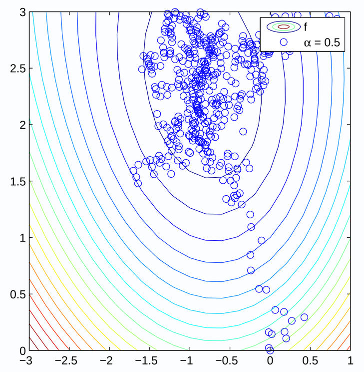

Tracking the gradients using the Hessian

Robert M. Gower, Nicolas Le Roux, and Francis Bach

robert.gower@telecom-paristech.fr, nicolas.le.roux@gmail.com, francis.bach@inria.fr

1. The problem

Minimize the average loss over N samples

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{N}\sum_{j=1}^{N} f_j(\theta), \quad (1)$$

where $f_i(\theta)$ is the loss incurred by parameters θ for the i-th sample. We assume each f_i is twice differentiable. We use the abbreviations

$$H_i(\theta) \equiv \nabla^2 f_i(\theta) \quad \text{and} \quad H(\theta) \equiv \frac{1}{N}\sum_{j=1}^{N} \nabla^2 f_j(\theta).$$

2. SGD with covariates

We solve (1) using an iterative 1st order method

$$\theta_{t+1} = \theta_t + \alpha g_t,$$

where g_t is an unbiased estimator of the gradient

$$\mathbb{E}[g_t] = \frac{1}{N}\sum_{j=1}^{N} \nabla f_j(\theta_t).$$

Using stochastic gradients with covariates $z_i(\theta_t) \in \mathbb{R}^d$ we can design an efficient method and control the variance

$$g_t = \nabla f_i(\theta_t) - z_i(\theta_t) + \frac{1}{N}\sum_{j=1}^{N} z_j(\theta_t)\,, \quad (2)$$

Specifically, if $z_i \approx \nabla f_i(\theta_t)$ then

$$\mathbb{VAR}[g_t] \leq \mathbb{VAR}[\nabla f_i(\theta_t)].$$

3. Building covariates using the Taylor expansion

Fix a reference point $\tilde{\theta} \in \mathbb{R}^d$ which is close to θ_t.

Zero order Taylor. Using $z_i(\theta_t) = \nabla f_i(\tilde{\theta}) \approx \nabla f_i(\theta_t)$ in (2) gives the SVRG gradient estimate:

$$g_t = \nabla f_i(\theta_t) - \nabla f_i(\bar{\theta}) + \frac{1}{N}\sum_{j=1}^{N} \nabla f_i(\bar{\theta}).$$

First order Taylor. Using $z_i(\theta_t) = \nabla f_i(\tilde{\theta}) + H_i(\tilde{\theta})(\theta_t - \tilde{\theta})$ in (2) gives SVRG2:

$$g_t = \nabla f_i(\theta_t) - \nabla f_i(\bar{\theta}) + \frac{1}{N}\sum_{j=1}^{N} \nabla f_i(\bar{\theta}) + \left(H(\bar{\theta}) - H_i(\bar{\theta}_k)\right)(\theta_t - \bar{\theta}).$$

4. SVRG algorithm

Parameter: Functions f_i for $i = 1, \ldots, N$
Choose $\bar{\theta} \in \mathbb{R}^d$ and stepsize $\gamma > 0$
for $k = 0, \ldots, K-1$ **do**
  Calculate $\frac{1}{N}\sum_{j=1}^{N} \nabla f_j(\bar{\theta})$, $\theta_0 = \bar{\theta}$
  for $t = 0, 1, 2, \ldots, T-1$ **do**
    $i \sim \mathcal{U}[1, N]$
    $g_t = \nabla f_i(\theta_t) - \nabla f_i(\bar{\theta}) + \frac{1}{N}\sum_{j=1}^{N} \nabla f_j(\bar{\theta})$
    $\theta_{t+1} = \theta_t - \gamma g_t$
  $\bar{\theta} = \theta_T$
Output $\bar{\theta}$

5. SVRG2 algorithm

Parameter: Functions f_i for $i = 1, \ldots, N$
Choose $\bar{\theta} \in \mathbb{R}^d$ and stepsize $\gamma > 0$
for $k = 0, \ldots, K-1$ **do**
  Calculate $\frac{1}{N}\sum_{j=1}^{N} \nabla f_j(\bar{\theta})$, $\theta_0 = \bar{\theta}$
  Calculate $H(\bar{\theta}) = \frac{1}{N}\sum_{j=1}^{N} H_j(\bar{\theta})$
  for $t = 0, 1, 2, \ldots, T-1$ **do**
    $i \sim \mathcal{U}[1, N]$
    $g_t = \nabla f_i(\theta_t) - \nabla f_i(\bar{\theta}) + \frac{1}{N}\sum_{j=1}^{N} \nabla f_j(\bar{\theta}) + \left(H(\bar{\theta}) - H_i(\bar{\theta})\right)(\theta_t - \bar{\theta})$
    $\theta_{t+1} = \theta_t - \gamma g_t$
  $\bar{\theta} = \theta_T$
Output $\bar{\theta}$

6. Costs and approximations

SVRG2 uses the following quantities:

- Full Hessian $\frac{1}{N}\sum_{j=1}^{N} H_j(\bar{\theta})$ costs $O(nd \times \text{eval}(f_i))$
- Hessian vector product $\frac{1}{N}\sum_{j=1}^{N} H_j(\bar{\theta})(\theta_t - \bar{\theta})$ costs $O(d^2)$

To bring down costs use approximations

$$\tilde{H}_i(\bar{\theta}) \approx H_i(\bar{\theta}) =: H_i$$

We use **Diagonal**, **rank-1 secant equation** and **low rank** sketching based approximations.

7. Diagonal Approximations

Robust secant equation: We can robustify the the secant equation

$$\hat{H}_i(\theta_t - \bar{\theta}) = \nabla f_i(\theta_t) - \nabla f_i(\bar{\theta}),$$

by minimizing the average squared-ℓ_2 distance within a small ball around the previous direction.

$$\hat{H}_i = \arg\min_{X \in \mathbb{R}^{d\times d}} \int_{\xi} \left\|(X - H_i)(\theta_t - \bar{\theta} + \xi)\right\|^2 p(\xi) d\xi.$$

Assuming $\xi \sim \mathcal{N}(0, \sigma^2 I)$, we get

$$\hat{H}_i = \frac{(\theta_t - \bar{\theta}) \odot \left(\nabla f_i(\theta_t) - \nabla f_i(\bar{\theta})\right) + \sigma^2 \text{diag}(H_i(\bar{\theta}))}{(\theta_t - \bar{\theta}) \odot (\theta_t - \bar{\theta}) + \sigma^2}$$

where we used $H_i(\bar{\theta})(\theta_t - \bar{\theta}) \approx \nabla f_i(\theta_t) - \nabla f_i(\bar{\theta})$.

8. Low-rank Action Matching

Use a *sketch* of the true Hessian to form an approximate Hessian. Let $S \in \mathbb{R}^{d\times\tau}$ with $\tau \ll d$ be a sketching matrix sampled $S \sim \mathcal{D}$ from a distribution over matrices.

$$\hat{H}_i = \arg\min_{X \in \mathbb{R}^{d\times d}} \|X\|^2_{F(H)}$$
$$\text{subject to} \quad XS = H_i S, \quad X = X^\top. \quad (3)$$

The solution is a rank 2τ matrix given by

$$\hat{H}_i = HS(S^T HS)^{-1}S^\top H_i\left(I - S(S^T HS)^{-1}S^\top H\right) + H_i S(S^T HS)^{-1}S^\top H.$$

9. Numerics

We experiment with two sketching matrices. Let

$$\bar{g}_i = \frac{\tau}{T}\sum_{j=\frac{T}{\tau}i}^{\frac{T}{\tau}(i+1)-1} g_j,$$

for $i = 0, \ldots, \tau-1$, be the the inner gradients averaged into τ buckets.

Legend	Description
AMprev−τ	$S = [\bar{g}_0, \ldots, \bar{g}_{\tau-1}]$.
AMgauss−τ	$S \sim \mathcal{N}(0, I)$ Gaussian entries
2D	$\hat{H}_i = \text{diag}(H_i)$
2Dsec	Secant+diagonal with $\sigma = 1$

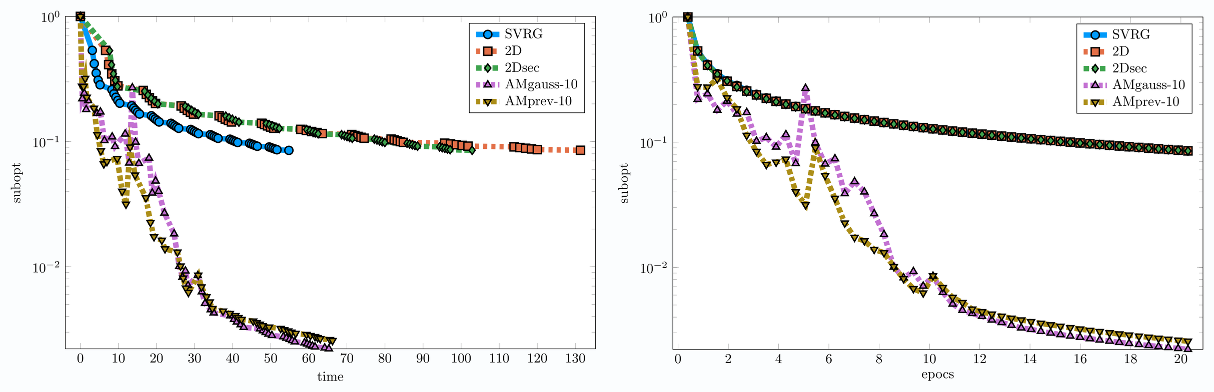

Figure 1: gisette_scale $(N; d) = (6000; 5000)$

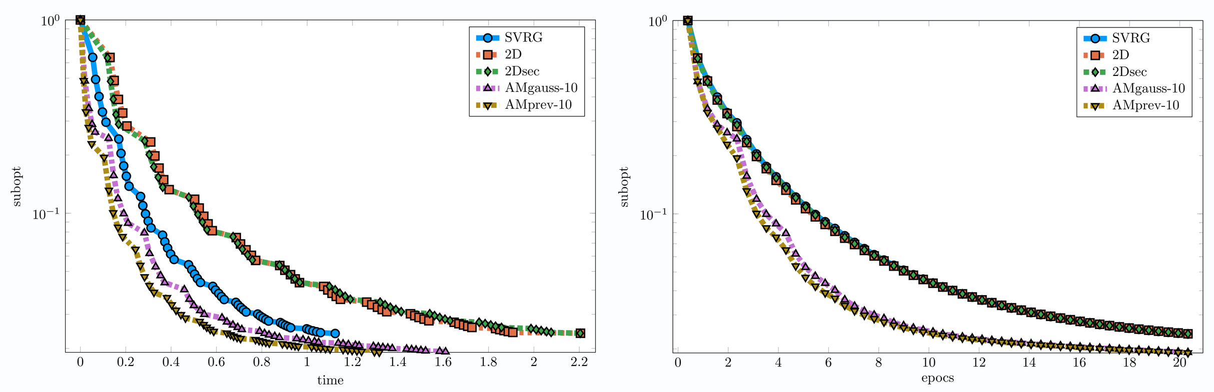

Figure 2: madelon $(N; d) = (2000; 200)$

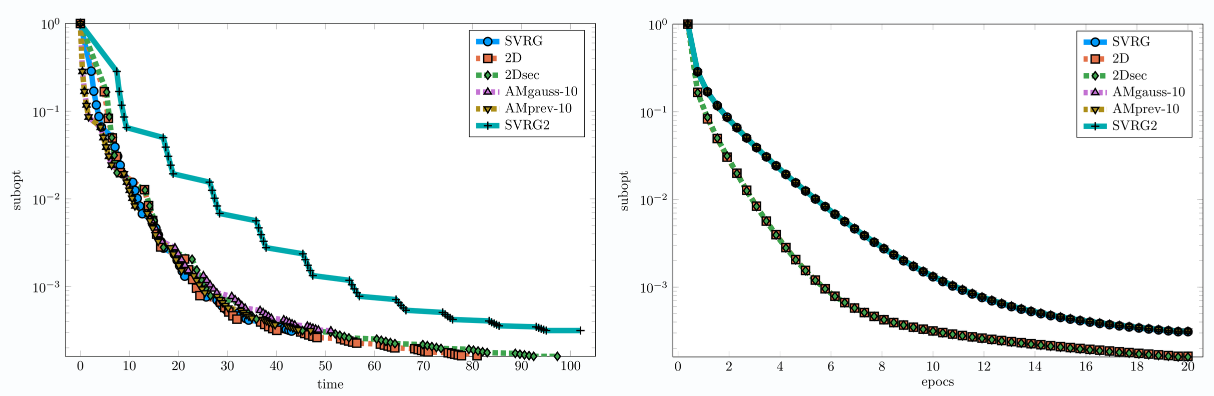

Figure 3: covtype $(N; d) = (581012; 54)$

References

[1] Rie Johnson and Tong Zhang (2014). In: Advances in Neural Information Processing Systems. Accelerating stochastic gradient descent using predictive variance reduction

[2] RMG, P. Richtarik (2015), SIAM. J. Matrix Anal. & Appl. Randomized Iterative Methods for Linear Systems

[3] Bruce Christianson (1992), IMA Journal of Numerical Analysis, Automatic Hessians by reverse accumulation