

Introduction to stochastic optimization

Dmitry A. Kropotov



Gradient descent

Optimization problem:

$$F(\boldsymbol{x}) \rightarrow \min_{\boldsymbol{x}}, \quad \boldsymbol{x} \in \mathbb{R}^d, \quad F - \text{some smooth function.}$$

Optimization problem:

$$F(\boldsymbol{x}) \rightarrow \min_{\boldsymbol{x}}, \quad \boldsymbol{x} \in \mathbb{R}^d, \quad F - \text{some smooth function.}$$

Gradient descent method:

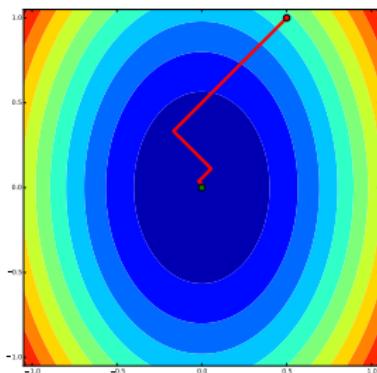
- Iteration $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla_{\boldsymbol{x}} F(\boldsymbol{x}_k);$
- Adapt stepsize α_k using some function decreasing criterion, e.g. Armijo rule:

$$\alpha_k : \quad F(\boldsymbol{x}_{k+1}) \leq F(\boldsymbol{x}_k) - \alpha_k c_1 \|\nabla_{\boldsymbol{x}} F(\boldsymbol{x}_k)\|^2, \quad c_1 = 10^{-4};$$

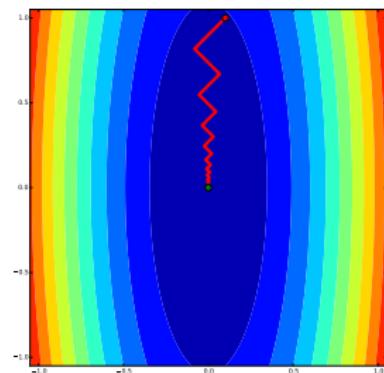
- Stop when $\|\nabla_{\boldsymbol{x}} F(\boldsymbol{x}_k)\|^2 \leq \varepsilon.$

Trajectories of Gradient Descent

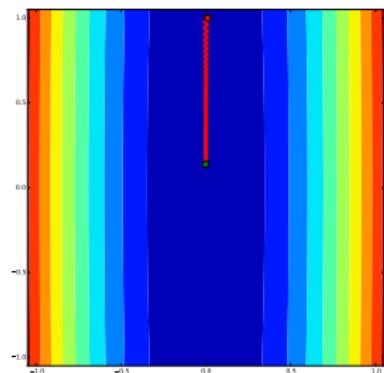
$$F(x, y) = \frac{1}{2}x^2 + \frac{\rho}{2}y^2 \rightarrow \min_{x,y}$$



$$\rho = 0.5$$



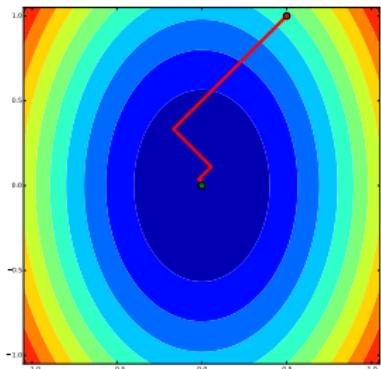
$$\rho = 0.1$$



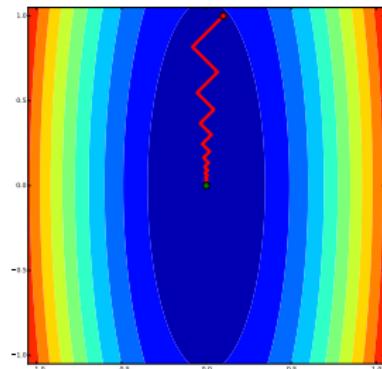
$$\rho = 0.01$$

Trajectories of Gradient Descent

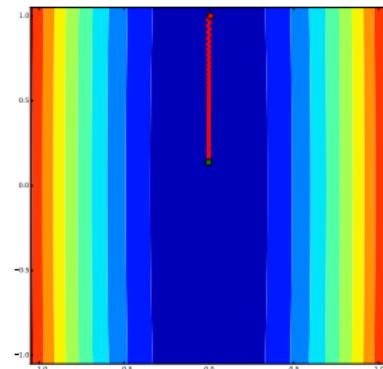
$$F(x, y) = \frac{1}{2}x^2 + \frac{\rho}{2}y^2 \rightarrow \min_{x,y}$$



$$\rho = 0.5$$



$$\rho = 0.1$$



$$\rho = 0.01$$

Gradient descent is sensitive to poor function scaling!

Newton method

$$F(\mathbf{x}) \approx m_k(\mathbf{x}) = F(\mathbf{x}_k) + \mathbf{g}_k^T(\mathbf{x} - \mathbf{x}_k) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_k)^T H_k(\mathbf{x} - \mathbf{x}_k) \rightarrow \min_{\mathbf{x}}$$
$$\mathbf{g}_k = \nabla_{\mathbf{x}} F(\mathbf{x}_k), \quad H_k = \nabla_{\mathbf{x}}^2 F(\mathbf{x}_k),$$
$$\arg \min_{\mathbf{x}} m_k(\mathbf{x}) = \mathbf{x}_k - H_k^{-1} \mathbf{g}_k.$$

Newton method

$$F(\mathbf{x}) \approx m_k(\mathbf{x}) = F(\mathbf{x}_k) + \mathbf{g}_k^T(\mathbf{x} - \mathbf{x}_k) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_k)^T H_k(\mathbf{x} - \mathbf{x}_k) \rightarrow \min_{\mathbf{x}}$$
$$\mathbf{g}_k = \nabla_{\mathbf{x}} F(\mathbf{x}_k), \quad H_k = \nabla_{\mathbf{x}}^2 F(\mathbf{x}_k),$$
$$\arg \min_{\mathbf{x}} m_k(\mathbf{x}) = \mathbf{x}_k - H_k^{-1} \mathbf{g}_k.$$

Newton method:

- Iteration $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k H_k^{-1} \mathbf{g}_k$;
- Adapt stepsize α_k using some function decreasing criterion, e.g. Armijo rule:

$$\alpha_k : \quad F(\mathbf{x}_{k+1}) \leq F(\mathbf{x}_k) - \alpha_k c_1 \mathbf{g}_k^T H_k^{-1} \mathbf{g}_k, \quad c_1 = 10^{-4};$$

- Stop when $\|\nabla_{\mathbf{x}} F(\mathbf{x}_k)\|^2 \leq \varepsilon$.

H_k^{-1} gives proper gradient scaling!

Limitations of Newton method:

- Hard to calculate Hessian H_k on each iteration;
- Hard to store Hessian in memory;
- Hard to inverse Hessian;
- Needs some Hessian modifications in case of non-convex function F .

Limitations of Newton method:

- Hard to calculate Hessian H_k on each iteration;
- Hard to store Hessian in memory;
- Hard to inverse Hessian;
- Needs some Hessian modifications in case of non-convex function F .

Advanced first-order optimization strategies:

- Conjugate gradients;
- Hessian-free Newton;
- Quasi-Newton methods (e.g. L-BFGS);
- etc.

Stochastic optimization problem

General stochastic optimization problem:

$$F(\mathbf{x}) = \mathbb{E}_{q(\mathbf{y})} f(\mathbf{x}, \mathbf{y}) \rightarrow \min_{\mathbf{x}}, \quad q - \text{some probability distribution.}$$

Stochastic optimization problem

General stochastic optimization problem:

$$F(\mathbf{x}) = \mathbb{E}_{q(\mathbf{y})} f(\mathbf{x}, \mathbf{y}) \rightarrow \min_{\mathbf{x}}, \quad q - \text{some probability distribution.}$$

Examples:

- Empirical risk minimization:

$$F(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N f_i(\mathbf{x}) = \mathbb{E}_{i \sim \text{Unif}\{1, \dots, N\}} f_i(\mathbf{x}) \rightarrow \min_{\mathbf{x}},$$

f_i – loss function for i-th training object.

Stochastic optimization problem

General stochastic optimization problem:

$$F(\mathbf{x}) = \mathbb{E}_{q(\mathbf{y})} f(\mathbf{x}, \mathbf{y}) \rightarrow \min_{\mathbf{x}}, \quad q - \text{some probability distribution.}$$

Examples:

- Empirical risk minimization:

$$F(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N f_i(\mathbf{x}) = \mathbb{E}_{i \sim \text{Unif}\{1, \dots, N\}} f_i(\mathbf{x}) \rightarrow \min_{\mathbf{x}},$$

f_i – loss function for i-th training object.

- Evidence lower bound (ELBO) maximization:

$$\mathbb{E}_{q(\mathbf{t}|\boldsymbol{\lambda})} \log p(\mathbf{x}, \mathbf{t}|\boldsymbol{\theta}) - \mathbb{E}_{q(\mathbf{t}|\boldsymbol{\lambda})} \log q(\mathbf{t}|\boldsymbol{\lambda}) \rightarrow \max_{\boldsymbol{\theta}, \boldsymbol{\lambda}}.$$

Computational challenges

$$F(\boldsymbol{x}) = \frac{1}{N} \sum_{i=1}^N f_i(\boldsymbol{x}) \rightarrow \min_{\boldsymbol{x}}.$$

Value	Calculation costs:
$f_i(\boldsymbol{x})$	$O(s)$
$\nabla_{\boldsymbol{x}} f_i(\boldsymbol{x})$	

Computational challenges

$$F(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N f_i(\mathbf{x}) \rightarrow \min_{\mathbf{x}}.$$

Value	Calculation costs:
$f_i(\mathbf{x})$	$O(s)$
$\nabla_{\mathbf{x}} f_i(\mathbf{x})$	$O(s)$
$F(\mathbf{x})$	$O(Ns)$
$\nabla F(\mathbf{x})$	$O(Ns)$

Can be challenging when $N \gg 1!$

Computational challenges

$$F(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N f_i(\mathbf{x}) \rightarrow \min_{\mathbf{x}}.$$

Value	Calculation costs:
$f_i(\mathbf{x})$	$O(s)$
$\nabla_{\mathbf{x}} f_i(\mathbf{x})$	$O(s)$
$F(\mathbf{x})$	$O(Ns)$
$\nabla F(\mathbf{x})$	$O(Ns)$

Can be challenging when $N \gg 1!$

$$F(\mathbf{x}) = \mathbb{E}_{q(\mathbf{y})} f(\mathbf{x}, \mathbf{y}) = \int f(\mathbf{x}, \mathbf{y}) q(\mathbf{y}) d\mathbf{y} \rightarrow \min_{\mathbf{x}}.$$

Exact calculation may not exist for complex distribution $q!$

Stochastic estimates

$$F(\mathbf{x}) = \mathbb{E}_{q(\mathbf{y})} f(\mathbf{x}, \mathbf{y}),$$

$$g(\mathbf{x}) = \nabla_{\mathbf{x}} F(\mathbf{x}) = \mathbb{E}_{q(\mathbf{y})} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}).$$

Stochastic estimates:

$$\mathbf{y}_1, \dots, \mathbf{y}_m \sim q(\mathbf{y}),$$

$$\hat{F}(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^m f(\mathbf{x}, \mathbf{y}_i), \quad \hat{g}(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}_i).$$

Estimates \hat{F} and \hat{g} :

- unbiased, i.e. $\mathbb{E}_{q(\mathbf{y})} \hat{F}(\mathbf{x}) = F(\mathbf{x})$, $\mathbb{E}_{q(\mathbf{y})} \hat{g}(\mathbf{x}) = g(\mathbf{x})$;
- become exact when $m \rightarrow +\infty$;
- easy to calculate for small m .

Stochastic Gradient Descent

Using stochastic estimates in gradient descent method leads to Stochastic Gradient Descent (SGD):

- Generate a small batch of samples: $y_1, \dots, y_m \sim q(y)$;

Stochastic Gradient Descent

Using stochastic estimates in gradient descent method leads to Stochastic Gradient Descent (SGD):

- Generate a small batch of samples: $\mathbf{y}_1, \dots, \mathbf{y}_m \sim q(\mathbf{y})$;
- Calculate $\hat{g}(\mathbf{x}_k) = \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_i)$;

Using stochastic estimates in gradient descent method leads to Stochastic Gradient Descent (SGD):

- Generate a small batch of samples: $\mathbf{y}_1, \dots, \mathbf{y}_m \sim q(\mathbf{y})$;
- Calculate $\hat{g}(\mathbf{x}_k) = \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_i)$;
- Make a step: $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \hat{g}(\mathbf{x}_k)$;

Using stochastic estimates in gradient descent method leads to Stochastic Gradient Descent (SGD):

- Generate a small batch of samples: $\mathbf{y}_1, \dots, \mathbf{y}_m \sim q(\mathbf{y})$;
- Calculate $\hat{g}(\mathbf{x}_k) = \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_i)$;
- Make a step: $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \hat{g}(\mathbf{x}_k)$;
- Iterate until convergence.

Using stochastic estimates in gradient descent method leads to Stochastic Gradient Descent (SGD):

- Generate a small batch of samples: $\mathbf{y}_1, \dots, \mathbf{y}_m \sim q(\mathbf{y})$;
- Calculate $\hat{g}(\mathbf{x}_k) = \frac{1}{m} \sum_{i=1}^m \nabla_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_i)$;
- Make a step: $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \hat{g}(\mathbf{x}_k)$;
- Iterate until convergence.

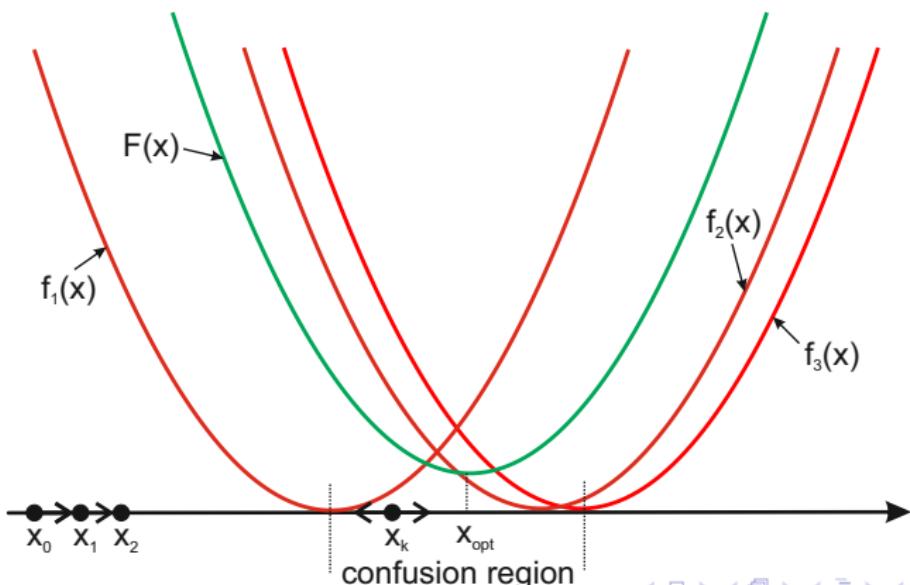
Questions for SGD:

- Convergence rate comparing to full gradient descent?
- How to choose step size α_k ?
- When to stop?

SGD trajectory

Fitting linear regression in 1D space:

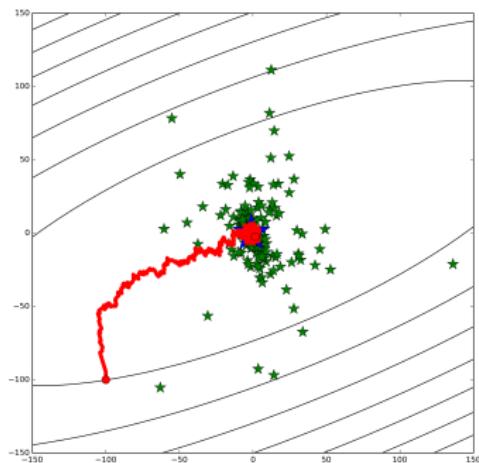
$$F(x) = \frac{1}{N} \sum_{i=1}^N \underbrace{\frac{1}{2}(b_i - a_i x)^2}_{f_i(x)} \rightarrow \min_{x \in \mathbb{R}}$$



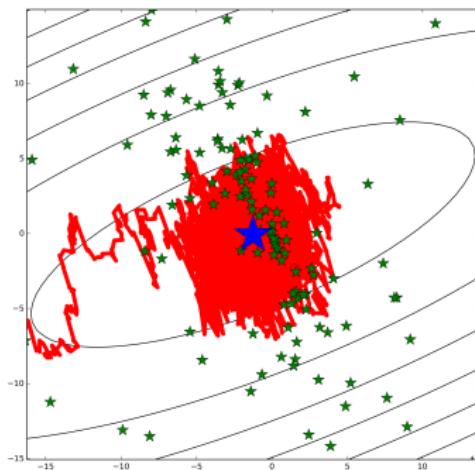
SGD trajectory

Fitting linear regression in 2D space:

$$F(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N (b_i - \mathbf{a}_i^T \mathbf{x})^2 \rightarrow \min_{\mathbf{x} \in \mathbb{R}^2} .$$



SGD trajectory



optimum vicinity

SGD properties:

- stochastic gradient doesn't generate descent direction;
- stochastic gradient doesn't equal zero in optimum of $F(\mathbf{x}) \Rightarrow$ SGD can't converge with constant step size;
- we don't have access to $F(\mathbf{x}) \Rightarrow$ can't use Armijo-like rules for adapting step size;
- we don't have access to $\nabla_{\mathbf{x}}F(\mathbf{x}) \Rightarrow$ can't use standard termination condition.

SGD convergence analysis

Optimization problem: $F(\mathbf{x}) = \mathbb{E}_{q(\mathbf{y})} f(\mathbf{x}, \mathbf{y}) \rightarrow \min_{\mathbf{x}}$

SGD iteration: $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \hat{\mathbf{g}}_k$, $\mathbb{E} \hat{\mathbf{g}}_k = \mathbf{g}_k = \nabla F(\mathbf{x}_k)$.

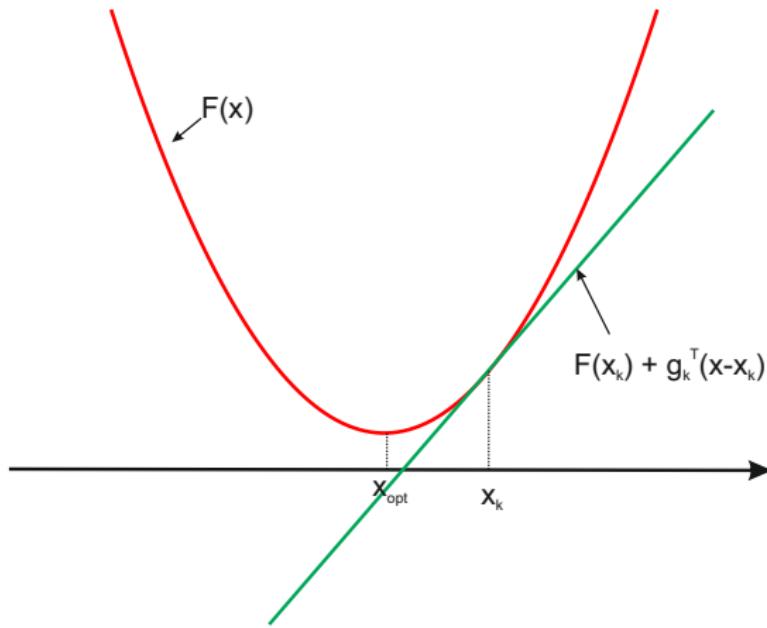
$$\begin{aligned}\|\mathbf{x}_{k+1} - \mathbf{x}_{opt}\|^2 &= \|\mathbf{x}_k - \mathbf{x}_{opt} - \alpha_k \hat{\mathbf{g}}_k\|^2 = \\ &= \|\mathbf{x}_k - \mathbf{x}_{opt}\|^2 - 2\alpha_k \hat{\mathbf{g}}_k^T (\mathbf{x}_k - \mathbf{x}_{opt}) + \alpha_k^2 \|\hat{\mathbf{g}}_k\|^2\end{aligned}$$

$$\mathbb{E} \|\mathbf{x}_{k+1} - \mathbf{x}_{opt}\|^2 = \|\mathbf{x}_k - \mathbf{x}_{opt}\|^2 - 2\alpha_k \mathbf{g}_k^T (\mathbf{x}_k - \mathbf{x}_{opt}) + \alpha_k^2 \mathbb{E} \|\hat{\mathbf{g}}_k\|^2$$

SGD convergence analysis

Suppose F is a convex function, i.e.

$$F(\mathbf{x}_{opt}) \geq F(\mathbf{x}_k) + \mathbf{g}_k^T (\mathbf{x}_{opt} - \mathbf{x}_k)$$



SGD convergence analysis

Таким образом имеем:

$$\mathbb{E}\|\mathbf{x}_{k+1} - \mathbf{x}_{opt}\|^2 = \|\mathbf{x}_k - \mathbf{x}_{opt}\|^2 - 2\alpha_k \mathbf{g}_k^T (\mathbf{x}_k - \mathbf{x}_{opt}) + \alpha_k^2 \mathbb{E}\|\hat{\mathbf{g}}_k\|^2 \quad (1)$$

$$F(\mathbf{x}_{opt}) \geq F(\mathbf{x}_k) + \mathbf{g}_k^T (\mathbf{x}_{opt} - \mathbf{x}_k). \quad (2)$$

Тогда

$$\begin{aligned} \alpha_k(F(\mathbf{x}_k) - F_{opt}) &\stackrel{(2)}{\leq} \alpha_k \mathbf{g}_k^T (\mathbf{x}_k - \mathbf{x}_{opt}) \stackrel{(1)}{=} \frac{1}{2} \|\mathbf{x}_k - \mathbf{x}_{opt}\|^2 + \\ &+ \frac{1}{2} \alpha_k^2 \mathbb{E}\|\hat{\mathbf{g}}_k\|^2 - \frac{1}{2} \mathbb{E}\|\mathbf{x}_{k+1} - \mathbf{x}_{opt}\|^2. \end{aligned}$$

Берём мат. ожидание от левой части и суммируем по k от нуля:

$$\begin{aligned} \sum_{i=0}^k \alpha_i (\mathbb{E}F(\mathbf{x}_i) - F_{opt}) &\leq \frac{1}{2} \|\mathbf{x}_0 - \mathbf{x}_{opt}\|^2 + \frac{1}{2} \sum_{i=0}^k \alpha_i^2 \mathbb{E}\|\hat{\mathbf{g}}_i\|^2 - \\ &- \frac{1}{2} \mathbb{E}\|\mathbf{x}_{k+1} - \mathbf{x}_{opt}\|^2 \leq \frac{1}{2} \|\mathbf{x}_0 - \mathbf{x}_{opt}\|^2 + \frac{1}{2} \sum_{i=0}^k \alpha_i^2 \mathbb{E}\|\hat{\mathbf{g}}_i\|^2. \end{aligned}$$

SGD convergence analysis

$$\mathbb{E}F\left(\underbrace{\frac{\sum_{i=0}^k \alpha_i \mathbf{x}_i}{\sum_{i=0}^k \alpha_i}}_{\hat{\mathbf{x}}_k}\right) - F_{opt} \leq \{\text{convexity of } F\} \leq$$
$$\frac{\sum_{i=0}^k \alpha_i (\mathbb{E}F(\mathbf{x}_i) - F_{opt})}{\sum_{i=0}^k \alpha_i} \leq \frac{\frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}_{opt}\|^2 + \frac{1}{2}\sum_{i=0}^k \alpha_i^2 \mathbb{E}\|\hat{\mathbf{g}}_i\|^2}{\sum_{i=0}^k \alpha_i}.$$

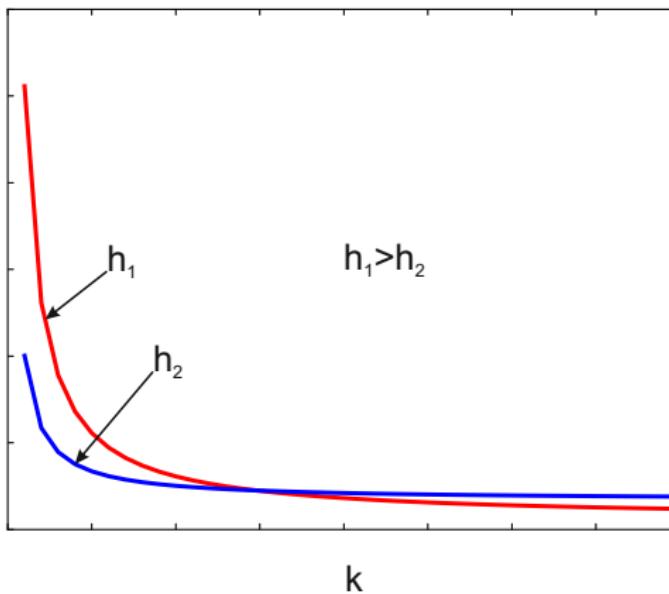
Suppose $\|\mathbf{x}_0 - \mathbf{x}_{opt}\|^2 \leq R^2$ and $\mathbb{E}\|\hat{\mathbf{g}}_k\|^2 \leq G^2$. Then

$$\boxed{\mathbb{E}F(\hat{\mathbf{x}}_k) - F_{opt} \leq \frac{R^2 + G^2 \sum_{i=0}^k \alpha_i^2}{2 \sum_{i=0}^k \alpha_i}}$$

Choosing step size

Strategy 1: $\alpha_i = h \forall i$. Then

$$\mathbb{E} F(\hat{\mathbf{x}}_k) - F_{opt} \leq \frac{R^2 + G^2 \sum_{i=0}^k \alpha_i^2}{2 \sum_{i=0}^k \alpha_i} = \frac{R^2}{2h(k+1)} + \frac{G^2 h}{2} \xrightarrow{k \rightarrow \infty} \frac{G^2 h}{2}.$$



Choosing step size

$$\mathbb{E}F(\hat{\mathbf{x}}_k) - F_{opt} \leq \frac{R^2 + G^2 \sum_{i=0}^k \alpha_i^2}{2 \sum_{i=0}^k \alpha_i}$$

Strategy 2: $\sum_{i=0}^{\infty} \alpha_i = \infty$, $\sum_{i=0}^{\infty} \alpha_i^2 < \infty$. This is true for the following choice:

$$\alpha_i = \frac{\eta}{(i+1)^{\tau}}, \quad \tau \in (1/2, 1].$$

If $\tau = 1$, then $\sum_{i=0}^k \frac{1}{i} \sim \log k$ and $\mathbb{E}F(\hat{\mathbf{x}}_k) - F_{opt} = O(1/\log k)$.

Strategy 3: $\sum_{i=0}^{\infty} \alpha_i = \infty$, $\sum_{i=0}^k \alpha_i^2 / \sum_{i=0}^k \alpha_i \xrightarrow[k \rightarrow \infty]{} 0$. This is true for the following choice:

$$\alpha_i = \frac{\eta}{(i+1)^{\tau}}, \quad \tau \leq 1/2.$$

If $\tau = 1/2$, then $\mathbb{E}F(\hat{\mathbf{x}}_k) - F_{opt} = O(\log k / \sqrt{k})$. This is optimal choice.

GD and SGD convergence rates

Function F	GD	SGD
Smooth and strongly convex	$O(c^k)$	$O(1/k)$
Smooth and convex	$O(1/k)$	$O(1/\sqrt{k})$

- Full convergence is very slow \Rightarrow we can't expect solution with high accuracy;
- Valid stopping criterion is absent \Rightarrow run the method for appropriate time period and measure the quality of solution using testing or control sample;
- In practice it is usually enough to reach confusion region, so choose strategy 1 for step size and possibly decrease the step size when testing quality stabilizes;
- It is important to have low G .

Using SGD with momentum

SGD with momentum:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \hat{\mathbf{g}}_k + \beta_k (\mathbf{x}_k - \mathbf{x}_{k-1}).$$

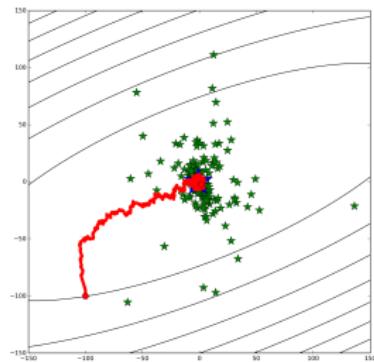


without momentum

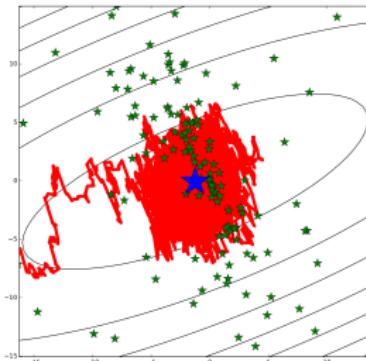


with momentum

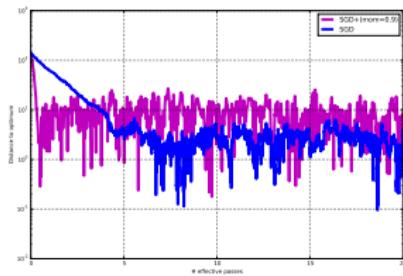
Using SGD with momentum



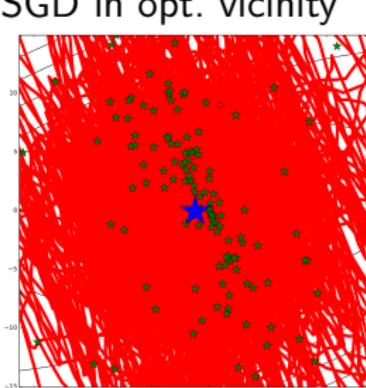
SGD



SGD in opt. vicinity



SGD+mom.



SGD+mom. in opt.

$$\begin{aligned}\mathbf{m}_{k+1} &= \beta_1 \mathbf{m}_k + (1 - \beta_1) \hat{\mathbf{g}}_k, \\ \mathbf{v}_{k+1} &= \beta_2 \mathbf{v}_k + (1 - \beta_2) \hat{\mathbf{g}}_k \odot \hat{\mathbf{g}}_k, \\ \mathbf{x}_{k+1} &= \mathbf{x}_k - \alpha_k \frac{\mathbf{m}_{k+1}}{\sqrt{\mathbf{v}_{k+1} + \varepsilon}}.\end{aligned}$$