Tracking the gradients using the Hessian: A new look at variance reducing stochastic methods

Robert M. Gower

Joint work with Nicolas Le Roux and Francis Bach

AISTATS 2018
Playa Blanca, Lanzarote, Canary Islands, April 9 - 11, 2018

Solve Empirical Risk Minimization

$$\min_{\theta \in \mathbf{R}^d} f(\theta) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(\theta),$$

where n is the num of data points and $d$the num of features.

Datum functions $f_i(\theta)$ is twice differentiable

Ridge Regression $f_i(\theta) = (y^i - \langle \theta, x^i \rangle)^2 + \lambda ||\theta||_2^2$

Logistic regression $f_i(\theta) = \ln(1 + e^{-y^i \langle \theta, x^i \rangle}) + \lambda ||\theta||_2^2$

Some neural nets $f_i(\theta) = \ldots$

Using a first order gradient method

$$\theta_{t+1} = \theta_t - \gamma g_t$$

Stepsize $\gamma > 0$

Unbiased

$$\mathbb{E}[g_t] = \nabla f(\theta_t)$$

Using a first order gradient method

$$\theta_{t+1} = \theta_t - \gamma g_t$$

Stepsize $\gamma > 0$

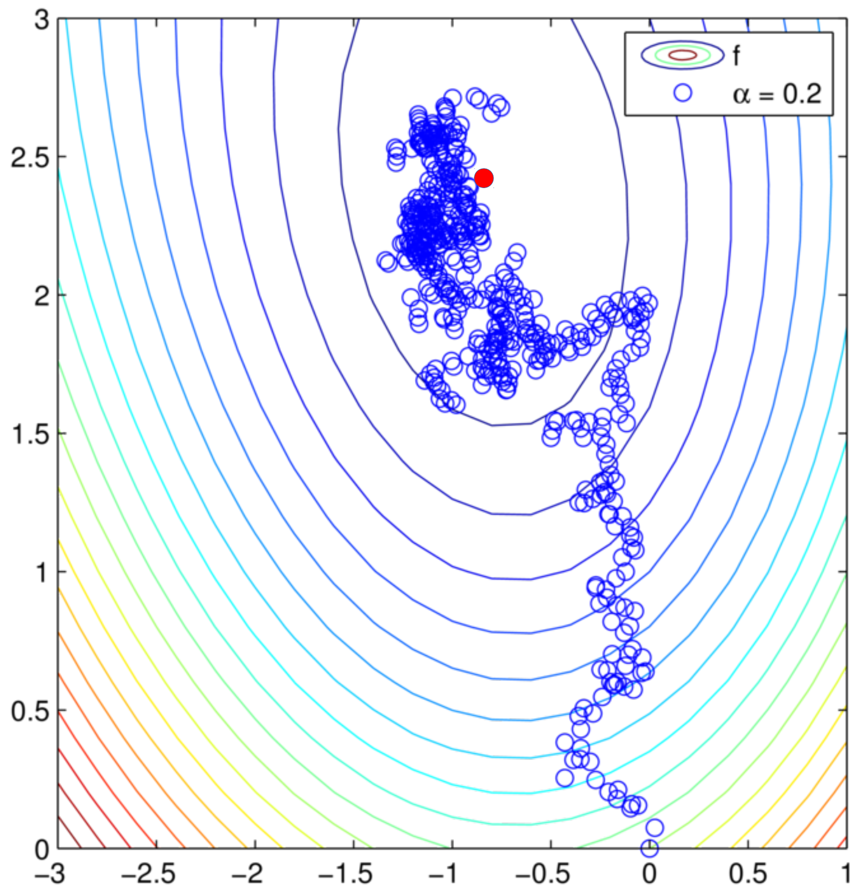
Unbiased

$$\mathbb{E}[g_t] = \nabla f(\theta_t)$$

EXE: Stochastic Gradient descent (SGD)

$$g_t = \nabla f_i(\theta_t), \quad \text{where } i \sim \mathcal{U}\{1, \ldots, n\}$$

Stochastic Gradient Descent γ = 0.2



Using a first order gradient method

$$\theta_{t+1} = \theta_t - \gamma g_t$$

Stepsize $\gamma > 0$

Unbiased

$$\mathbb{E}[g_t] = \nabla f(\theta_t)$$

EXE: Stochastic Gradient descent (SGD)

$$g_t = \nabla f_i(\theta_t), \quad \text{where } i \sim \mathcal{U}\{1, \ldots, n\}$$

Using a first order gradient method

$$\theta_{t+1} = \theta_t - \gamma g_t$$

Stepsize $\gamma > 0$

Unbiased

$$\mathbb{E}[g_t] = \nabla f(\theta_t)$$

EXE: Stochastic Gradient descent (SGD)

$$g_t = \nabla f_i(\theta_t), \quad \text{where } i \sim \mathcal{U}\{1, \ldots, n\}$$

EXE: SGD with covariates

$$g_t = \nabla f_i(\theta_t) - z_i + \frac{1}{n}\sum_{j=1}^{n} z_j, \; \text{where } i \sim \mathcal{U}\{1, \ldots, n\}$$

$$z_i \in \mathbb{R}^d, \text{ for } i = 1, \ldots, n$$

Choosing the covariates

SGD with covariates: $$g_t = \nabla f_i(\theta_t) - z_i + \frac{1}{n}\sum_{j=1}^{n} z_j$$

1) Correlated to the stochastic gradients

$$\text{If } \nabla f_i(\theta_t) \approx z_i \text{ then } \mathbb{VAR}(g_t) \leq \mathbb{VAR}(\nabla f_i(\theta_t))$$

Choosing the covariates

SGD with covariates:

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

1) Correlated to the stochastic gradients

$$\text{If } \nabla f_i(\theta_t) \approx z_i \text{ then } \mathbb{VAR}(g_t) \leq \mathbb{VAR}(\nabla f_i(\theta_t))$$

2) Cheap to compute

$$\text{cost}(g_t) \quad \leq \quad \text{cost}\Big(\frac{1}{n}\sum_{j=1}^{n}\nabla f_j(\theta_t)\Big)$$

Choosing the covariates

SGD with covariates:

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

1) Correlated to the stochastic gradients

$$\text{If } \nabla f_i(\theta_t) \approx z_i \text{ then } \mathbb{VAR}(g_t) \leq \mathbb{VAR}(\nabla f_i(\theta_t))$$

2) Cheap to compute

$$\text{cost}(g_t) \quad \leq \quad \text{cost}\Big(\frac{1}{n}\sum_{j=1}^{n} \nabla f_j(\theta_t)\Big)$$

EXE: Too costly

$$z_i = \nabla f_i(\theta_t)$$

$$g_t = \nabla f(\theta_t)$$

Choosing the covariates

SGD with covariates: $$g_t = \nabla f_i(\theta_t) - z_i + \frac{1}{n}\sum_{j=1}^n z_j$$

1) Correlated to the stochastic gradients

$$\text{If } \nabla f_i(\theta_t) \approx z_i \text{ then } \mathbb{VAR}(g_t) \leq \mathbb{VAR}(\nabla f_i(\theta_t))$$

2) Cheap to compute

$$\text{cost}(g_t) \quad \leq \quad \text{cost}\Big(\frac{1}{n}\sum_{j=1}^n \nabla f_j(\theta_t)\Big)$$

EXE: Too costly

$$z_i = \nabla f_i(\theta_t)$$

$$g_t = \nabla f(\theta_t)$$

EXE: High variance

$$z_i = 0$$

$$g_t = \nabla f_i(\theta_t)$$

Choosing the covariates

SGD with covariates:

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

1) Correlated to the stochastic gradients

$$\text{If } \nabla f_i(\theta_t) \approx z_i \text{ then } \mathbb{VAR}(g_t) \leq \mathbb{VAR}(\nabla f_i(\theta_t))$$

2) Cheap to compute

$$\text{cost}(g_t) \quad \leq \quad \text{cost}\Big(\frac{1}{n}\sum_{j=1}^{n} \nabla f_j(\theta_t)\Big)$$

EXE: Too costly

$$z_i = \nabla f_i(\theta_t)$$

$$g_t = \nabla f(\theta_t)$$

Want something in between

EXE: High variance

$$z_i = 0$$

$$g_t = \nabla f_i(\theta_t)$$

SVRG: Stochastic Variance Reduced Gradients

$$\theta_{t+1} = \theta_t - \gamma g_t$$

Reference point $\tilde{\theta} \in \mathbb{R}^d$

SVRG: Stochastic Variance Reduced Gradients

$$\theta_{t+1} = \theta_t - \gamma g_t$$

Reference point	$\tilde{\theta} \in \mathbb{R}^d$
Sample	$\nabla f_i(\theta_t), \quad i \in \{1, \ldots, n\}$ uniformly

SVRG: Stochastic Variance Reduced Gradients

$$\theta_{t+1} = \theta_t - \gamma g_t$$

Reference point	$\tilde{\theta} \in \mathbb{R}^d$
Sample	$\nabla f_i(\theta_t), \quad i \in \{1, \ldots, n\}$ uniformly
0th order Taylor	$\lvert\lvert\tilde{\theta} - \theta_t\rvert\rvert$ is small $\Rightarrow \nabla f_i(\theta_t) \approx \nabla f_i(\tilde{\theta})$

SVRG: Stochastic Variance Reduced Gradients

$$\theta_{t+1} = \theta_t - \gamma g_t$$

Reference point	$\tilde{\theta} \in \mathbb{R}^d$
Sample	$\nabla f_i(\theta_t), \quad i \in \{1, \ldots, n\}$ uniformly
0th order Taylor	$\|\|\tilde{\theta} - \theta_t\|\|$ is small $\Rightarrow \nabla f_i(\theta_t) \approx \nabla f_i(\tilde{\theta})$ $=: z_i$

SVRG: Stochastic Variance Reduced Gradients

$$\theta_{t+1} = \theta_t - \gamma g_t$$

Reference point	$\tilde{\theta} \in \mathbb{R}^d$
Sample	$\nabla f_i(\theta_t), \quad i \in \{1, \ldots, n\}$ uniformly
0th order Taylor	$\|\|\tilde{\theta} - \theta_t\|\|$ is small $\Rightarrow \nabla f_i(\theta_t) \approx \nabla f_i(\tilde{\theta}) =: z_i$
SVRG	$g_t = \nabla f_i(\theta_t) - \nabla f_i(\tilde{\theta}) + \nabla f(\tilde{\theta})$

SVRG: Stochastic Variance Reduced Gradients

$$\theta_{t+1} = \theta_t - \gamma g_t$$

Reference point	$\tilde{\theta} \in \mathbb{R}^d$
Sample	$\nabla f_i(\theta_t), \quad i \in \{1, \ldots, n\}$ uniformly
0th order Taylor	$\|\|\tilde{\theta} - \theta_t\|\|$ is small $\Rightarrow \nabla f_i(\theta_t) \approx \nabla f_i(\tilde{\theta})$ $=: z_i$
SVRG	$g_t = \nabla f_i(\theta_t) - \nabla f_i(\tilde{\theta}) + \nabla f(\tilde{\theta})$

$$g_t = \nabla f_i(\theta_t) - z_i \qquad + \frac{1}{n} \sum_{j=1} z_j$$

SVRG: Stochastic Variance Reduced Gradients

Set $\theta_0 = 0$, choose $\gamma > 0, m \in \mathbb{N}$

$\tilde{\theta}_0 = \theta_0$

for $k = 0, 1, 2, \ldots, T-1$

  calculate $\nabla f(\tilde{\theta}_k)$

  $\theta_0 = \tilde{\theta}_k$

  for $t = 0, 1, 2, \ldots, m-1$

    sample $i \in \{1, \ldots, n\}$

    $g_t = \nabla f_i(\theta_t) - \nabla f_i(\tilde{\theta}_k) + \nabla f(\tilde{\theta}_k)$

    $\theta_{t+1} = \theta_t - \gamma g_t$

  $\tilde{\theta}_{k+1} = \theta_m$

Output $\tilde{\theta}_T$

Johnson, R. & Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction Advances in Neural Information Processing Systems, 2013, 315-323

SVRG: Stochastic Variance Reduced Gradients

Set $\theta_0 = 0$, choose $\gamma > 0, m \in \mathbb{N}$

$\tilde{\theta}_0 = \theta_0$

for $k = 0, 1, 2, \ldots, T-1$

    calculate $\nabla f(\tilde{\theta}_k)$

    $\theta_0 = \tilde{\theta}_k$

    for $t = 0, 1, 2, \ldots, m-1$

        sample $i \in \{1, \ldots, n\}$

        $g_t = \nabla f_i(\theta_t) - \nabla f_i(\tilde{\theta}_k) + \nabla f(\tilde{\theta}_k)$

        $\theta_{t+1} = \theta_t - \gamma g_t$

    $\tilde{\theta}_{k+1} = \theta_m$

Output $\tilde{\theta}_T$

Freeze reference point for m iterations

Johnson, R. & Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction Advances in Neural Information Processing Systems, 2013, 315-323

SVRG: Stochastic Variance Reduced Gradients

Set $\theta_0 = 0$, choose $\gamma > 0, m \in \mathbb{N}$

$\tilde{\theta}_0 = \theta_0$

for $k = 0, 1, 2, \ldots, T-1$

  calculate $\nabla f(\tilde{\theta}_k)$

  $\theta_0 = \tilde{\theta}_k$

  for $t = 0, 1, 2, \ldots, m-1$

    sample $i \in \{1, \ldots, n\}$

    $g_t = \nabla f_i(\theta_t) - \nabla f_i(\tilde{\theta}_k) + \nabla f(\tilde{\theta}_k)$

    $\theta_{t+1} = \theta_t - \gamma g_t$

  $\tilde{\theta}_{k+1} = \theta_m$

Output $\tilde{\theta}_T$

Freeze reference point for m iterations

Why not 1^{st} Taylor?

Johnson, R. & Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction Advances in Neural Information Processing Systems, 2013, 315-323

SVRG2: Second order tracking

$$\theta_{t+1} = \theta_t - \gamma g_t$$

Reference point	$\tilde{\theta} \in \mathbb{R}^d$
1st order Taylor exp.	$\nabla f_i(\theta_t) \approx \nabla f_i(\tilde{\theta}) + H_i(\tilde{\theta})(\theta_t - \tilde{\theta})$

SVRG2: Second order tracking

$$\theta_{t+1} = \theta_t - \gamma g_t$$

Reference point

$$\tilde{\theta} \in \mathbb{R}^d$$

$$H_i(\tilde{\theta}) := \nabla^2 f_i(\tilde{\theta})$$

1st order Taylor exp.

$$\nabla f_i(\theta_t) \approx \nabla f_i(\tilde{\theta}) + H_i(\tilde{\theta})(\theta_t - \tilde{\theta}) =: z_i$$

SVRG2: Second order tracking

$$\theta_{t+1} = \theta_t - \gamma g_t$$

Reference point	$\tilde{\theta} \in \mathbb{R}^d$	$H_i(\tilde{\theta}) := \nabla^2 f_i(\tilde{\theta})$
1st order Taylor exp.	$\nabla f_i(\theta_t) \approx \nabla f_i(\tilde{\theta}) + H_i(\tilde{\theta})(\theta_t - \tilde{\theta})$	$=: z_i$
Expected covariate	$\frac{1}{n}\sum_{j=1} z_j = \nabla f(\tilde{\theta}) + \frac{1}{n}\sum_{i=1} H_i(\tilde{\theta})(\theta_t - \tilde{\theta})$	

SVRG2: Second order tracking

$$\theta_{t+1} = \theta_t - \gamma g_t$$

Reference point	$\tilde{\theta} \in \mathbb{R}^d$	$H_i(\tilde{\theta}) := \nabla^2 f_i(\tilde{\theta})$
1st order Taylor exp.	$\nabla f_i(\theta_t) \approx \nabla f_i(\tilde{\theta}) + H_i(\tilde{\theta})(\theta_t - \tilde{\theta})$	$=: z_i$
Expected covariate	$\frac{1}{n}\sum_{j=1} z_j = \nabla f(\tilde{\theta}) + \frac{1}{n}\sum_{i=1} H_i(\tilde{\theta})(\theta_t - \tilde{\theta})$	
SVRG2	$g_t = \nabla f_i(\theta_t) - z_i + \frac{1}{n}\sum_{j=1} z_j$ $= \nabla f_i(\theta_t) - \nabla f_i(\tilde{\theta}) + \nabla f(\tilde{\theta})$ $+(\frac{1}{n}\sum_{j=1}^{n} H_j(\tilde{\theta}) - H_i(\tilde{\theta}))(\theta_t - \tilde{\theta})$	

SVRG2: Second order tracking

$$\theta_{t+1} = \theta_t - \gamma g_t$$

Reference point

$$\tilde{\theta} \in \mathbb{R}^d$$

$$H_i(\tilde{\theta}) := \nabla^2 f_i(\tilde{\theta})$$

1st order Taylor exp.

$$\nabla f_i(\theta_t) \approx \nabla f_i(\tilde{\theta}) + H_i(\tilde{\theta})(\theta_t - \tilde{\theta}) =: z_i$$

Expected covariate

$$\frac{1}{n}\sum_{j=1} z_j = \nabla f(\tilde{\theta}) + \frac{1}{n}\sum_{i=1} H_i(\tilde{\theta})(\theta_t - \tilde{\theta})$$

SVRG2

$$\begin{aligned} g_t &= \nabla f_i(\theta_t) - z_i + \frac{1}{n}\sum_{j=1} z_j \\ &= \nabla f_i(\theta_t) - \nabla f_i(\tilde{\theta}) + \nabla f(\tilde{\theta}) \\ &\quad + \left(\frac{1}{n}\sum_{j=1}^{n} H_j(\tilde{\theta}) - H_i(\tilde{\theta})\right)(\theta_t - \tilde{\theta}) \end{aligned}$$

PDF Adobe H. T. Wai, W. Shi, A. Nedic, and A. Scaglione. Curvature-aided incremental aggregated gradient method, Allerton. IEEE, 2017,

SVRG2: Stochastic Variance Reduced Gradients with tracking

Set $\theta_0 = 0$, choose $\gamma > 0, m \in \mathbb{N}$

$\tilde{\theta} = \theta_0$

for $k = 0, 1, 2, \ldots, T-1$

  calculate $\nabla f(\tilde{\theta}), \underline{H = \nabla^2 f(\tilde{\theta})}$

  $\theta_0 = \tilde{\theta}$

  for $t = 0, 1, 2, \ldots, m-1$

    sample $i \in \{1, \ldots, n\}$

    $$g_t = \nabla f_i(\theta_t) - \nabla f_i(\tilde{\theta}) + \nabla f(\tilde{\theta}) + \underline{(H - H_i(\tilde{\theta}))(\theta_t - \tilde{\theta})}$$

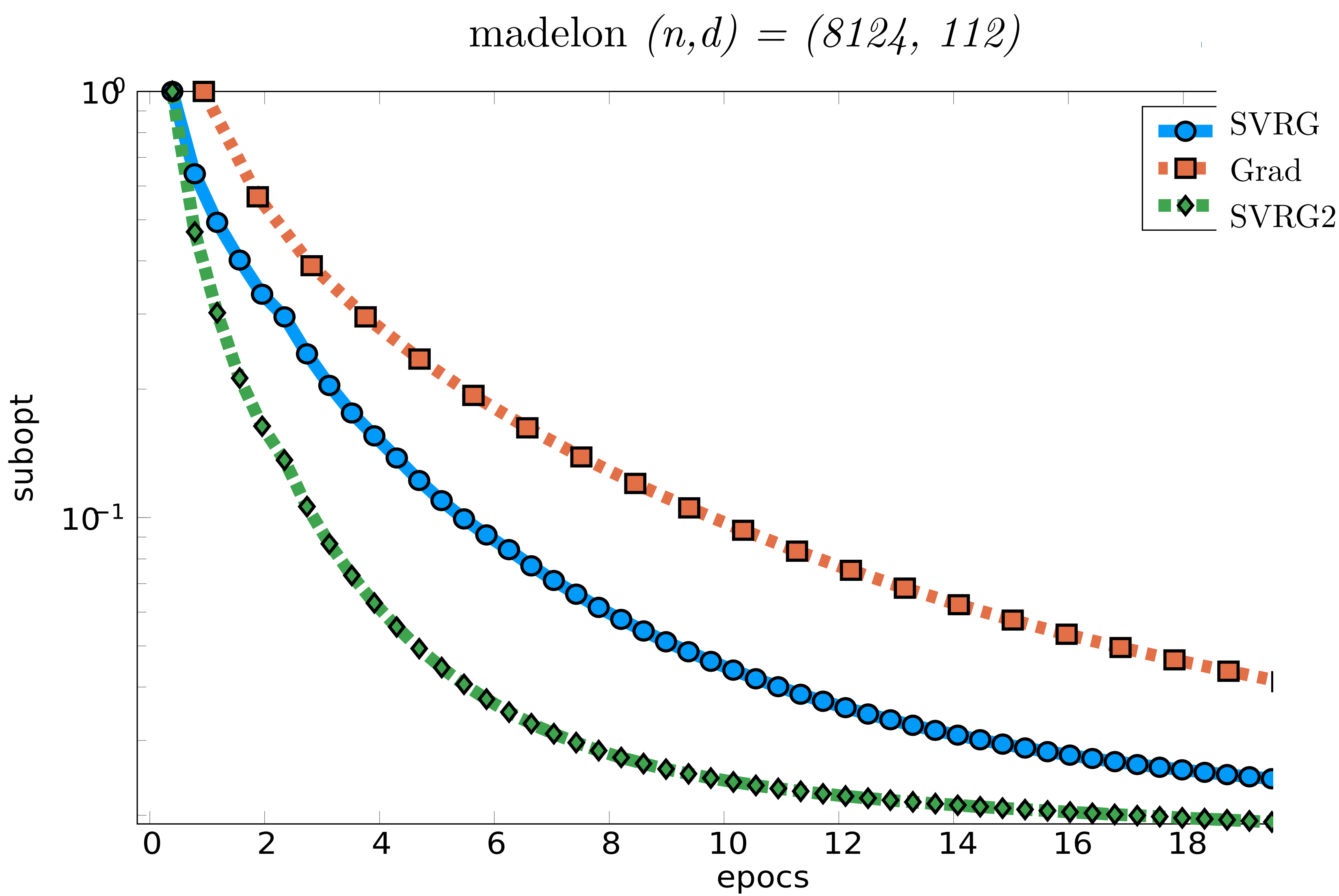
    $\theta_{t+1} = \theta_t - \gamma g_t$

  $\tilde{\theta} = \theta_m$

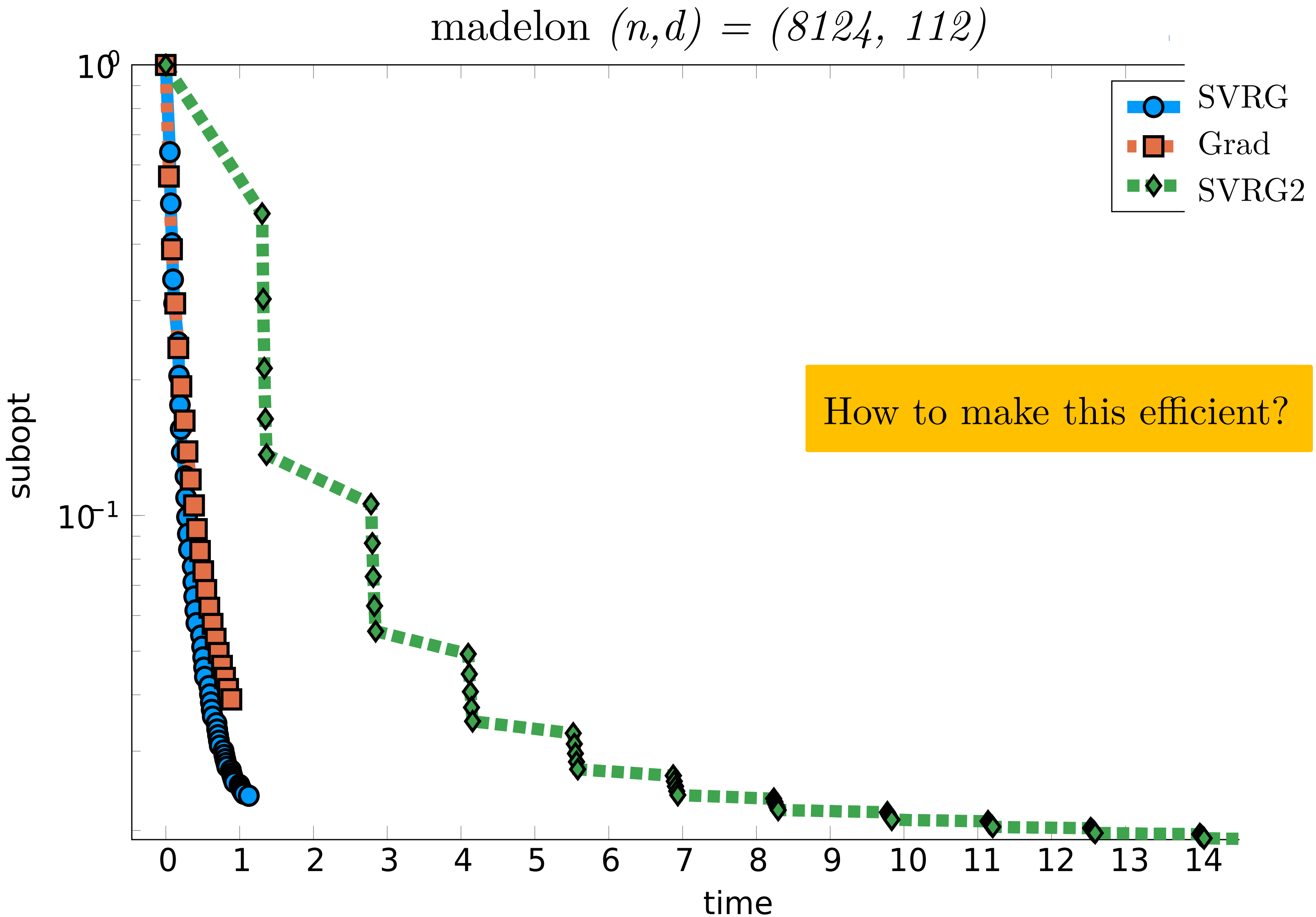
Output $\tilde{\theta}$

Does this actually work?

SVRG2: first experiment



SVRG2: first experiment



How to make this efficient?

SVRG2: Stochastic Variance Reduced Gradients with tracking

$$
\begin{array}{l}
\text{Set } \theta_0 = 0, \text{ choose } \gamma > 0, m \in \mathbb{N} \\
\tilde{\theta} = \theta_0 \\
\text{for } k = 0, 1, 2, \ldots, T-1 \\
\qquad \text{calculate } \nabla f(\tilde{\theta}), \boxed{\underline{H = \nabla^2 f(\tilde{\theta})}} \\
\qquad \theta_0 = \tilde{\theta} \\
\qquad \text{for } t = 0, 1, 2, \ldots, m-1 \\
\qquad\qquad \text{sample } i \in \{1, \ldots, n\} \\
\qquad\qquad g_t = \nabla f_i(\theta_t) - \nabla f_i(\tilde{\theta}) + \nabla f(\tilde{\theta}) \\
\qquad\qquad\qquad + \boxed{\underline{(H - H_i(\tilde{\theta}))(\theta_t - \tilde{\theta})}} \\
\qquad\qquad \theta_{t+1} = \theta_t - \gamma g_t \\
\qquad \tilde{\theta} = \theta_m \\
\text{Output } \tilde{\theta}
\end{array}
$$

Cost of SVRG2

$$g_t = \nabla f_i(\theta_t) - \nabla f_i(\tilde{\theta}) + \nabla f(\tilde{\theta})$$
$$+ \left(\frac{1}{n}\sum_{j=1}^{n} H_j(\tilde{\theta}) - H_i(\tilde{\theta})\right)(\theta_t - \tilde{\theta})$$

- Full Hessian $H = \frac{1}{n}\sum_{j=1}^{n} H_j(\tilde{\theta})$ costs $O(nd \times \text{eval}(f_i))$
- Hessian vector product $H(\theta_t - \tilde{\theta})$ costs $O(d^2)$
- Directional derivative $H_i(\tilde{\theta})(\theta_t - \tilde{\theta})$ costs $O(\text{eval}(f_i))$

Cost of SVRG2

$$g_t = \nabla f_i(\theta_t) - \nabla f_i(\tilde{\theta}) + \nabla f(\tilde{\theta})$$
$$+ (\frac{1}{n}\sum_{j=1}^{n} H_j(\tilde{\theta}) - H_i(\tilde{\theta}))(\theta_t - \tilde{\theta})$$

- Full Hessian $H = \frac{1}{n}\sum_{j=1}^{n} H_j(\tilde{\theta})$ costs $O(nd \times \text{eval}(f_i))$
- Hessian vector product $H(\theta_t - \tilde{\theta})$ costs $O(d^2)$
- Directional derivative $H_i(\tilde{\theta})(\theta_t - \tilde{\theta})$ costs $O(\text{eval}(f_i))$

Cost of SVRG2

$$
\begin{aligned}
g_t = \nabla f_i(\theta_t) - \nabla f_i(\tilde{\theta}) + \nabla f(\tilde{\theta}) \\
+\left(\frac{1}{n}\sum_{j=1}^{n} H_j(\tilde{\theta}) - H_i(\tilde{\theta})\right)(\theta_t - \tilde{\theta})
\end{aligned}
$$

- Full Hessian $H = \frac{1}{n}\sum_{j=1}^{n} H_j(\tilde{\theta})$ costs $O(nd \times \mathrm{eval}(f_i))$
- Hessian vector product $H(\theta_t - \tilde{\theta})$ costs $O(d^2)$
- Directional derivative $H_i(\tilde{\theta})(\theta_t - \tilde{\theta})$ costs $O(\mathrm{eval}(f_i))$

Build approximations $\hat{H}_j(\theta) \approx H_j(\theta)$

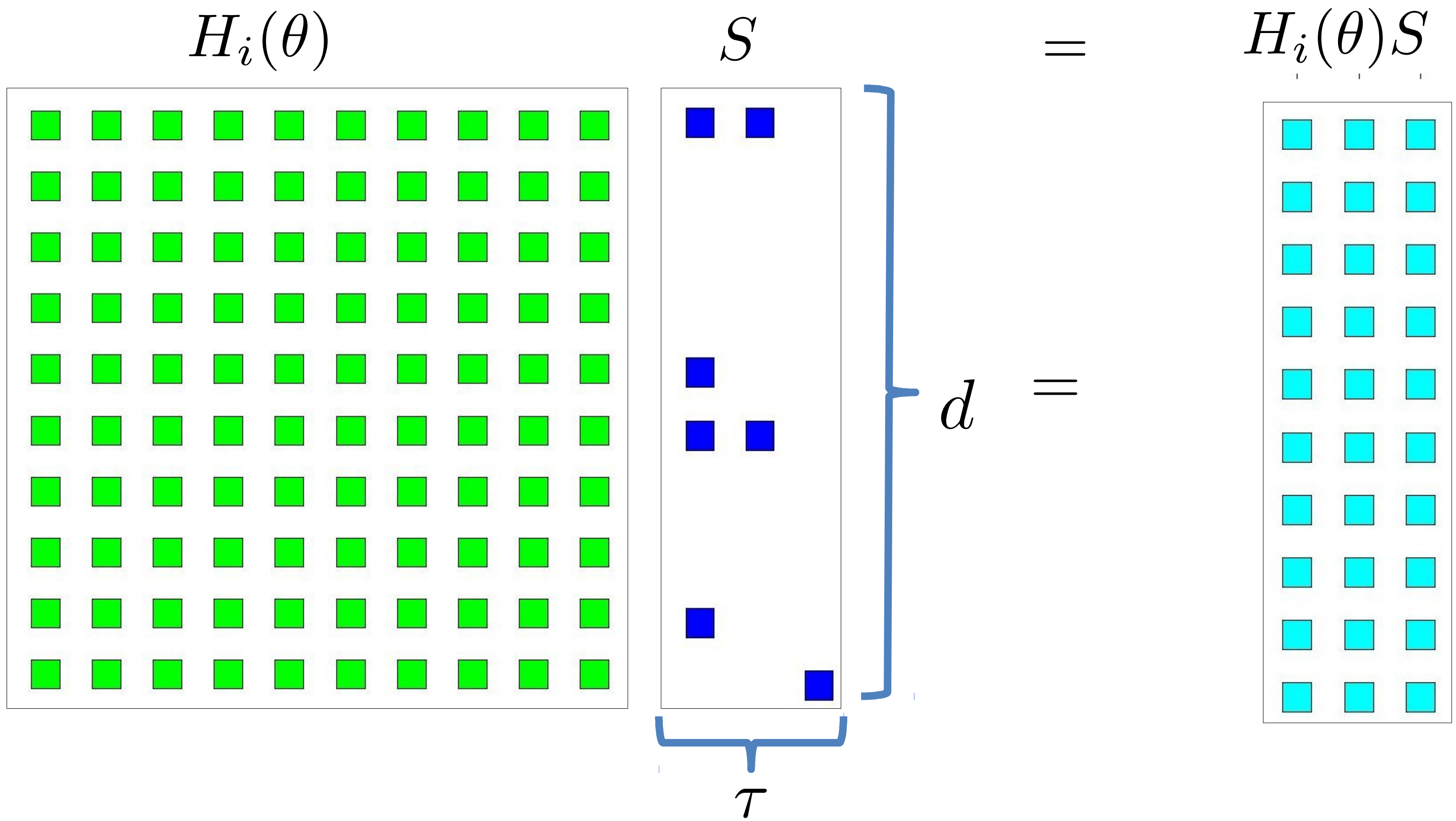
Different ways to approximate the Hessian

$$\hat{H}_i(\theta) \approx H_i(\theta)$$

We tried:

- Diagonal approximations
- Rank-1 approximation based on secant equation
- **Low rank approximations using Sketching and projecting**

Sketching the stochastic Hessian



Sketching matrix
$S \sim \mathcal{D}$ fixed distribution $S \in \mathbb{R}^{d\times\tau}$

Costs $\tau \times O(\text{eval}(f_i))$
to evaluate $H_i(\theta)S$

Sketching and Projecting the Hessian: Action Matching (AM) approximation

find X such that

$$XS = H_i S$$

Sketching and Projecting the Hessian: Action Matching (AM) approximation

$$\text{find } X \text{ such that}$$

$$XS = H_i S, \quad X = X^\top$$

Sketching and Projecting the Hessian: Action Matching (AM) approximation

$$\hat{H}_i = \arg\min_{X \in \mathbb{R}^{d \times d}} ||X||^2_{F(H)}$$

$$\text{subject to} \quad XS = H_i S, \quad X = X^\top$$

where $||X||^2_{F(H)} \stackrel{\text{def}}{=} \mathbf{Tr}\left(XHX^\top H\right)$ and $H = \nabla^2 f(\tilde{\theta})$

Sketching and Projecting the Hessian: Action Matching (AM) approximation

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

where $||X||^2_{F(H)} \stackrel{\text{def}}{=} \mathbf{Tr}\left(XHX^\top H\right)$ and $H = \nabla^2 f(\tilde{\theta})$

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

Total inner iteration costs: $O\left(\tau \times \text{eval}(f_i) + \tau^2 d + \tau^3\right)$

Sketching and Projecting the Hessian: Action Matching (AM) approximation

$$\hat{H}_i = \arg\min_{X \in \mathbb{R}^{d\times d}} ||X||^2_{F(H)}$$

$$\text{subject to} \quad XS = H_i S, \quad X = X^\top$$

where $||X||^2_{F(H)} \overset{\text{def}}{=} \mathbf{Tr}\left(XHX^\top H\right)$ and $H = \nabla^2 f(\tilde{\theta})$

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

rank 2τ

Total inner iteration costs: $O\left(\tau \times \text{eval}(f_i) + \tau^2 d + \tau^3\right)$

Sketching and Projecting the Hessian: Action Matching (AM) approximation

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

where $\|X\|^2_{F(H)} \stackrel{\text{def}}{=} \mathbf{Tr}\left(XHX^\top H\right)$ and $H = \nabla^2 f(\tilde{\theta})$

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

rank 2τ

Total inner iteration costs: $O\left(\tau \times \text{eval}(f_i) + \tau^2 d + \tau^3\right)$

$$\frac{1}{n}\sum_{j=1}^{n} \hat{H}_j = HS(S^T HS)^{-1}S^\top H.$$

Total outer costs: $O(n\tau \times \text{eval}(f_i))$

Sketching and Projecting the Hessian: Action Matching (AM) approximation

$$\hat{H}_i = \arg \min_{X \in \mathbb{R}^{d\times d}} ||X||^2_{F(H)}$$

$$\text{subject to} \quad XS = H_i S, \quad X = X^\top$$

where $||X||^2_{F(H)} \stackrel{\text{def}}{=} \mathbf{Tr}\left(XHX^\top H\right)$ and $H = \nabla^2 f(\tilde{\theta})$

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

rank 2τ

Total inner iteration costs: $O\left(\tau \times \text{eval}(f_i) + \tau^2 d + \tau^3\right)$

$$\frac{1}{n}\sum_{j=1}^{n} \hat{H}_j = HS(S^T HS)^{-1} S^\top H.$$

Total outer costs: $O(n\tau \times \text{eval}(f_i))$

Sketching and Projecting the Hessian: Action Matching (AM) approximation

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

where $||X||^2_{F(H)} \overset{\text{def}}{=} \mathbf{Tr}\left(XHX^\top H\right)$ and $H = \nabla^2 f(\tilde{\theta})$

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

rank 2τ

Total inner iteration costs: $O\big(\tau \times \text{eval}(f_i) + \tau^2 d + \tau^3\big)$

$$\frac{1}{n}\sum_{j=1}^n \hat{H}_j = HS(S^T HS)^{-1} S^\top H.$$

What about S ?

Total outer costs: $O(n\tau \times \text{eval}(f_i))$

Choosing the sketch matrix

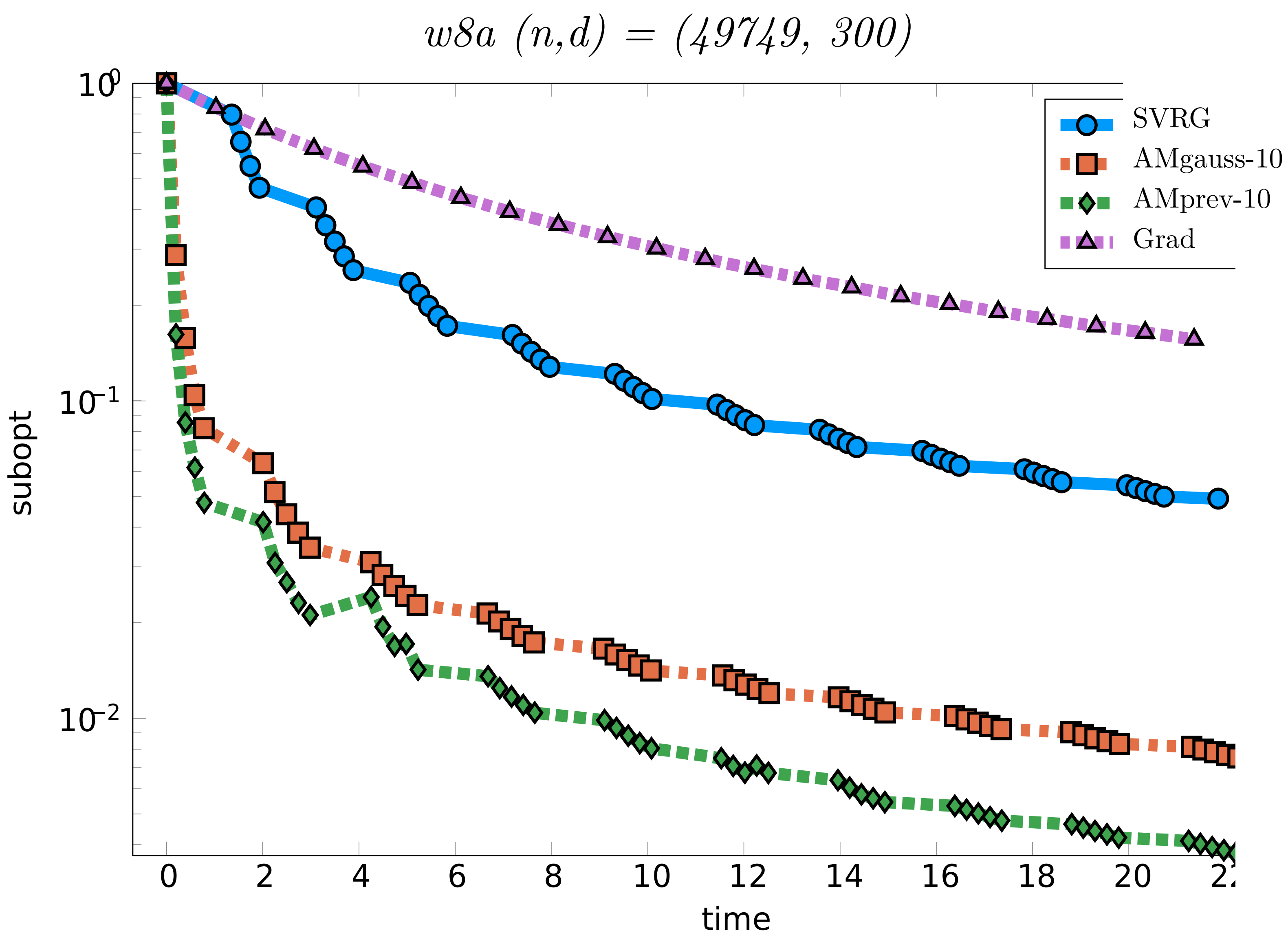
$$\hat{H}_i = \arg \min_{X \in \mathbb{R}^{d \times d}} ||X||^2_{F(H)}$$
$$\text{subject to} \quad XS = H_i S, \quad X = X^\top$$

AMgauss: $S \sim \mathcal{N}(0, I)$ has Gaussian entries sampled i.i.d at each iteration

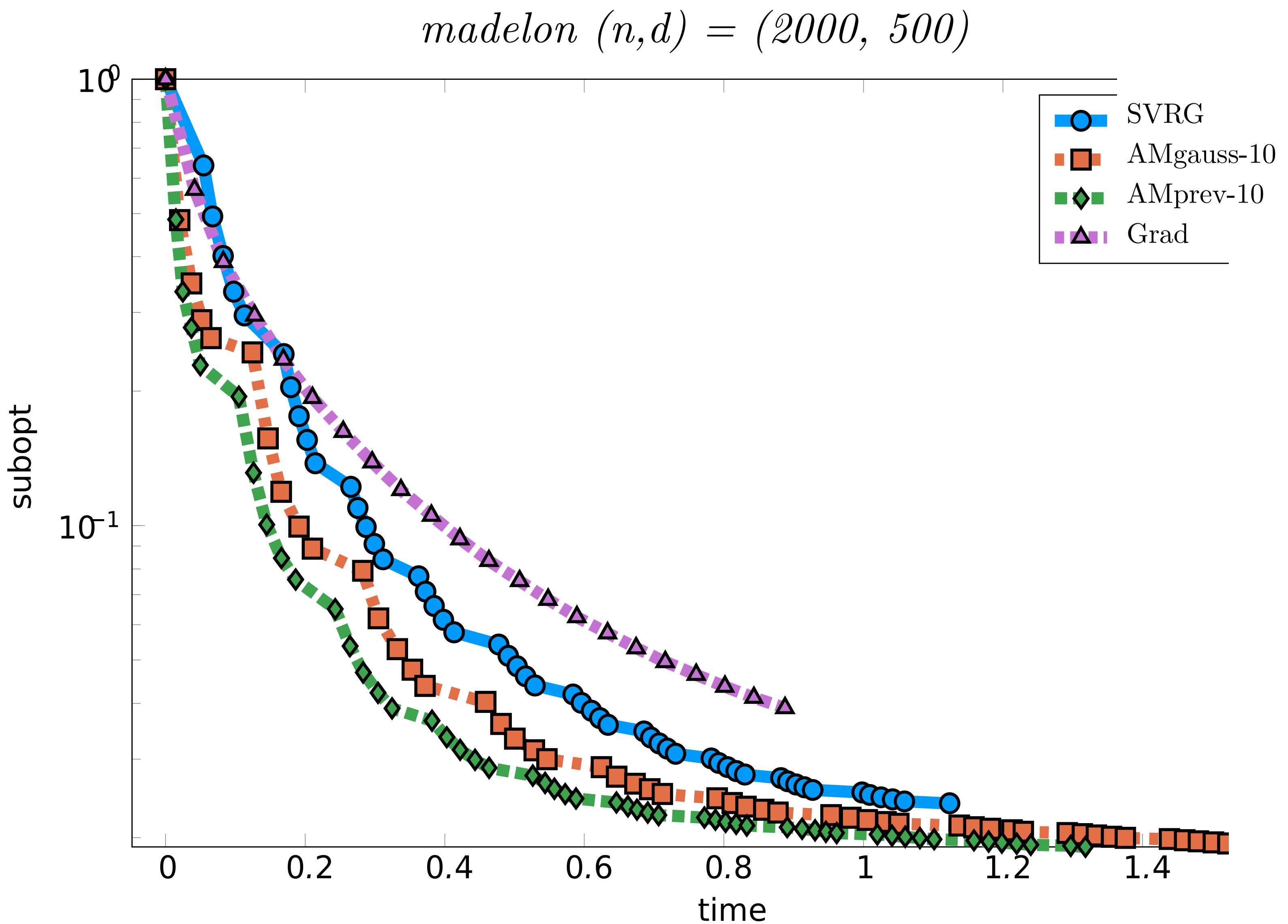
AMprev: Averages of **previous** search directions

$$S = [\bar{g}_0, \ldots, \bar{g}_{\tau-1}] \quad \text{where} \quad \bar{g}_i = \frac{\tau}{m} \sum_{j=\frac{m}{\tau} i}^{\frac{m}{\tau}(i+1)-1} g_j,$$

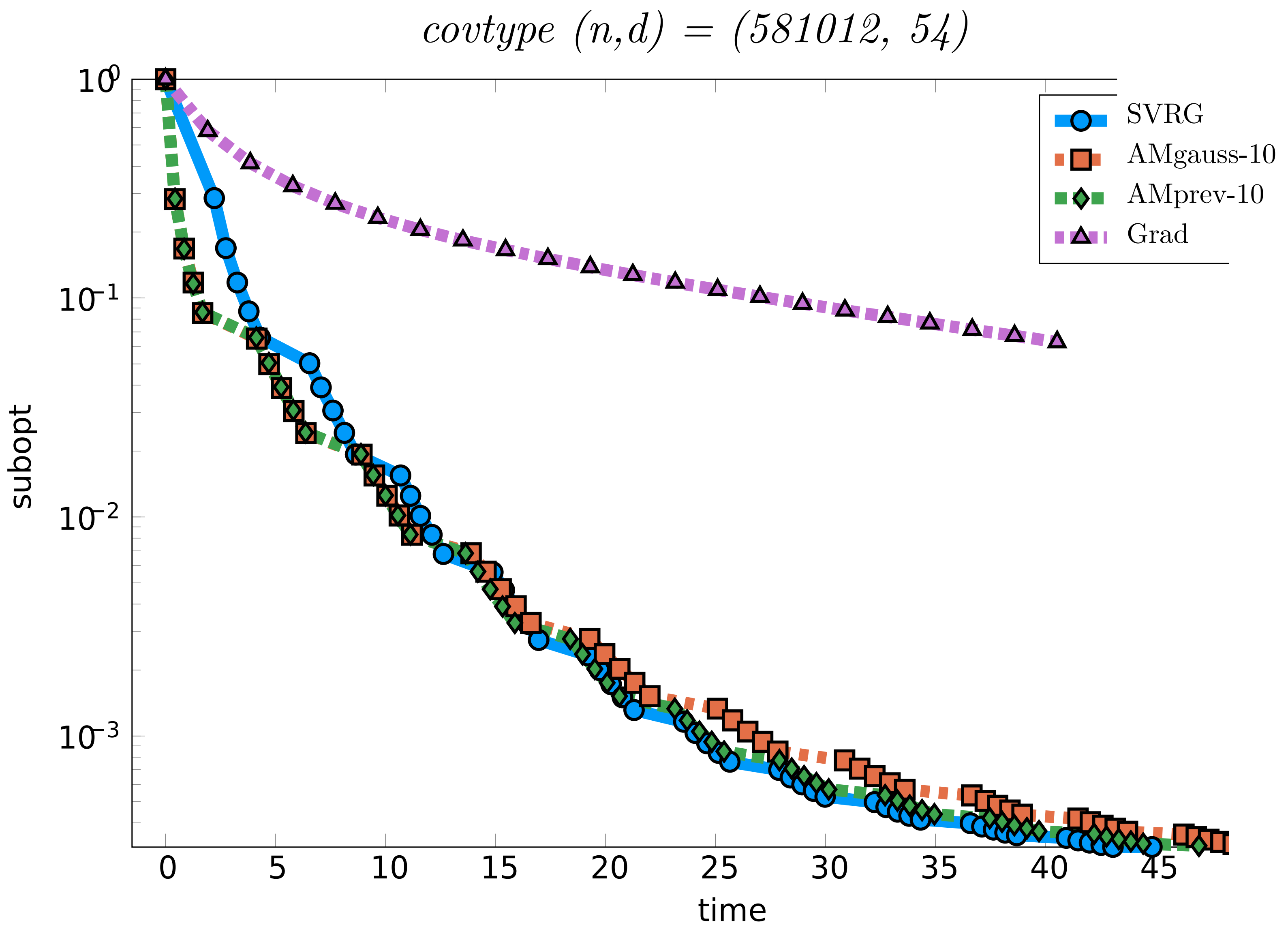
AM: Experiment works well



AM: Experiment works ok



AM: Experiment works badly



Take home:

Can use Hessian to diminish variance

Speed-ups with less gain and risk compared to Newton type methods.

New compressed Hessian estimates using sketching and projecting

gowerrobert/StochOpt.jl

Bruce Christianson. **Automatic Hessians by reverse accumulation**. In: IMA Journal of Numerical Analysis 12.2 (1992), pp. 135–150.

RMG and P. Richtárik, **Randomized quasi-Newton updates are linearly convergent matrix inversion algorithms,** SIAM Journal on Matrix Analysis and Applications , 38(4), 1380-1409, 2017

D.. Goldfarb, (1970). **A Family of Variable-Metric Methods Derived by Variational Means**. Mathematics of Computation, 24(109), 23.