Optimization methods

Simon Grimm

MCMC methods

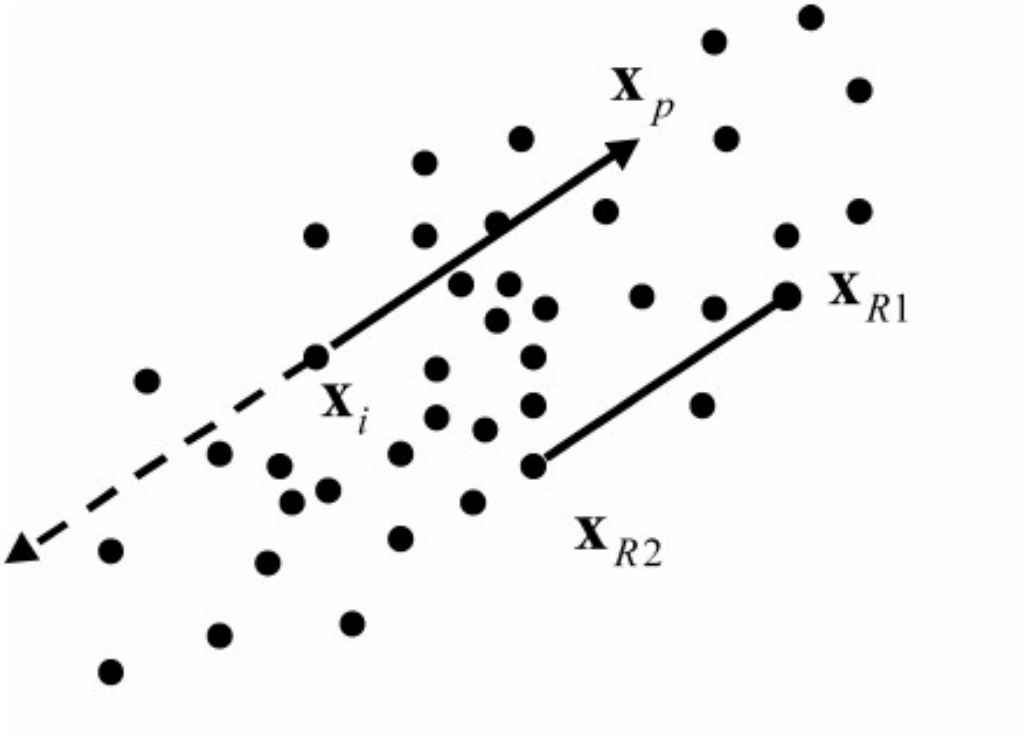
- Single chain methods need more than 500000 interatios
- Often too slow

Parallel MCMC methods

- Affine invariant method
- DEMCMC

- Use parallel chains to speed up calculation

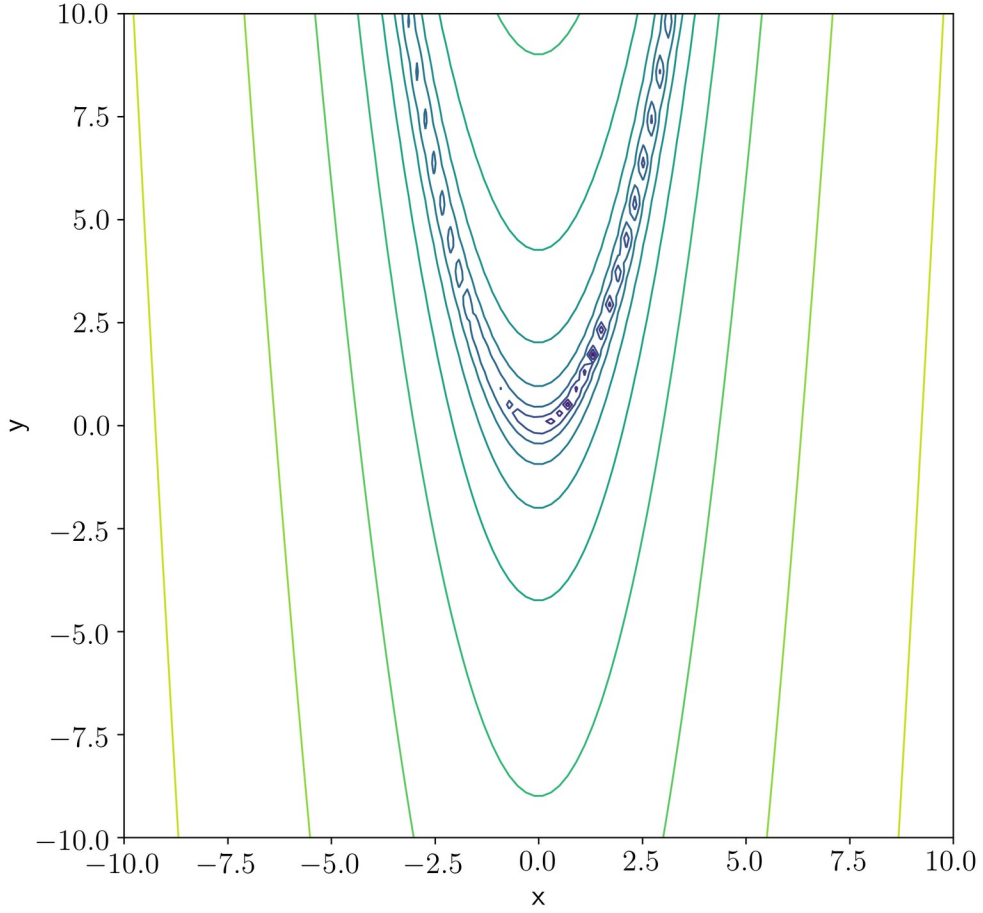
DEMCMC, emcee



Problem with curvatures

Ter Braak 2006

Rosenbrock function



f = (a - x) * (a - x) + b * (y - x * x) * (y - x * x)

a = 1.0
b = 100.0

SVGD

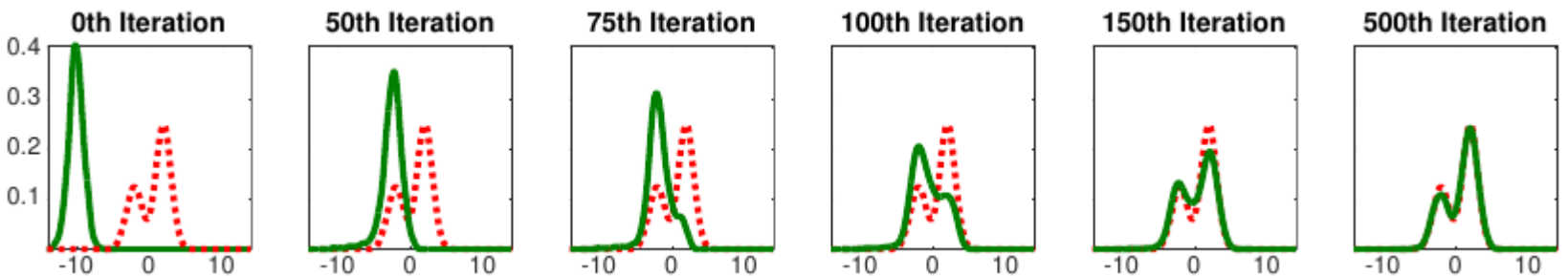
- Stein Variational gradient descent

Algorithm 1 Bayesian Inference via Variational Gradient Descent

Input: A target distribution with density function $p(x)$ and a set of initial particles $\{x_i^0\}_{i=1}^n$.
Output: A set of particles $\{x_i\}_{i=1}^n$ that approximates the target distribution.
for iteration ℓ **do**

$$x_i^{\ell+1} \leftarrow x_i^\ell + \epsilon_\ell \hat{\phi}^*(x_i^\ell) \quad \text{where} \quad \hat{\phi}^*(x) = \frac{1}{n}\sum_{j=1}^{n}\left[k(x_j^\ell, x)\nabla_{x_j^\ell}\log p(x_j^\ell) + \nabla_{x_j^\ell} k(x_j^\ell, x)\right], \quad (8)$$

where ϵ_ℓ is the step size at the ℓ-th iteration.
end for

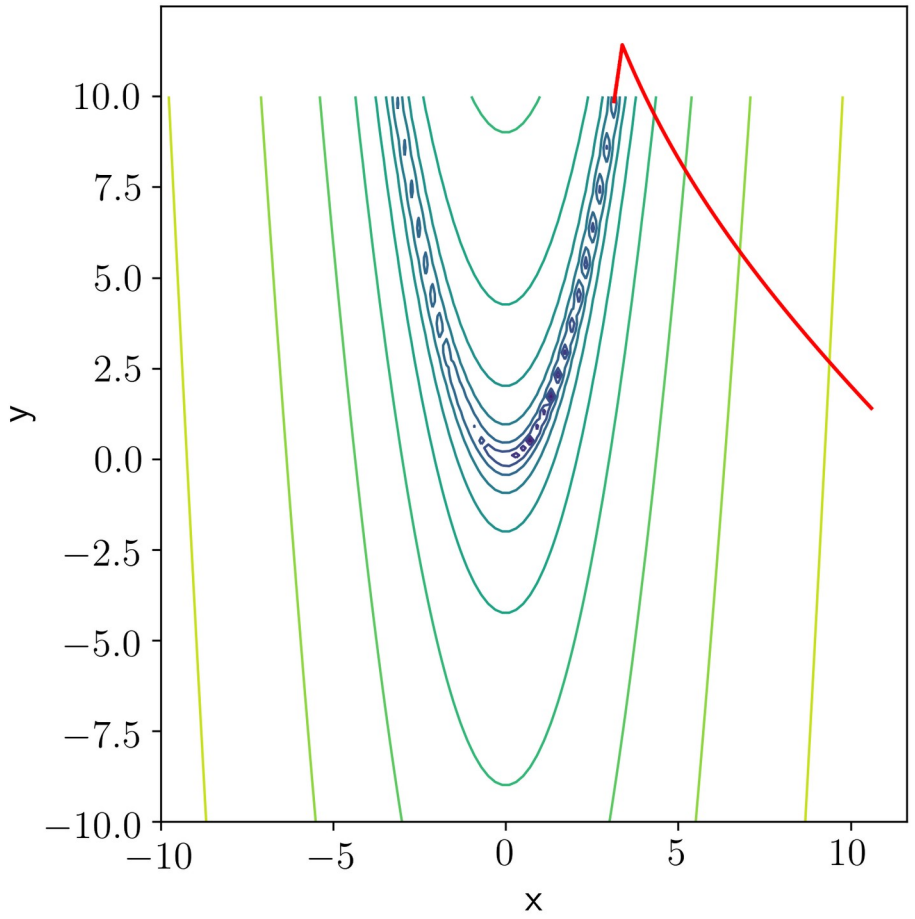


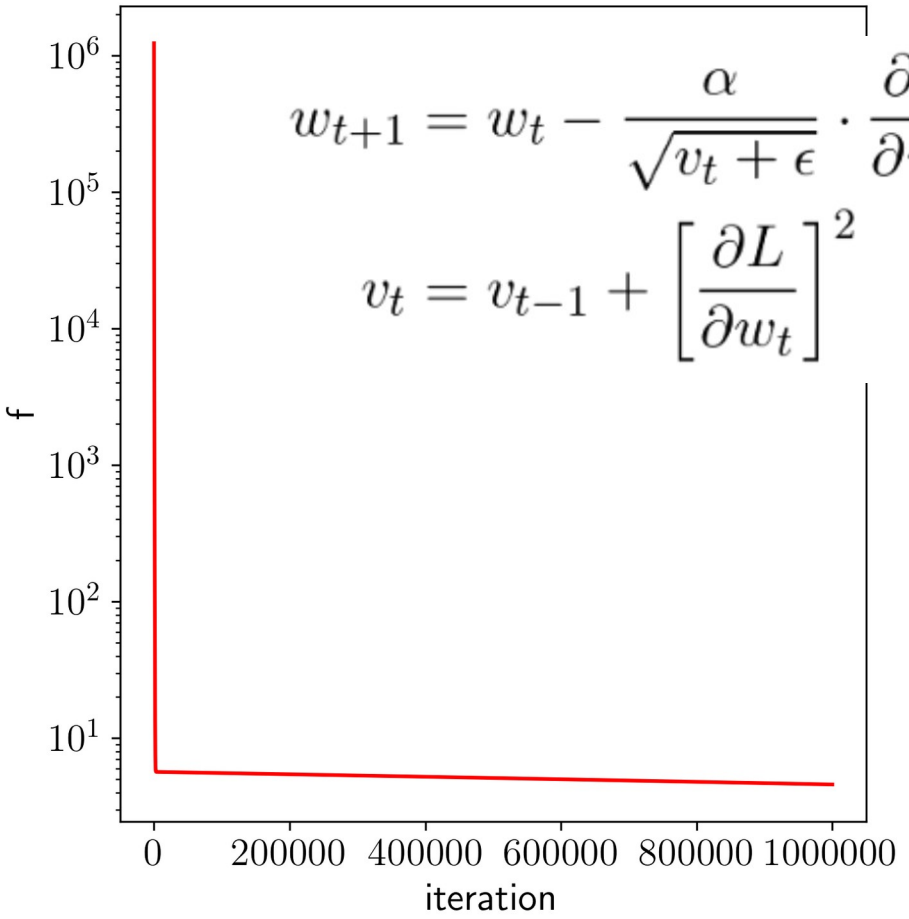
Liu & Wang 2019

SVGD

- Particle method
- Can be parallelzed easily

AdaGrad

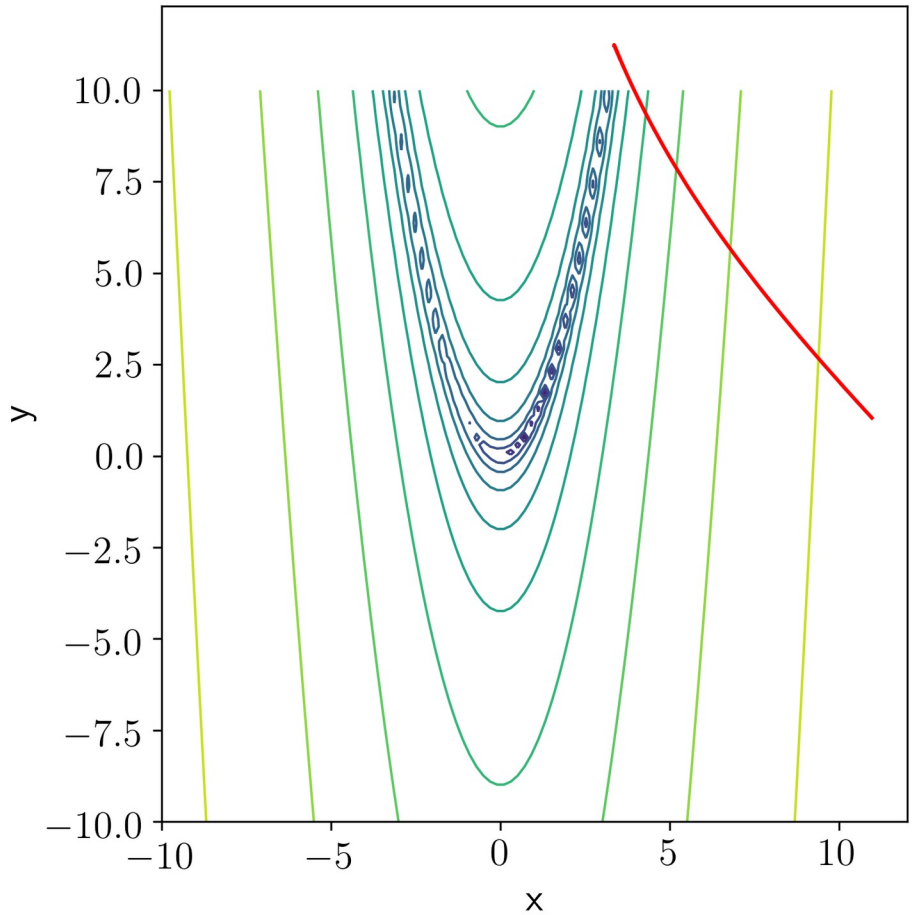


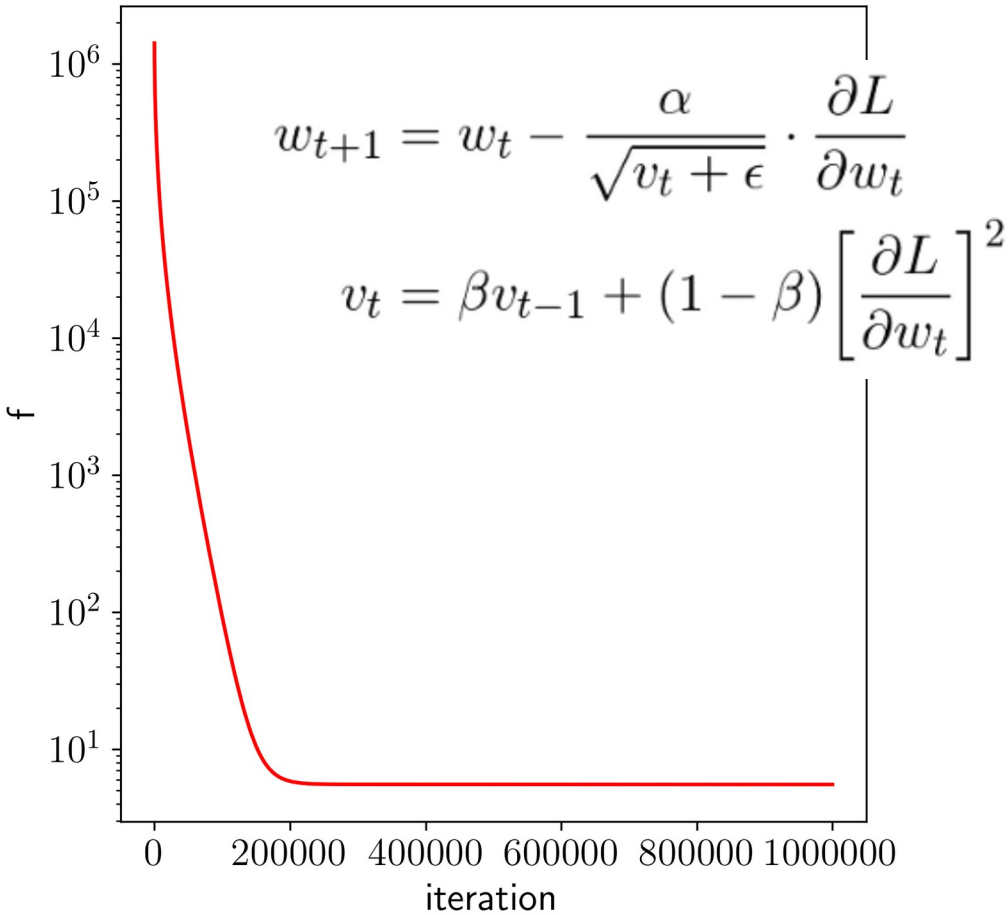


$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{v_t + \epsilon}} \cdot \frac{\partial L}{\partial w_t}$$

$$v_t = v_{t-1} + \left[\frac{\partial L}{\partial w_t}\right]^2$$

RMSprop

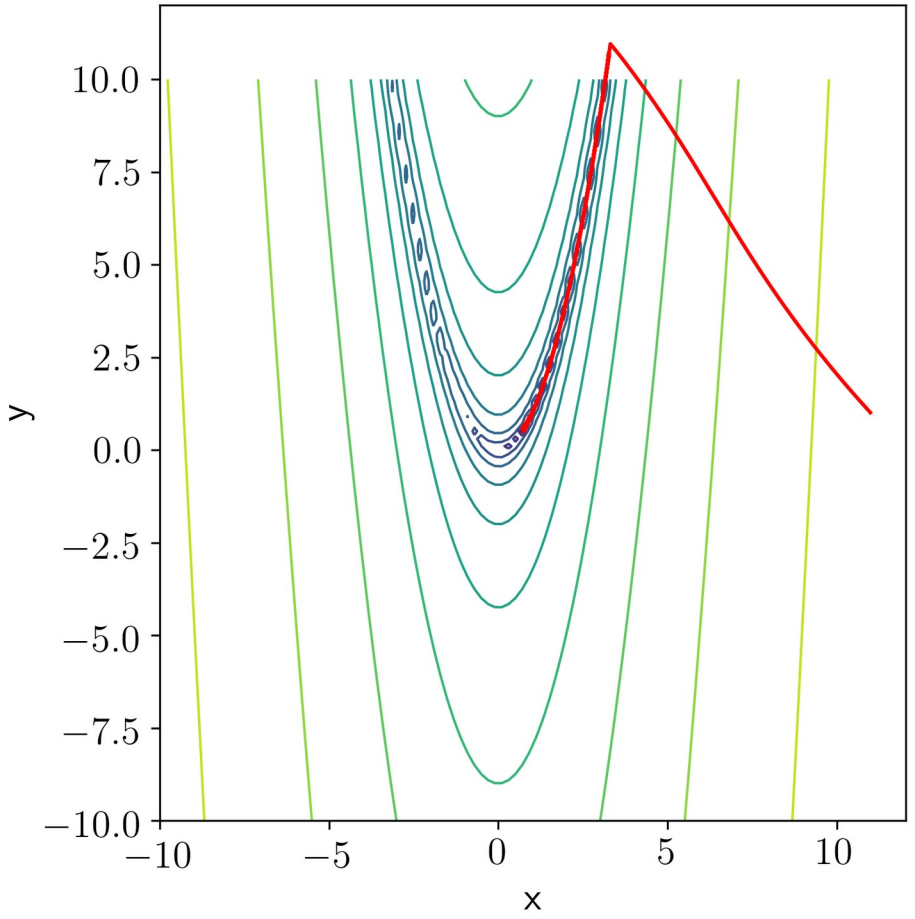


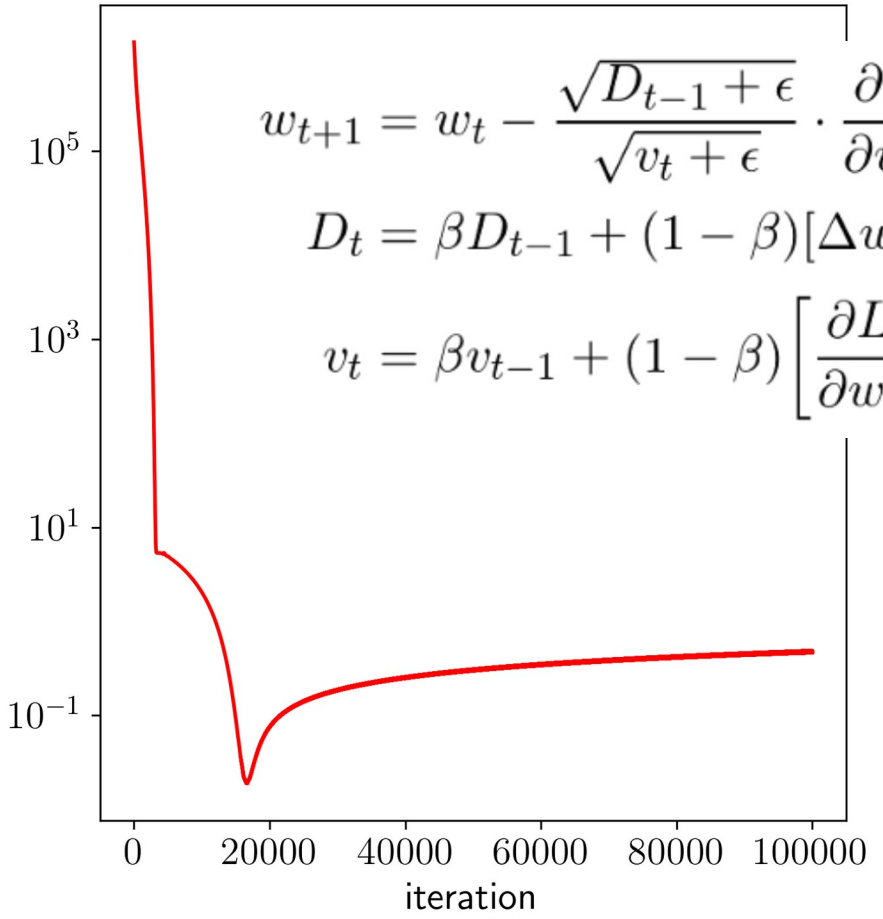


$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{v_t + \epsilon}} \cdot \frac{\partial L}{\partial w_t}$$

$$v_t = \beta v_{t-1} + (1 - \beta) \left[\frac{\partial L}{\partial w_t} \right]^2$$

Adadelta





$$w_{t+1} = w_t - \frac{\sqrt{D_{t-1} + \epsilon}}{\sqrt{v_t + \epsilon}} \cdot \frac{\partial L}{\partial w_t}$$

$$D_t = \beta D_{t-1} + (1 - \beta)[\Delta w_t]^2$$

$$v_t = \beta v_{t-1} + (1 - \beta)\left[\frac{\partial L}{\partial w_t}\right]^2$$

Adam

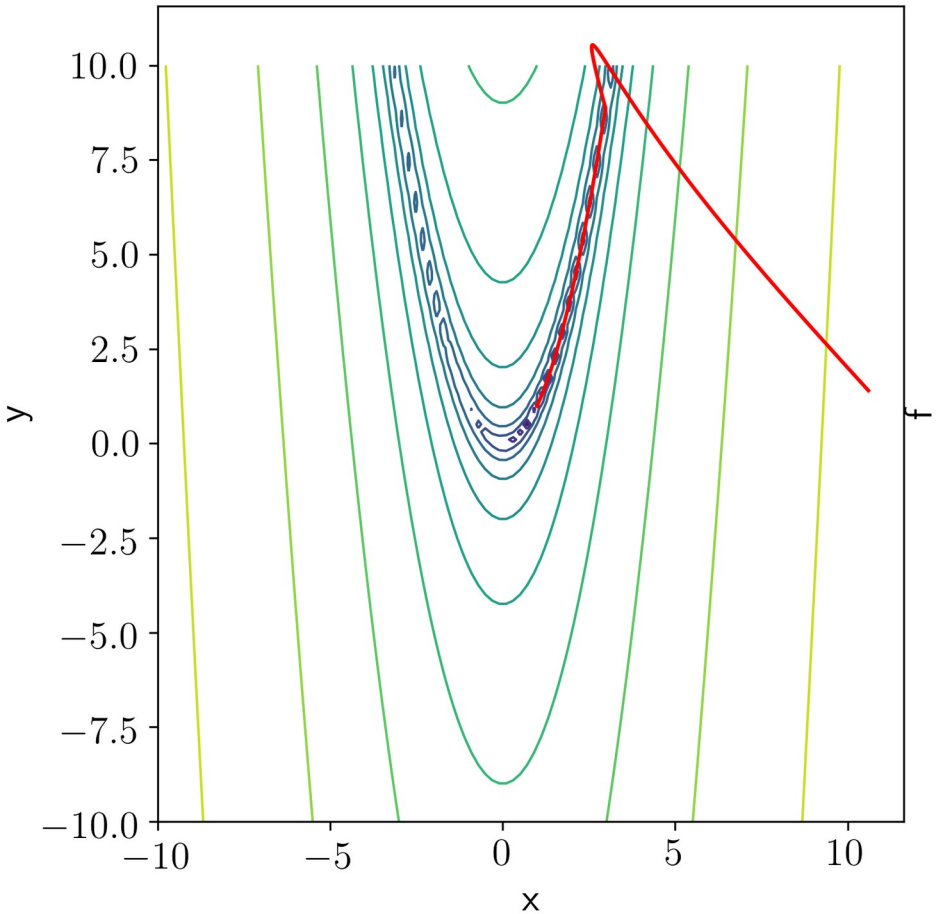
$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{\hat{v}_t} + \epsilon} \cdot \hat{m}_t$$

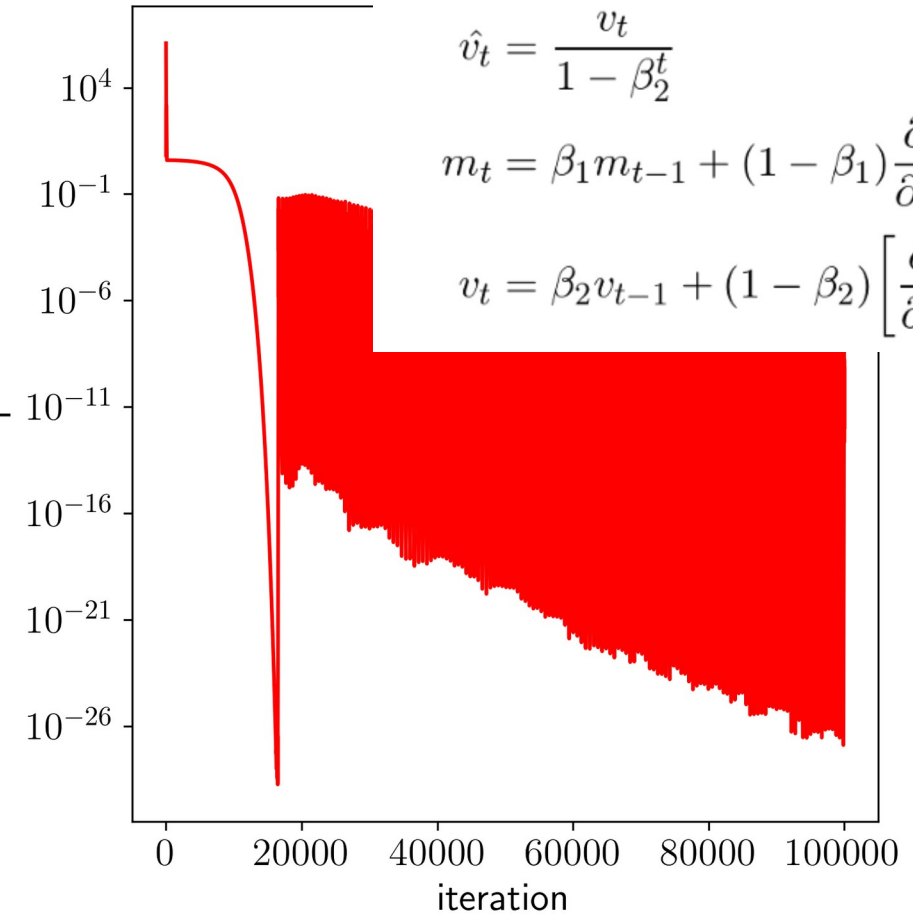
$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

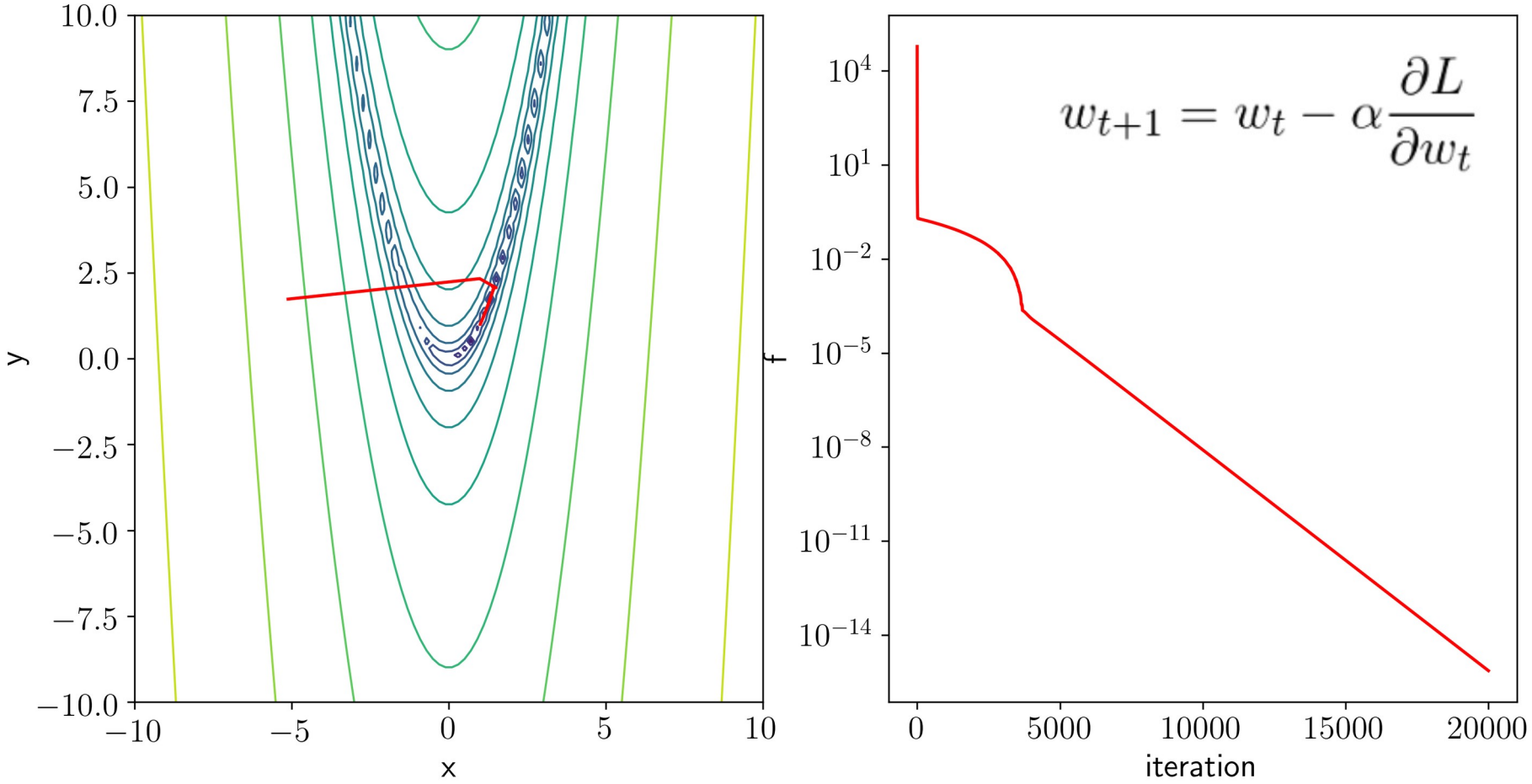
$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \frac{\partial L}{\partial w_t}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) \left[\frac{\partial L}{\partial w_t} \right]^2$$

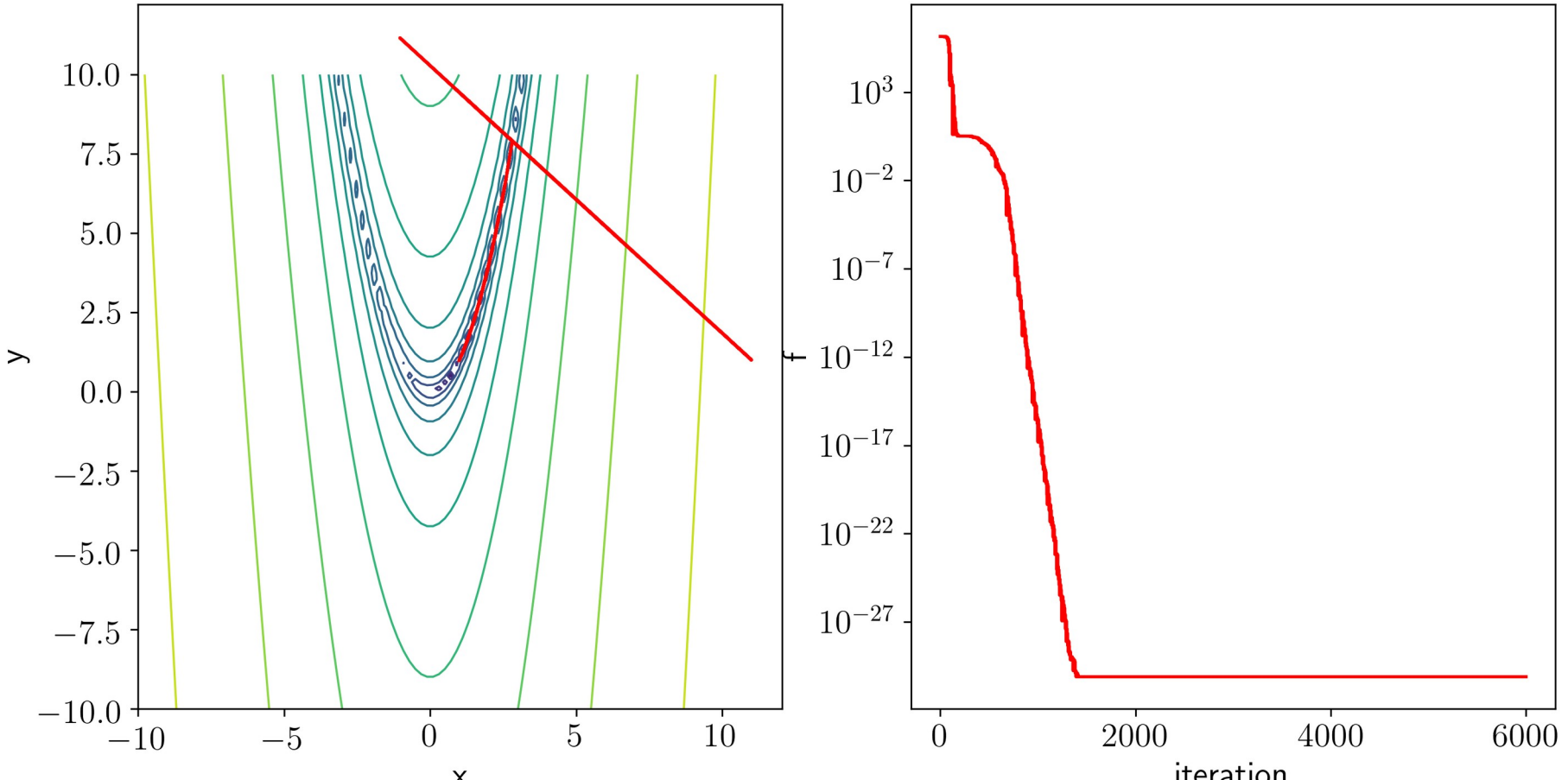




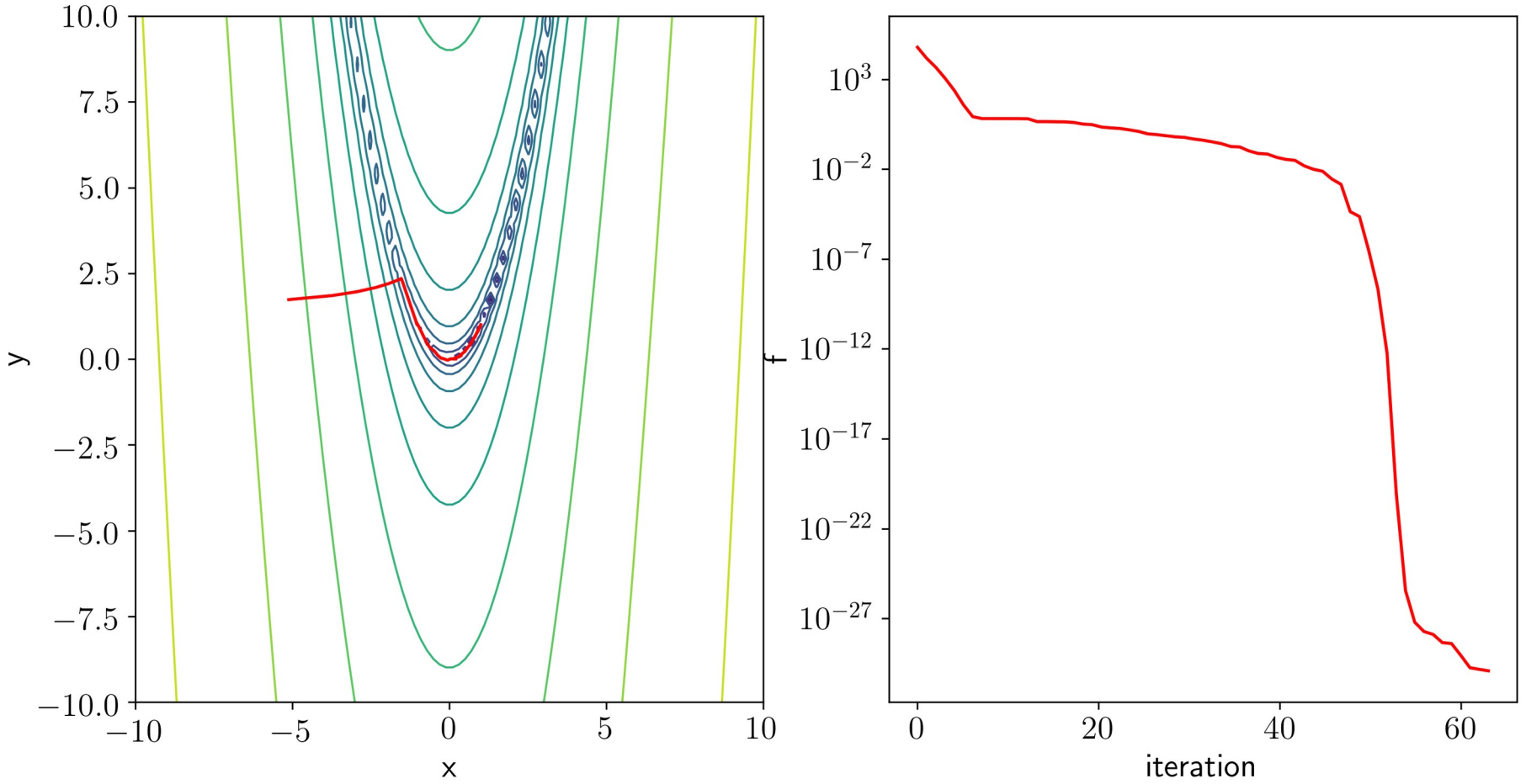
Steepest descent with line search



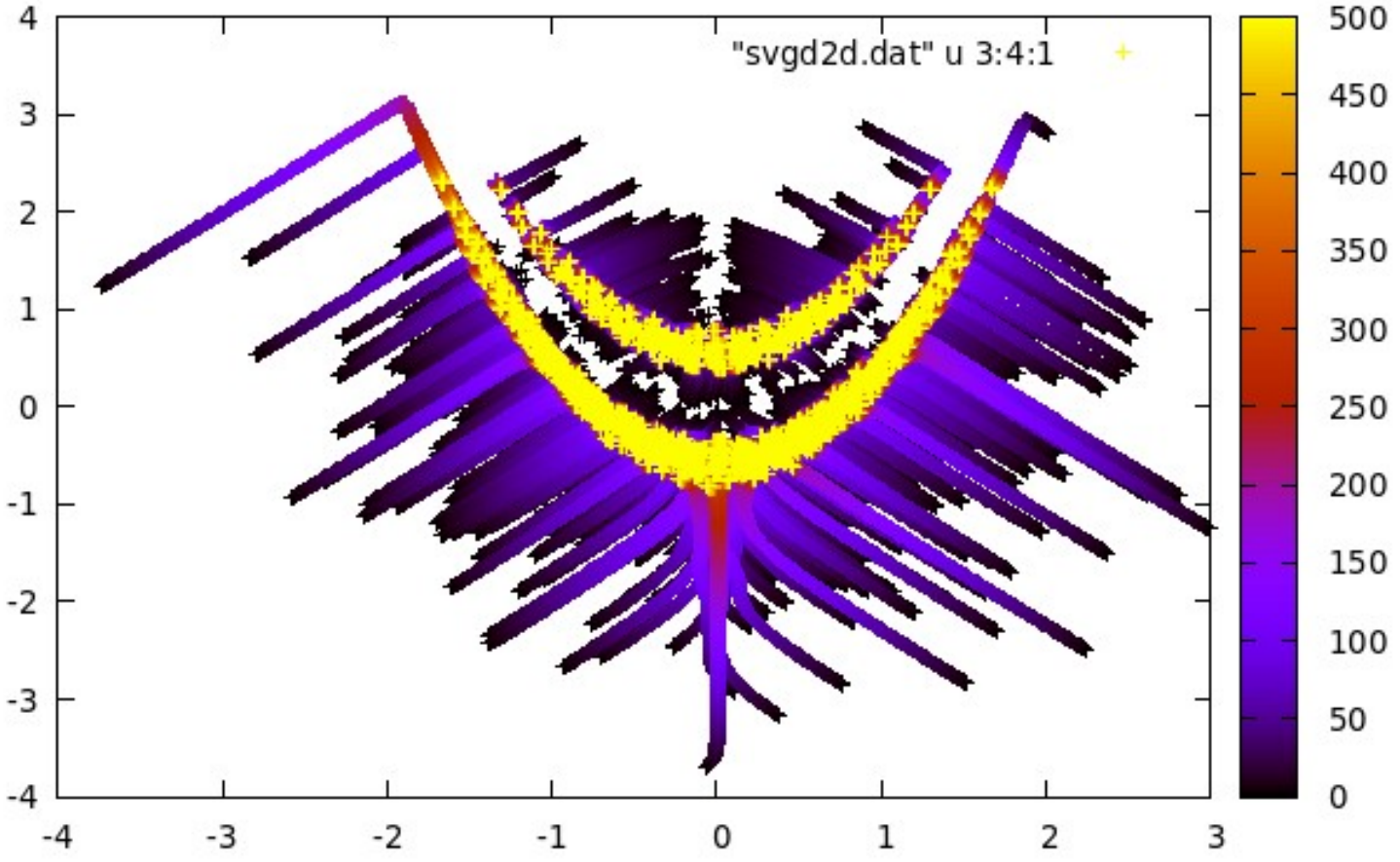
Nelder Mead, downhill simplex

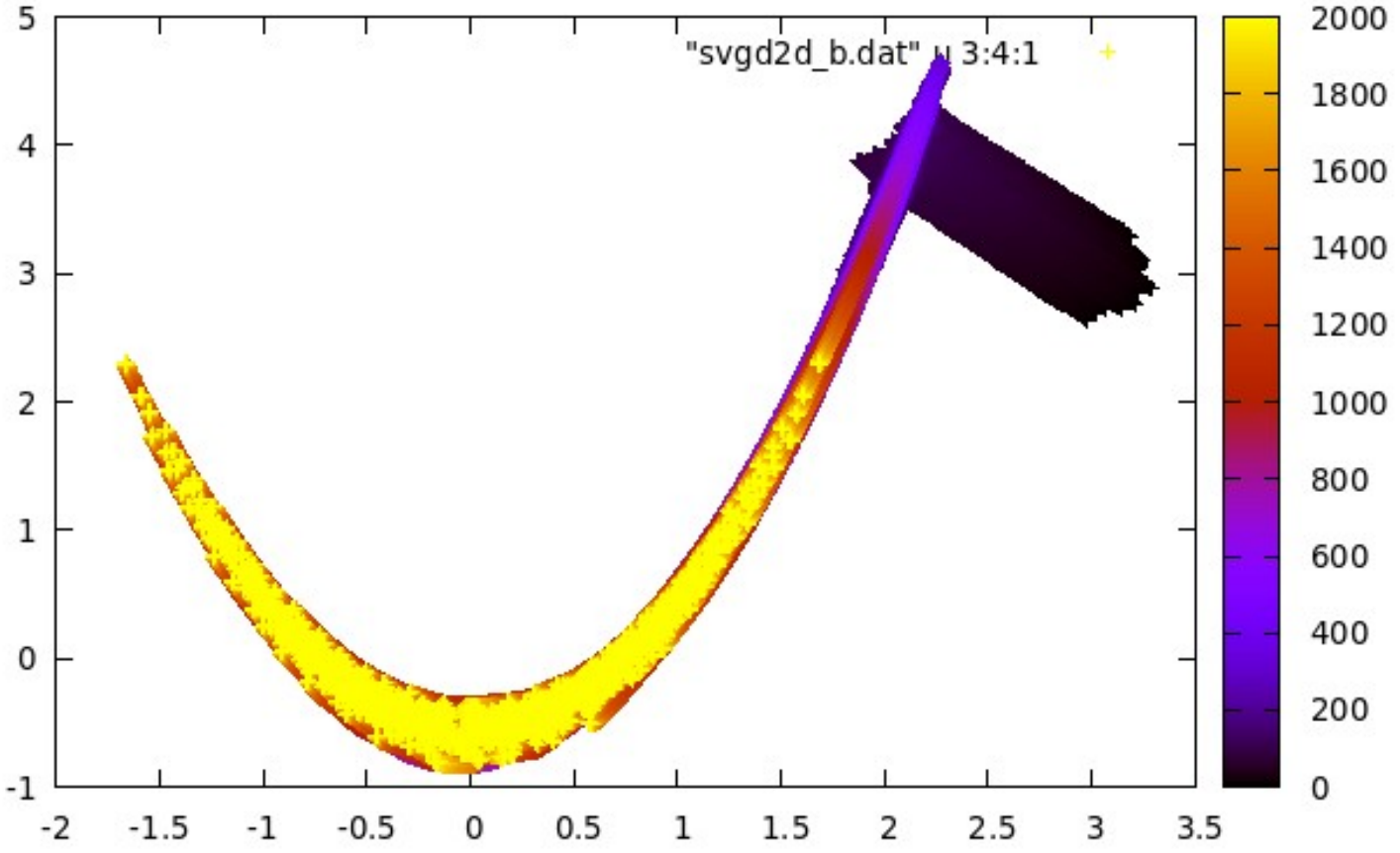


LBFGS, second order

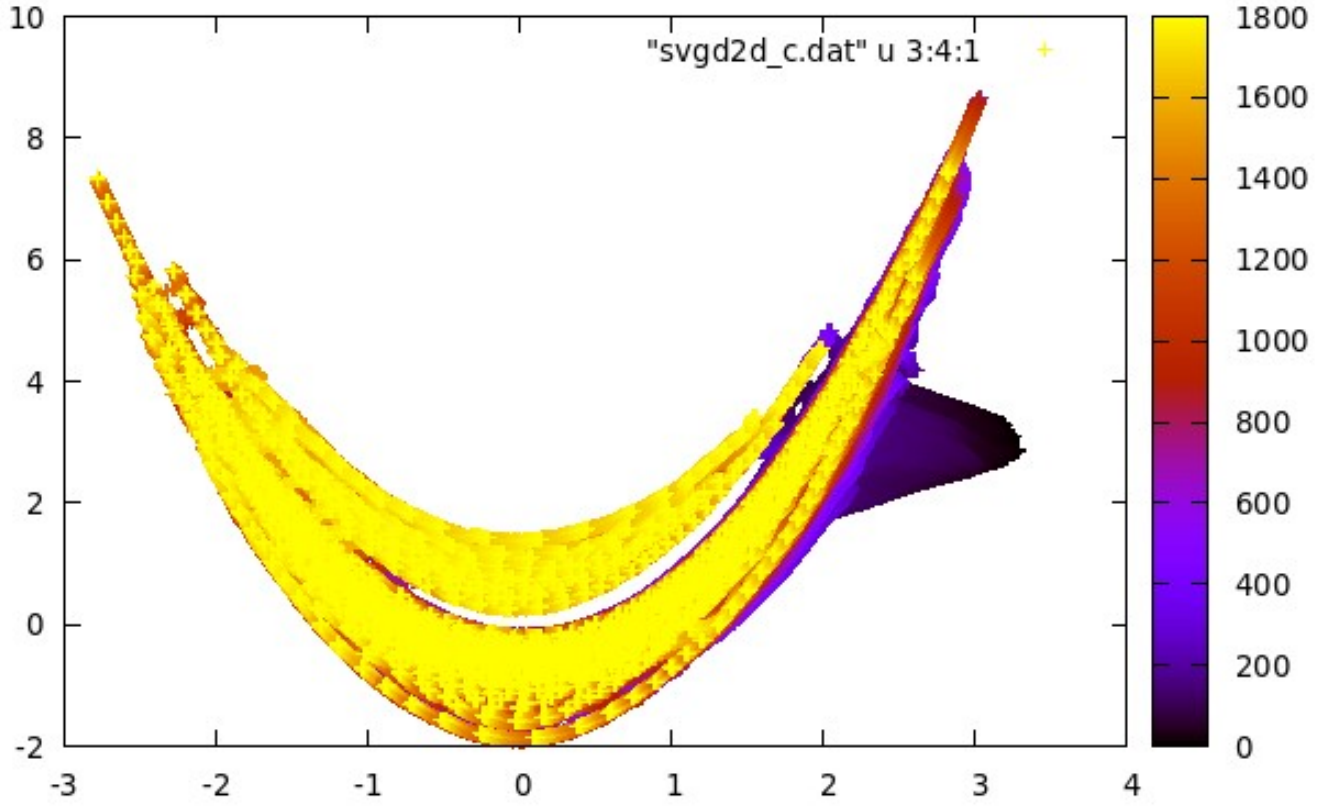


SVGD with RMSprop

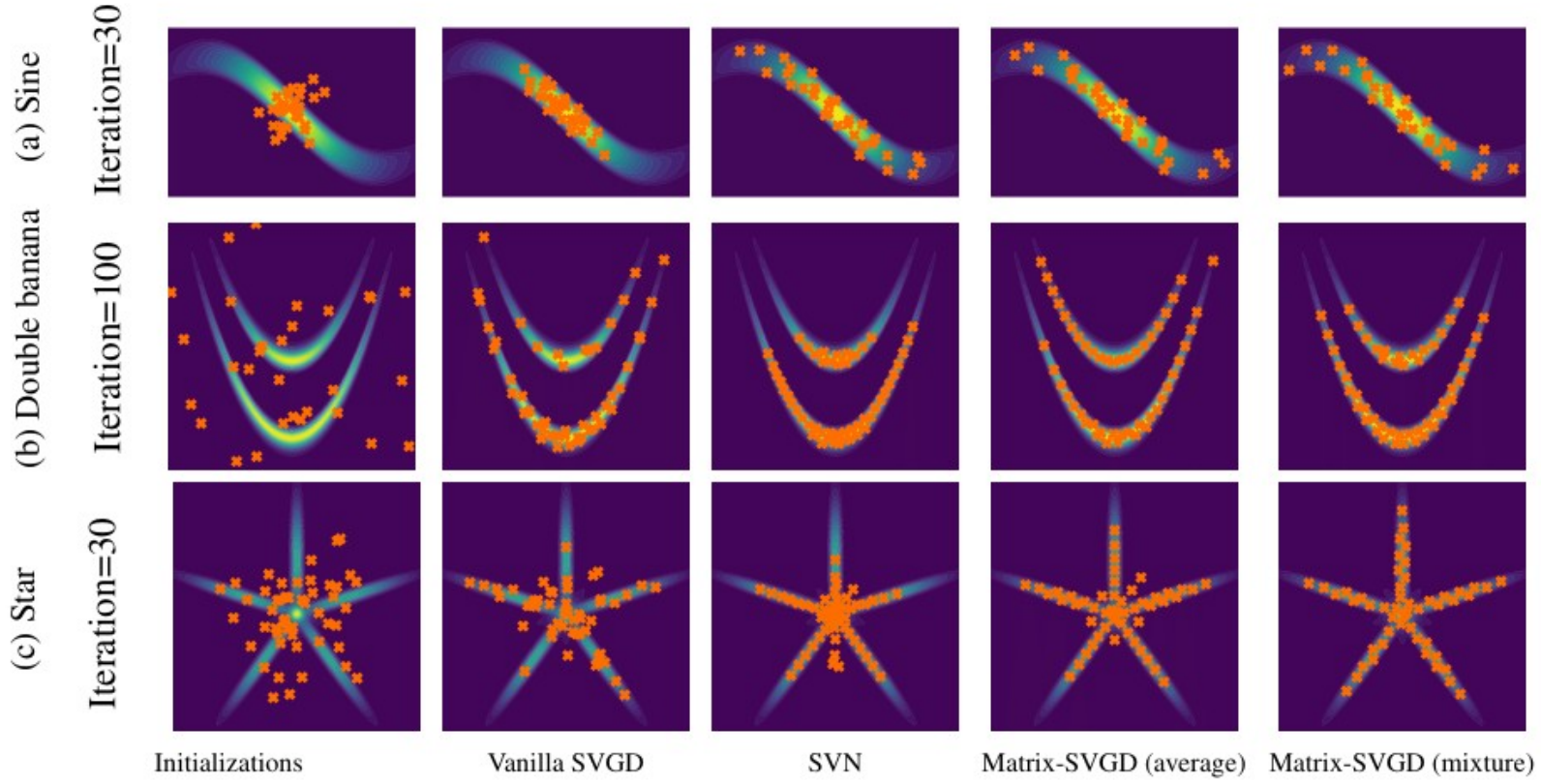



"svgd2d_b.dat" u 3:4:1
5
4
3
2
1
0
-1
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5
2000
1800
1600
1400
1200
1000
800
600
400
200
0

Increase repulsive term



SVGD with higher orders



Wang et al. 2019

Annealed SVGD

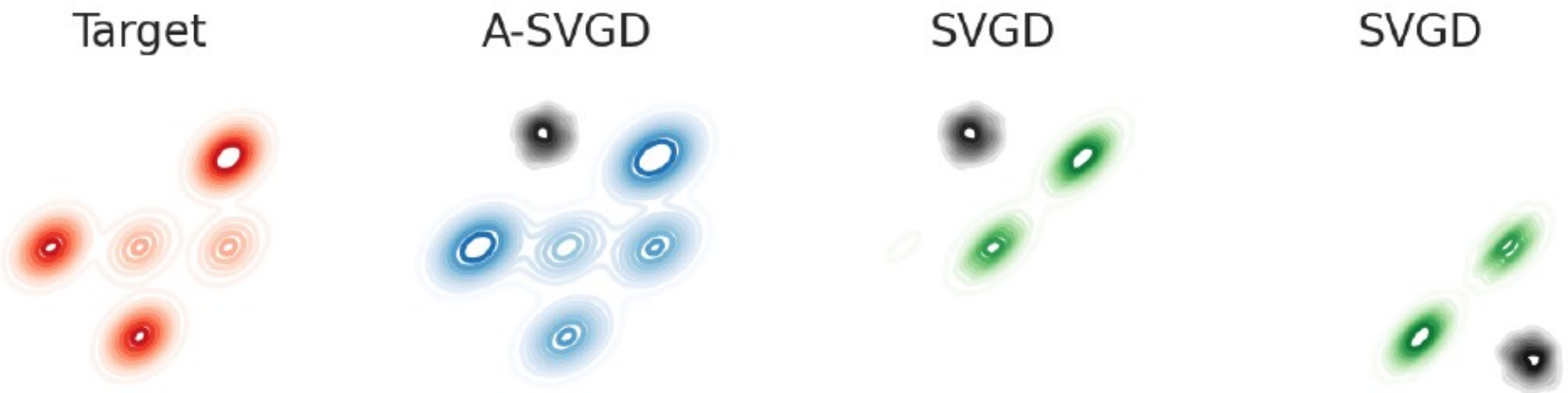


Figure 2: **Mode covering of SVGD.** We compare the final stationary distribution of standard SVGD (green) from two different initialization (black) and A-SVGD (blue) to approximate a mixture of Gaussians (red).

D'Angelo & Fortuin 2021