Stochastic gradient descent proof

Shuvomoy Das Gupta

December 25, 2022

We study a simple proof of stochastic gradient descent.

Algorithm description

Problem setup. We are interested in solving the problem

$$p^\star = \left(\begin{array}{ll} \underset{x \in \mathbb{R}^d}{\text{minimize}} & f(x) \\ \text{subject to} & x \in C, \end{array} \right) \qquad (\mathcal{P})$$

where we have the following assumptions regarding the nature of the problem.

Notation
Π_C: projection onto the set C

Assumption 1 (Structure of $\mathcal{P}$)**.** *We assume:*

- $f : \mathbb{R}^d \to (-\infty, \infty]$ *is a closed, proper, and convex function,*
- C *is a nonempty, closed, convex set, with* $C \subseteq \operatorname{int}\operatorname{dom} f$*, and*
- *argmin*$f(C) = X^\star \neq \emptyset$.

Stochastic gradient descent. The stochastic gradient descent (SGD) algorithm to solve ($\mathcal{P}$) is described by Algorithm 1, where we make the following assumption regarding the nature of the oracle.

Algorithm 1 SGD to solve ($\mathcal{P}$)

input: f, C, iteration limit K

algorithm:
1. initialization:
pick $x_0 \in C$ arbitrarily
2. main iteration:
for $k = 0, 1, 2, \ldots, K-1$
  pick stepsizes $\alpha_k > 0$ and random $g_k \in \mathbb{R}^d$ satisfying Assumption 2
  $x_{k+1} \leftarrow \Pi_C\left(x_k - \alpha_k g_k\right)$
end for
3. return x_K

Assumption 2 (Stochastic oracle)**.** *We assume that given an iterate* x_k*, the stochastic oracle is capable of producing a random vector* g_k*with the following properties:*

- *(unbiased)* $\forall_{k \geq 0}\ \mathbf{E}\left[g_k \mid x_k\right] \in \partial f(x_k)$*, and*
- *(bounded variance)* $\exists_{G>0}\ \forall_{k \geq 0}\ \mathbf{E}\left[\|g_k\|^2 \mid x_k\right] \leq G^2$.

Convergence analysis. First, note that, for all $k \geq 0$:

$$
\begin{aligned}
&\mathbf{E}\left[\| \overbrace{x_{k+1}}^{=\Pi_C(x_k-\alpha_k g_k)} - \overbrace{x_\star}^{=\Pi_C(x_\star)} \|^2 \mid x_k\right] \\
=&\mathbf{E}\left[\overbrace{\|\Pi_C\left(x_k-\alpha_k g_k\right)-\Pi_C(x_\star)\|^2}^{\leq \|x_k-\alpha_k g_k - x_\star\|^2} \mid x_k\right] \\
\leq&\mathbf{E}\left[\overbrace{\|x_k-\alpha_k g_k-x_\star\|^2}^{=\|x_k-x_\star\|^2+\alpha_k^2\|g_k\|^2-2\alpha_k\langle x_k-x_\star; g_k\rangle} \mid x_k\right] \\
=&\mathbf{E}\left[\|x_k-x_\star\|^2+\alpha_k^2\|g_k\|^2-2\alpha_k\left\langle x_k-x_\star; g_k\right\rangle \mid x_k\right] \\
=&\mathbf{E}\left[\|x_k-x_\star\|^2 \mid x_k\right]+\alpha_k^2\mathbf{E}\left[\|g_k\|^2 \mid x_k\right]-2\alpha_k\mathbf{E}\left[\left\langle x_k-x_\star; g_k\right\rangle \mid x_k\right] && \triangleright \text{ using linearity of expectation} \\
=&\|x_k-x_\star\|^2+\alpha_k^2\mathbf{E}\left[\|g_k\|^2 \mid x_k\right]-2\alpha_k\left\langle x_k-x_\star; \mathbf{E}\left[g_k \mid x_k\right]\right\rangle && \triangleright \text{ using "taking out what's known" rule } \mathbf{E}\left[h(X)Y \mid X\right]=h(X)\mathbf{E}\left[Y \mid X\right] \\
\leq&\|x_k-x_\star\|^2+\alpha_k^2G^2-2\alpha_k\left\langle x_k-x_\star; \mathbf{E}\left[g_k \mid x_k\right]\right\rangle \\
&\quad /* \\
&\quad \text{we have } \mathbf{E}\left[\|g_k\|^2 \mid x_k\right] \leq G^2 \\
&\quad \Leftrightarrow \forall_y\ f(y) \geq f(x_k)+\left\langle \mathbf{E}\left[g_k \mid x_k\right]; y-x_k\right\rangle \\
&\quad \overset{y\leftarrow x_\star}{\Rightarrow}\ f(x_\star) \geq f(x_k)-\left\langle \mathbf{E}\left[g_k \mid x_k\right]; x_k-x_\star\right\rangle \\
&\quad \Rightarrow -\left\langle \mathbf{E}\left[g_k \mid x_k\right]; x_k-x_\star\right\rangle \leq f(x_\star)-f(x_k) \\
&\quad */ \\
=&\|x_k-x_\star\|^2+\alpha_k^2G^2-2\alpha_k\left(f(x_k)-f(x_\star)\right),
\end{aligned}
$$

So, we have proved

$$
\mathbf{E}\left[\|x_{k+1}-x_\star\|^2 \mid x_k\right] \leq \|x_k-x_\star\|^2+\alpha_k^2G^2-2\alpha_k\left(f(x_k)-f(x_\star)\right),
$$

so taking expectation with respect to x_k on both sides, we get:

$$
\begin{aligned}
&\mathbf{E}\left[\mathbf{E}\left[\|x_{k+1}-x_\star\|^2 \mid x_k\right]\right] \\
=&\mathbf{E}\left[\|x_{k+1}-x_\star\|^2\right] && \triangleright \text{ using Adam's law} \mathbf{E}\left[\mathbf{E}\left[Y \mid X\right]\right]=\mathbf{E}\left[Y\right] \\
\leq&\mathbf{E}\left[\|x_k-x_\star\|^2+\alpha_k^2G^2-2\alpha_k\left(f(x_k)-f(x_\star)\right)\right] \\
=&\mathbf{E}\left[\|x_k-x_\star\|^2\right]-2\alpha_k\mathbf{E}\left[f(x_k)-f(x_\star)\right]+\alpha_k^2G^2,
\end{aligned}
$$

so

$$
\mathbf{E}\left[\|x_{k+1}-x_\star\|^2\right]-\mathbf{E}\left[\|x_k-x_\star\|^2\right] \leq -2\alpha_k\mathbf{E}\left[f(x_k)-f(x_\star)\right]+\alpha_k^2G^2.
$$

Now, let us do a telescoping sum:

$$\begin{aligned}
\mathbf{E}\left[\|x_{k+1}-x_\star\|^2\right]-\mathbf{E}\left[\|x_k-x_\star\|^2\right] &\leq -2\alpha_k\mathbf{E}\left[f(x_k)-f(x_\star)\right]+\alpha_k^2G^2\\
\mathbf{E}\left[\|x_k-x_\star\|^2\right]-\mathbf{E}\left[\|x_{k-1}-x_\star\|^2\right] &\leq -2\alpha_k\mathbf{E}\left[f(x_{k-1})-f(x_\star)\right]+\alpha_{k-1}^2G^2\\
&\vdots\\
\mathbf{E}\left[\|x_{m+1}-x_\star\|^2\right]-\mathbf{E}\left[\|x_m-x_\star\|^2\right] &\leq -2\alpha_m\mathbf{E}\left[f(x_m)-f(x_\star)\right]+\alpha_m^2G^2,
\end{aligned}$$

and adding the equations above, we get:

$$\begin{aligned}
&\mathbf{E}\left[\|x_{k+1}-x_\star\|^2\right]-\mathbf{E}\left[\|x_m-x_\star\|^2\right] \leq -2\sum_{i=m}^{k}\alpha_i\mathbf{E}\left[f(x_i)-f(x_\star)\right]+G^2\sum_{i=m}^{k}\alpha_i^2\\
\Leftrightarrow & 0\leq\mathbf{E}\left[\|x_{k+1}-x_\star\|^2\right]\leq\mathbf{E}\left[\|x_m-x_\star\|^2\right]-2\sum_{i=m}^{k}\alpha_i\mathbf{E}\left[f(x_i)-f(x_\star)\right]+G^2\sum_{i=m}^{k}\alpha_i^2\\
\Rightarrow & 0\leq\mathbf{E}\left[\|x_m-x_\star\|^2\right]-2\sum_{i=m}^{k}\alpha_i\mathbf{E}\left[f(x_i)-f(x_\star)\right]+G^2\sum_{i=1}^{m}\alpha_i^2\\
\Leftrightarrow & \sum_{i=m}^{k}\alpha_i\mathbf{E}\left[f(x_i)-f(x_\star)\right]\leq\frac{1}{2}\left(\mathbf{E}\left[\|x_m-x_\star\|^2\right]+G^2\sum_{i=m}^{k}\alpha_i^2\right)\\
\Rightarrow & \left(\sum_{i=m}^{k}\alpha_i\right)\left(\min_{i\in\{m,\ldots,k\}}\mathbf{E}\left[f(x_i)-f(x_\star)\right]\right)\\
&\quad\leq\frac{1}{2}\left(\mathbf{E}\left[\|x_m-x_\star\|^2\right]+G^2\sum_{i=m}^{k}\alpha_i^2\right) && \triangleright \text{ for } b_k\geq 0\text{, we have } \left(\min_k a_k\right)\textstyle\sum_k b_k\leq\sum_k a_kb_k\\
\Rightarrow & \mathbf{E}\left[\min_{i\in\{m,\ldots,k\}}\left\{f(x_i)-f(x_\star)\right\}\right]\leq\min_{i\in\{m,\ldots,k\}}\mathbf{E}\left[f(x_i)-f(x_\star)\right]\\
&\quad\leq\frac{\mathbf{E}\left[\|x_m-x_\star\|^2\right]+G^2\sum_{i=m}^{k}\alpha_i^2}{2\sum_{i=m}^{k}\alpha_i}. && \triangleright \text{ using } \mathbf{E}\left[\min_i X_i\right]\leq\min_i\mathbf{E}\left[X_i\right]
\end{aligned}$$

In the last inequality, m is arbitrary, so set $m\leftarrow 0$, which leads to:

$$\mathbf{E}\left[\min_{i\in\{0,\ldots,k\}}f(x_i)\right]-f(x_\star)\leq\frac{\mathbf{E}\left[\|x_0-x_\star\|^2\right]+G^2\sum_{i=0}^{k}\alpha_i^2}{2\sum_{i=0}^{k}\alpha_i},$$

so if we have $\sum_{i=0}^{k}\alpha_i^2<\infty$ and $\sum_{i=0}^{k}\alpha_i=\infty$, then we have

$$\mathbf{E}\left[\min_{i\in\{0,\ldots,k\}}f(x_i)\right]\to f(x_\star).$$