Uncon. Opt.

$$\boldsymbol{d}_t = \operatorname*{argmin}_{\boldsymbol{d}} M_t(\boldsymbol{d})$$
$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta_t \boldsymbol{d}_t$$

1st-order method

$$\mathcal{L}(\boldsymbol{\theta}) \approx \mathcal{L}(\boldsymbol{\theta}_t) + \boldsymbol{g}_t^\top \boldsymbol{d} =: M_t(\boldsymbol{d})$$

- steepest descent
  - gradient for direction: $\boldsymbol{d}_t = -\boldsymbol{g}_t$
  - line search for step size: $\eta_t = \operatorname*{argmin}_{\eta > 0} \mathcal{L}(\boldsymbol{\theta}_t + \eta \boldsymbol{d}_t)$
- momentum methods
  - vanilla momentum:
    $$\boldsymbol{m}_{t+1} = \beta \boldsymbol{m}_t + \boldsymbol{g}_t$$
    $$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \boldsymbol{m}_t$$
  - Nestrov momentum:
    $$\tilde{\boldsymbol{\theta}}_{t+1} = \boldsymbol{\theta}_t + \beta(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1})$$
    $$\boldsymbol{\theta}_{t+1} = \tilde{\boldsymbol{\theta}}_{t+1} - \eta_t \nabla \mathcal{L}(\tilde{\boldsymbol{\theta}}_{t+1})$$

2st-order method

$$\mathcal{L}(\boldsymbol{\theta}) \approx \mathcal{L}(\boldsymbol{\theta}_t) + \boldsymbol{g}_t^\top \boldsymbol{d} + \frac{1}{2} \boldsymbol{d}^\top \mathbf{H}_t \boldsymbol{d} =: M_t(\boldsymbol{d})$$

- Newton's method: $\boldsymbol{d}_t = -\mathbf{H}_t^{-1} \boldsymbol{g}_t$
- quasi-Newton methods
  - BFGS:
    $$\mathbf{H}_t \approx \mathbf{B}_t$$
    $$\mathbf{B}_{t+1} = \mathbf{B}_t + \frac{\Delta \boldsymbol{g}_t \Delta \boldsymbol{g}_t^\top}{\Delta \boldsymbol{g}_t^\top \Delta \boldsymbol{\theta}_t} - \frac{[\mathbf{B}_t \Delta \boldsymbol{\theta}_t][\mathbf{B}_t \Delta \boldsymbol{\theta}_t]^\top}{\Delta \boldsymbol{\theta}_t^\top \mathbf{B}_t \Delta \boldsymbol{\theta}_t}$$
  - L-BFGS:
    $$\mathbf{H}_t^{-1} \approx \mathbf{C}_t$$
    $$\mathbf{C}_{t+1} = \left[\mathbf{I} - \frac{\Delta \boldsymbol{\theta}_t \Delta \boldsymbol{g}_t^\top}{\Delta \boldsymbol{g}_t^\top \Delta \boldsymbol{\theta}_t}\right] \mathbf{C}_t \left[\mathbf{I} - \frac{\Delta \boldsymbol{g}_t \Delta \boldsymbol{\theta}_t^\top}{\Delta \boldsymbol{g}_t^\top \Delta \boldsymbol{\theta}_t}\right] + \frac{\Delta \boldsymbol{\theta}_t \Delta \boldsymbol{\theta}_t^\top}{\Delta \boldsymbol{g}_t^\top \Delta \boldsymbol{\theta}_t}$$
- Trust region method: $\boldsymbol{d}_t = \operatorname*{argmin}_{\boldsymbol{d} \in \mathcal{R}_t} M_t(\boldsymbol{d})$

SGD

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{q(\boldsymbol{z})} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z})$$
$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \nabla \mathcal{L}(\boldsymbol{\theta}_t, \boldsymbol{z}_t) = \boldsymbol{\theta}_t - \eta_t \boldsymbol{g}_t$$

- interate averaging: $\bar{\boldsymbol{\theta}}_t = \frac{1}{t} \sum_{i=1}^{t} \boldsymbol{\theta}_i = \frac{1}{t} \boldsymbol{\theta}_t + \frac{t-1}{t} \bar{\boldsymbol{\theta}}_{t-1}$
- variance reduction
  - SVRG:
    $$\tilde{\boldsymbol{\theta}}: \text{ any snapshot of } \boldsymbol{\theta} \text{ (fixed within each epoch)}$$
    $$\boldsymbol{g}_t = \nabla \mathcal{L}_t(\boldsymbol{\theta}_t) - \nabla \mathcal{L}_t(\tilde{\boldsymbol{\theta}}) + \nabla \mathcal{L}(\tilde{\boldsymbol{\theta}})$$
  - SAGA:
    $$\boldsymbol{g}_{t-1,n}^{\text{local}} = \nabla \mathcal{L}_n(\boldsymbol{\theta}_{t-1}), \ \boldsymbol{g}_{t-1}^{\text{avg}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{g}_{t-1,n}^{\text{local}}$$
    $$\boldsymbol{g}_t = \nabla \mathcal{L}_n(\boldsymbol{\theta}_t) - \boldsymbol{g}_{t-1,n}^{\text{local}} + \boldsymbol{g}_{t-1}^{\text{avg}}$$
- Preconditioned SGD: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \mathbf{M}_t^{-1} \boldsymbol{g}_t$
  - AdaGrad:
    $$\mathbf{M}_t = \operatorname{diag}(\boldsymbol{s}_t + \boldsymbol{\epsilon})^{\frac{1}{2}}, \text{ where } \boldsymbol{s}_t \in \mathbb{R}^D,$$
    $$s_{t,d} = \sum_{\tau=1}^{t} g_{\tau,d}^2$$
  - RMSProp:
    $$s_{t,d} = \beta s_{t-1,d} + (1-\beta) g_{t-1,d}^2$$
    $$\approx \frac{1}{t} \sum_{\tau=1}^{t} g_{\tau,d}^2 \text{ if } \beta \approx 1$$
  - AdaDelta:
    $$\mathbf{M}_t^{-1} = \frac{\sqrt{\boldsymbol{\delta}_{t-1} + \boldsymbol{\epsilon}}}{\sqrt{\boldsymbol{s}_t + \boldsymbol{\epsilon}}}$$
    $$\boldsymbol{s}_t \text{ same as RMSProp}, \ \boldsymbol{\delta}_{t-1} = \beta \boldsymbol{\delta}_{t-2} + (1-\beta)[\Delta \boldsymbol{\theta}_{t-1}]^2$$
  - Adam:
    $$\boldsymbol{m}_t = \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1) \boldsymbol{g}_t, \ \boldsymbol{s}_t = \beta_2 \boldsymbol{s}_{t-1} + (1-\beta_2) \boldsymbol{g}_t^2$$
    $$\Delta \boldsymbol{\theta}_t = -\eta_t \frac{1}{\sqrt{\boldsymbol{s}_t} + \boldsymbol{\varepsilon}} \boldsymbol{m}_t$$