Optimization for Data Science

F. Rinaldi[1]

[1] Università degli Studi di Padova | Dipartimento Matematica

Padova
2020

Outline

Optimization for Data Science

1 Stochastic Optimization

2 Sample Average Approximation

3 Stochastic Gradient Approximation

4 Why Using SG?

Uncertainty and Optimization

- Decision makers often have to deal with uncertainty when making decisions.
- Many decision problems are formulated as optimization problems with uncertain parameters.
- It is usually quite difficult to formulate and solve such problems, both conceptually and numerically.

Conceptual Stage

- There is a variety of ways in which the uncertainty can be formalized.
- GOAL: Trade-off between realism of model and tractability.
- A large number of ways to model uncertainty.

Uncertainty and Optimization

- Decision makers often have to deal with uncertainty when making decisions.
- Many decision problems are formulated as optimization problems with uncertain parameters.
- It is usually quite difficult to formulate and solve such problems, both conceptually and numerically.

Conceptual Stage

- There is a variety of ways in which the uncertainty can be formalized.
- GOAL: Trade-off between realism of model and tractability.
- A large number of ways to model uncertainty.

Uncertainty and Data Science

- Here we both give an overview of some classic approaches and describe some recent methods.
- We will focus on specific techniques that are widely used in the big data community.
- In particular, we first overview classic methods like sample average approximation and stochastic approximation.
- Then we will focus on finite sum problems and on specific techniques that help us to deal with those problems.

Uncertainty and Data Science

- Here we both give an overview of some classic approaches and describe some recent methods.
- We will focus on specific techniques that are widely used in the big data community.
- In particular, we first overview classic methods like sample average approximation and stochastic approximation.
- Then we will focus on finite sum problems and on specific techniques that help us to deal with those problems.

Problem Formulation

Problem Formulation

We consider the following problem:

$$\min_{x \in \mathbb{R}^n} \quad f(x) = \mathbb{E}_\xi[F(x, \xi)] \tag{1}$$

- $F(x, \xi)$ is a function that involves our set of decision variables x and a random variable ξ
- ξ has given sample space Ω and probability distribution P.

Problem Formulation

Problem Formulation

We consider the following problem:

$$\min_{x \in \mathbb{R}^n} \quad f(x) = \mathbb{E}_\xi [F(x, \xi)] \tag{1}$$

- $F(x, \xi)$ is a function that involves our set of decision variables x and a random variable ξ
- ξ has given sample space Ω and probability distribution P.

Example in Data Science

Expected Risk Minimization

- Given two spaces of objects X and Y learn a function (often called hypothesis) which outputs an object $y \in Y$ given $x \in X$.
- x and y are random input/output data.
- **Prediction function** $h(x; w)$ has fixed form and is parameterized by a vector w over which our optimization will be performed.
- GOAL: Find the prediction function $h(x; w)$ (i.e., the parameters w defining it) that minimizes the losses incurred by inaccurate predictions (also called *prediction losses* or *prediction errors*).
- Losses measured via *loss function* (it measures the difference between predicted and real outputs).
- Loss function indicated with $\ell(h(x; w), y)$, where $h(x; w)$ and y respectively represent predicted and true outputs.

Example in Data Science

Expected Risk Minimization

- Given two spaces of objects X and Y learn a function (often called hypothesis) which outputs an object $y \in Y$ given $x \in X$.
- x and y are random input/output data.
- **Prediction function** $h(x; w)$ has fixed form and is parameterized by a vector w over which our optimization will be performed.
- GOAL: Find the prediction function $h(x; w)$ (i.e., the parameters w defining it) that minimizes the losses incurred by inaccurate predictions (also called *prediction losses* or *prediction errors*).
- Losses measured via *loss function* (it measures the difference between predicted and real outputs).
- Loss function indicated with $\ell(h(x; w), y)$, where $h(x; w)$ and y respectively represent predicted and true outputs.

Example in Data Science

Expected Risk Minimization

- Given two spaces of objects X and Y learn a function (often called hypothesis) which outputs an object $y \in Y$ given $x \in X$.
- x and y are random input/output data.
- **Prediction function** $h(x; w)$ has fixed form and is parameterized by a vector w over which our optimization will be performed.
- GOAL: Find the prediction function $h(x; w)$ (i.e., the parameters w defining it) that minimizes the losses incurred by inaccurate predictions (also called *prediction losses* or *prediction errors*).
- Losses measured via *loss function* (it measures the difference between predicted and real outputs).
- Loss function indicated with $\ell(h(x; w), y)$, where $h(x; w)$ and y respectively represent predicted and true outputs.

Expected Risk Minimization: Formulation

Formulation

We then want to solve the following stochastic optimization problem:

$$\min_{w} \quad R(w) = \mathbb{E}_{xy}[\ell(h(x;w),y)]. \tag{2}$$

- We want to describe two classic approaches that can be considered for solving this class of problems.

Expected Risk Minimization: Formulation

Formulation

We then want to solve the following stochastic optimization problem:

$$\min_{w} \quad R(w) = \mathbb{E}_{xy}[\ell(h(x;w),y)]. \tag{2}$$

- We want to describe two classic approaches that can be considered for solving this class of problems.

Sample Average Approximation Approach

Sample average approximation

We consider N random samples for the random variable ξ and build the approximation of the expected value by considering the *sample average*:

$$\min_{x \in \mathbb{R}^n} \quad f^N(x) = \frac{1}{N} \sum_{i=1}^{N} F(x, \xi_i). \tag{3}$$

PROs and CONs

PROs

- $f^N(x)$ converges to $f(x)$ with probability one when $N \to \infty$.
- Once we build up problem (3), we can use any method from classic deterministic optimization for solving it.

CONs

- Hard to determine a priori the sample size that guarantees good accuracy for the model.
- Obviously, the larger N the better the model.
- Choosing a very large N might be very expensive.

PROs and CONs

PROs

- $f^N(x)$ converges to $f(x)$ with probability one when $N \to \infty$.
- Once we build up problem (3), we can use any method from classic deterministic optimization for solving it.

CONs

- Hard to determine a priori the sample size that guarantees good accuracy for the model.
- Obviously, the larger N the better the model.
- Choosing a very large N might be very expensive.

Why Choosing Large N is Bad

- Assume the function F is continuously differentiable with respect to x for any given ξ_i.
- Once you build up sample average approximation problem (3) you can use, e.g., gradient method to solve it.

Remark

- Computing the gradient $\nabla f^N(x)$ is highly expensive in Data Science applications.
- It corresponds to calculate $\mathcal{O}(N)$ gradients in practice!

Why Choosing Large N is Bad

- Assume the function F is continuously differentiable with respect to x for any given ξ_i.
- Once you build up sample average approximation problem (3) you can use, e.g., gradient method to solve it.

Remark

- Computing the gradient $\nabla f^N(x)$ is highly expensive in Data Science applications.
- It corresponds to calculate $\mathcal{O}(N)$ gradients in practice!

Stochastic Gradient Approximation

- We now describe the stochastic gradient method by Robbins and Monro (1951).
- We again assume that the function F is continuously differentiable with respect to x for any given ξ.
- The stochastic gradient method generates a new iterate as follows:

$$x_{k+1} = x_k - \alpha_k \nabla F(x_k, \xi_k).$$

- ξ_k a sample realization of ξ and α_k a suitably chosen stepsize.

Algorithmic Scheme

Algorithm 1 `Stochastic Gradient (SG) method`

```
Choose a point $x_1 \in \mathbb{R}^n$
For $k = 1, \dots$
    If $x_k$ satisfies some specific condition, then STOP
    Choose $\xi_k$ a sample realization of $\xi$
    Set $x_{k+1} = x_k - \alpha_k \nabla F(x_k, \xi_k)$, with $\alpha_k > 0$
    a suitably chosen stepsize
End for
```

Comments

- It is easy to see that the stochastic gradient is *unbiased*, i.e.,

$$\mathbb{E}[\nabla F(x,\xi)] = \nabla f(x).$$

- In the algorithm we need a diminishing stepsize α_k in order to ensure convergence.
- We need a sequence $\{\alpha_k\}$ such that $\alpha_k \to 0$ when k goes to infinity.
- **Intuitive Idea:** at optimality we have

$$x^* = x^* - \alpha\nabla F(x^*,\xi)$$

and since $\nabla F(x^*,\xi)$ is random, we cannot guarantee $\nabla F(x^*,\xi) = 0$ for all $\xi \in \Omega$.

Comments

- It is easy to see that the stochastic gradient is *unbiased*, i.e.,

$$\mathbb{E}[\nabla F(x,\xi)] = \nabla f(x).$$

- In the algorithm we need a diminishing stepsize α_k in order to ensure convergence.
- We need a sequence $\{\alpha_k\}$ such that $\alpha_k \to 0$ when k goes to infinity.
- **Intuitive Idea:** at optimality we have

$$x^* = x^* - \alpha \nabla F(x^*, \xi)$$

and since $\nabla F(x^*, \xi)$ is random, we cannot guarantee $\nabla F(x^*, \xi) = 0$ for all $\xi \in \Omega$.

Law of Total Expectation

Law of Total Expectation

If all the expectations are finite, then for any random variables X and Y, we have:

- $\mathbb{E}[X] = \mathbb{E}_Y\left[\mathbb{E}[X|Y]\right];$
- $\mathbb{E}[g(X)] = \mathbb{E}_Y\left[\mathbb{E}[g(X)|Y]\right]$ for any function g.

Note that we can pick any r.v. Y, to make the expectation as easy as we can.

Main Convergence Result

Convergence for SG

Let $f:\mathbb{R}^n\rightarrow\mathbb{R}$ be a σ-strongly convex function with continuous Lipschitz gradient. assume that there exists $M>0$ s.t.

$$\mathbb{E}[\|\nabla F(x,\xi)\|^2]\leq M^2, \forall\, x\in\mathbb{R}^n.$$

Stochastic gradient method with $\alpha_k=\frac{\gamma}{k+\delta}$, $\delta>0$ and $\gamma>1/2\sigma$ satisfies:

$$\mathbb{E}\left[f(x_k)-f(x^*)\right]\leq\frac{LC(\gamma)}{2(k+\delta)},$$

where $C(\gamma)$ is

$$C(\gamma)=\max\{\gamma^2M^2(2\sigma\gamma-1)^{-1},\ (1+\delta)\|x_1-x^*\|^2\}.$$

Proof I

Remark

Iterate x_k is a function of the generated random process $\xi_{[k-1]} = (\xi_1, \ldots, \xi_{k-1})$.

At an iteration k of the SG algorithm, given x_k and a sample ξ_k, we have that the distance of the new iterate x_{k+1} from the optimal value x^* is such that

$$\begin{aligned} ||x_{k+1} - x^*||^2 &= ||x_k - \alpha_k \nabla F(x_k, \xi_k) - x^*||^2 \\ &= ||x_k - x^*||^2 - 2\alpha_k (\nabla F(x_k, \xi_k)^\top (x_k - x^*)) + \alpha_k^2 ||\nabla F(x_k, \xi_k)||^2. \end{aligned} \tag{4}$$

Taking expectation on both sides and keeping in mind properties of the gradient, we can write

$$\begin{aligned} \mathbb{E}[||x_{k+1} - x^*||^2] &= \mathbb{E}[||x_k - x^*||^2] - 2\alpha_k \mathbb{E}[\nabla F(x_k, \xi_k)^\top (x_k - x^*)] \\ &+ \alpha_k^2 \mathbb{E}[||\nabla F(x_k, \xi_k)||^2] \\ &\leq \mathbb{E}[||x_k - x^*||^2] - 2\alpha_k \mathbb{E}[\nabla F(x_k, \xi_k)^\top (x_k - x^*)] + \alpha_k^2 M^2 \end{aligned} \tag{5}$$

Proof I

Remark

Iterate x_k is a function of the generated random process $\xi_{[k-1]} = (\xi_1, \ldots, \xi_{k-1})$.

At an iteration k of the SG algorithm, given x_k and a sample ξ_k, we have that the distance of the new iterate x_{k+1} from the optimal value x^* is such that

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \|x_k - \alpha_k \nabla F(x_k, \xi_k) - x^*\|^2 \qquad (4) \\ &= \|x_k - x^*\|^2 - 2\alpha_k (\nabla F(x_k, \xi_k)^\top (x_k - x^*)) + \alpha_k^2 \|\nabla F(x_k, \xi_k)\|^2. \end{aligned}$$

Taking expectation on both sides and keeping in mind properties of the gradient, we can write

$$\begin{aligned} \mathbb{E}[\|x_{k+1} - x^*\|^2] &= \mathbb{E}[\|x_k - x^*\|^2] - 2\alpha_k \mathbb{E}[\nabla F(x_k, \xi_k)^\top (x_k - x^*)] \qquad (5) \\ &+ \alpha_k^2 \mathbb{E}[\|\nabla F(x_k, \xi_k)\|^2] \\ &\leq \mathbb{E}[\|x_k - x^*\|^2] - 2\alpha_k \mathbb{E}[\nabla F(x_k, \xi_k)^\top (x_k - x^*)] + \alpha_k^2 M^2 \end{aligned}$$

Proof I

Remark

Iterate x_k is a function of the generated random process $\xi_{[k-1]} = (\xi_1, \ldots, \xi_{k-1})$.

At an iteration k of the SG algorithm, given x_k and a sample ξ_k, we have that the distance of the new iterate x_{k+1} from the optimal value x^* is such that

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \|x_k - \alpha_k \nabla F(x_k, \xi_k) - x^*\|^2 \qquad (4) \\ &= \|x_k - x^*\|^2 - 2\alpha_k (\nabla F(x_k, \xi_k)^\top (x_k - x^*)) + \alpha_k^2 \|\nabla F(x_k, \xi_k)\|^2. \end{aligned}$$

Taking expectation on both sides and keeping in mind properties of the gradient, we can write

$$\begin{aligned} \mathbb{E}[\|x_{k+1} - x^*\|^2] &= \mathbb{E}[\|x_k - x^*\|^2] - 2\alpha_k \mathbb{E}[\nabla F(x_k, \xi_k)^\top (x_k - x^*)] \qquad (5) \\ &+ \alpha_k^2 \mathbb{E}[\|\nabla F(x_k, \xi_k)\|^2] \\ &\leq \mathbb{E}[\|x_k - x^*\|^2] - 2\alpha_k \mathbb{E}[\nabla F(x_k, \xi_k)^\top (x_k - x^*)] + \alpha_k^2 M^2 \end{aligned}$$

Proof II

Now, using law of total expectation and taking into account the fact that x_k is independent with respect to ξ_k, we can write

$$
\begin{aligned}
\mathbb{E}[\nabla F(x_k,\xi_k)^\top (x_k - x^*)] &= \mathbb{E}_{\xi_{[k-1]}}[\mathbb{E}_{\xi_k}[\nabla F(x_k,\xi_k)^\top (x_k - x^*)|\xi_{[k-1]}]] \\
&= \mathbb{E}_{\xi_{[k-1]}}[\mathbb{E}_{\xi_k}[\nabla F(x_k,\xi_k)|\xi_{[k-1]}]^\top (x_k - x^*)] \\
&\quad \text{(by independence of sample } \xi_k\text{)} \\
&= \mathbb{E}_{\xi_{[k-1]}}[\nabla f(x_k)^\top (x_k - x^*)] \\
&= \mathbb{E}[\nabla f(x_k)^\top (x_k - x^*)],
\end{aligned}
$$

that is

$$
\mathbb{E}[\nabla F(x_k,\xi_k)^\top (x_k - x^*)] = \mathbb{E}[\nabla f(x_k)^\top (x_k - x^*)]. \tag{6}
$$

Proof II

Now, using law of total expectation and taking into account the fact that x_k is independent with respect to ξ_k, we can write

$$\begin{aligned}
\mathbb{E}[\nabla F(x_k,\xi_k)^\top(x_k-x^*)] &= \mathbb{E}_{\xi_{[k-1]}}[\mathbb{E}_{\xi_k}[\nabla F(x_k,\xi_k)^\top(x_k-x^*)|\xi_{[k-1]}]] \\
&= \mathbb{E}_{\xi_{[k-1]}}[\mathbb{E}_{\xi_k}[\nabla F(x_k,\xi_k)|\xi_{[k-1]}]^\top(x_k-x^*)] \\
&\quad \text{(by independence of sample } \xi_k\text{)} \\
&= \mathbb{E}_{\xi_{[k-1]}}[\nabla f(x_k)^\top(x_k-x^*)] \\
&= \mathbb{E}[\nabla f(x_k)^\top(x_k-x^*)],
\end{aligned}$$

that is

$$\mathbb{E}[\nabla F(x_k,\xi_k)^\top(x_k-x^*)] = \mathbb{E}[\nabla f(x_k)^\top(x_k-x^*)]. \tag{6}$$

Proof II

Now, using law of total expectation and taking into account the fact that x_k is independent with respect to ξ_k, we can write

$$\begin{aligned}
\mathbb{E}[\nabla F(x_k,\xi_k)^\top(x_k - x^*)] &= \mathbb{E}_{\xi_{[k-1]}}[\mathbb{E}_{\xi_k}[\nabla F(x_k,\xi_k)^\top(x_k - x^*)|\xi_{[k-1]}]] \\
&= \mathbb{E}_{\xi_{[k-1]}}[\mathbb{E}_{\xi_k}[\nabla F(x_k,\xi_k)|\xi_{[k-1]}]^\top(x_k - x^*)] \\
&\quad \text{(by independence of sample } \xi_k\text{)} \\
&= \mathbb{E}_{\xi_{[k-1]}}[\nabla f(x_k)^\top(x_k - x^*)] \\
&= \mathbb{E}[\nabla f(x_k)^\top(x_k - x^*)],
\end{aligned}$$

that is

$$\mathbb{E}[\nabla F(x_k,\xi_k)^\top(x_k - x^*)] = \mathbb{E}[\nabla f(x_k)^\top(x_k - x^*)]. \tag{6}$$

Proof III

Using σ-strong convexity for f, we can write, for all $x \in \mathbb{R}^n$, the following:

$$(\nabla f(x) - \nabla f(x^*))^\top (x - x^*) \geq \sigma \|x - x^*\|^2,$$

which can be rewritten as

$$\nabla f(x)^\top (x - x^*) \geq \sigma \|x - x^*\|^2 + \nabla f(x^*)^\top (x - x^*).$$

Thus, keeping in mind that $\nabla f(x^*) = 0$, we get, by taking expectation, the following:

$$\mathbb{E}[\nabla f(x)^\top (x - x^*)] \geq \sigma \mathbb{E}[\|x - x^*\|^2].$$

Combining the last one with

$$\mathbb{E}[\nabla F(x_k, \xi_k)^\top (x_k - x^*)] = \mathbb{E}[\nabla f(x_k)^\top (x_k - x^*)],$$

we get

$$\mathbb{E}[\nabla F(x_k, \xi_k)^\top (x_k - x^*)] = \mathbb{E}[\nabla f(x_k)^\top (x_k - x^*)] \geq \sigma \mathbb{E}[\|x_k - x^*\|^2].$$

Proof III

Using σ-strong convexity for f, we can write, for all $x \in \mathbb{R}^n$, the following:

$$(\nabla f(x) - \nabla f(x^*))^\top (x - x^*) \geq \sigma \|x - x^*\|^2,$$

which can be rewritten as

$$\nabla f(x)^\top (x - x^*) \geq \sigma \|x - x^*\|^2 + \nabla f(x^*)^\top (x - x^*).$$

Thus, keeping in mind that $\nabla f(x^*) = 0$, we get, by taking expectation, the following:

$$\mathbb{E}[\nabla f(x)^\top (x - x^*)] \geq \sigma \mathbb{E}[\|x - x^*\|^2].$$

Combining the last one with

$$\mathbb{E}[\nabla F(x_k, \xi_k)^\top (x_k - x^*)] = \mathbb{E}[\nabla f(x_k)^\top (x_k - x^*)],$$

we get

$$\mathbb{E}[\nabla F(x_k, \xi_k)^\top (x_k - x^*)] = \mathbb{E}[\nabla f(x_k)^\top (x_k - x^*)] \geq \sigma \mathbb{E}[\|x_k - x^*\|^2].$$

Proof III

Using σ-strong convexity for f, we can write, for all $x \in \mathbb{R}^n$, the following:

$$(\nabla f(x) - \nabla f(x^*))^\top (x - x^*) \geq \sigma \|x - x^*\|^2,$$

which can be rewritten as

$$\nabla f(x)^\top (x - x^*) \geq \sigma \|x - x^*\|^2 + \nabla f(x^*)^\top (x - x^*).$$

Thus, keeping in mind that $\nabla f(x^*) = 0$, we get, by taking expectation, the following:

$$\mathbb{E}[\nabla f(x)^\top (x - x^*)] \geq \sigma \mathbb{E}[\|x - x^*\|^2].$$

Combining the last one with

$$\mathbb{E}[\nabla F(x_k, \xi_k)^\top (x_k - x^*)] = \mathbb{E}[\nabla f(x_k)^\top (x_k - x^*)],$$

we get

$$\mathbb{E}[\nabla F(x_k, \xi_k)^\top (x_k - x^*)] = \mathbb{E}[\nabla f(x_k)^\top (x_k - x^*)] \geq \sigma \mathbb{E}[\|x_k - x^*\|^2].$$

Proof IV

Now, taking into account last inequality

$$\mathbb{E}[\nabla F(x_k, \xi_k)^\top (x_k - x^*)] \geq \sigma \mathbb{E}[\|x_k - x^*\|^2].$$

and

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \quad \leq \quad \mathbb{E}[\|x_k - x^*\|^2] - 2\alpha_k \mathbb{E}[\nabla F(x_k, \xi_k)^\top (x_k - x^*)] + \alpha_k^2 M^2,$$

we get

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq (1 - 2\alpha_k \sigma) \mathbb{E}[\|x_k - x^*\|^2] + \alpha_k^2 M^2.$$

Keeping in mind that $\alpha_k = \gamma/(k+\delta)$ and $\gamma \geq 1/2\sigma$, we have

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq \left(1 - \frac{2\sigma\gamma}{k+\delta}\right) \mathbb{E}[\|x_k - x^*\|^2] + \frac{\gamma^2 M^2}{(k+\delta)^2}. \tag{7}$$

Now we use induction to prove that

$$\mathbb{E}[\|x_k - x^*\|^2] \leq \frac{C(\gamma)}{k+\delta}. \tag{8}$$

Proof IV

Now, taking into account last inequality

$$\mathbb{E}[\nabla F(x_k, \xi_k)^\top (x_k - x^*)] \geq \sigma \mathbb{E}[\|x_k - x^*\|^2].$$

and

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \quad \leq \quad \mathbb{E}[\|x_k - x^*\|^2] - 2\alpha_k \mathbb{E}[\nabla F(x_k, \xi_k)^\top (x_k - x^*)] + \alpha_k^2 M^2,$$

we get

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq (1 - 2\alpha_k \sigma)\mathbb{E}[\|x_k - x^*\|^2] + \alpha_k^2 M^2.$$

Keeping in mind that $\alpha_k = \gamma/(k+\delta)$ and $\gamma \geq 1/2\sigma$, we have

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq \left(1 - \frac{2\sigma\gamma}{k+\delta}\right) \mathbb{E}[\|x_k - x^*\|^2] + \frac{\gamma^2 M^2}{(k+\delta)^2}. \tag{7}$$

Now we use induction to prove that

$$\mathbb{E}[\|x_k - x^*\|^2] \leq \frac{C(\gamma)}{k+\delta}. \tag{8}$$

Proof IV

Now, taking into account last inequality

$$\mathbb{E}[\nabla F(x_k, \xi_k)^\top (x_k - x^*)] \geq \sigma \mathbb{E}[\|x_k - x^*\|^2].$$

and

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \quad \leq \quad \mathbb{E}[\|x_k - x^*\|^2] - 2\alpha_k \mathbb{E}[\nabla F(x_k, \xi_k)^\top (x_k - x^*)] + \alpha_k^2 M^2,$$

we get

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq (1 - 2\alpha_k \sigma)\mathbb{E}[\|x_k - x^*\|^2] + \alpha_k^2 M^2.$$

Keeping in mind that $\alpha_k = \gamma/(k+\delta)$ and $\gamma \geq 1/2\sigma$, we have

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq \left(1 - \frac{2\sigma\gamma}{k+\delta}\right) \mathbb{E}[\|x_k - x^*\|^2] + \frac{\gamma^2 M^2}{(k+\delta)^2}. \tag{7}$$

Now we use induction to prove that

$$\mathbb{E}[\|x_k - x^*\|^2] \leq \frac{C(\gamma)}{k+\delta}. \tag{8}$$

Proof V

Taking into account expression

$$C(\gamma) = \max\{\gamma^2 M^2 (2\sigma\gamma - 1)^{-1},\ (1+\delta)\|x_1 - x^*\|^2\},$$

it is easy to see that the inequality considered before holds for $k = 1$:

$$\mathbb{E}[\|x_1 - x^*\|^2] = \|x_1 - x^*\|^2 \leq \frac{C(\gamma)}{1+\delta}. \tag{9}$$

Now we assume that inequality holds for some $k \geq 1$. By

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq \left(1 - \frac{2\sigma\gamma}{k+\delta}\right)\mathbb{E}[\|x_k - x^*\|^2] + \frac{\gamma^2 M^2}{(k+\delta)^2}.$$

and calling $\hat{k} = k + \delta$, we have

Proof V

Taking into account expression

$$C(\gamma) = \max\{\gamma^2 M^2 (2\sigma\gamma - 1)^{-1},\ (1+\delta)\|x_1 - x^*\|^2\},$$

it is easy to see that the inequality considered before holds for $k = 1$:

$$\mathbb{E}[\|x_1 - x^*\|^2] = \|x_1 - x^*\|^2 \leq \frac{C(\gamma)}{1+\delta}. \tag{9}$$

Now we assume that inequality holds for some $k \geq 1$. By

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \leq \left(1 - \frac{2\sigma\gamma}{k+\delta}\right) \mathbb{E}[\|x_k - x^*\|^2] + \frac{\gamma^2 M^2}{(k+\delta)^2}.$$

and calling $\hat{k} = k + \delta$, we have

Proof VI

$$\begin{aligned}
\mathbb{E}[\|x_{k+1} - x^*\|^2] &\leq \left(1 - \frac{2\sigma\gamma}{\hat{k}}\right)\frac{C(\gamma)}{\hat{k}} + \frac{\gamma^2 M^2}{\hat{k}^2} \\
&\leq \left(\frac{\hat{k} - 2\sigma\gamma}{\hat{k}^2}\right)C(\gamma) + \frac{\gamma^2 M^2}{\hat{k}^2} \\
&\leq \left(\frac{\hat{k} - 1}{\hat{k}^2}\right)C(\gamma) - \left(\frac{2\sigma\gamma - 1}{\hat{k}^2}\right)C(\gamma) + \frac{\gamma^2 M^2}{\hat{k}^2} \\
&\quad \text{(we use definition of } C(\gamma) \text{ to get } -C(\gamma) \leq -\tfrac{\gamma^2 M^2}{2\sigma\gamma - 1}) \\
&\leq \left(\frac{\hat{k} - 1}{\hat{k}^2}\right)C(\gamma) - \frac{\gamma^2 M^2}{\hat{k}^2} + \frac{\gamma^2 M^2}{\hat{k}^2} \leq \frac{C(\gamma)}{\hat{k} + 1}
\end{aligned}$$

Thus we get the result (last inequality comes from $\hat{k}^2 \geq (\hat{k} - 1)(\hat{k} + 1)$).

Proof VII

Now exploiting Lipschitz continuity of the gradient, and the fact that $\nabla f(x^*) = 0$, we can write:

$$f(x_k) - f(x^*) \leq \nabla f(x^*)^\top (x_k - x^*) + \frac{L}{2}\|x_k - x^*\|^2 \leq \frac{L}{2}\|x_k - x^*\|^2.$$

Taking expectations on both sides of inequality and using (8), we get

$$\mathbb{E}[f(x_k) - f(x^*)] \leq \frac{L}{2}\mathbb{E}[\|x_k - x^*\|^2] \leq \frac{LC(\gamma)}{2(k+\delta)}.$$

Comments

- Here we use Markov inequality to get

$$P(f(x_k) - f(x^*) \geq \epsilon) \leq \frac{\mathbb{E}[f(x_k) - f(x^*)]}{\epsilon}.$$

- It is easy to see that,

$$P(f(x_k) - f(x^*) \geq \epsilon) \leq \frac{\mathbb{E}[f(x_k) - f(x^*)]}{\epsilon} \leq \frac{c}{k\epsilon} \leq \beta.$$

- We need $\mathcal{O}(1/\epsilon\beta)$ iterations to get

$$P(f(x_k) - f(x^*) < \epsilon) \geq 1 - \beta.$$

Comments II

Remark

- In the SG method we need strong convexity to get a sublinear convergence rate of $\mathcal{O}(1/k)$
- In the gradient method only needed Lipschitz continuity of the gradient to get the same rate.

- Stochastic gradient method seems not to be as good as the classic gradient method!
- Why the method has lately re-gained popularity among researchers in data science?

Comments II

Remark

- In the SG method we need strong convexity to get a sublinear convergence rate of $\mathcal{O}(1/k)$
- In the gradient method only needed Lipschitz continuity of the gradient to get the same rate.

- Stochastic gradient method seems not to be as good as the classic gradient method!
- Why the method has lately re-gained popularity among researchers in data science?

Back to Expected Risk Minimization

Expected Risk Minimization problem

Finding the prediction function $h(x; w)$ that minimizes losses from inaccurate predictions.

- Ideally, w minimizes the expected loss for any input-output pair (x, y).
- We assume to know the probability distribution P describing the relationship between input and outputs.
- In practice, we never have that P.
- This is the reason why we try to just estimate the expected risk R.

Back to Expected Risk Minimization

Expected Risk Minimization problem

Finding the prediction function $h(x; w)$ that minimizes losses from inaccurate predictions.

- Ideally, w minimizes the expected loss for any input-output pair (x, y).
- We assume to know the probability distribution P describing the relationship between input and outputs.
- In practice, we never have that P.
- This is the reason why we try to just estimate the expected risk R.

Back to Expected Risk Minimization

Expected Risk Minimization problem

Finding the prediction function $h(x; w)$ that minimizes losses from inaccurate predictions.

- Ideally, w minimizes the expected loss for any input-output pair (x, y).
- We assume to know the probability distribution P describing the relationship between input and outputs.
- In practice, we never have that P.
- This is the reason why we try to just estimate the expected risk R.

Empirical Risk

Supervised Learning and Empirical Risk

- In supervised learning goal is inferring a function from labeled data.
- We hence have the so called *training set*, that is m independently picked input-output samples $(x_i, y_i) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, with $i = 1, \ldots, m$ (describing a real phenomenon we want to somehow represent)
- We have the *empirical risk* function

$$R_m(w) = \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i; w), y_i), \tag{10}$$

where ℓ, as we already said, is a given loss function.

- In practice, we try to minimize R_m, which represents the so-called *misclassification error* over the training set, with respect to w.

Empirical Risk

Supervised Learning and Empirical Risk

- In supervised learning goal is inferring a function from labeled data.
- We hence have the so called *training set*, that is m independently picked input-output samples $(x_i, y_i) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, with $i = 1, \dots, m$ (describing a real phenomenon we want to somehow represent)
- We have the *empirical risk* function

$$R_m(w) = \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i; w), y_i), \tag{10}$$

  where ℓ, as we already said, is a given loss function.
- In practice, we try to minimize R_m, which represents the so-called *misclassification error* over the training set, with respect to w.

How to Simplify Notations

- We now simplify notations:
    - Let us represent a sample (or set of samples) by a random seed ξ (e.g., just imagine a realization of ξ as a single sample (x, y) or a set of p samples (x_i, y_i), with $i = 1, \ldots, p$).
    - Let us indicate with x the parameters representing the model.
    - let us refer to the loss incurred for a given ξ as $F(x, \xi)$.
- We have that expected risk $R(x) = \mathbb{E}[F(x, \xi)]$.
- When given a set of realizations $\{\xi_1, \ldots, \xi_m\}$, corresponding to a sample set $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ we define the loss incurred by the parameter vector x with respect to the i-th sample as $f_i(x) = F(x, \xi_i)$.
- Empirical risk minimization problem takes the form

$$\min_{x \in \mathbb{R}^n} \quad \frac{1}{m} \sum_{i=1}^{m} f_i(x). \tag{11}$$

How to Simplify Notations

- We now simplify notations:
  - Let us represent a sample (or set of samples) by a random seed ξ (e.g., just imagine a realization of ξ as a single sample (x, y) or a set of p samples (x_i, y_i), with $i = 1, \ldots, p$).
  - Let us indicate with x the parameters representing the model.
  - let us refer to the loss incurred for a given ξ as $F(x, \xi)$.
- We have that expected risk $R(x) = \mathbb{E}[F(x, \xi)]$.
- When given a set of realizations $\{\xi_1, \ldots, \xi_m\}$, corresponding to a sample set $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ we define the loss incurred by the parameter vector x with respect to the i-th sample as $f_i(x) = F(x, \xi_i)$.
- Empirical risk minimization problem takes the form

$$\min_{x \in \mathbb{R}^n} \quad \frac{1}{m} \sum_{i=1}^{m} f_i(x). \tag{11}$$

How to Simplify Notations

- We now simplify notations:
  - Let us represent a sample (or set of samples) by a random seed ξ (e.g., just imagine a realization of ξ as a single sample (x, y) or a set of p samples (x_i, y_i), with $i = 1, \ldots, p$).
  - Let us indicate with x the parameters representing the model.
  - let us refer to the loss incurred for a given ξ as $F(x, \xi)$.
- We have that expected risk $R(x) = \mathbb{E}[F(x, \xi)]$.
- When given a set of realizations $\{\xi_1, \ldots, \xi_m\}$, corresponding to a sample set $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ we define the loss incurred by the parameter vector x with respect to the i-th sample as $f_i(x) = F(x, \xi_i)$.
- Empirical risk minimization problem takes the form

$$\min_{x \in \mathbb{R}^n} \quad \frac{1}{m} \sum_{i=1}^{m} f_i(x). \tag{11}$$

Empirical Risk Problem and Sample Average Approximation

Empirical Risk Problem

Empirical risk minimization problem takes the form

$$\min_{x\in\mathbb{R}^n} \quad \frac{1}{m}\sum_{i=1}^{m} f_i(x). \tag{12}$$

- It is directly connected to Sample Average Approximation!

Empirical Risk Problem and Sample Average Approximation

Empirical Risk Problem

Empirical risk minimization problem takes the form

$$\min_{x \in \mathbb{R}^n} \quad \frac{1}{m} \sum_{i=1}^{m} f_i(x). \tag{12}$$

- It is directly connected to Sample Average Approximation!

Optimization Methods for Minimizing Risk - Part I

Batch/Deterministic Approaches

- The gradient method belongs to this class.
- Its iteration becomes

$$x_{k+1} = x_k - \alpha_k \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(x_k).$$

- Iteration is not cheap...cost depends on m!
- Convergence Rate: $\mathcal{O}\left(\frac{\eta-1}{\eta+1}\right)^{2k}$, with $\eta = L/\sigma$.
- Cost per iteration: $\mathcal{O}(m)$ (m gradient calculations here).
- Overall complexity: $\mathcal{O}(m \log(1/\epsilon))$.

Optimization Methods for Minimizing Risk - Part II

Stochastic Approaches

- The stochastic class obviously include the SG method.
- SG iteration
$$x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k),$$
where i_k is a random number related to the sample (x_{i_k}, y_{i_k}).
- Iteration is very cheap: involves only the computation of the gradient related to sample i_k!
- Convergence Rate: $\mathcal{O}(1/k)$.
- Cost per iteration: $\mathcal{O}(1)$ (gradient calculation is the unit here).
- Overall complexity: $\mathcal{O}(1/\epsilon\beta)$.

Why do data scientists need stochastic gradient?

Example

- Training set S consisting of 100 copies of a set S'.
- Minimizing the empirical risk over S is basically the same as minimizing it over S'.
- Batch approach: iteration 100 times more expensive than if one only had S'.
- SG performs the same computations in both scenarios (S and S').
- In many huge-scale applications the data does involve a good number of (approximate) redundant samples.

Why do data scientists need stochastic gradient?

	Stochastic	Batch
Convergence Rate	$\mathcal{O}(1/k)$	$\mathcal{O}\left(\frac{\eta-1}{\eta+1}\right)^{2k}$
Cost per iteration	$\mathcal{O}(1)$	$\mathcal{O}(m)$
Overall complexity	$\mathcal{O}(1/\epsilon\beta)$	$\mathcal{O}(m\log(1/\epsilon))$

- SG uses information in a more efficient way than a batch method!
- The overall complexity of SG can be larger than the one of classic gradient for moderate values of m.
- Comparison favors SG when one moves to the big data regime where m is large and one is constrained by a computational time budget.

Why do data scientists need stochastic gradient?

	Stochastic	Batch
Convergence Rate	$\mathcal{O}(1/k)$	$\mathcal{O}\left(\frac{\eta-1}{\eta+1}\right)^{2k}$
Cost per iteration	$\mathcal{O}(1)$	$\mathcal{O}(m)$
Overall complexity	$\mathcal{O}(1/\epsilon\beta)$	$\mathcal{O}(m\log(1/\epsilon))$

- SG uses information in a more efficient way than a batch method!
- The overall complexity of SG can be larger than the one of classic gradient for moderate values of m.
- Comparison favors SG when one moves to the big data regime where m is large and one is constrained by a computational time budget.

Why do data scientists need stochastic gradient?

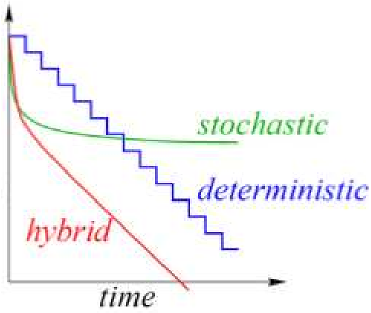

Figure: Comparison between stochastic, deterministic and hybrid gradient method.

Why do data scientists need stochastic gradient?

- Batch method has big cost per iteration in a huge-scale framework (notice the stair-step behavior of the picture).
- Stochastic method has a small cost per iteration.
- Even if the deterministic gradient method guarantees a linear reduction, stochastic gradient method reduces faster than the gradient!!!
- The best would be developing algorithms with a linear convergence rate and cheap iteration cost (hybrid methods).