12 Stochastic Gradient

Goal

Noisy (sub)gradient, stochastic gradient-descent-ascent, convergence in mean, convergence in high probability, effect of averaging

Alert 12.1: Convention

A nice survey for this lecture and the next (and much more beyond) is Bottou et al. (2018).

Gray boxes are not required hence can be omitted for unenthusiastic readers.

This note is likely to be updated again soon.

Bottou, L., F. E. Curtis, and J. Nocedal (2018). "Optimization Methods for Large-Scale Machine Learning". *SIAM Review*, vol. 60, no. 2, pp. 223–311.

Definition 12.2: Problem

We continue our discussion of the minimax problem

$$\min_{\mathbf{x}\in\mathsf{X}\subseteq\mathbb{R}^p}\ \max_{\mathbf{y}\in\mathsf{Y}\subseteq\mathbb{R}^d}\ f(\mathbf{x},\mathbf{y}),$$

but we are now restricted to a noisy (sub)gradient map $\hat{\mathsf{T}}\mathbf{w}$, such that for any $\mathbf{w} = (\mathbf{x}, \mathbf{y})$,

$$\hat{\mathsf{T}}\mathbf{w} = \mathsf{T}\mathbf{w} + \boldsymbol{\varepsilon}(\mathbf{w}), \quad \text{where} \quad \mathsf{T}\mathbf{w} := (\partial_{\mathbf{x}} f(\mathbf{x},\mathbf{y}), \partial_{\mathbf{y}}-f(\mathbf{x},\mathbf{y})) \tag{12.1}$$

and $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}(\mathbf{w})$ represents some random noise (e.g. numerical error) at $\mathbf{w}$. For simplicity we assume

$$\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}(\mathbf{w}) \overset{i.i.d.}{\sim} \mathsf{p}(\cdot), \qquad \text{and} \qquad \mathbb{E}\boldsymbol{\varepsilon} \equiv \mathbf{0}, \ \ \mathbb{E}\|\boldsymbol{\varepsilon}\|_2^2 =: \varsigma^2 < \infty,$$

i.e., the noisy (sub)gradient is unbiased and has bounded variance. **??** essentially dealt with the "trivial" case where $\varsigma = 0$.

We note that our general formulation here also includes the minimization problem:

$$\min_{\mathbf{x}\in\mathsf{X}}\ g(\mathbf{x}).$$

Indeed, we need only set $\mathsf{Y} = \{\mathbf{y}_0\}$ as a singleton and let $f(\mathbf{x}, \mathbf{y}_0) = g(\mathbf{x})$. Therefore, everything we discuss below immediately applies to stochastically minimizing a function.

Example 12.3: Where does the noise come from?

The noise in (12.1) may come from a variety of sources:

- measurement error: most physical devices are accurate up to a certain noise level;
- numerical error: any computation we perform on a finite precision computer is subject to truncation, cancellation, underflow and overflow;
- problem scale: in ML we typically minimize the empirical loss

$$f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\ell_i(\mathbf{w}), \quad \partial f(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\partial\ell_i(\mathbf{w}), \tag{12.2}$$

where n is on the scale of millions or billions (or even larger). It is thus unrealistic to compute the average in the (sub)gradient above. Instead, we randomly sample a minibatch $I = \{i_1, \ldots, i_m\}$ of size

$m \ll n$ and compute

$$\hat{\partial} f(\mathbf{w}) = \frac{1}{m}\sum_{k=1}^{m} \partial \ell_{i_k}(\mathbf{w}) = \partial f(\mathbf{w}) + [\overbrace{\hat{\partial} f(\mathbf{w}) - \partial f(\mathbf{w})}^{\boldsymbol{\varepsilon}(\mathbf{w})}].$$

- convenience: sometimes our objective function f can be written as an expectation (either by nature or by reformulation):

  $$f(\mathbf{w}) = \mathbb{E}F(\mathbf{w}, \boldsymbol{\zeta}), \qquad \text{hence under mild regularity} \qquad \partial f(\mathbf{w}) = \mathbb{E}\partial F(\mathbf{w}, \boldsymbol{\zeta}),$$

  where computing the expectation exactly is either very costly or impossible. Upon taking a few samples $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_m \overset{i.i.d.}{\sim} \boldsymbol{\zeta}$ we may use the empirical average as an approximation:

  $$\hat{\partial} f(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m} \partial F(\mathbf{w}, \boldsymbol{\zeta}_i) = \partial f(\mathbf{w}) + [\overbrace{\hat{\partial} f(\mathbf{w}) - \partial f(\mathbf{w})}^{\boldsymbol{\varepsilon}(\mathbf{w})}].$$

  If we only draw the samples once, we arrive at essentially the empirical risk in (12.2).

- privacy: when the functions ℓ_i are user-specific while computation is conducted on a cloud, it is important to not leak user-specific information through communicating gradients. A typical approach is to corrupt the gradient with random noise (e.g. Chaudhuri et al., 2011; Bassily et al., 2014; Wang et al., 2017, 2019) before submitting to a server where aggregation and parameter update are taken place.

- regularization: adding noise to the input or output is a standard training trick in ML and can be connected to regularization (e.g. Bishop, 1995). Adding noise in the gradient likely has similar regularizing effects.

Chaudhuri, K., C. Monteleoni, and A. D. Sarwate (2011). "Differentially Private Empirical Risk Minimization". *Journal of Machine Learning Research*, vol. 12, no. 29, pp. 1069–1109.

Bassily, R., A. Smith, and A. Thakurta (2014). "Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds". In: *Proceedings of the IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 464–473.

Wang, D., M. Ye, and J. Xu (2017). "Differentially Private Empirical Risk Minimization Revisited: Faster and More General". In: *Advances in Neural Information Processing Systems*, pp. 2722–2731.

Wang, D., C. Chen, and J. Xu (2019). "Differentially Private Empirical Risk Minimization with Non-convex Loss Functions". In: *Proceedings of the 36th International Conference on Machine Learning*, pp. 6526–6535.

Bishop, C. M. (1995). "Training with Noise is Equivalent to Tikhonov Regularization". *Neural Computation*, vol. 7, no. 1, pp. 108–116.

Remark 12.4: Sampling vs. optimization

Ma et al., 2019

Ma, Y.-A., Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan (2019). "Sampling can be faster than optimization". *Proceedings of the National Academy of Sciences*, vol. 116, no. 42, pp. 20881–20885.

Lemma 12.5: Inexact forward-backward splitting

Let $C \subseteq \mathbb{R}^d$ be closed convex, $\mathbf{w}_0 \in C$ and for all $t \geq 0$ define

$$\mathbf{w}_{t+1} := \mathrm{P}_C(\mathbf{w}_t - \eta_t \hat{\mathbf{w}}_t^*), \quad \textit{where} \quad \hat{\mathbf{w}}_t^* = \mathbf{w}_t^* + \boldsymbol{\varepsilon}_t,\ \eta_t \geq 0,\ b_{t,k} := \gamma_k \eta_k / \Gamma_t,\ \Gamma_t := \sum_{k=0}^{t} \gamma_k \eta_k,\ \bar{\boldsymbol{\varepsilon}}_t := \sum_{k=0}^{t} b_{t,k} \boldsymbol{\varepsilon}_k.$$

The following estimate holds for any $\mathbf{w} \in C$:

$$\sum_{k=0}^{t} \langle \mathbf{w}_k - \mathbf{w}, b_{t,k}\mathbf{w}_k^* \rangle \le \frac{\gamma_0\|\mathbf{w}_0-\mathbf{w}\|_2^2 + \sum_{k=1}^{t}(\gamma_k-\gamma_{k-1})\|\mathbf{w}_k-\mathbf{w}\|_2^2 + \sum_{k=0}^{t}\gamma_k\|\eta_k\hat{\mathbf{w}}_k^*\|_2^2}{2\Gamma_t} + \langle \mathbf{w}, \bar{\boldsymbol{\varepsilon}}_t\rangle - \sum_{k=0}^{t}\langle \mathbf{w}_k, b_{t,k}\boldsymbol{\varepsilon}_k\rangle$$

Proof: For any $\mathbf{w} \in K \subseteq C$:

$$\begin{aligned}\|\mathbf{w}_{k+1} - \mathbf{w}\|_2^2 = \|\mathrm{P}_C(\mathbf{w}_k - \eta_k\hat{\mathbf{w}}_k^*) - \mathrm{P}_C(\mathbf{w})\|_2^2 &\le \|\mathbf{w}_k - \mathbf{w} - \eta_k\hat{\mathbf{w}}_k^*\|_2^2 \\ &= \|\mathbf{w}_k - \mathbf{w}\|_2^2 + \|\eta_k\hat{\mathbf{w}}_k^*\|_2^2 - 2\langle \mathbf{w}_k - \mathbf{w}, \eta_k\hat{\mathbf{w}}_k^*\rangle \\ &= \|\mathbf{w}_k - \mathbf{w}\|_2^2 + \|\eta_k\hat{\mathbf{w}}_k^*\|_2^2 - 2\langle \mathbf{w}_k - \mathbf{w}, \eta_k\mathbf{w}_k^* + \eta_k\boldsymbol{\varepsilon}_k\rangle .\end{aligned}$$

Multiplying both sides by γ_k/Γ_t, rearranging and summing from $k = 0$ to $k = t$:

$$\sum_{k=0}^{t}\langle \mathbf{w}_k - \mathbf{w}, b_{t,k}\mathbf{w}_k^*\rangle \le \frac{\sum_{k=0}^{t}\gamma_k[\|\mathbf{w}_k - \mathbf{w}\|_2^2 - \|\mathbf{w}_{k+1} - \mathbf{w}\|_2^2 + \|\eta_k\hat{\mathbf{w}}_k^*\|_2^2]}{2\Gamma_t} - \sum_{k=0}^{t}\langle \mathbf{w}_k - \mathbf{w}, b_{t,k}\boldsymbol{\varepsilon}_k\rangle .$$

Simplifying and rearranging completes the proof. ■

Theorem 12.6: Stochastic gradient-descent-ascent (Nemirovskii and Judin, 1978)

Let $f : \mathsf{X}\times\mathsf{Y} \to \mathbb{R}$ be (closed) convex in $\mathbf{x}$ and (closed) concave in $\mathbf{y}$, where $\mathsf{X} \subseteq \mathbb{R}^p$ and $\mathsf{Y} \subseteq \mathbb{R}^d$ are closed convex. Starting from $\mathbf{w}_0 := (\mathbf{x}_0, \mathbf{y}_0) \in \mathsf{X}\times\mathsf{Y}$, define

$$\begin{cases}\mathbf{x}_{t+1} := \mathrm{P}_{\mathsf{X}}(\mathbf{x}_t - \eta_t\hat{\mathbf{x}}_t^*), \\ \mathbf{y}_{t+1} := \mathrm{P}_{\mathsf{Y}}(\mathbf{y}_t - \eta_t\hat{\mathbf{y}}_t^*),\end{cases} \quad \text{where} \quad \begin{cases}\hat{\mathbf{x}}_t^* = \mathbf{x}_t^* + \boldsymbol{\xi}_t, & \mathbf{x}_t^* \in \partial_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \quad \mathbb{E}[\boldsymbol{\xi}_t|\mathbf{x}_t,\mathbf{y}_t] = \mathbf{0} \\ \hat{\mathbf{y}}_t^* = \mathbf{y}_t^* + \boldsymbol{\zeta}_t, & \mathbf{y}_t^* \in \partial_{\mathbf{y}} \text{-} f(\mathbf{x}_t, \mathbf{y}_t), \quad \mathbb{E}[\boldsymbol{\zeta}_t|\mathbf{x}_t,\mathbf{y}_t] = \mathbf{0}\end{cases}$$

$$(\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t, \bar{\boldsymbol{\xi}}_t, \bar{\boldsymbol{\zeta}}_t) := \sum_{k=0}^{t} b_{t,k}(\mathbf{x}_k, \mathbf{y}_k, \boldsymbol{\xi}_k, \boldsymbol{\zeta}_k), \quad \text{where} \quad b_{t,k} := \gamma_k\eta_k/\Gamma_t, \quad \eta_t, \gamma_t \ge 0, \quad \Gamma_t := \sum_{k=0}^{t}\gamma_k\eta_k.$$

Let $\mathbf{w} = (\mathbf{x}, \mathbf{y})$ (along with similar quantities), the following estimates hold for any compact set $K \subseteq \mathsf{X}\times\mathsf{Y}$:

$$\mathbb{E}[\overline{f}_K(\bar{\mathbf{x}}_t) - \underline{f}_K(\bar{\mathbf{y}}_t)] \le \frac{\mathbb{E}\Big[\max_{\mathbf{w}\in K}\sum_{k=0}^{t}(\gamma_k - \gamma_{k-1})\|\mathbf{w}_k - \mathbf{w}\|_2^2 + \sum_{k=0}^{t}\gamma_k\|\eta_k\hat{\mathbf{w}}_t^*\|_2^2\Big]}{2\Gamma_t} + \mathbb{E}\sigma_K(\bar{\boldsymbol{\varepsilon}}_t) \tag{12.3}$$

$$\le \frac{\mathbb{E}\Big[\max_{\mathbf{w}\in K}\sum_{k=0}^{t}(\gamma_k - \gamma_{k-1})\|\mathbf{w}_k - \mathbf{w}\|_2^2 + \varrho^2(K) + \sum_{k=0}^{t}\gamma_k\eta_k^2[\|\mathbf{w}_k^*\|_2^2 + (1+\gamma_k)\varsigma_k^2]\Big]}{2\Gamma_t}, \tag{12.4}$$

where $\overline{f}_K(\mathbf{x}) := \max_{(\mathbf{x},\mathbf{y})\in K} f(\mathbf{x}, \mathbf{y})$ and similarly $\underline{f}_K(\mathbf{y}) := \min_{(\mathbf{x},\mathbf{y})\in K} f(\mathbf{x}, \mathbf{y})$, $\varrho(K) := \min_{\mathbf{z}}\max_{\mathbf{w}\in K}\|\mathbf{z} - \mathbf{w}\|_2$ is the radius of K, $\gamma_{-1} := 0$, and $\varsigma_k^2 := \mathbb{E}[\|\boldsymbol{\xi}_k\|_2^2 + \|\boldsymbol{\zeta}_k\|_2^2]$ is the variance.

Proof: We set $C = \mathsf{X}\times\mathsf{Y}$, $\mathbf{w} = (\mathbf{x}, \mathbf{y})$ and $\boldsymbol{\varepsilon} = (\boldsymbol{\xi}, \boldsymbol{\zeta})$ and apply Lemma 12.5:

$$\begin{aligned}\max_{\mathbf{w}\in K}\sum_{k=0}^{t} b_{t,k}\langle \mathbf{w}_k - \mathbf{w}, \mathbf{w}_k^*\rangle &= \max_{\mathbf{w}\in K}\sum_{k=0}^{t} b_{t,k}[\langle \mathbf{x}_k - \mathbf{x}, \partial_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k)\rangle - f(\mathbf{x}_k, \mathbf{y}_k) + f(\mathbf{x}_k, \mathbf{y}_k) + \langle \mathbf{y}_k - \mathbf{y}, \partial_{\mathbf{y}} \text{-} f(\mathbf{x}_k, \mathbf{y}_k)\rangle] \\ &\ge \max_{\mathbf{w}\in K}\sum_{k=0}^{t} b_{t,k}[f(\mathbf{x}_k, \mathbf{y}) - f(\mathbf{x}, \mathbf{y}_k)] \ge \max_{\mathbf{w}\in K}[f(\bar{\mathbf{x}}_t, \mathbf{y}) - f(\mathbf{x}, \bar{\mathbf{y}}_t)] = \overline{f}_K(\bar{\mathbf{x}}_t) - \underline{f}_K(\bar{\mathbf{y}}_t).\end{aligned}$$

To obtain the first claim (12.3), we need only observe that,

$$\mathbb{E}\sum_{k=0}^{t}\langle \mathbf{w}_k, b_{t,k}\boldsymbol{\varepsilon}_k\rangle = \sum_{k=0}^{t}\mathbb{E}\langle \mathbf{w}_k, b_{t,k}\boldsymbol{\varepsilon}_k\rangle = \sum_{k=0}^{t}\mathbb{E}\big[\langle \mathbf{w}_k, b_{t,k}\mathbb{E}(\boldsymbol{\varepsilon}_k|\mathbf{w}_k)\rangle\big] = 0,$$

thanks to the unbiased assumption on the noise $\boldsymbol{\varepsilon}$ and the law of total expectation.

To obtain the second claim (12.4), we verify that

$$\mathbb{E}\|\hat{\mathbf{w}}_k^*\|_2^2 = \mathbb{E}\|\mathbf{w}_k^* + \boldsymbol{\varepsilon}_k\|_2^2 = \mathbb{E}\|\mathbf{w}_k^*\|_2^2 + \mathbb{E}\|\boldsymbol{\varepsilon}_k\|_2^2 + 2\mathbb{E}\langle \mathbf{w}_k^*, \boldsymbol{\varepsilon}_k\rangle = \mathbb{E}\|\mathbf{w}_k^*\|_2^2 + \varsigma_k^2,$$

using again unbiasedness and law of total expectation. Moreover, fix any $\mathbf{z}$:

$$\mathbb{E}\sigma_K(\bar{\boldsymbol{\varepsilon}}_t) = \mathbb{E}[\max_{\mathbf{w}\in K}\langle \mathbf{w}-\mathbf{z}, \bar{\boldsymbol{\varepsilon}}_t\rangle + \langle \mathbf{z}, \bar{\boldsymbol{\varepsilon}}_t\rangle] \le \max_{\mathbf{w}\in K}\|\mathbf{w}-\mathbf{z}\|_2 \cdot \mathbb{E}\|\bar{\boldsymbol{\varepsilon}}_t\|_2 + \mathbb{E}\langle \mathbf{z}, \bar{\boldsymbol{\varepsilon}}_t\rangle,$$

where the last term vanishes due to unbiasedness and the middle term can be further bounded as:

$$(\mathbb{E}\|\bar{\boldsymbol{\varepsilon}}_t\|_2)^2 \le \mathbb{E}\|\bar{\boldsymbol{\varepsilon}}_t\|_2^2 = \sum_{k=0}^{t} b_{t,k}^2\mathbb{E}\|\boldsymbol{\varepsilon}_k\|_2^2 + 2\sum_{i}\sum_{j>i} b_{t,i}b_{t,j}\underbrace{\mathbb{E}\langle \boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j\rangle}_{=0} = \sum_{k=0}^{t}\gamma_k^2\eta_k^2\varsigma_k^2/\Gamma_t^2.$$

Since $\mathbf{z}$ was arbitrary, we obtain

$$\mathbb{E}\sigma_K(\bar{\boldsymbol{\varepsilon}}_t) \le \varrho(K)\cdot\frac{\sqrt{\sum_{k=0}^{t}\gamma_k^2\eta_k^2\varsigma_k^2}}{\Gamma_t} \le \frac{\varrho^2(K) + \sum_{k=0}^{t}\gamma_k^2\eta_k^2\varsigma_k^2}{2\Gamma_t}, \tag{12.5}$$

which leads immediately to the bound (12.4). ■

The bound (12.4), with slightly worse constants, appeared first in Nemirovskii and Judin (1978) and later with more details in Nemirovski et al. (2009).

Nemirovskii, A. S. and D. B. Judin (1978). "Cesari convergence of the gradient method of approximating saddle points of convex-concave functions". *Soviet Mathematics Doklady*, vol. 19, no. 2, pp. 482–486.

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro (2009). "Robust Stochastic Approximation Approach to Stochastic Programming". *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1574–1609.

Remark 12.7: Parsing the previous result

Let us first assume X and Y are bounded and f is L-Lipschitz continuous (jointly in $\mathbf{x}$ and $\mathbf{y}$). We can thus take $K = C := \mathsf{X}\times\mathsf{Y}$ and we know a saddle point exists. So, the bound (12.4) simplifies to:

$$\mathbb{E}[\underbrace{\overline{f}(\bar{\mathbf{x}}_t) - \mathfrak{p}_\star}_{\text{primal gap}} + \underbrace{\mathfrak{d}^\star - \underline{f}(\bar{\mathbf{y}}_t)}_{\text{dual gap}}] \le \frac{\mathbb{E}\Big[\sum_{k=0}^{t}(\gamma_k-\gamma_{k-1})_+\operatorname{diam}^2(C) + \varrho^2(C) + \sum_{k=0}^{t}\gamma_k\eta_k^2[\mathsf{L}^2 + (1+\gamma_k)\varsigma_k^2]\Big]}{2\Gamma_t},$$

where recall that $\varrho(C) = \min_{\mathbf{z}}\max_{\mathbf{w}\in C}\|\mathbf{z}-\mathbf{w}\|_2$ is the radius while $\operatorname{diam}(C) = \max_{\mathbf{w},\mathbf{z}\in C}\|\mathbf{w}-\mathbf{z}\|_2$ is the diameter. Obviously, $\varrho(C) \le \operatorname{diam}(C) \le 2\varrho(C)$.

- If we set $\gamma_k \equiv 1$ for $k \ge 0$ and assume the noise variance is uniformly bounded, i.e. $\varsigma_k \le \varsigma$, we obtain

  $$\mathbb{E}[\underbrace{\overline{f}(\bar{\mathbf{x}}_t) - \mathfrak{p}_\star}_{\text{primal gap}} + \underbrace{\mathfrak{d}^\star - \underline{f}(\bar{\mathbf{y}}_t)}_{\text{dual gap}}] \le \frac{2\operatorname{diam}^2(C) + \sum_{k=0}^{t}\eta_k^2[\mathsf{L}^2 + 2\varsigma^2]}{2H_t}, \quad \text{where} \quad H_t = \sum_{k=0}^{t}\eta_k.$$

  Thus, the expected primal and dual gaps go to 0 if $H_t \to \infty$ and $\eta_t \to 0$. Note also that if $\varsigma \equiv 0$, then our bound reduces to the one in Remark 11.4, where the constant $2\operatorname{diam}^2(C)$ may be halved (the looseness came from (12.5)).

- To maximize the convergence rate, let us set $\eta_k = \frac{c}{\sqrt{k+1}}$ and $\gamma_k = 0$ if $k < s$ and 1 otherwise. Thus,

$$\begin{aligned}\mathbb{E}[\overline{f}(\bar{\mathbf{x}}_t) - \mathfrak{p}_\star + \mathfrak{d}^\star - \underline{f}(\bar{\mathbf{y}}_t)] &\le \frac{2\operatorname{diam}^2(C) + \sum_{k=s}^t \frac{c^2}{k+1}[\mathsf{L}^2 + 2\varsigma^2]}{2\sum_{k=s}^t \frac{c}{\sqrt{k+1}}} \\ &\le \frac{2\operatorname{diam}^2(C)}{2\sum_{k=s}^t \frac{c}{\sqrt{t+1}}} + \frac{\sum_{k=s}^t \frac{1}{\sqrt{k+1}} \frac{c}{\sqrt{s+1}}[\mathsf{L}^2 + 2\varsigma^2]}{2\sum_{k=s}^t \frac{1}{\sqrt{k+1}}} \\ &= \frac{\operatorname{diam}^2(C)\sqrt{t+1}}{(t-s+1)c} + \frac{c(\mathsf{L}^2 + 2\varsigma^2)}{2\sqrt{s+1}}.\end{aligned}$$

If we optimize c we obtain

$$\mathbb{E}[\overline{f}(\bar{\mathbf{x}}_t) - \mathfrak{p}_\star + \mathfrak{d}^\star - \underline{f}(\bar{\mathbf{y}}_t)] \le \operatorname{diam}(C)\sqrt{2(\mathsf{L}^2 + 2\varsigma^2)}\sqrt[4]{\tfrac{t+1}{s+1}} \cdot \tfrac{1}{\sqrt{t-s+1}}.$$

Thus, with $s \propto t$, e.g. $s = \lceil t/2 \rceil$, the expected primal and dual gaps converge to 0 at rate $O(1/\sqrt{t})$.

Remark 12.8: Convergence of SGD

Let us now consider the special case $\mathsf{Y} = \{\mathbf{y}_0\}$, i.e. we are only interested in minimizing $f(\mathbf{x}) := f(\mathbf{x}, \mathbf{y}_0)$ w.r.t. $\mathbf{x}$. The stochastic GDA algorithm reduces to the popular stochastic gradient descent (SGD) algorithm that can be traced back to Robbins and Monro (1951).

Take $K = \{(\mathbf{x}, \mathbf{y}_0)\}$ with $\mathbf{x} \in \mathsf{X}$ arbitrary and assume the noise variance is uniformly bounded by ς^2. Then,

$$\begin{gathered}\overline{f}_K(\bar{\mathbf{x}}_t) = f(\bar{\mathbf{x}}_t), \quad \underline{f}_K(\bar{\mathbf{y}}_t) = f(\mathbf{x}), \quad \varrho(K) = 0, \quad \varsigma_k^2 = \mathbb{E}\|\boldsymbol{\xi}_k\|_2^2, \quad \mathbb{E}\sigma_K(\bar{\boldsymbol{\xi}}_t) = 0 \quad \gamma_k \equiv 1 \quad \Longrightarrow \\ \mathbb{E}[f(\bar{\mathbf{x}}_t) - f(\mathbf{x})] \le \frac{\|\mathbf{x}_0 - \mathbf{x}\|_2^2 + \sum_{k=0}^t \eta_k^2[\mathsf{L}^2 + \varsigma^2]}{2H_t},\end{gathered}$$

assuming $\mathbf{x}_0$ is non-random (or being conditioned on). Therefore, if $H_t \to \infty$ and $\eta_t \to 0$, the expected function value at the averaged iterate $\bar{\mathbf{x}}_t$ eventually crosses below that of any feasible $\mathbf{x} \in \mathsf{X}$.

As above we can set $\gamma_k = 0$ if $k < s \approx t/2$ and 1 otherwise, so that we obtain an $O(1/\sqrt{t})$ convergence rate of the expected function value. This rate is optimal.

Robbins, H. and S. Monro (1951). "A Stochastic Approximation Method". *Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407.

Remark 12.9: Lagrangian dual

More generally, if say Y is bounded and f is L-Lipschitz continuous, then according to the minimax Theorem 8.15 strong duality holds. Let us assume there exists a saddle point $(\mathbf{x}_\star, \mathbf{y}^\star)$ so we choose $K = \{\mathbf{x}_\star\} \times \mathsf{Y}$ to arrive at:

$$\begin{gathered}\overline{f}_K(\bar{\mathbf{x}}_t) = \overline{f}(\bar{\mathbf{x}}_t), \quad \underline{f}_K(\bar{\mathbf{y}}_t) = f(\mathbf{x}_\star, \bar{\mathbf{y}}_t) \le \mathfrak{p}_\star, \quad \varrho(K) = \varrho(\mathsf{Y}), \quad \mathbb{E}\sigma_K(\bar{\boldsymbol{\varepsilon}}_t) = \mathbb{E}\sigma_{\mathsf{Y}}(\bar{\boldsymbol{\zeta}}_t), \quad \gamma_k \equiv 1 \quad \Longrightarrow \\ \mathbb{E}\overline{f}(\bar{\mathbf{x}}_t) - \mathfrak{p}_\star \le \frac{\|\mathbf{x}_0 - \mathbf{x}_\star\|_2^2 + 2\operatorname{diam}^2(\mathsf{Y}) + \sum_{k=0}^t \eta_k^2(\mathsf{L}^2 + 2\varsigma^2)}{2H_t}.\end{gathered}$$

Thus, the expected primal gap goes to 0 if $H_t \to \infty$ and $\eta_t \to 0$. Needless to say, we can set $\gamma_k = 0$ if $k < s \approx t/2$ to derive the now familiar $O(1/\sqrt{t})$ convergence rate.

The above result is particularly applicable to the Lagrangian when Slater's condition holds (for the primal problem).

Remark 12.10: From expectation to high probability

So far, our results are about the expected behaviour, i.e., if we repeat the optimization algorithm many times, then the averaged result will converge to optimum. In practice, a single run is usually enough (unless we are extremely unlucky). This is due to the phenomenon known as concentation of measure, i.e. average of almost i.i.d. random variables is very likely concentrated around their mean. Indeed, our main performance measure

$$\mathfrak{M}_t := \max_{\mathbf{w}\in K} \sum_{k=0}^{t} b_{t,k} \left\langle \mathbf{w}_k - \mathbf{w}, \boldsymbol{\varepsilon}_k \right\rangle,$$

after a covering argument to union bound the max, we are left with averages of a martingale difference, which behaves probabilistically as i.i.d. random variables. Thus, with high probability (and moderately large t), $\mathfrak{M}_t \approx 0$ hence a single run will be close to the expected performance. Complete details can be found in Nemirovski et al. (2009).

Nemirovski, A., A. Juditsky, G. Lan, and A. Shapiro (2009). "Robust Stochastic Approximation Approach to Stochastic Programming". *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1574–1609.