requires the evaluation of n gradients per iteration.

Stochastic Gradient Descent

Computing the full gradient is expensive for large n, so we can replace $\nabla f(x_k)$ with an estimate of the gradient, $\widetilde{\nabla} f(x_k)$, where

$$\widetilde{\nabla}_{\text{SGD}} f(x_k) = \nabla f_j(x_k), \qquad \text{(SGD)}$$

$$\widetilde{\nabla}_{\text{SAGA}} f(x_k) = \nabla f_j(x_k) - \nabla f_j(\varphi_k^j) + \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\varphi_k^i), \qquad \text{(SAGA)}$$

$$\widetilde{\nabla}_{\text{SVRG}} f(x_k) = \nabla f_j(x_k) - \nabla f_j(\tilde{x}) + \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{x}). \qquad \text{(SVRG)}$$

The index j is chosen uniformly at random.

φ_k^j — The gradient $\nabla f_j(\varphi_k^j)$ is stored for future iterates.

$\tilde{x}$ — The full gradient $\frac{1}{n}\sum_{i=1}^{n} \nabla f(\tilde{x})$ is computed every $2n$ iterations and stored for future iterates.

Convergence Rates

With x^* the minimiser of $F(x) := \frac{1}{n}\sum_{i=1}^n f_i(x) + g(x)$, the suboptimality at iteration k is $F(x_k) - F(x^*)$.

For proximal gradient descent on convex objectives,

$$F(x_k) - F(x^*) \leq \mathcal{O}\left(\frac{1}{k}\right)$$

For SVRG and SAGA,

$$\mathbb{E}\left[F(x_k) - F(x^*)\right] \leq \mathcal{O}\left(\frac{1}{k}\right)$$

Accelerating Full Gradient Methods

Nesterov's momentum trick is a slight modification that offers enormous acceleration:

$$y_{k+1} = x_k + \alpha(x_k - x_{k-1})$$
$$x_{k+1} = \text{prox}_{\gamma g}\left(y_{k+1} - \gamma \nabla f(y_{k+1})\right)$$

With α and γ chosen appropriately,

$$F(x_k) - F(x^*) \leq \mathcal{O}\left(\frac{1}{k^2}\right),$$

and this rate is optimal.

Visualising Momentum

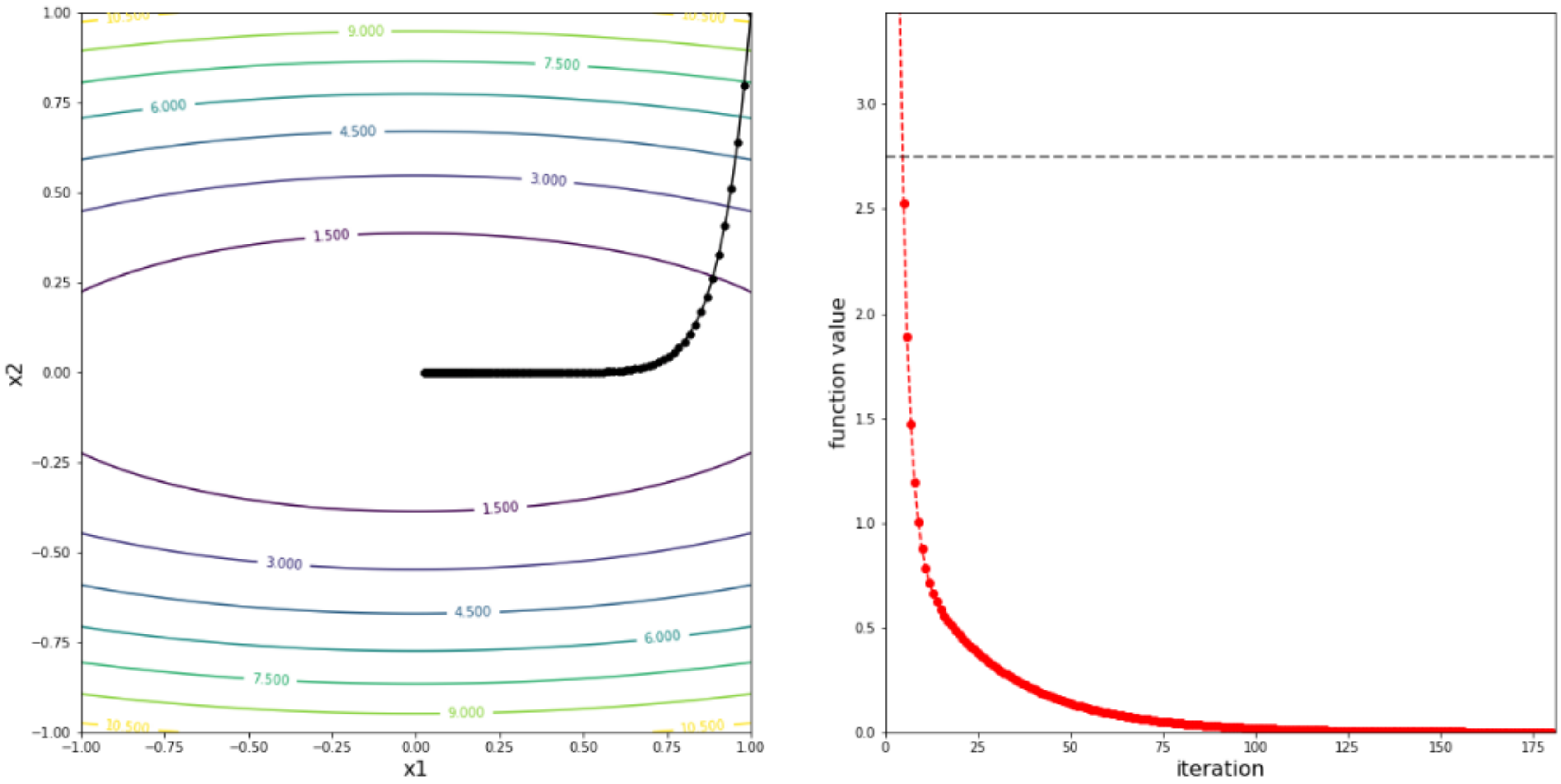

Figure: Using gradient descent to minimize $x_1^2 + 10x_2^2$ without momentum.

Visualising Momentum

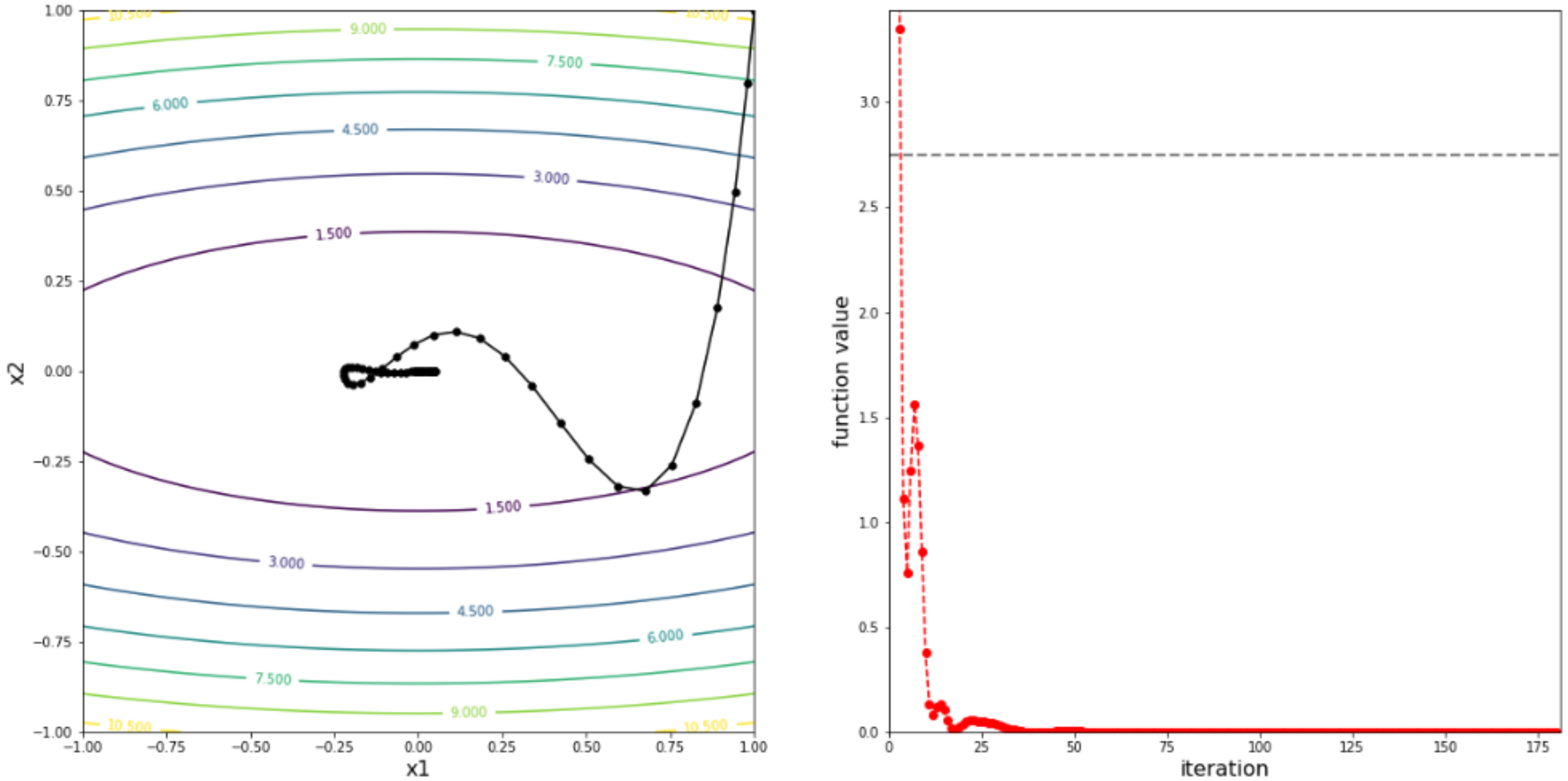

Figure: Using gradient descent to minimize $x_1^2 + 10x_2^2$ with momentum.

Accelerating Stochastic Gradient Methods

In the stochastic setting, momentum propagates "bad" gradient evaluations, so it is unclear whether momentum-based methods improve performance.

"Katyusha" (Allen-Zhu, 2017) achieves the optimal $\mathcal{O}\left(\frac{1}{k^2}\right)$-rate using **negative momentum** and **linear coupling**:

$$
\begin{aligned}
x_{k+1} &= \alpha_1 z_k + \boxed{\alpha_2 \tilde{x}} + (1 - \alpha_1 - \alpha_2) y_k \\
y_{k+1} &= \text{prox}\left(x_{k+1} - \gamma \alpha_1 \widetilde{\nabla}_{\text{SVRG}} f(x_{k+1})\right) \\
z_{k+1} &= \text{prox}\left(z_k - \gamma \widetilde{\nabla}_{\text{SVRG}} f(x_{k+1})\right)
\end{aligned}
$$

The term $\alpha_2 \tilde{x}$ "attracts" x_{k+1}, supposedly limiting the effects of detrimental gradient evaluations.

Our Results

Using linear-coupling analysis, we prove the following:

Theorem

Consider the algorithm

$$\begin{aligned} x_{k+1} &= \alpha z_k + (1-\alpha) y_k \\ y_{k+1} &= \mathsf{prox}\left(x_{k+1} - \gamma\alpha\widetilde{\nabla} f(x_{k+1})\right) \\ z_{k+1} &= \mathsf{prox}\left(z_k - \gamma\widetilde{\nabla} f(x_{k+1})\right) \end{aligned}$$

with $\widetilde{\nabla} = \widetilde{\nabla}_{\mathsf{SAGA}}$ or $\widetilde{\nabla}_{\mathsf{SVRG}}$. Choosing α and γ appropriately,

$$F(x_k) - F(x^*) \leq \mathcal{O}\left(\frac{1}{k^2}\right)$$

Our analysis can be easily extended to many other (variance-reduced) stochastic gradient estimators.

Our Results

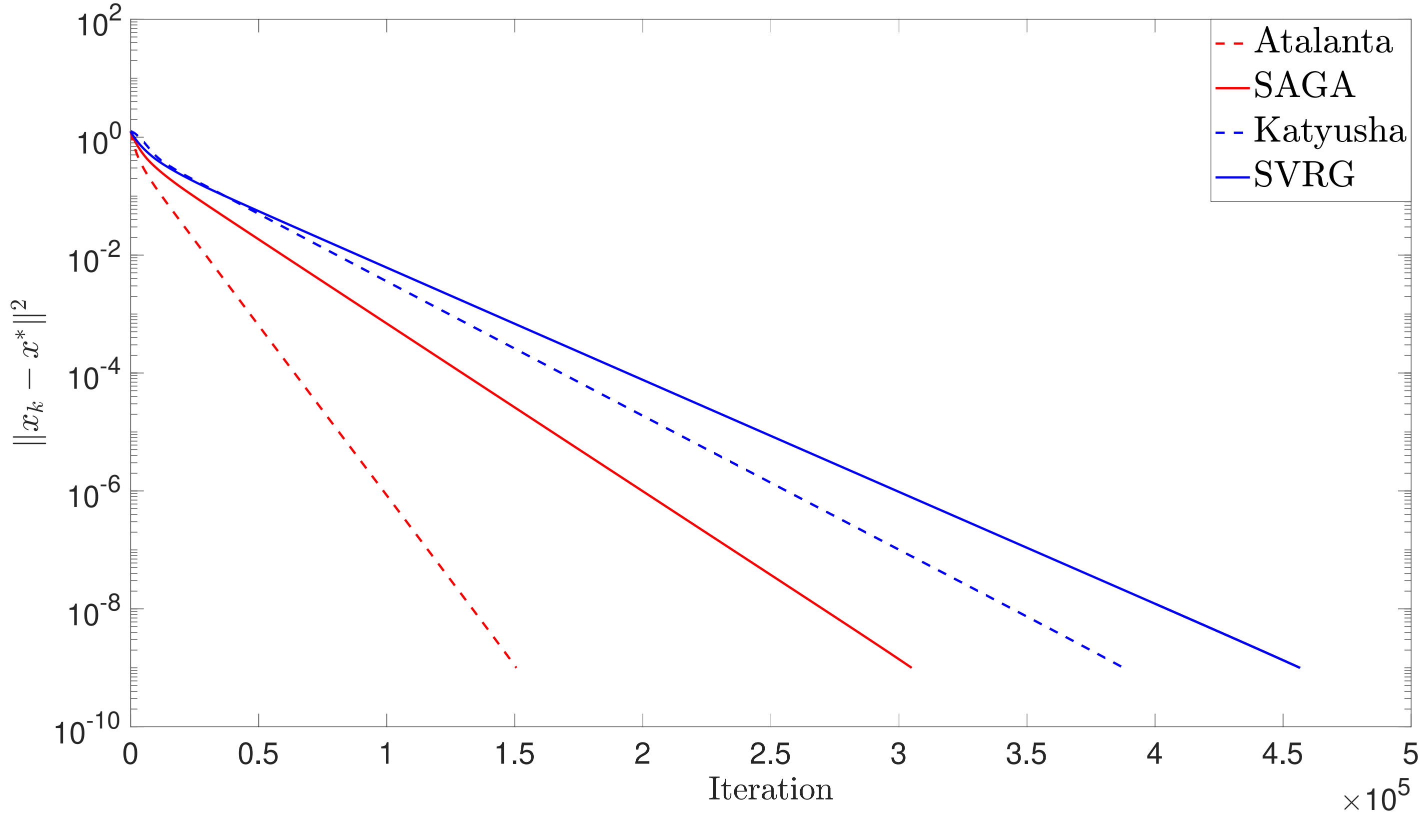

Figure: Comparing stochastic gradient methods on a sparse logistic regression problem (see poster for details).

"Atalanta" uses our acceleration framework with the SAGA gradient estimate, and is extremely fast.